Semiparametric graph neural networks for energy regression of hadron showers in the CMS High Granularity Calorimeter

A Thesis

submitted to
Indian Institute of Science Education and Research Pune
in partial fulfillment of the requirements for the
BS-MS Dual Degree Programme

by

Nitish Kumar K V

Indian Institute of Science Education and Research Pune
Dr. Homi Bhabha Road,
Pashan, Pune 411008, INDIA.

May, 2023

Supervisor: Dr. Seema Sharma

© Nitish Kumar K V 2023

All rights reserved

Certificate

This is to certify that this dissertation entitled **Semiparametric graph neural networks for energy regression of hadron showers in the CMS High Granularity Calorimeter**, towards the partial fulfilment of the BS-MS dual degree programme at the Indian Institute of Science Education and Research, Pune represents study/work carried out by Nitish Kumar K V at Indian Institute of Science Education and Research under the supervision of Dr. Seema Sharma, Associate Professor, Department of Physics , during the academic year 2022-2023.

Nitish Kumar K V

Dr. Seema Sharma

Committee:

Dr. Seema Sharma

Dr. Aditee Rane

This thesis is dedicated to my parents, Mr. K V Gowda and Mrs. Sujatha H M.

Declaration

I hereby declare that the matter embodied in the report entitled **Semiparametric graph neural networks for energy regression of hadron showers in the CMS High Granularity Calorimeter**, are the results of the work carried out by me at the Department of Physics, Indian Institute of Science Education and Research, Pune, under the supervision of Dr. Seema Sharma and the same has not been submitted elsewhere for any other degree.

Dr. Seema Sharma

Nitish Kumar K V

Acknowledgments

I would like to begin by expressing my sincere gratitude to my supervisor, Dr. Seema Sharma, for her invaluable guidance and support throughout my research project. Dr. Seema Sharma challenged me to overcome my comfort zone in research and helped me become a better researcher every day. Thanks to Dr. Shubham Pandey and Alpana Sirohi for sharing their knowledge and skills in the field of high energy physics with me and helping me with both the technical and conceptual difficulties I have faced. The expertise and assistance provided by Dr. Rajdeep Chatterjee, Dr. Roger Rusack, and Simon Rotham during our GNN group meetings were invaluable in advancing the project. Special thanks to my thesis expert, Dr. Aditee Rane, for all the suggestions and discussions regarding my master's project. I am grateful to our group members in the IISER-HEP group, Bhumika Kansal, Chirayu Gupta, Pradyun Hebbar and Chetan Agrawal, for all the suggestions and discussions during the group meetings. I want to thank the Minnesota Supercomputing Institute for the provision of computing resources for my research project. Also, the support and the resources provided by 'PARAM Brahma Facility' under the National Supercomputing Mission (NSM), Government of India at the Indian Institute of Science Education and Research (IISER) Pune are gratefully acknowledged. I also thank Nisha Kurkure and Goldi Misra for their prompt technical support regarding the NSM cluster.

Lastly, I am grateful to my parents and friends for their love and support. Thank you for being there for me.

Abstract

The high luminosity phase of the LHC (HL-LHC) poses significant challenges of radiation damage to the components of the CMS detector expected from the increased integrated luminosity and high event pileup. In view of HL-LHC phase, the CMS collaboration has opted for the High Granularity Calorimeter (HGCAL) to replace the current electromagnetic and hadronic calorimeters. The HGCAL features high radiation tolerance, unprecedented transverse and longitudinal segmentation for both electromagnetic and hadronic compartments, and high-precision timing capabilities, facilitating efficient particle flow reconstruction, energy rejection from the pileup, and particle identification. To validate the proposed design of the HGCAL, a series of beam tests have been carried out using silicon and scintillator based sampling calorimeter prototype. The prototype was exposed to beams of high-energy e^+ and π^- of momentum ranging from 20 to 300 GeV/c, and μ^- of momentum 200 GeV/c. This thesis focuses on the energy regression of charged pions in the HGCAL test beam prototype using semiparametric graph neural networks. Given the complexity of the pion showers, advanced machine learning models like graph neural networks can fully utilize the reconstructed hit information, i.e., the spatial coordinates and the energy information from the fine lateral and longitudinal granularity of the HGCAL, to reconstruct pion energy efficiently.

Contents

Abstract **vi**

1 Introduction **6**

2 Particle interactions with matter **8**

2.1 Energy loss by charged particles 8

2.1.1 Energy loss by ionization processes 8

2.1.2 Energy loss by radiation 10

2.2 Energy loss by photons 11

2.3 Energy loss by hadrons 12

2.4 Shower development 13

2.4.1 Electromagnetic shower 13

2.4.2 Hadronic shower 14

2.5 Calorimeters 15

2.5.1 Semiconductors as the active material 16

2.5.2 Scintillators as the active material 18

2.5.3 Energy response and resolution of calorimeters 18

2.6 Detector simulation using GEANT4 toolkit 19

3 The CMS High Granularity Calorimeter **21**

3.1 HGCAL test beam setup October-2018 21

3.1.1 CE-E prototype 22

3.1.2 CE-H prototype 23

3.1.3 AHCAL prototype 23

3.1.4 Source of particle beams 24

3.1.5 MIP calibration 24

3.2 Test beam simulation setup 24

3.3 Full HGCAL simulation setup 26

3.4 Particles created in pion showers 26

3.4.1 Particles produced in the first hadronic interaction of pion 26

3.4.2 π^0 produced in pion showers 29

3.5 Measurement of energy of charged pions 32

4 Energy regression of charged pions using semiparametric graph neural networks **34**

4.1 Basic building blocks of machine learning algorithm 35

4.1.1 Fully connected neural network 35

4.1.2 Optimization algorithm 36

4.1.3 Activation functions 37

4.2 Dynamic Reduction Network 39

4.3 Training 40

4.3.1 Simulation samples used for training 40

4.3.2 Input features 41

4.3.3 Target 41

4.4 DRN performance on Simulation 44

4.5 DRN performance on HGCAL-TB data 47

4.6 Resolving Edge Effects 51

4.7 DRN performance as a function of π_f^0 54

4.8 Performance of the DRN when providing reduced information as inputs 58

4.8.1 Performance of DRN when the rechit level information is combined along the transverse direction 58

4.8.2 Performance of DRN when the rechit level information is combined along the longitudinal direction 62

5 Summary **65**

Appendix A Plots and graph **66**

List of Figures

2.1 Mean energy loss by charged particles in different materials [1]. 9

2.2 A strong nuclear interaction of proton with the nucleus in a photographic emulsion [2]. 12

2.3 A schematic of an electromagnetic shower induced by a photon [3]. 13

2.4 Simulation of the longitudinal development of 10 GeV electron showers in aluminum (AL), iron (Fe) and lead (Pb) [3]. 14

2.5 A schematic of development of a hadronic shower [3]. 14

2.6 The simulation of energy deposition of two different 270 GeV pion in a block of copper [3]. 15

2.7 A schematic of p-n junction diode. 17

2.8 A depiction of a charged particle passing through a silicon sensor. 17

3.1 The longitudinal cross-section of the CMS HGCAL [4]. 22

3.2 HGCAL-TB setup, October-2018 [5]. 22

3.3 An image of mini-cassette of CE-E section (left), a CE-H layer with seven modules arranged in a daisy-like structure (middle), AHCAL layer showing scintillator mounted on the SiPM [6]. 23

3.4 The ADC spectrum of the muon in a 200μm thick silicon cell fitted with a Landau distribution convoluted with a Gaussian distribution [7]. 25

3.5 A simulation of the HGCAL prototype showing the active layers and the absorber material [6]. 25

3.6 Distribution of the number (left) and the fraction of energy carried by the neutrons (right) created in the first hadronic interaction occurring in CE-E section as predicted by QGSP_BERT hadronic physics list. 27

3.7 Average particle multiplicity (top row) and the average energy carried by the secondary particles (bottom row) created in the first hadronic interaction occurring in CE-E section (left column) and CE-H section (right column) as predicted by QGSP_BERT hadronic physics list. 28

3.8 The fraction of energy carried by π^0 in first hadronic interaction (π^0_f), in the later part of the shower (π^0_l) and all π^0 (π^0_{all}) as predicted by QGSP_BERT hadronic physics list. 30

3.9 Mean (sample mean) of π^0_f, π^0_l and π^0_{all} as a function of incident pion energy as predicted by QGSP_BERT hadronic physics list. 30

3.10 The fraction of energy carried by π^0 in first hadronic interaction (π^0_f) versus the fraction of energy carried by π^0 in the later part of the shower (π^0_l) as predicted by QGSP_BERT hadronic physics list. 31

3.11 Sum of energy measured (MIP units) in CE-E versus CE-H+AHCAL. 32

4.1 A fully connected neural network with 1 hidden layer [8]. 35

4.2 A depiction of the gradient descent algorithm for different values of learning rate (α) [9]. 37

4.3 Activation functions (left) and their derivatives (right). 37

4.4 A schematic of dynamic reduction network. 39

4.5 Distributions of input features and the number of rechits. 42

4.6 Distribution of different targets used for training: E_{true} (left) and $\frac{E_{true}}{E_{raw}}$ (right). . . . 42

4.7 The DSCB probability distribution (left) and the negative logarithm of DSCB distribution (right) for different values of parameters. 43

4.8 Loss versus epoch plot for model trained using E_{true} target (left) and $\frac{E_{true}}{E_{raw}}$ target (right). 44

4.9 The predicted energy distributions and the fitted gaussian function is shown for the prediction made by the DRN-TrueE model on the validation dataset. 45

4.10 The comparison of resolution (a) and response (b) for the validation and training dataset for the model DRN-TrueE. 45

4.11 The predicted energy distributions and the fitted gaussian function is shown for the prediction made by the DRN-Ratio model on the validation dataset. 46

4.12 The comparison of resolution (a) and response (b) for the validation and training dataset for the model DRN-Ratio. 46

4.13 Predicted energy distribution comparison of the validation dataset for three different models: DRN-Ratio (blue curve), DRN-TrueE (magenta curve) and χ^2 Method (black curve). 48

4.14 The comparison of resolution (a) and response (b) of the validation dataset for three different models: DRN-Ratio (blue curve), DRN-TrueE (magenta curve) and χ^2 Method (black curve). 48

4.15 Predicted energy distribution comparison of the validation dataset (red curve) and test beam data (black curve) for the DRN-TrueE model. 49

4.16 Comparison of resolution (a) and response (b) of the validation dataset (red) and test beam data (black) for the DRN-TrueE model. 49

4.17 Predicted energy distribution comparison of the validation dataset (red curve) and test beam data (black curve) for the DRN-Ratio model. 50

4.18 Comparison of resolution (a) and response (b) of the validation dataset (red) and test beam data (black) for the DRN-Ratio model. 50

4.19 Pion energy distribution: true energy (black curve), predicted energy for the DRN-TrueE model (magenta curve) and predicted energy for the DRN-Ratio model. . . . 51

4.20 True (black curve) and predicted (red curve) energy distribution (training set) for a model trained with a guassian edge in pion energy distribution. 52

4.21 Comparison of resolution (a) and response (b) of the models DRN-GaussEdge (green curve) and DRN-Flat60to300 (brown curve). 53

4.22 Predicted energy as a function of π_f^0: χ^2 prediction (left) and DRN prediction (right). 55

4.23 Distribution of π_f^0 for 100 GeV pion and the different bins used for analysis. 56

4.24 Predicted energy distribution for different bins of π_f^0 for χ^2 (left) and DRN (right). 57

4.25 The number of hits recorded for different pion events in the reduced configuration versus the original configuration of the HGCAL prototype 59

4.26 Predicted energy distribution comparison of the training dataset for three different configurations: HGCAL-Original (blue curve), HGCAL-ReducedBy4 (red curve) and HGCAL-ReducedBy16 (green curve). 60

4.27 The comparison of resolution (a) and response (b) of the training dataset for three different configurations: HGCAL-Original (blue curve), HGCAL-ReducedBy4 (red curve) and HGCAL-ReducedBy16 (green curve). 60

4.28 Predicted energy distribution comparison of the training dataset for three different configurations: HGCAL-Original (blue curve), HGCAL-Layerwise_Ez (magenta curve). 61

4.29 The comparison of resolution (a) and response (b) of the training dataset for three different configurations: HGCAL-Original (blue curve), HGCAL-Layerwise_Ez (magenta curve). 61

4.30 Predicted energy distribution comparison of the training dataset for four different configurations: HGCAL-Original (blue curve), HGCAL-RedLong_CEE (red curve), HGCAL-RedLong_CEH (green curve) and HGCAL-RedLong_All (magenta curve). 63

4.31 The comparison of resolution (a) and response (b) of the training dataset for four different configurations: HGCAL-Original (blue curve), HGCAL-RedLong_CEE (red curve), HGCAL-RedLong_CEH (green curve) and HGCAL-RedLong_All (magenta curve). 64

A.1 The predicted energy distributions and the fitted gaussian function is shown for the prediction made by the DRN-TrueE model on the validation dataset. 67

A.2 The predicted energy distributions and the fitted gaussian function is shown for the prediction made by the DRN-Ratio model on the validation dataset. 68

A.3 Predicted energy distribution comparison of the validation dataset for three different configurations: original (blue curve), reduced by a factor of 4 (red curve) and 16 (green curve). 69

A.4 The comparison of resolution (a) and response (b) of the validation dataset for three different configurations: original (blue curve), reduced by a factor of 4 (red curve) and 16 (green curve). 69

A.5 Predicted energy distribution comparison of the validation dataset for three different configurations: HGCAL-Original (blue curve), HGCAL-Layerwise_Ez (magenta curve). 70

A.6 The comparison of resolution (a) and response (b) of the validation dataset for three different configurations: HGCAL-Original (blue curve), HGCAL-Layerwise_Ez (magenta curve). 70

A.7 Predicted energy distribution comparison of the validation dataset for four different configurations: HGCAL-Original (blue curve), HGCAL-RedLong_CEE (red curve), HGCAL-RedLong_CEH (green curve) and HGCAL-RedLong_All (magenta curve). 71

A.8 The comparison of resolution (a) and response (b) of the validation dataset for four different configurations: HGCAL-Original (blue curve), HGCAL-RedLong_CEE (red curve), HGCAL-RedLong_CEH (green curve) and HGCAL-RedLong_All (magenta curve). 71

Chapter 1

Introduction

The Large Hadron Collider (LHC) [10] at CERN resumed proton-proton (pp) collisions at center-of-mass energy $\sqrt{s}$=13.6 TeV in July 2022 after the long shutdown (LS2) of three years. At the end of the ongoing operations in 2025, called Run-3, the CMS [11] and ATLAS [12] experiments are expected to accumulate pp collision data corresponding to a total integrated luminosity of approximately 300 fb^{-1} (including data collected during Run-2 at $\sqrt{s}$ = 13 TeV). The LHC is presently operating at the instantaneous luminosity of 2.1×10^{34} cm^{-2} s^{-1}.

Following the next long shutdown (LS3), the LHC will commence its high luminosity phase, called HL-LHC, with pp collisions at much higher instantaneous luminosity (5×10^{34} cm^{-2} s^{-1}) [4]. It is expected to start in 2029 and aims to collect data corresponding to an integrated luminosity of 3000 fb^{-1} during a span of ten years of operations. Such a large dataset will open opportunities for precision measurements of the standard model (SM) predictions and searches for new physics beyond the SM (BSM).

The HL-LHC phase poses significant challenges of radiation damage to the components of the CMS detector coming from the increased integrated luminosity and a large number of pp collisions overlapping with the collision of interest (pileup) owing to the substantial increase in the instantaneous luminosity. The current CMS is not designed to operate under such harsh conditions, and various detector components must be upgraded to maintain optimal efficiency, resolution and signal-to-noise ratio of the detector during the HL-LHC phase. As a part of the HL-LHC upgrade, the CMS collaboration plans to replace the current endcap calorimeters with the High Granularity Calorimeter (HGCAL) [4], which is a sampling calorimeter with silicon and scintillator as the active materials.

The high amount of radiation in the scintillators creates absorber bands in the materials, which significantly affects light propagation and thus, makes them less transparent. The current endcap calorimeters, which are based on scintillators, the $PbWO_4$-based electromagnetic calorimeter and the plastic scintillator-based hadronic calorimeter, are designed to handle an integrated luminosity of 500 fb^{-1} [4]. The detector replacing the current endcap calorimeters in the HL-LHC should tolerate integrated radiation that will be ten times larger than the LHC. Radiation hard silicon sensors are chosen as the active material for the region of the HGCAL calorimeter, where the cumulative radiation doses are expected to be high. Whereas in the other regions of the HGCAL, plastic scintillators directly read out by silicon photomultipliers (SiPMs) are chosen as the active material [4].

The thesis begins with an introduction to the interactions of particles with matter, electromagnetic and hadronic showers, calorimeters and detector simulation using the GEANT4 toolkit in chapter 2. Chapter 3 describes the details of the proposed CMS HGCAL detector and the test beam setup of HGCAL. The detector calibration techniques and the studies related to particles produced in the hadronic shower are also described in chapter 3.

This thesis focuses on the energy reconstruction of charged pions in the HGCAL test beam setup using semiparametric graph neural networks. Chapter 4 discusses the details of the graph neural network-based dynamic reduction network (DRN) model and its performance on energy reconstruction of pions in the HGCAL test beam setup. The performance of the DRN is also studied as a function of the energy carried by π^0 in the first hadronic interaction. At the end of the chapter, we study the performance of the DRN when providing reduced information as inputs. Chapter 5 summarizes the main results and outcome of the thesis.

Chapter 2

Particle interactions with matter

The particles produced in the collider-based experiments carry high energy. Heavy unstable particles decay into lighter particles that interact with the materials of the detector via different processes. The particles produced in the collisions, such as hadrons and electrons, deposit their energy in the detector by developing a cascade of shower particles. The various processes by which the energetic particles interact and lose their energy in matter are described in the following sections.

2.1 Energy loss by charged particles

2.1.1 Energy loss by ionization processes

A charged particle passing through a medium interacts with the atomic electrons through Coulombic interaction [3]. This causes the electrons of the atom to either excite (excitation) or to be ejected from the parent atom (ionization) and thus, results in the energy loss of the incident charged particle. The energy loss through the ionization process by the charged particle per unit path length (x) is given by the Bethe-Bloch equation:

$$-\frac{dE}{dx}|_{ion} = \rho\ N_A\ \frac{Z}{A}\ \frac{4\pi\alpha^2\hbar^2}{m_e}\ \frac{q^2}{\beta^2}\left[ln\left(\frac{2m_ec^2\beta^2\gamma^2}{I}\right) - \beta^2 - \frac{\delta}{2}\right] \tag{2.1}$$

where E is the kinetic energy of the incident particle of charge q traveling with velocity v, β is v/c, γ is the Lorentz factor, N_A is the Avogadro's number, ρ is the density of the medium, Z/A

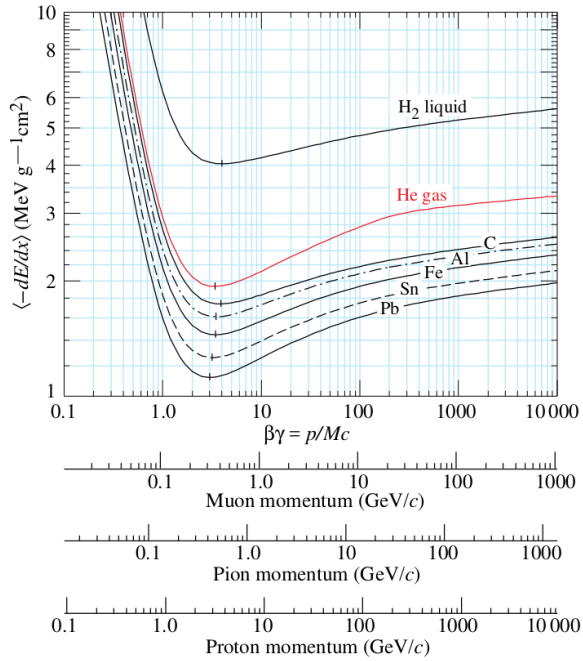


Figure 2.1: Mean energy loss by charged particles in different materials [1].

is the atomic number/mass of the medium, I $(\approx 10 \times Z\ eV)$ is the mean ionization potential of the medium, m_e is the mass of the electron, α is the fine-structure constant and δ is the correction for dielectric screening effect for a highly relativistic particle.

There are a few notable features that can be observed from the formula:

- The energy loss is inversely proportional to β^2 as the slower particle has more time to spend in the vicinity of the Coulombic field of atoms and loses more energy.
- The more charge the particle carries, the more will be the energy loss ($\propto q^2$).
- Higher the density of the medium, the higher the energy loss ($\propto \rho$).
- At relativistic speeds ($\beta \approx 1$), the logarithmic term of the equation dominates and increases the energy loss by the charged particle. This is due to the increase in the transverse electric field of the relativistic charged particle and is called the relativistic rise of energy loss.

The energy loss is usually expressed in units of $MeV\ g^{-1}\ cm^2$ by dividing the equation 2.1 by ρ, which is roughly the same in a wide range of materials. Figure 2.1 shows the energy loss of

charged particles as a function of particle momentum. The figure shows that the energy loss rapidly decreases with increasing β and reaches a minimum value at around $\beta\gamma$ =3–4. The particle at this point is considered *minimum ionizing particle* or MIP since it loses minimum energy through ionization. On the further increase of the particle momentum, the energy loss increases due to the relativistic rise.

2.1.2 Energy loss by radiation

A charged particle passing through a medium undergoes acceleration or deceleration due to force from the electric fields of the nucleus of an atom. The change in momentum of the incident charged particle results in radiation in the form of photons called bremsstrahlung. The emitted photon carries a fraction of the energy of the incident charged particle. The probability of this process for a charged particle of mass m and atomic mass (Z) of the material in which the particle is passing through is given by:

$$\sigma_{brem} \propto \frac{Z^2\alpha^3}{m^2c^4} \tag{2.2}$$

The probability of bremsstrahlung is inversely proportional to the mass of the charged particle ($\propto \frac{1}{m^2}$). If we compare the muon ($m_\mu = 105.6$ MeV) and electron ($m_e = 0.5$ MeV), an electron is approximately 40000 times more probable to radiate bremsstrahlung photon than a muon. Muon loses very less energy through radiation and rarely initiates particle showers. Muon loses its energy mainly through the ionization process and it behaves as a minimum ionizing particle for a wide range of particle momentum. Usually, the energy loss of electrons at various depths it traverses in different materials is specified in terms of radiation length (X_0), which is the mean distance over which the electron loses 63.2% ($1-1/e$) of its energy.

The electron loses energy mainly through the ionization processes at lower energies and radiation processes at higher energies. The critical energy (ε) of a medium with the atomic number Z is the energy at which the energy loss due to ionization and radiation are equal [3], and has an approximate value of $\frac{560}{Z}$ MeV. The higher the atomic number (Z) of the material lower will be the critical energy. This follows from Equation 2.2, as the probability of bremsstrahlung is directly proportional to Z^2.

2.2 Energy loss by photons

As a photon traverses a material, it loses its energy mainly through the photoelectric effect, compton scattering and pair production. The photoelectric effect dominates at lower energies, the compton scattering dominates in the intermediate regime and the pair production dominates at higher photon energies.

- **Photoelectric effect:** It is a process in which an atom emits an electron after absorbing a photon. The photoelectric cross-section is given by :

$$\sigma photelectric \propto Z^5 \alpha^4 \left(\frac{m_e c^2}{E_\gamma}\right)^n \tag{2.3}$$

  where $n = \frac{7}{2}$ at $E_\gamma << m_e c^2$ and $n \rightarrow 1$ at $E_\gamma >> m_e c^2$ [3]. The cross-section of the photoelectric effect is highly dependent on the number of electrons available in the medium [2] and thus, strongly depends on the Z value of the medium ($\propto Z^5$). For high Z materials such as lead (Pb) and uranium (U), the energy loss for low-energy photons is mainly through the photoelectric effect.

- **Compton scattering:** It is a process in which an electron scatters a photon and a fraction of the energy of the photon is lost to the struck electron [2]. The compton scattering cross-section is given by:

$$\sigma_{compton} \propto \frac{lnE_\gamma}{E_\gamma} \tag{2.4}$$

- **Pair production:** In the presence of the electric field of the nucleus of an atom, if the energy carried by a photon is larger than two times the rest mass of the electron, it creates an electron-positron pair. This process is called pair production. The electron-positron pairs can lose energy by ionization or through bremsstrahlung radiation. The cross-section of the pair production is given by:

$$\sigma_{pair} \propto \frac{7}{9}\frac{A}{N_A}\frac{1}{X_0} \tag{2.5}$$

  The energy loss of photons through pair production starts to dominate at much lower energy for material with high atomic mass ($\propto A$).

2.3 Energy loss by hadrons

A highly energetic hadron (charged or neutral) deposits its energy in a material by interacting with its nuclei via strong nuclear interactions. A charged hadron continues to lose its energy through ionization until it undergoes such a hadronic interaction. In dealing with hadrons, it is convenient to specify the depth of different materials in terms of interaction length (λ_{int}), which is defined as the mean distance traveled by a hadron before it undergoes a strong hadronic interaction. The inelastic strong interaction of the incoming hadron with the atomic nucleus leads to spallation. The spallation process can be described in two steps, a fast intranuclear cascade followed by evaporation of the excited nucleus. Figure 2.2 shows the interaction of a 30 GeV proton with the nucleus in a photographic emulsion.

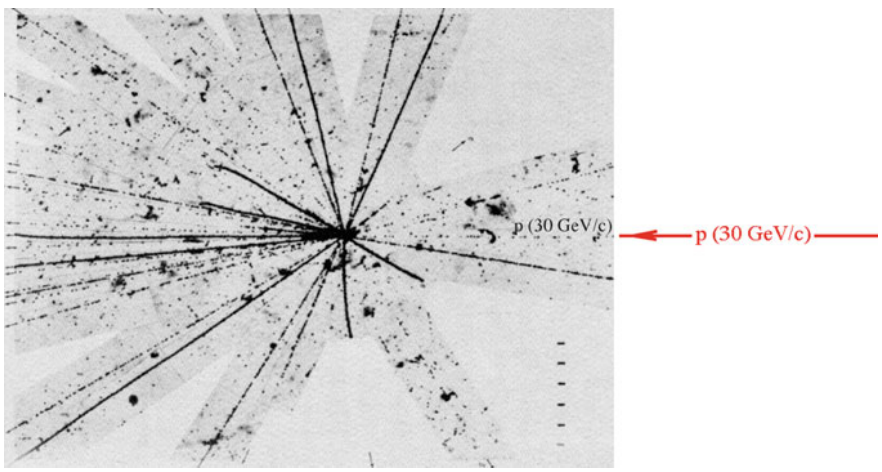


Figure 2.2: A strong nuclear interaction of proton with the nucleus in a photographic emulsion [2].

In the fast intranuclear cascade, the incoming hadron has a quasi-free collision [2] with the nucleons of the atomic nucleus. When some fraction of the energy of incident hadron is transferred to the nucleons, they start traveling within the nucleus and collide with other nucleons, which leads to the development of a cascade of fast nucleons. If the energy transferred by the incoming hadron is sufficiently high, pions and other unstable hadrons can be created [2]. The particles having enough kinetic energy to overcome the nuclear binding energy will escape the nuclear boundary and other particles with lower kinetic energy distribute their energy in the nucleus to form an excited nucleus.

The process following the fast intranuclear cascade is relatively slow, where the excited nucleus returns to the ground state by emitting certain particles (evaporation). The emitted particles are

predominantly the free nucleons, i.e., protons and neutrons. Even α particles and photons are emitted in the process. The heavier nucleus, such as uranium, can also undergo nuclear fission reaction. The particles are emitted more or less isotropically in this process [2].

In Figure 2.2, the tracks appearing in the direction of the incident proton are mostly the charged pions and protons coming from the fast intranuclear cascade. We also see that a few charged particles travel in the positive x direction. These particles are emitted during the evaporation stage.

2.4 Shower development

2.4.1 Electromagnetic shower

A cascade of photons and electrons develops when a highly energetic photon or electron is incident on an absorber material. Figure 2.3 shows a schematic of an electromagnetic shower produced by a photon. The photon pair produce electron-positron and these, in turn, produce photons via bremsstrahlung. With an increasing depth of the absorber material, the number of particles in the electromagnetic shower increases till some depth (called the shower maximum) while the mean energy carried by the particles decreases.

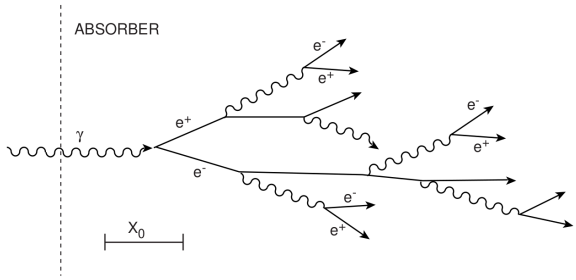


Figure 2.3: A schematic of an electromagnetic shower induced by a photon [3].

Figure 2.4 shows the longitudinal development of the electron shower in different materials. The multiplication of the shower particles continues until the energy carried by the particles is below the critical energy (ε) of the material and the energy deposition increases accordingly. After the shower maximum has been reached, the energy deposition falls off exponentially. At this point, the photons lose energy mainly through the compton scattering or the photoelectric effect and the e^-, e^+ lose energy through ionization of the absorber material. The shower maximum is deeper for the high Z material (low critical energy) as the multiplication of shower particles continues down to lower energies.

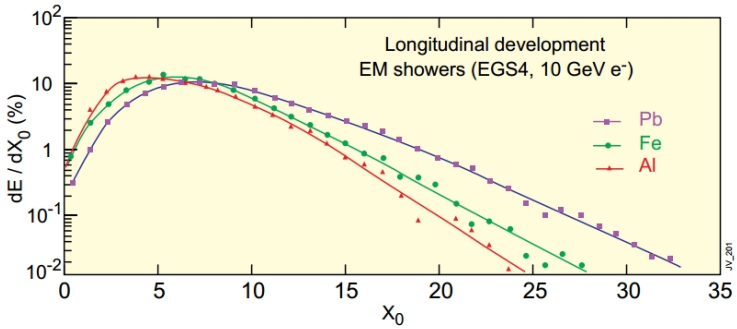


Figure 2.4: Simulation of the longitudinal development of 10 GeV electron showers in aluminum (AL), iron (Fe) and lead (Pb) [3].

The electromagnetic shower spreads in the lateral direction due to multiple scattering of the secondary e^-, e^+ and emission of photons away from the shower axis. The lateral extent of the electromagnetic shower for different materials can be specified in terms of Moliére radius (R_M), which is the radius of the cylinder of infinite length, which contains 90% of the shower energy.

2.4.2 Hadronic shower

As mentioned in the section 2.3, the strong nuclear interactions produce multiple secondary hadrons. These secondary hadrons can further interact with the absorber nuclei leading to an increase in the number of particles in the cascade. The particle multiplication continues until the energy carried by the secondary hadron goes below the threshold for producing other hadrons. The cascade contains two distinct components (Figure 2.5), namely the electromagnetic one ($\eta/\pi^0 \rightarrow \gamma\gamma$) and the hadronic one ($\pi^-$, π^+, n, p, etc.) [3].

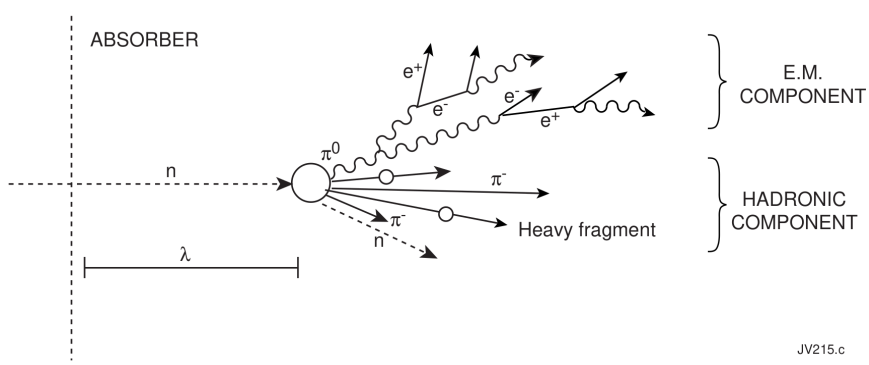


Figure 2.5: A schematic of development of a hadronic shower [3].

The particles emitted during the nuclear spallation process have to overcome the nuclear binding energy to escape the nuclear boundary. The energy needed for this process does not contribute

to the detector signal and is termed invisible energy. Hence, in general, the energy deposited, which is available to convert to a signal, from a pion shower is less than an electron shower of the same energy.

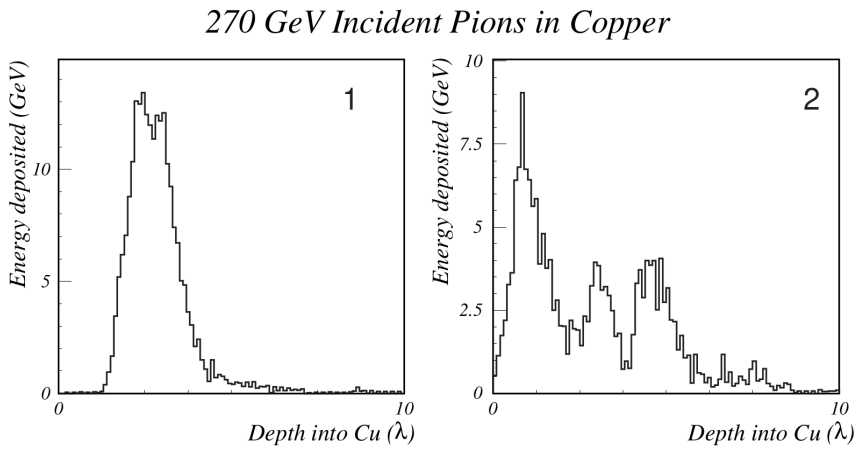


Figure 2.6: The simulation of energy deposition of two different 270 GeV pion in a block of copper [3].

For hadronic showers, there is also high shower-to-shower fluctuation in energy carried by the electromagnetic component (mainly coming from π^0) for a given pion energy. Also, the depth at which the π^0 is produced varies from shower-to-shower. Figure 2.6 shows the simulation of energy deposition of two different 270 GeV pion in a block of copper [3]. The peaks in the distribution represent π^0 production, which leads to higher energy deposition through electromagnetic interaction. I will elaborate more on the π^0 produced in the hadronic showers in section 3.4.2. The high-energy hadronic showers have a dense electromagnetic core that is surrounded by an exponentially decreasing non-electromagnetic halo [3]. The longitudinal spread of hadronic showers is more than an electromagnetic shower of the same energy. For example, $\sim$9 λ_{int} ($\sim$151 cm) of iron is needed to contain 99% of 100 GeV π^- shower [2]. In comparison, $\sim$25 X_0 ($\sim$44 cm) of iron is needed to contain 99% of 100 GeV e^- shower [7].

2.5 Calorimeters

Calorimeters measure the energy and position of the particles that pass through them and are designed to fully contain the showers that develop from e^-, e^+, photons and hadrons. The absorber medium is where the particle shower develops and deposits most of its energy and the active medium is where the deposited energy is actually measured. Based on the active and absorber material construction, calorimeters can be of two types: homogenous and sampling calorimeters.

- In **homogeneous calorimeters**, a single medium serves both the purpose of shower development (absorber) and energy measurements (active). For example, the current CMS electromagnetic calorimeter (ECAL) is a homogenous calorimeter made up of $PbWO_4$ scintillating crystals.

- In **sampling calorimeters**, dense absorber layers are sandwiched between light active layers. Most of the energy deposition occurs in the absorber layers and the active layers take a snapshot of energy deposition at intermediate depths. These calorimeters are mainly used to measure the energy from hadronic showers, which need a larger depth for containment. For example, the current CMS hadronic calorimeter (HCAL) is a sampling calorimeter with brass as the absorber material and plastic scintillators as the active material.

2.5.1 Semiconductors as the active material

A semiconductor is a material that has an electrical conductivity value between a conductor and an insulator. When a charged particle passes through a semiconductor crystal, the electrons of the semiconductor atoms excite from the valence band to the conduction band, forming an electron-hole pair. The pure crystal of semiconductor material (intrinsic semiconductor) is not suitable for particle detection due to the low signal-to-noise. The conductivity of the pure semiconductor material can be increased by adding impurities. Introducing donor atoms (phosphorous) forms the n-type semiconductors and adding acceptor atoms (aluminum) forms the p-type semiconductors. The majority charge carriers in n-type and p-type semiconductors are electrons and holes, respectively.

The p-type and n-type semiconductors can be joined to form a p-n junction diode, as shown in Figure 2.7. The majority charge carriers diffuse and recombine, creating a net electric field and potential difference across the junction, forming a depletion zone. Forward biasing the p-n junction diode can increase the noise in the detector due to the thermal generation of electron-hole pairs in the p-n junction. In reverse bias mode, the depletion region is free of charge carriers, which reduces the noise in the detector. Reverse biasing the p-n junction diode increases the electric field and the width of the depletion zone across the junction. Figure 2.8 depicts a charged particle passing through a silicon sensor. When an ionizing particle passes through the depletion zone of the silicon sensor, electron-hole pairs are created along its path. The electric field drifts the electron and holes to opposite electrodes and a current is generated proportional to the energy deposited by the ionizing particle. The more particles pass through the semiconductor sensor, the more will the

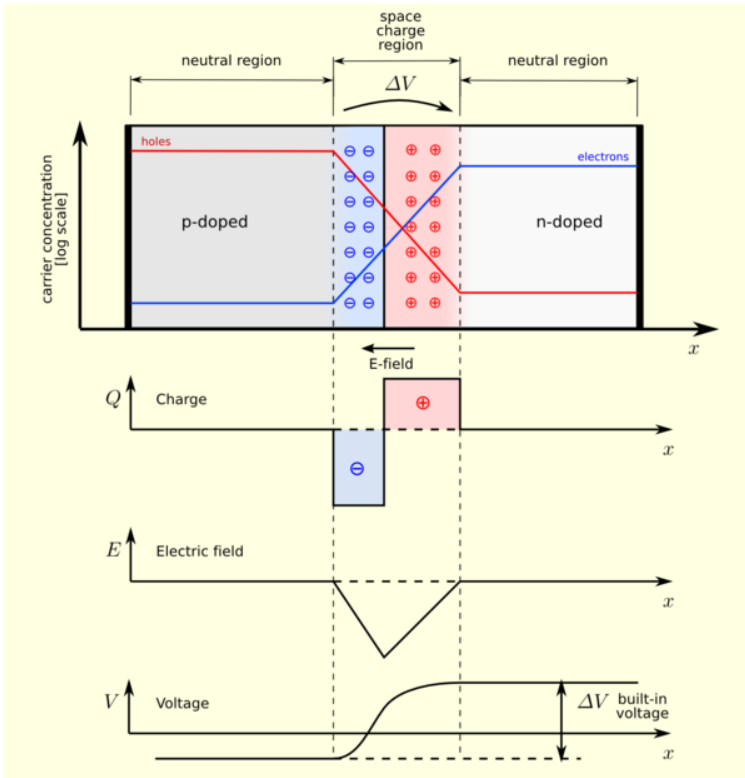


Figure 2.7: A schematic of p-n junction diode.

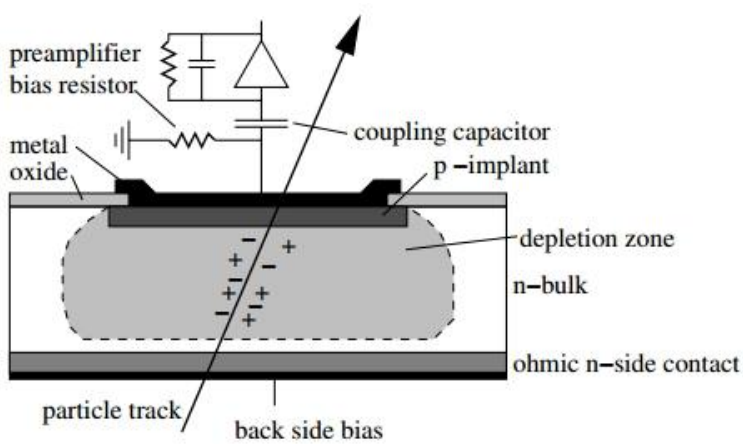


Figure 2.8: A depiction of a charged particle passing through a silicon sensor.

generation of electron-hole pairs. Thus, the current generated in the semiconductor cell is also indicative of the number of particles passing through it.

2.5.2 Scintillators as the active material

When a charged particle passes through scintillating material, it excites the molecules in the material. The excited molecules emit photons when returning to the ground state. The photons emitted are called scintillating photons. Usually, the scintillators are optically coupled with the photo-detector, which produces an electrical pulse proportional to the number of scintillating photons, and hence, the energy deposited in the scintillating material. The scintillating photons are transmitted to the photo-detector directly or by a light guide [13]. The photo-detector used in the HGCAL detector is the silicon photomultiplier (SiPM). A SiPM is made up of an array of small, electrically and optically isolated avalanche photodiodes, each of which can detect a single photon of light. The signals obtained from the SiPM are a sum of the signals obtained from each avalanche photodiode. The SiPM outputs an electrical signal proportional to the number of scintillating photons striking it. The scintillators are used to measure the particle energy and because of their fast response time, they are also used as trigger generators.

2.5.3 Energy response and resolution of calorimeters

The performance of the calorimeter can be measured in terms of response and resolution. The response is the average energy measured by the calorimeter for particles of a given energy, whereas the resolution is the measure of fluctuation in the calorimetric response.

Energy response

The response for the electrons and photons is linear with energy because the electromagnetic shower deposits all of its energy in the calorimeter and the energy deposited is completely converted to a detectable signal. However, non-linearity might arise due to the saturation of the device that produces the signal or the leakage of shower energy outside the calorimeter.

As discussed earlier, the hadronic shower has two components, the electromagnetic and the hadronic components. The energy of the electromagnetic component is completely converted to a detectable signal. In the hadronic component, a part of the energy is utilized for breaking up the nuclei and does not contribute to a detectable signal (invisible energy). Hence, the response of the

calorimeter is different for the hadronic and electromagnetic shower of the same energy. Also, the energy carried by electromagnetic component increases with an increase in the energy of incident hadron. This leads to non-linearity in the calorimetric response for hadrons.

Energy resolution

The resolution of the calorimeter is affected by different types of fluctuations, each with its own energy dependence. The energy resolution of the calorimeters can be parametrized by [3]:

$$\frac{\sigma}{E} = \frac{a}{\sqrt{E}} \oplus \frac{b}{E} \oplus c \tag{2.6}$$

Here, a is a stochastic term that accounts for fluctuation in the shower particle multiplicity that produces the detectable signal. The particle multiplicity follows Poissonian statistics resulting in a $E^{-1/2}$ dependence on the relative error. The b is a noise term that accounts for the fluctuations coming from the electronic noise in the detector system and the fluctuation in energy carried by particles coming from the pileup. The c is a constant term that accounts for the fluctuation coming from energy leakage and imperfect cell-to-cell inter-calibration error. The effects mentioned in the three terms are uncorrelated and therefore added in quadrature. Because of the different energy dependencies of the resolution, the noise term is the dominant contributor at lower energies (below a few GeV), the stochastic term is the dominant contributor at the intermediate regime ($\sim$10–100 GeV) and the constant term is the dominant contributor at very high energies (above $\sim$100 GeV) [2].

2.6 Detector simulation using GEANT4 toolkit

The GEANT4 is a toolkit that uses Monte Carlo methods to simulate the propagation and interaction of particles through matter [14]. It was originally built for high-energy physics experiments but also finds its application in space physics and medical sciences. In high energy physics, the GEANT4 simulation can be used for designing the detectors during the developmental stage and to study the detector response for physics studies [15]. The user can implement the detector geometry, define the physics processes and save information like various secondary particles created, energy deposited in the material, etc.

The detector simulation in the GEANT4 starts with the detector description. The GEANT4 has various handles through which the geometric shape, size, position and material attributes of the

detector components can be defined. It is also possible to define specific volumes of the detector as the *sensitive* material through which the simulated *hit* information of the detector can be obtained. For example, the energy deposited in the sensitive material can be used to compute the simulated hit (*simHit*) energy and as the position of the sensitive material is defined by the user, the spatial information of the hit can also be obtained.

Based on the knowledge of all possible particle interactions with matter, the GEANT4 offers a variety of physics processes that can be divided into electromagnetic and hadronic interactions, decays, etc. [16]. The user can choose the desired physics process for different particles involved in the shower. The physics processes decide how particles interact and traverse the detector material. The GEANT4 also has the *physics lists* with predefined particle interaction settings, especially when a particular setting is not optimal for a wide range of energy of particles to be simulated. For example, FTFP_BERT physics list includes all standard electromagnetic physics, uses Bertini (BER) style cascade for hadrons < 5 GeV and Fritiof (FTF) model for high energy hadrons ($>$ 4 GeV). QGSP_BERT physics list includes all standard electromagnetic physics, uses Bertini style cascaed for hadrons < 9.9 GeV and Quark gluon string model (QGS) model for high energy hadrons (> 18 GeV) and Fritiof in between.

The event simulation can start after specifying the detector geometry and physics processes. In the context of GEANT4, a *run* consists of a collection of events. The user cannot change the specified detector setup and physics list within a run. The *event* starts with the generation of one or more primary particles. The user has to define the attributes of the primary particle generator, such as the particle type, kinetic energy, etc. At the beginning of the event simulation, the primary tracks are generated, which are pushed into a stack. Each *track* present in the stack is traced through the detector and any secondary track generated will also be added to the same stack. The tracking is stopped when no track is left in the stack [15]. The *track* is a snapshot of the particle and it is deleted from the stack when it goes out of the outer volume, if the particle decays or has an inelastic collision, the kinetic energy of the particle goes to zero, for which no rest interaction is defined or if the user decides to terminate it. A *step* is the delta information of the track from which the user can access information such as the energy deposited by the particles, the secondary particles created while tracking the current track, the position and energy of the particles to name a few [17]. The processing of the event is stopped when the stack of tracks becomes empty.

Chapter 3

The CMS High Granularity Calorimeter

The proposed CMS HGCAL is a sampling calorimeter consisting of an electromagnetic calorimeter (CE-E) followed by a hadronic calorimeter (CE-H). The longitudinal cross-section of the HGCAL is shown in Figure 3.1. The CE-E section comprises 28 sampling layers with silicon as the active material. The silicon wafers are placed between Cu, CuW and Pb absorber material. The total thickness of the CE-E section is 34 cm covering a depth of approximately 26 X_0 and 1.7 λ_{int}. The CE-H section comprises 12 sampling layers with 35mm thick stainless steel plates (CE-H-fine) followed by 12 sampling layers with 68mm thick stainless steel plates (CE-H-coarse) [4]. The active material in the CE-H section are chosen according to the expected radiation level, i.e., the silicon sensors in the high radiation region (CE-H-Si) and scintillators directly read out by silicon photomultipliers (SiPM) in the low radiation region (CE-H-Sci). The total thickness of the CE-H section is approximately 157 cm covering a depth of approximately 9.5 λ_{int}. To lower the dark currents in the silicon sensors and the SiPMs, the HGCAL calorimeter will be placed in a cold volume maintained at -30°C [5].

3.1 HGCAL test beam setup October-2018

The HGCAL-TB (test beam) setup consists of Si-based electromagnetic (CE-E) and hadronic (CE-H) sections, followed by a scintillator-based CALICE AHCAL prototype section. Figure 3.2 shows the experimental setup of the beam test experiment. The prototype was exposed to high-energy beams of e^+, π^- with momenta ranging from 20 to 300 GeV/c and μ^- with momentum 200 GeV/c. The studies presented in this thesis are mostly based on the π^- data.

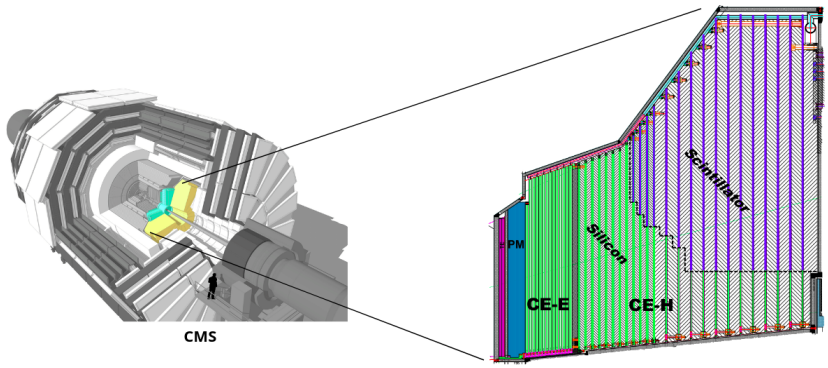


Figure 3.1: The longitudinal cross-section of the CMS HGCAL [4].

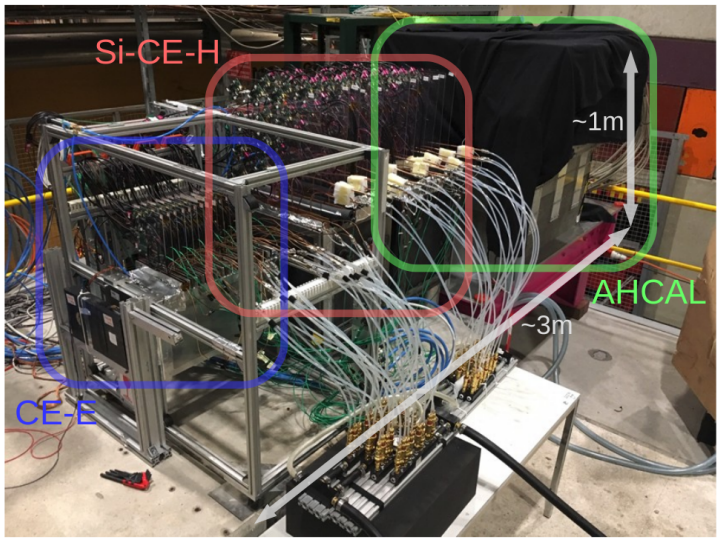


Figure 3.2: HGCAL-TB setup, October-2018 [5].

3.1.1 CE-E prototype

The CE-E section consists of 14 mini-cassettes, each with two silicon sensor modules. One such mini-cassette is shown in Figure 3.3 (left). This totals about 28 sampling layers, each with one silicon module. In the CE-E section, each sampling layer will have either Cu+CuW or Pb absorber in an alternative fashion and the physical lengths of Cu+CuW and Pb are such that there is a depth of $1X_0$ between each active layer. The CE-E section covers a lateral depth of $\sim 27.8\ X_0$ and $\sim 1.4\ \lambda_{int}$. The area of each silicon sensor module is 12×14 cm^2 and is divided into 128 independently readout channels.

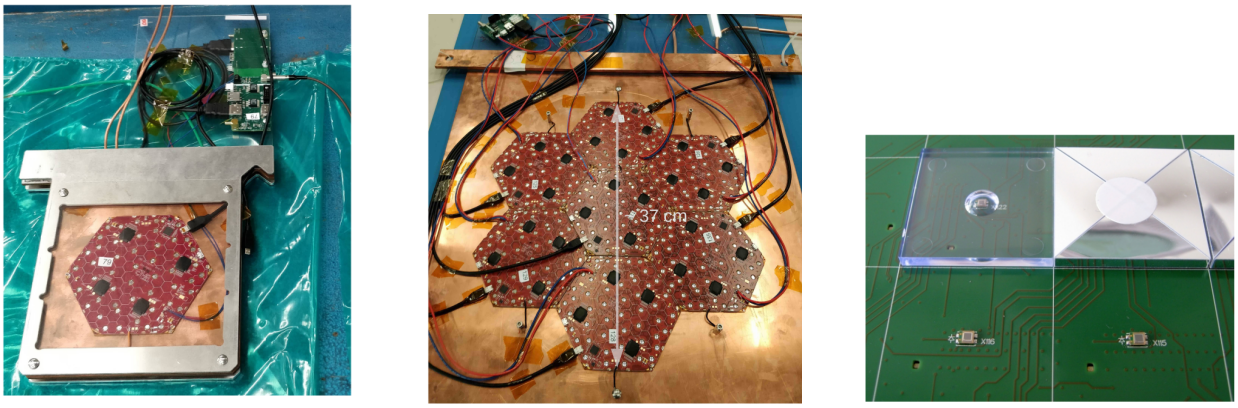


Figure 3.3: An image of mini-cassette of CE-E section (left), a CE-H layer with seven modules arranged in a daisy-like structure (middle), AHCAL layer showing scintillator mounted on the SiPM [6].

3.1.2 CE-H prototype

The CE-H follows the CE-E prototype and is comprised of 12 sampling layers with 4 cm thick steel absorbers in between each sensitive layer. The first 9 sampling layers consist of 7 silicon sensor modules per layer (896 channels per layer) arranged in a daisy-like structure, as shown in Figure 3.3 (middle), which increases the transverse coverage. The last three layers of CE-H have only one silicon sensor module per layer (128 channels per layer). The transverse coverage of the first 9 layers is about 37×35 cm^2 and for the last 3 layers is 12×14 cm^2. The depth of the CE-H section is $\sim 37.6\ X_0$ and $\sim 3.8\ \lambda_{int}$.

3.1.3 AHCAL prototype

The AHCAL prototype is placed downstream of CE-H. There are 39 sampling layers with 17.7 mm thick steel absorbers in between each sensitive layer. The active material for this section is the plastic scintillator directly read out by the silicon photomultiplier (SiPM). Each layer has 576 3×3 cm^2 scintillator tile mounted on the SiPM as shown in Figure 3.3 (right). The AHCAL section has a transverse coverage of 72×72 cm^2 and longitudinal depth of $\sim 41.3\ X_0$ and $\sim 4.3\ \lambda_{int}$.

The Tabel 3.1 shows the depth of each section of the HGCAL in terms of X_0 and λ_{int}. The CE-E and CE-H combined have approximately 12k channels and the AHCAL section has approximately 22k channels.

Section	Total X_0	Total λ_{int}
CE-E	$\sim$27.8	$\sim$1.4
CE-H	$\sim$37.6	$\sim$3.8
AHCAL	$\sim$41.3	$\sim$4.3

Table 3.1: Depth of different sections of the HGCAL prototype in terms of X_0 and λ_{int}.

3.1.4 Source of particle beams

The single particle beams provided at the CERN test beam site are obtained by making the high energy proton beam of momentum 400 GeV/c from Super Proton Synchrotron (SPS) interact with a 500 mm thick beryllium target. The secondary muons, electrons and pions generated in the interaction are separated according to the required momentum and are used as a particle beam source for the experiment [6]. Several particle detectors, such as delay wire chambers (DWC) and plastic scintillators, were also placed upstream of the HGCAL prototype to study the purity, timing and direction of the incoming beam of particles.

3.1.5 MIP calibration

The signals obtained from different sensitive cells of the detector may vary for the same ionizing particle traversing through it [7]. The response from the different sensitive cells is equalized using the minimum ionizing particles (MIPs). For this purpose muon beam of momentum, 200 GeV/c is used and this method is referred to as MIP calibration. The signals obtained from the sensitive cells are digitized using an analog-to-digital (ADC) converter. The ADC spectrum of a muon is obtained for each cell and is fitted with a Landau distribution convoluted with a Gaussian distribution (refer to Figure 3.4) [6]. The most probable value (MPV) of the fitted distribution is extracted and the ADC spectrum is normalized such that the MPV of the distribution is at 1 MIP for each cell. The energy measured in each of the sensitive cells in units of MIPs is referred to as the *reconstructed hits* or the *rechits*. Since the physical location of the sensitive cells is known, rechits also have the spatial information of the shower particle passing through it.

3.2 Test beam simulation setup

The HGCAL test beam setup is simulated using the GEANT4 software in the CMSSW framework. The pions (π^-) are shot from 33m upstream of the HGCAL prototype and the particle interaction

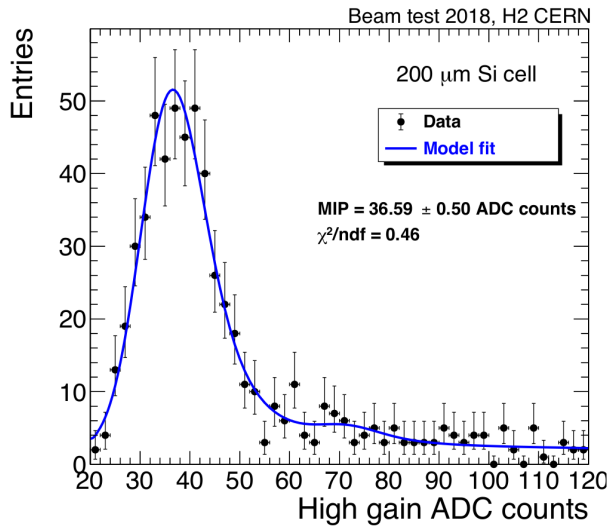


Figure 3.4: The ADC spectrum of the muon in a 200μm thick silicon cell fitted with a Landau distribution convoluted with a Gaussian distribution [7].

with the matter is simulated using the GEANT4 physics list FTFP_BERT_EMN. The simulation of the active layers and absorber materials of different sections of the HGCAL prototype is shown in Figure 3.5. In the simulation, the energy spectrum of the 200 GeV/c muons in each cell is also fitted with a Landau distribution convoluted with a Gaussian distribution. The MPV of the fitted distribution is used to convert the energy measured to the number of MIPs. To account for the electronic noise, the MIP spectra is smeared with a gaussian distribution in simulation [6].

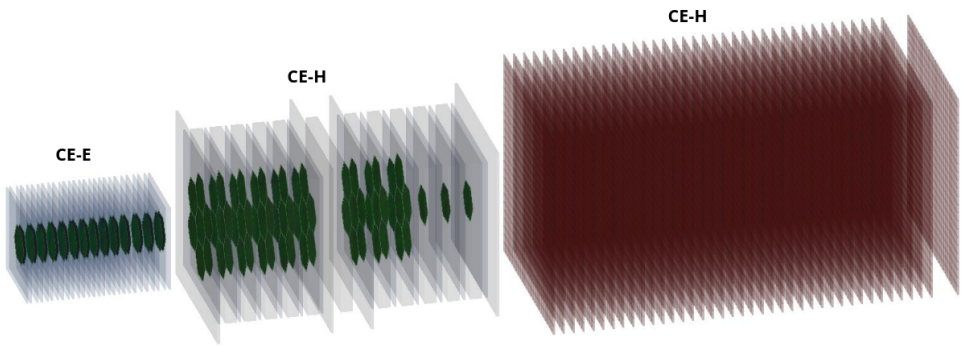


Figure 3.5: A simulation of the HGCAL prototype showing the active layers and the absorber material [6].

3.3 Full HGCAL simulation setup

The full HGCAL setup is simulated using a standalone GEANT4 simulation framework [18]. Unlike the proposed design of the full HGCAL detector, this simulation framework uses 26, 11 and 10 sampling layers in the CE-E, CE-H-fine and CE-H-coarse sections, respectively. But the detector depth in terms of X_0 and λ_{int} in each of the sections is maintained. More details can be found here [19]. The pion (π^-) gun is fixed at η=2.0 and ϕ ranges from 0-2π. The particle interaction with the matter is simulated using the GEANT4 physics list QGSP_BERT.

3.4 Particles created in pion showers

Given the complexity and diverse nature of hadronic showers, we wanted to study the various particles and their kinematics to understand their behavior as a function of incident energy. π^0 is particularly interesting as the difference in response of the detector to the electromagnetic and hadronic component of the hadronic shower is a key component driving the resolution. We tried to study some of these details to understand the results we obtained using advanced machine learning techniques like graph neural networks.

As mentioned in the section 2.6, GEANT4 allows the user to save information regarding the secondary particles created while tracking a specific step. Starting with the simulation setup provided to me by more experienced collaborators, I learned the GEANT4 simulation machinery and its key concepts in the context of interaction processes, and added my contributions to the existing standalone full HGCAL simulation setup to extract and save the information regarding the secondary particles created in the pion showers. I would like to thank Dr. Pedro Ferreira da Silva for discussing the necessary addition to the simulation code. The information on the secondary particles can be accessed through the secondary list of the GEANT4 track vector. The details on extracting the secondary particle information can be found on GitHub [20]. Approximate 10k pion events with the additional secondary information are saved for the energy points [20, 50, 80, 100, 150, 200, 250, 300] GeV.

3.4.1 Particles produced in the first hadronic interaction of pion

The particles produced in the first hadronic interaction are separated by specifying the parent ID of the secondary particles to be less than or equal to zero and the creator process to be a hadronic

interaction, i.e., `HadInelastic`, `HadAtRest`, `HadFission` or `HadCapture`. If this condition is satisfied, the interaction z value of the first hadronic interaction, the particle ID and the kinetic energy of the secondary particles are saved. It should be noted that the information on secondary particles created in the first hadronic interaction is specific to the physics list (QGSP_BERT) used for the simulation.

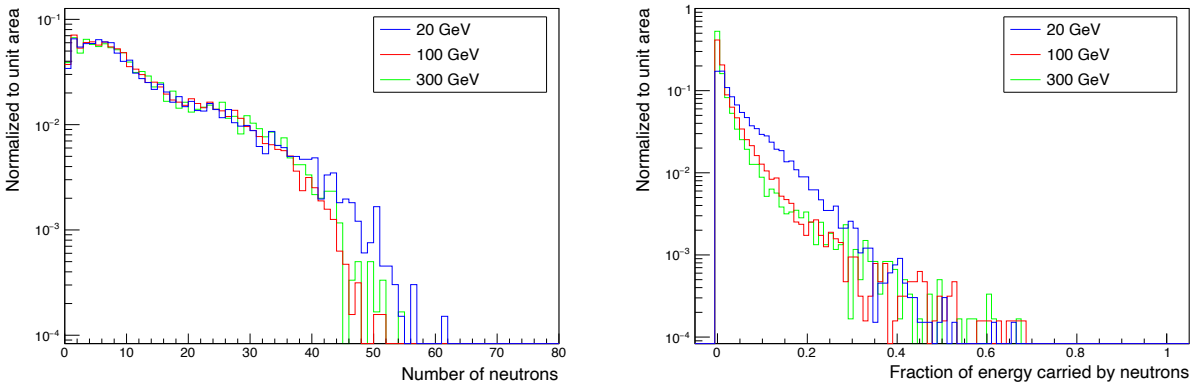
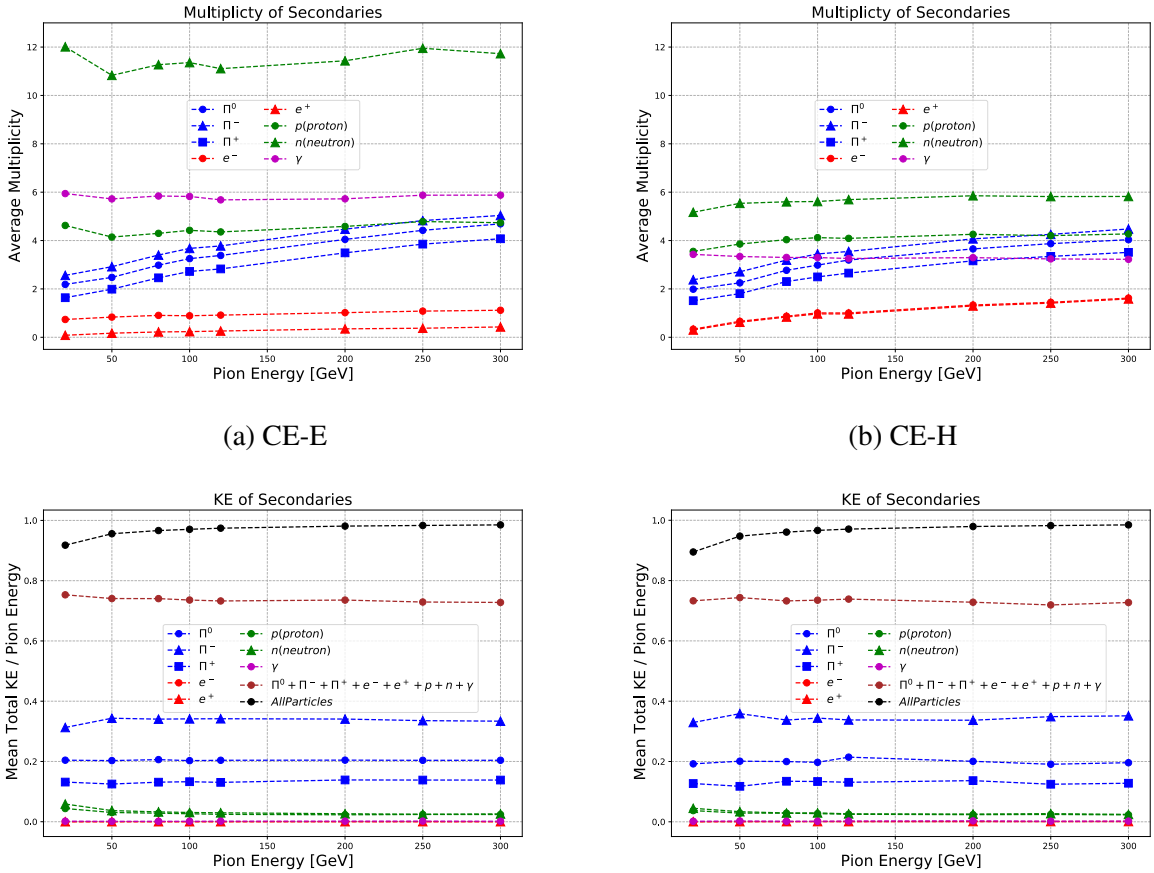


Figure 3.6: Distribution of the number (left) and the fraction of energy carried by the neutrons (right) created in the first hadronic interaction occurring in CE-E section as predicted by QGSP_BERT hadronic physics list.

Figure 3.6 shows the distribution of the number (left) and the fraction of energy carried by the neutrons (right) created in the first hadronic interaction occurring in the CE-E section. A sample mean is used to study the multiplicity and the kinematics of various particles created in the first hadronic interaction. Figure 3.7 shows the average particle multiplicity and average fractional energy carried by the secondary particles created in the first hadronic interaction occurring in CE-E and CE-H section. The pion events starting shower in the CE-E and CE-H are separated using the true information of interaction z value of first hadronic interaction. The number of π^-, π^+, π^0 produced increases with energy and they carry the most fraction of the energy of the incident pion. When there is a soft interaction of the pion, there is a very less number of secondaries produced and the momentum transfer to the secondaries is also less. These soft interactions are recorded as the first hadronic interaction and the incident π^- cannot be differentiated from other secondary particles in the GEANT4 simulation. Since the soft interacted π^- is also present in the secondary list, the average energy carried by the π^- is relatively higher than the π^+ and π^0. The number of protons and neutrons produced in the first hadronic interaction is high, but they carry very less energy. The average number of neutrons and protons created in the first hadronic interaction in the CE-E section differs largely. This is mainly because of the lead absorber in the CE-E section, for

which the probability of emitting a proton from an excited nucleus is significantly less due to the high coulomb barrier. For the steel absorber in the CE-H section, the coulomb barrier is lower and hence, the probability of emitting protons and neutrons from the excited nucleus is similar [2]. The average fractional energy carried by all particles is not equal to one as some of the energy of the incoming pion is utilized by the particles emitted during the nuclear spallation process to overcome the nuclear binding energy (invisible energy).



(a) CE-E

(b) CE-H

(c) CE-E

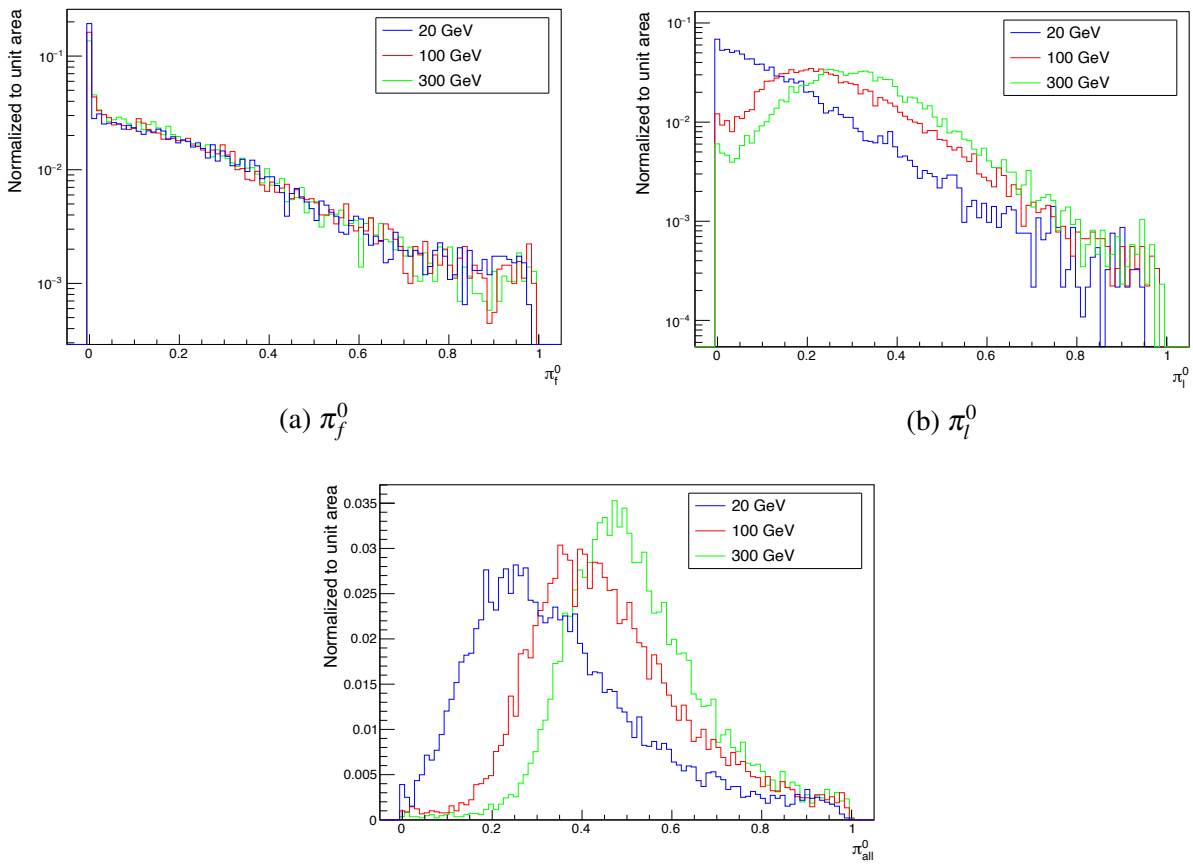
(d) CE-H

Figure 3.7: Average particle multiplicity (top row) and the average energy carried by the secondary particles (bottom row) created in the first hadronic interaction occurring in CE-E section (left column) and CE-H section (right column) as predicted by QGSP_BERT hadronic physics list.

3.4.2 π^0 produced in pion showers

The position and the kinetic energy of the π^0 are saved irrespective of where it was created in the pion shower and the creator process of the π^0 is restricted to be hadronic interactions or `Decay`. Figure 3.8 shows the fraction of energy carried by π^0 in first hadronic interaction (π^0_f), in the later part of the shower (π^0_l) and all π^0 (π^0_{all}) for different incident pion energy. The subscript f and l stand for 'first' and 'late', respectively. The distribution of π^0_f is similar, having a mean (sample mean) value of 20% (Figure 3.9) for all pion energy. With the increase in incident pion energy, the mean π^0_l increases as the later shower particles carry more energy. Hence, we see an overall π^0_{all} increase with an increase in incident pion energy as contributions of π^0_l increase. In Figure 3.10 we see that for a given pion energy, as the value of π^0_f increases, the value of π^0_l decreases. When π^0_f is higher, less energy is available to the rest of the shower particles. Hence, the probability of producing π^0 and the energy they carry will be less.

The $\pi^0 \rightarrow \gamma\gamma$ leads to an electromagnetic shower and the energy it carries is completely converted into a detectable signal. Figure 3.8c shows that the fraction of energy carried by π^0 highly fluctuates shower-to-shower for the same incident pion energy. Also, Figure 3.9 shows that the fraction of energy carried by π^0 increases on average, with an increase in incident pion energy. Along with these fluctuations, fluctuations also come from the particle multiplicity of the pion shower and the invisible energy. These fluctuations directly affect the signals obtained from the detector and hence, the energy resolution of the pion showers is always worse than the pure electromagnetic shower of the same energy.



(a) π^0_f

(b) π^0_l

(c) π^0_{all}

Figure 3.8: The fraction of energy carried by π^0 in first hadronic interaction (π^0_f), in the later part of the shower (π^0_l) and all π^0 (π^0_{all}) as predicted by QGSP_BERT hadronic physics list.

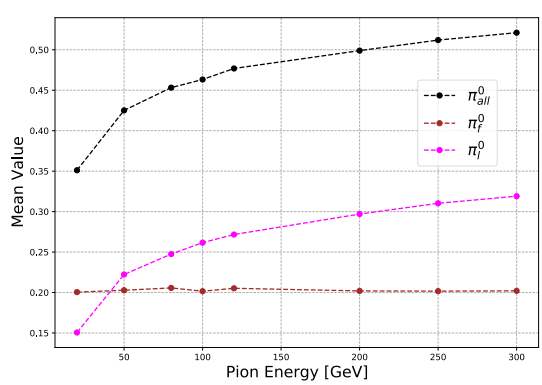
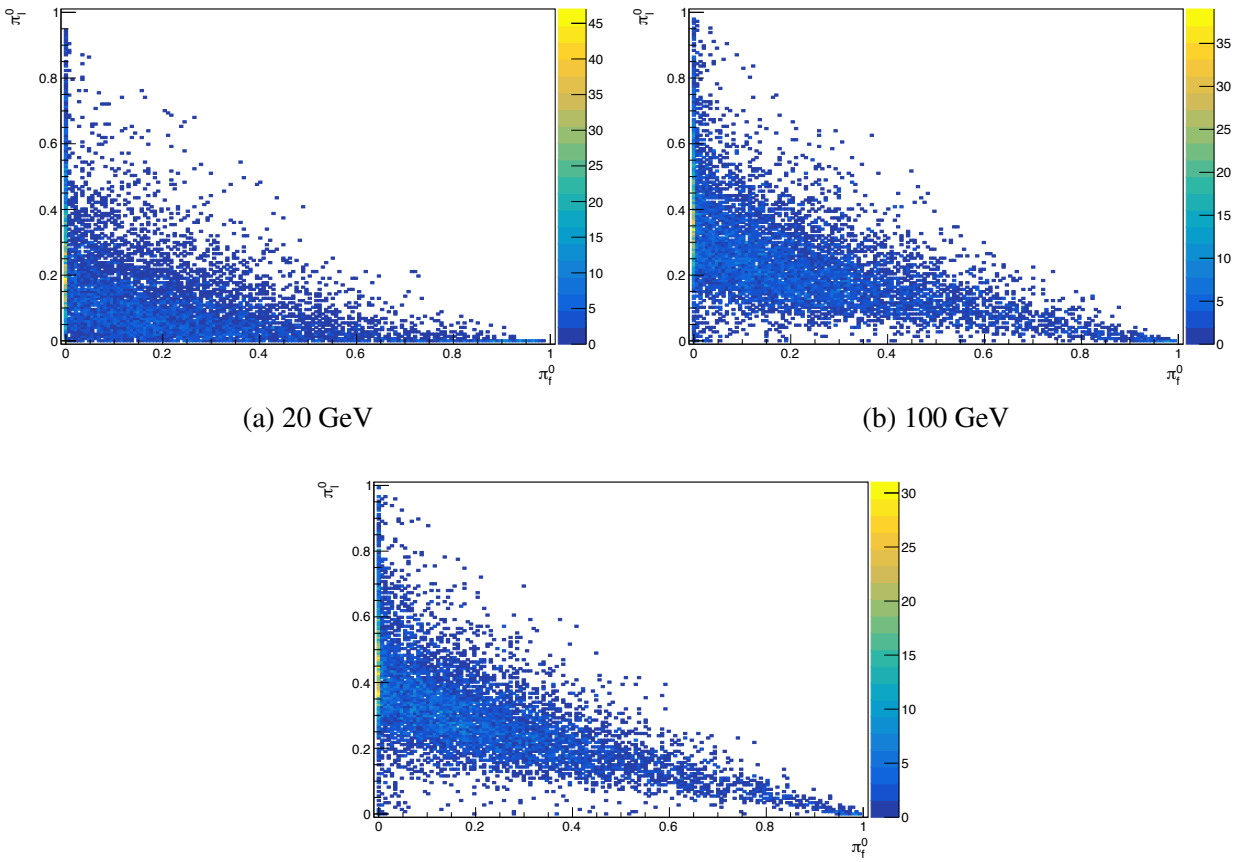


Figure 3.9: Mean (sample mean) of π^0_f, π^0_l and π^0_{all} as a function of incident pion energy as predicted by QGSP_BERT hadronic physics list.



(a) 20 GeV

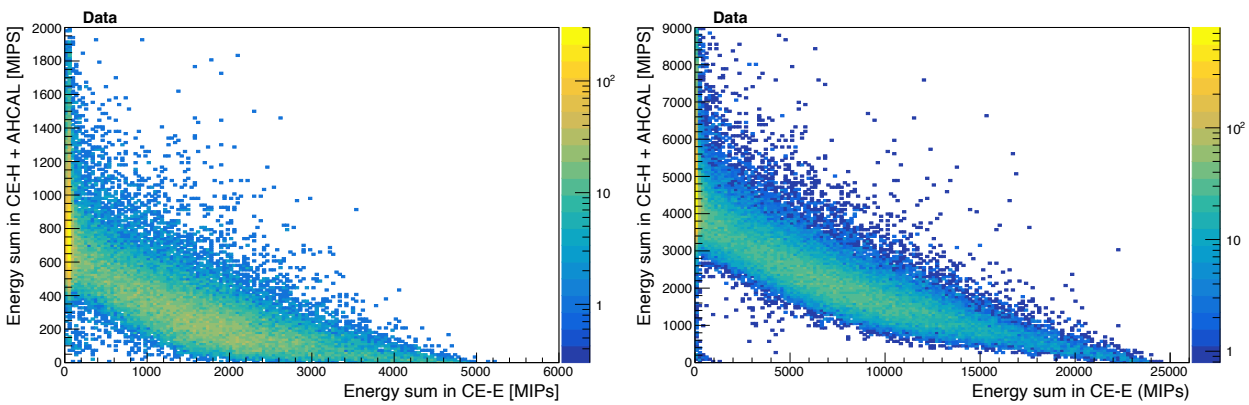
(b) 100 GeV

(c) 300 GeV

Figure 3.10: The fraction of energy carried by π^0 in first hadronic interaction (π_f^0) versus the fraction of energy carried by π^0 in the later part of the shower (π_l^0) as predicted by QGSP_BERT hadronic physics list.

3.5 Measurement of energy of charged pions

Due to the difference in absorber material and sampling fraction in different sections of the detector, the energy measured in units of MIPs does not correspond to the same energy deposited by the particle. Figure 3.11 shows the energy measured (MIP units) in CE-E versus the CE-H + AHCAL section for pion showers in the HGCAL prototype. If we look at the plot for 50 GeV pion, when the CE-E energy is close to zero (pion is MIP like in CE-E) the CE-H + AHCAL energy peaks around 700 MIPs. But when the CE-H + AHCAL energy is close to zero the CE-E energy peaks around 3500 MIPs. In order to calibrate the different sections of the prototype, the energy measured in units of MIPs is converted to GeV. This method is referred to as detector level calibration and the MIP-to-GeV conversion scales obtained for each section are referred to as fixed weights in the rest of the thesis. The details explained below are from previous studies on pion beams in the HGCAL prototype and are referred from the pion paper [6]. The results are used as a published reference for the results using ML techniques presented in this thesis.



(a) 50 GeV Pion

(b) 250 GeV Pion

Figure 3.11: Sum of energy measured (MIP units) in CE-E versus CE-H+AHCAL.

Let E_{MIPs}^{CE-E}, E_{MIPs}^{CE-H} and E_{MIPs}^{AHCAL} be the energy measure in units of MIPs in CE-E, CE-H and AHCAL section, respectively. Then the energy of the pion in units of GeV can be given by:

$$E_{fix}(GeV) = \alpha^{fix} \times E_{MIPs}^{CE-E} + \beta^{fix} \times (E_{MIPs}^{CE-H} + \delta^{fix} \times E_{MIPs}^{AHCAL}) \tag{3.1}$$

where α^{fix} and β^{fix} are the MIP-to-GeV conversion factors for electromagnetic and hadronic compartments, respectively and δ^{fix} is the relative weight factor between CE-H and AHCAL section. These coefficients are obtained by minimizing the relative resolution calculated as σ/μ of the

gaussian fit. The 50 GeV positron beam test data is used to find the α^{fix} and has an approximate value of 10.5 MeV per MIP. The 50 GeV pions, which are MIP-like in the CE-E section, are used to obtain the values of β^{fix} and δ^{fix}. The value of δ^{fix} is 0.4 and β^{fix} is approximately 80 MeV per MIP.

The MIP-to-GeV conversion scales are further optimized as a function of incident pion energy and separately for pions starting shower in the CE-E and the CE-H sections [6] [7]. The reconstructed energy for pions starting shower in the CE-E and the CE-H sections are given by Equation 3.2 and 3.3, respectively. In Equation 3.2 and 3.3, α_1, $\beta_{1,2}$ and $\gamma_{1,2}$ are the optimized parameters obtained by minimizing the χ^2 (Equation 3.4) for each beam energy (E_{beam}). The E_{fix}^{CE-E}, E_{fix}^{CE-H} and E_{fix}^{AHCAL} refer to the energy calculated using the fixed weights for different sections. This method is referred to as the χ^2 method in the rest of the thesis.

$$E_{\chi^2}(GeV) = \alpha_1(E_{beam}) \times E_{fix}^{CE-E} \; + \; \beta_1(E_{beam}) \times E_{fix}^{CE-H} \; + \; \gamma_1(E_{beam}) \times E_{fix}^{AHCAL} \tag{3.2}$$

$$E_{\chi^2}(GeV) = E_{fix}^{CE-E} \; + \; \beta_2(E_{beam}) \times E_{fix}^{CE-H} \; + \; \gamma_2(E_{beam}) \times E_{fix}^{AHCAL} \tag{3.3}$$

$$\chi^2 = \sum_i \frac{(E_{beam} - E_{\chi^2}^i)^2}{\sigma^2(E_{fix}^i)} \tag{3.4}$$

The methods mentioned above to reconstruct the energy of the pion do not utilize the high granular nature of the HGCAL detector and do not account for shower-to-shower fluctuations in pion showers discussed in the previous section. Further improvements in energy resolution would require much fine-tuning according to incident energies.

Chapter 4

Energy regression of charged pions using semiparametric graph neural networks

As discussed earlier, the high event-to-event fluctuations in the particle multiplicity, the energy carried by electromagnetic component and invisible energy worsen the energy resolution of the pion showers. To improve the resolution, it is necessary to account for the event-to-event fluctuations in the pion shower. There are many advantages of using Graph Neural Networks (GNNs) for reconstructing the energy of the pion in the HGCAL detector. Unlike the method discussed in section 3.5, where the sum of rechit energy in each section of the detector is taken as input, the GNNs can take individual rechit information as input from the detector. In this way, the GNN has access to full rechit information, i.e., the spatial coordinates and the energy, from which it can account for the event-to-event fluctuations in pion showers. Also, graphs are felicitous for representing highly sparse, unstructured, and high-dimensional data. These allow a variable number of inputs and work with a detector of arbitrary geometry. The GNNs also solve problems that arise in traditional techniques like CNNs, i.e., the need for zero-padding or truncation of input data. We will use a GNN-based model called Dynamic Reduction Network (DRN) [21] to reconstruct the energy of pion showers. Before getting into the details of the DRN model, a brief introduction to the basic building blocks of the machine learning algorithms is discussed in the following section.

4.1 Basic building blocks of machine learning algorithm

4.1.1 Fully connected neural network

A fully connected neural network (FCNN), also known as a multilayer perceptron (MLP), comprises a series of fully connected layers, where every node (neuron) in one layer is connected to every node in the next layer. A fully connected layer is a function from R^n to R^m, where n is the number of nodes in the current layer and m is the number of nodes in the next layer. The function mapping features from the current layer to the next layer is defined as:

$$X^{i+1}_{m\times 1} = \sigma(w^i_{m\times n} \cdot X^i_{n\times 1} + b^i_{n\times 1}) \tag{4.1}$$

Here, $X^i_{n\times 1}$ is the column of input features and $X^{i+1}_{n\times 1}$ is the column of output features. The matrices $w^i_{m\times n}$ and $b^i_{n\times 1}$ are generally called weight and bias respectively. The elements of the matrices $w^i_{m\times n}$ and $b^i_{n\times 1}$ are the trainable parameters that are learned during model training. σ is a non-linear activation function that is applied at the end of each layer of FCNN.

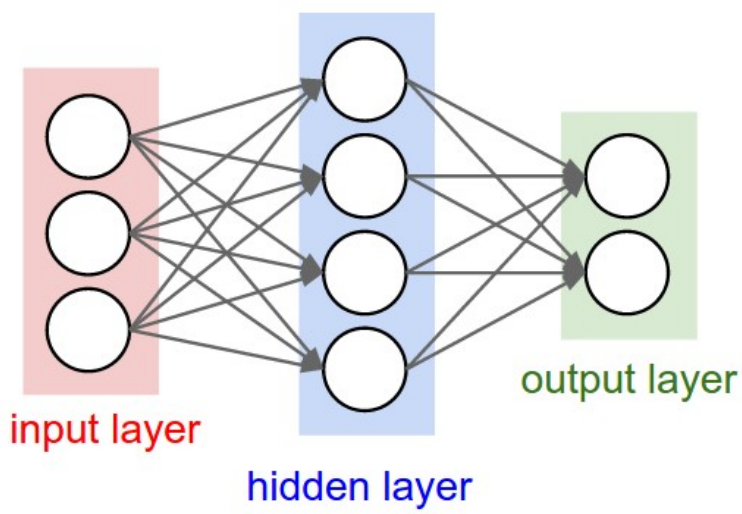


Figure 4.1: A fully connected neural network with 1 hidden layer [8].

Figure 4.1 shows a schematic of a FCNN consisting of one hidden layer, three input features and two output features. The dimension of the hidden layer is 4. The sequence of operations that map the input features to the output features (forward propagation) is as follows:

$$X^1_{4\times1} = \sigma(w^1_{4\times3} \cdot X^{input}_{3\times1} + b^1_{3\times1}) \tag{4.2}$$

$$y' = X^{output}_{2\times1} = \sigma(w^2_{2\times4} \cdot X^1_{4\times1} + b^2_{2\times1}) \tag{4.3}$$

4.1.2 Optimization algorithm

Gradient descent

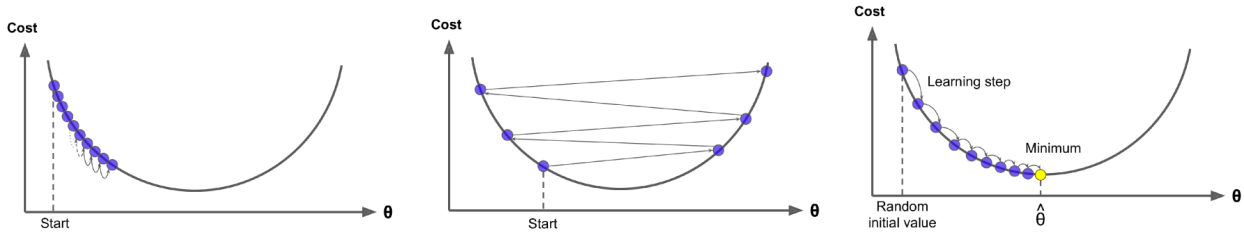
Gradient Descent is an iterative optimization algorithm that is commonly used for training machine learning models and neural networks. The general idea of gradient descent is to correct the trainable parameters iteratively to minimize the cost (loss) function. Here, the cost function is a function of the trainable parameters (θ) and the input features (X^{input}). It measures the difference or error between the true value (y) and the predicted (y') value. The choice of cost function depends on the type of task, for example, χ^2, *mean squared error*, *mean absolute error*, etc., for the regression tasks and *binary cross-entropy*, *categorical cross-entropy*, etc., for classification tasks.

In the gradient descent algorithm, the parameters of the model are iteratively updated according to the equation:

$$\theta_{i+1} = \theta_i - \alpha \cdot \nabla_\theta C(\theta, X^{input}) \tag{4.4}$$

Here, θ_i and θ_{i+1} are the values of parameters at i^{th} and $(i+1)^{th}$ iteration respectively and the gradient of the cost function (C) is taken with respect to the parameters. α is the learning rate determining the step size for each iteration while moving towards the minimum cost function. If the learning rate is too small (Figure 4.2a), the algorithm must go through many iterations to reach the minimum cost function (longer time for convergence). On the other hand, a larger learning rate may cause drastic updates, which leads to divergent behaviors (Figure 4.2b).

Historically, weights of the neural network layers were initialized using small random numbers, but over the years, more specific initialization has been used based on the type of activation function and the number of nodes in the layer for faster convergence. One such method for ReLU activation function is to initialize the weights using a zero-centered gaussian probability distribution with standard deviation $\sqrt{2/n}$. Here, n is the number of nodes in the previous layer. Gradient descent



(a) α is too small (b) α is too large (c) α has an optimum value

Figure 4.2: A depiction of the gradient descent algorithm for different values of learning rate (α) [9].

is a simple algorithm and there exist many other optimizers that help the machine learning model to converge faster. The optimizer used for training the DRN models in this thesis is the AdamW optimizer. In this algorithm, the weight updation at each step is adapted by calculating the first moment (mean) and the second moment (variance) of the gradients. It also uses a learning rate multiplier with a cosine policy, which decreases the weight updation step size after each iteration. Refer [22] for more details on the AdamW optimizer.

4.1.3 Activation functions

The activation function is an essential feature in a neural network, which introduces non-linearity to the neuron's output. The non-linearity helps the model learn complex representations and patterns from the input data that would not be possible using a simple linear network. Some commonly used activation functions are *sigmoid*, *tanh*, *ReLU* and *ELU*. Figure 4.3 shows these activation functions and their derivatives.

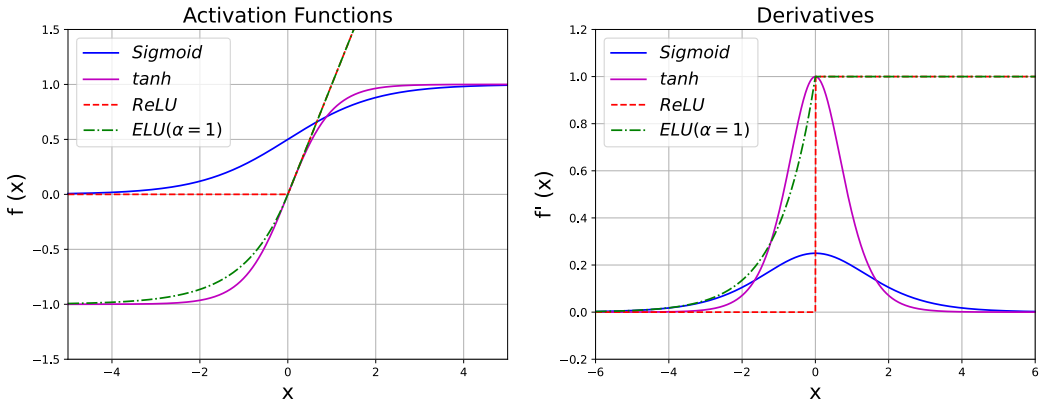


Figure 4.3: Activation functions (left) and their derivatives (right).

- **The sigmoid (σ) function:** The sigmoid activation function takes a real value and outputs a value between 0 and 1. As the output is between 0 and 1, it is specifically used for the models where we want to predict the probability of some quantity as output (e.g., binary classification). When x is very far from zero, the derivative (gradient) is very close to zero. Under these conditions, the model refuses to learn further or is drastically slow. This is known as the vanishing gradient problem.

$$\sigma(x) = \frac{1}{1+e^{-x}} \tag{4.5}$$

- **The hyperbolic tangent (*tanh*) function:** The hyperbolic tangent function is very similar to the sigmoid function, except that the output values range from -1 to 1. The derivative of the tanh function is greater than the sigmoid function and helps the model to learn faster [9] (faster convergence). However, the problem of vanishing gradient still exists in the tanh function.

$$tanh(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}} \tag{4.6}$$

- **The Rectified Linear Unit (*ReLU*) function:** *ReLU* is one of the most used activation functions in machine learning models. It is less computationally expensive [9] than the tanh and sigmoid, which requires computing exponents. This has a huge impact on reducing training time for big and more complex networks. *ReLU* also solves the problem of vanishing gradients for positive values of x, as the gradient is always greater than zero in this region. But for negative values of x, the gradient is zero and the neurons may stop learning.

$$ReLU(x) = max(0,x) \tag{4.7}$$

- **The Exponential Linear Unit (*ELU*) function:** *ELU* is very similiar to *RELU* except for negative inputs. For negative inputs, *ELU* becomes smooth slowly until its output is equal to $-\alpha$, whereas *ReLU* sharply goes to zero. Unlike *ReLU*, *ELU* can produce negative outputs. The hyperparameter α controls the value to which ELU saturates for negative inputs and it has a default value of one in both *TensorFlow* and *PyTorch* libraries.

$$ELU(x,\alpha) = \begin{cases} x, & x > 0 \\ \alpha \cdot (e^{x} - 1), & x <= 0 \end{cases} \tag{4.8}$$

4.2 Dynamic Reduction Network

The DRN architecture was derived in the context of the energy regression of hadrons in high energy physics [21] and a similar model has been used for the energy regression of electrons and photons in the current CMS electromagnetic calorimeter (ECAL) [23]. Pions are complex objects and such tools are expected to improve energy resolution as indicated in [24]. As mentioned in the section 3.1.5, the signals obtained from the HGCAL detector are in the form of rechits. Rechits have the information of where a shower particle passed through (position information) and the corresponding energy deposited by the particle in units of MIPs. Figure 4.4 shows a schematic of the DRN. The information flow from rechits to the required target [21] is as follows:

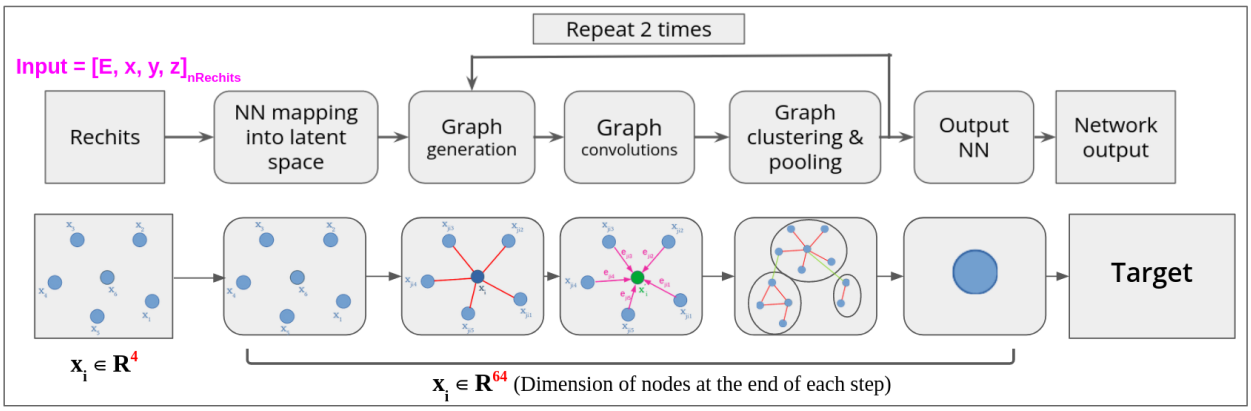


Figure 4.4: A schematic of dynamic reduction network.

1. The n number of rechits ([E,x,y,z]$_n$) are normalized by fixed values such that the input features are in the range $[0,1]$.

2. The normalized 4-dimensional input nodes (rechits) are mapped to 64-dimensional latent space using a fully connected neural network. This fully connected neural network has three layers with *ELU* activation function, each with 64 output features.

3. A graph is created by connecting the 16 nearest neighboring nodes in the 64-dimensional latent space. A self-loop is present during the graph creation, meaning every node is connected to itself and 15 other nearest neighbors. The constructed graph can be represented as $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \{1,...,n\}$ is the set of vertices and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ is the set of edges.

4. An edge convolution step [25] processes the graph where the messages from the neighboring nodes are calculated according to Equation 4.9 and aggregated according to Equation 4.10.

In the Equation 4.9, $\Theta = (\theta_1, ...,_M, \phi_1, ..., \phi_M)$ are the weights for M different filters (outputs). Here, each θ_m, ϕ_m has the same dimension as x (node features), and $\cdot$ represents the Euclidean inner product. There are two such edge convolution layers, with the first and the second layer having $M = 128$ and $M = 64$ outputs, respectively. In this way, the information from neighboring nodes is aggregated and the features of each node are updated.

$$e'_{ijm} = ELU(\theta_m \cdot (\mathrm{x}_j - \mathrm{x}_i) + \phi_m \cdot \mathrm{x}_i) \tag{4.9}$$

$$x'_{im} = \underset{j:(i,j)\in\mathcal{E}}{\mathrm{add}} e'_{ijm} \tag{4.10}$$

5. The nodes with updated features are weighted by distance and clustered using a greedy clustering algorithm [26]. Then the node features are pooled by taking the maximum of the clustered node.

6. The resulting set of nodes is passed again to the graph generation, edge convolution, clustering and pooling step.

7. The reduced set of nodes is globally max pooled and passed through a fully connected neural network with *ELU* activation function. The output network has two layers, the first layer has 64 output features and the number of output features of the second layer corresponds to the dimension of the required target (one if the target is the pion energy).

Using the dynamic graph convolution, graph clustering and the pooling step, the amount of information passed onto the next step is reduced in an optimal way to maintain essential relations between the data. Hence, the network is called a dynamic reduction network (DRN).

4.3 Training

4.3.1 Simulation samples used for training

As mentioned in the section 3.3, the HGCAL test beam setup is simulated using the GEANT4 software in the CMSSW framework. The simulated information of the detector signal was saved for approximately 5M pion events with energy ranging from 10–350 GeV following a uniform distribution. Before using the simulation samples for training, pion events are selected using various

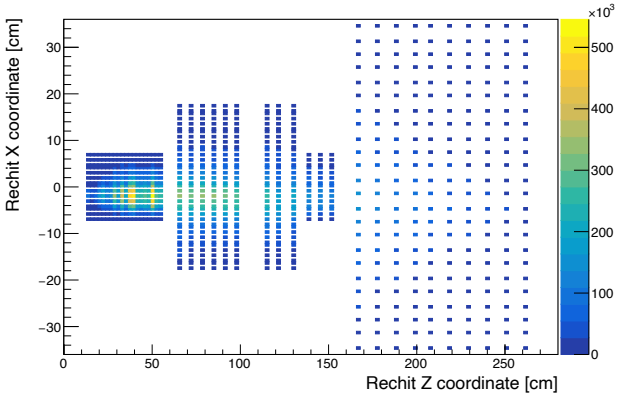
selection criteria to maintain the purity of the pion samples. These studies were previously done by Dr. Shubham Pandey and can be referred to the pion paper [6] for more details. While the samples used for studies presented in the previous chapter were generated by me, the cleaned pion samples for data and respective simulation were provided to me by Alpana Sirohi. The selection criteria include removing the rechit signals from faulty channels, rejecting muon-like events, rejecting pre-showering pions, rejecting pion events that are impinging far away from the HGCAL prototype, etc. After applying these selection criteria, we are left with 4.1M pion samples to train the DRN model. These cleaning cuts are applied to both data and simulation samples to make consistent comparisons.

4.3.2 Input features

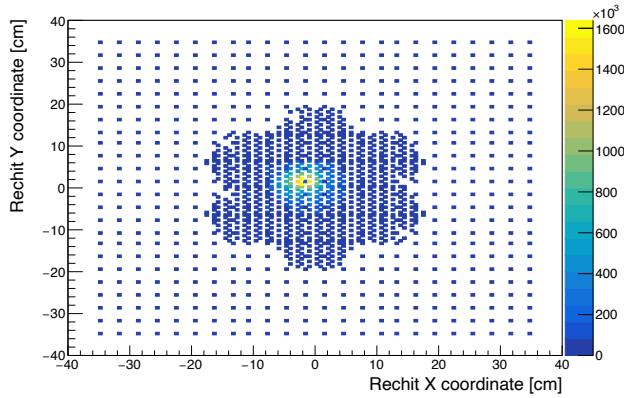
The position information of the rechits (x,y,z) and energy are the inputs to the model. Figure 4.5 shows the distributions of the input features and the number of rechits. For the AHCAL section, the rechit information from only 10 out of 39 layers is considered for input rechits. The layers are selected in such a way as to match the full HGCAL detector configuration, i.e., every fourth layer of the AHCAL section, starting from the third layer. The rechit energy is weighted using the fixed weights obtained from detector level calibration as mentioned in section 3.5 (the fixed weights are recalculated after considering only ten layers of AHCAL section).

4.3.3 Target

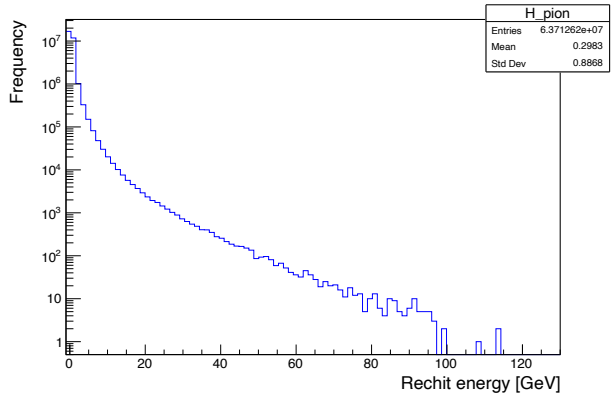
In the semiparametric approach, instead of the model predicting a single value of the required target, the model predicts the distribution shapes of the target. This is done by predicting the 6-dimensional parameters of the DSCB function for the target. The DSCB function helps to better model the response of the pion showers in the detector. The semiparametric DRN model is trained for two different types of targets, i.e., the true beam energy (E_{true}) and the ratio of true beam energy to the energy calibrated using the detector level calibration ($\frac{E_{true}}{E_{raw}}$). Here, E_{raw} is the same as the E_{fix} as mentioned in the section 3.5. Figure 4.6 shows the distribution of two different targets. The E_{true} target has sharp edges at the extreme ends of the distribution, whereas the edges are smoothened out in the $\frac{E_{true}}{E_{raw}}$ target distribution. A non-semiparametric version of the DRN model for pion energy regression has been studied earlier and the results can be found in [24].



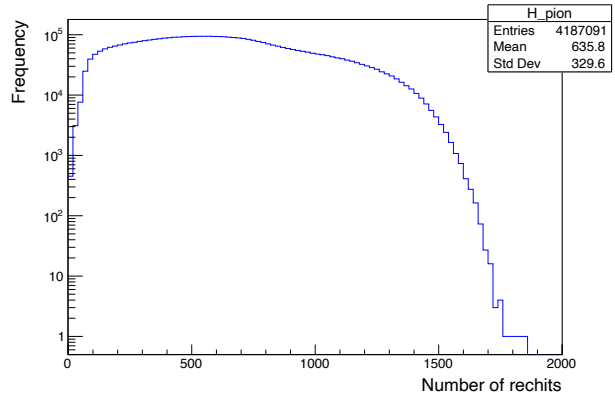
(a) Rechit X versus Rechit Z



(b) Rechit Y versus Rechit X

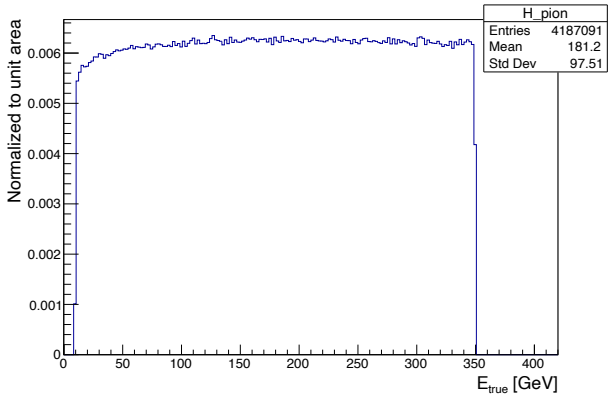


(c) Rechit energy in GeV

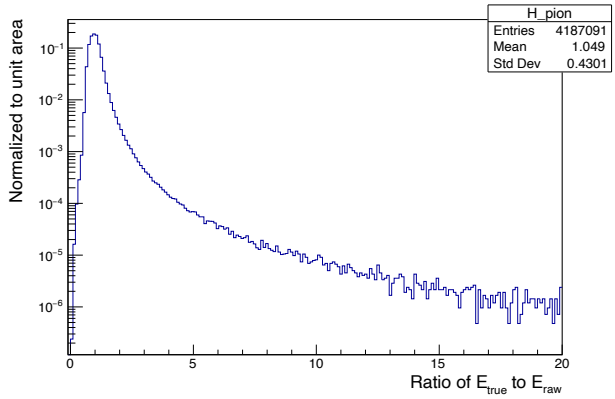


(d) Number of rechits

Figure 4.5: Distributions of input features and the number of rechits.



(a) True beam energy (E_{true})



(b) Ratio of E_{true} to E_{raw}

Figure 4.6: Distribution of different targets used for training: E_{true} (left) and $\frac{E_{true}}{E_{raw}}$ (right).

The DSCB function is as follows:

$$DSCB(x;\mu,\sigma,\alpha_L,n_L,\alpha_R,n_R,) = \begin{cases} Ne^{-\frac{\alpha_L^2}{2}}\frac{\alpha_L}{n_L}(\frac{n_L}{\alpha_L}-\alpha_L-\frac{x-\mu}{\sigma})^{-n_L} & \frac{x-\mu}{\sigma} \leq -\alpha_L \\ Ne^{-\frac{(x-\mu)^2}{2\sigma^2}} & -\alpha_L \leq \frac{x-\mu}{\sigma} \leq \alpha_R \\ Ne^{-\frac{\alpha_R^2}{2}}\frac{\alpha_R}{n_R}(\frac{n_R}{\alpha_R}-\alpha_R+\frac{x-\mu}{\sigma})^{-n_R} & \alpha_R \leq \frac{x-\mu}{\sigma} \end{cases} \tag{4.11}$$

where N is the normalizing factor. The parameters α_L and α_R decides the width from the mean (μ) up to which the function should behave gaussian like. The parameters n_L and n_R decide how quickly the tails should reduce. The higher the value of the n_L and n_R, the more quickly the tails will reduce. Figure 4.7 (left) shows the DSCB distribution for different values of parameters.

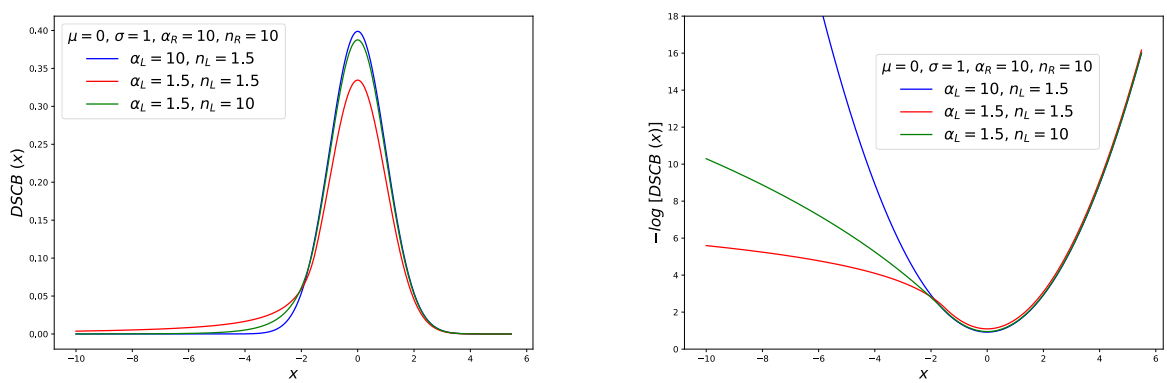
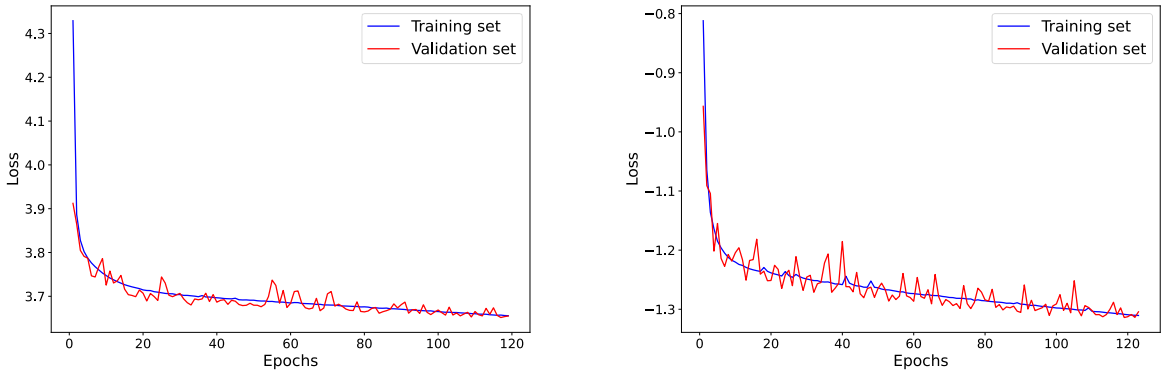


Figure 4.7: The DSCB probability distribution (left) and the negative logarithm of DSCB distribution (right) for different values of parameters.

The loss function for the model is the negative log-likelihood of the DSCB function. Figure 4.7 (right) shows the negative logarithm of DSCB distribution for different values of parameters. The loss for a given target (x') and the predicted DSCB parameters is calculated as $-log[DSCB(x';\mu,\sigma,\alpha_L,n_L,\alpha_R,n_R)]$. The farther the predicted μ is from the target, the more will be the loss value. The loss value also depends on other parameters of the DSCB function.

The model is trained on *NVIDIA V100 GPU* with cyclic learning rate schedule with cosine policy (starting value of $\alpha = 6 \cdot 10^{-4}$) using AdamW optimizer [22]. The above mentioned loss function, GPU configuration, learning rate schedule and optimizer are common to all the DRN



(a) DRN-TrueE

(b) DRN-Ratio

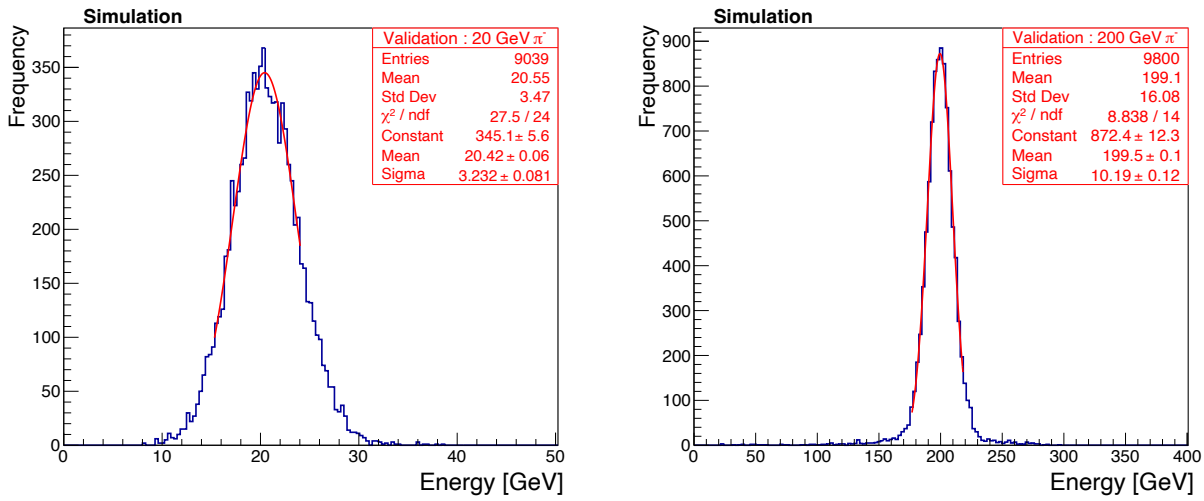
Figure 4.8: Loss versus epoch plot for model trained using E_{true} target (left) and $\frac{E_{true}}{E_{raw}}$ target (right).

models mentioned in the thesis. Out of 4.1M events 80% of the events were used for training and the rest of the events were used for validation. The approximate time taken for training one epoch is 48 minutes. The number of epochs trained for the model with E_{true} [DRN-TrueE] and $\frac{E_{true}}{E_{raw}}$ [DRN-Ratio] target is 120 and 124, respectively. The best epoch of the models is chosen to be the epoch with minimum validation loss, i.e., epochs 117 and 119 for the models DRN-TrueE and DRN-Ratio, respectively. Figure 4.8 shows the loss versus epoch plot for two models. In both the plots, the training and the validation loss overlap each other and decrease till the end. This suggested that there was no overtraining of any of the models.

4.4 DRN performance on Simulation

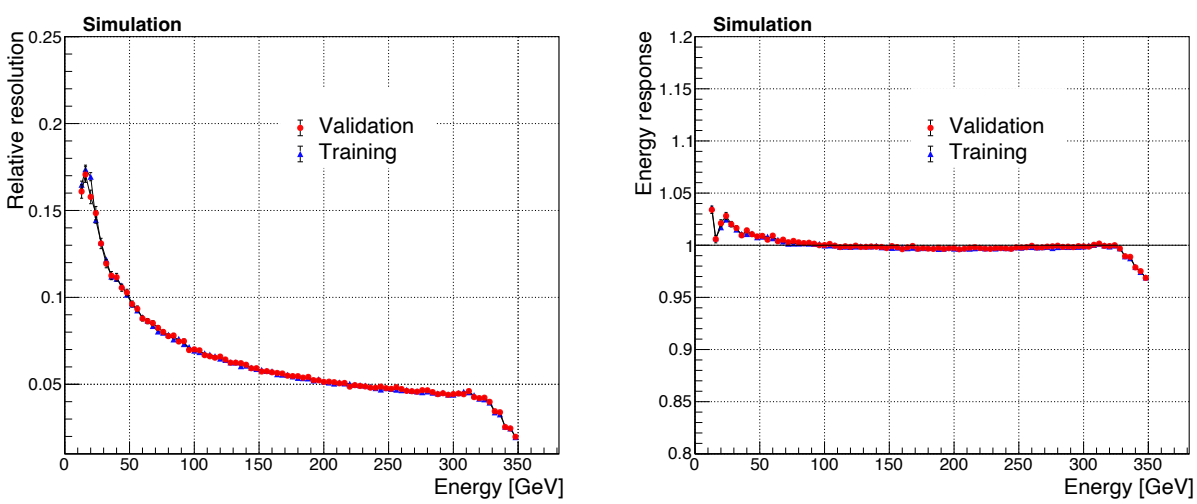
For the DRN-TrueE model, the mean value (μ) out of 6 predicted parameters of the DCSB function will be the predicted pion energy. For the DRN-Ratio model, the μ value multiplied by the E_{raw} will be the predicted pion energy. The performance of the models is measured in terms of the resolution and the response. The flat energy pion samples of 10–350 GeV are divided into 4 GeV energy bins, totaling 85 bins. For each bin, we plot the predicted energy distribution and fit a gaussian function around the core of the distribution. After extracting the fitted gaussian distribution's μ and σ, the relative resolution and energy response are defined as μ/σ and μ/E_{bin} respectively. Here E_{bin} is the central energy point of the corresponding energy bin. Figure 4.9 and 4.11 show the predicted energy distributions (20 GeV and 200 GeV) for the validation dataset and the fitted gaussian function for the models DRN-TrueE and DRN-Ratio, respectively. The

predicted energy distribution with the fitted gaussian function for a few more energy points is in the appendix.



(a) 20 GeV (b) 200 GeV

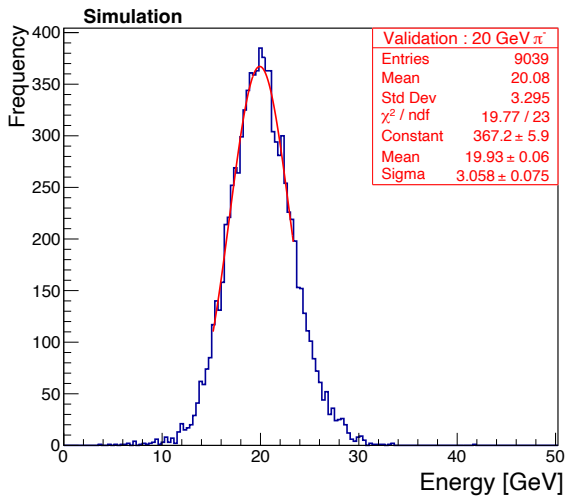
Figure 4.9: The predicted energy distributions and the fitted gaussian function is shown for the prediction made by the DRN-TrueE model on the validation dataset.



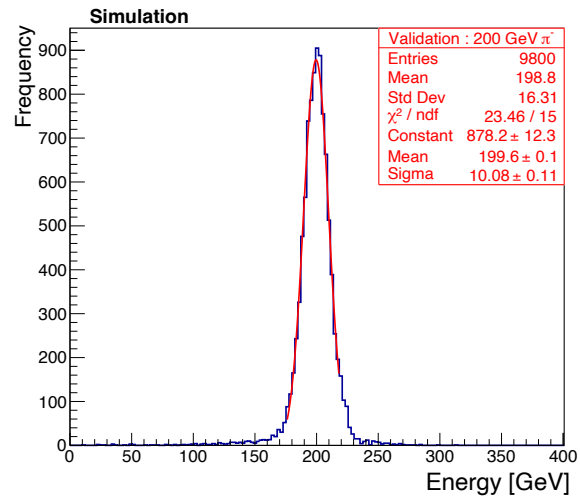
(a) Resolution (b) Response

Figure 4.10: The comparison of resolution (a) and response (b) for the validation and training dataset for the model DRN-TrueE.

Figure 4.10 and 4.12 show the performance in terms of the resolution and the response for the models DRN-TrueE and DRN-Ratio, respectively. For both models, the resolution and response of

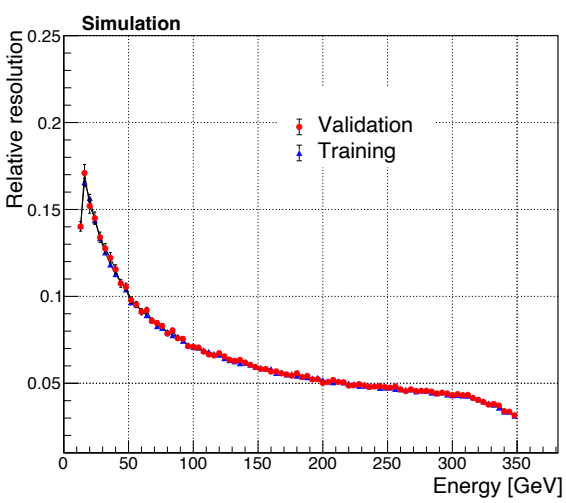


(a) 20 GeV

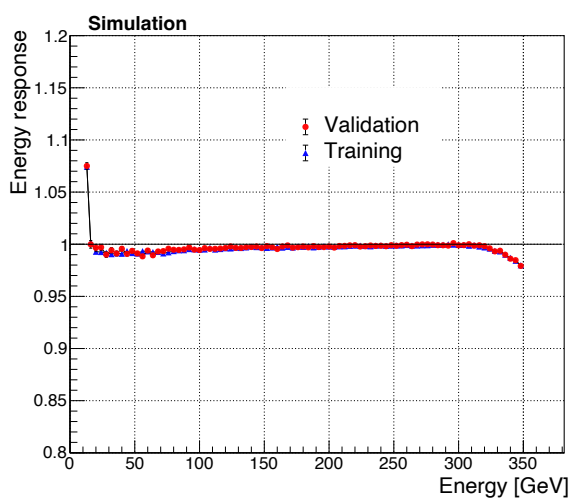


(b) 200 GeV

Figure 4.11: The predicted energy distributions and the fitted gaussian function is shown for the prediction made by the DRN-Ratio model on the validation dataset.



(a) Resolution



(b) Response

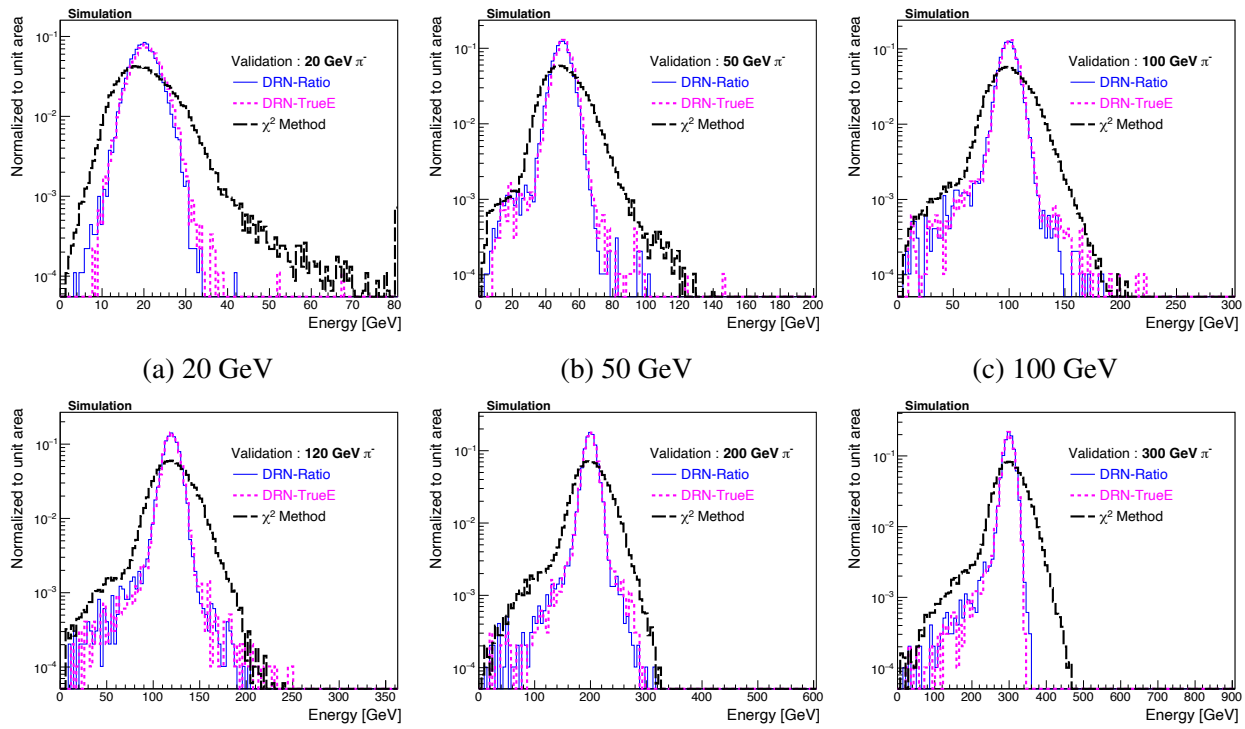
Figure 4.12: The comparison of resolution (a) and response (b) for the validation and training dataset for the model DRN-Ratio.

the training and validation dataset closely match (except for a few energy points below 20 GeV), indicating there is no overtraining.

To check the efficacy of the DRN model, we compare the performance of the model DRN-TrueE and DRN-Ratio with the χ^2 method. Figure 4.13 shows the predicted energy distribution comparison of the validation set for the three models. Compared to the χ^2 method, the DRN models have narrower peaks and significantly reduced tails. Figure 4.14 shows the performance comparison for the three models in terms of resolution and response. The error bars mentioned in the response and resolution plots are only statistical errors and do not include the systematic errors which might come from the pion beam impurity, uncertainty in the pion beam energy, low-level MIP calibration, etc. The bottom panel of Figure 4.14a plots the ratio of resolution values of the χ^2 method (black curve) and DRN-TrueE (magenta curve) with the DRN-Ratio model. There is a large improvement in pion energy resolution using DRN compared to the χ^2 method (resolution is improved by a factor 2). The resolution of the DRN-TrueE and DRN-Ratio are comparable in the energy range of 20–320 GeV. For the DRN models, the resolution seems to deviate from the standard resolution curve after 300 GeV, which is not physical. The response of the DRN-TrueE is linear in the range 100–320 GeV, beyond which it starts to deviate. The response of the DRN-Ratio is linear in the range 100–320 GeV, has a lower response (within 1%) in the range 20–100 GeV and starts to decrease up to 2% for energy greater than 320 GeV.

4.5 DRN performance on HGCAL-TB data

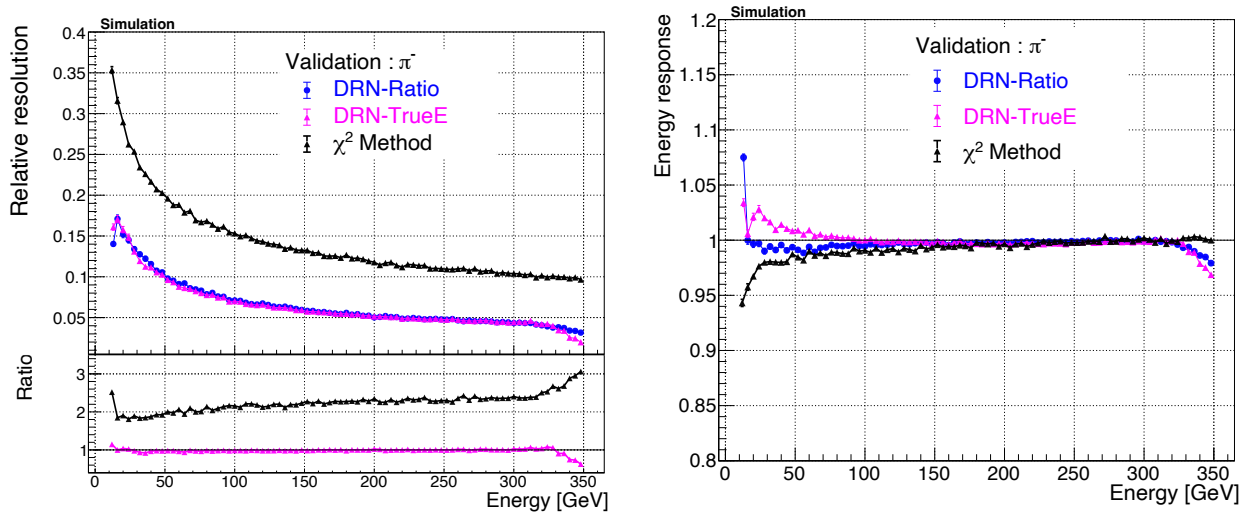
For getting the predictions for data, the rechit energy of the CE-E section is upscaled by 3.5%, and the CE-H/AHCAL section is upscaled by 9.5% to match the energy scale of data and simulation [27][6]. Figure 4.15 and 4.17 shows the predicted energy distribution comparison for the simulation (validation dataset) and test beam data for the DRN-TrueE and DRN-Ratio models, respectively. The resolution and response are calculated similarly as explained in the section 4.4 for eight energy points. There is good agreement between the resolution of data and simulation (within 1.5%) for both the models (refer to Figure 4.16 and 4.18). The difference in response of data and simulation is about 0–5% for the DRN-TrueE model and 0–3% for the DRN-Ratio model.



(a) 20 GeV (b) 50 GeV (c) 100 GeV

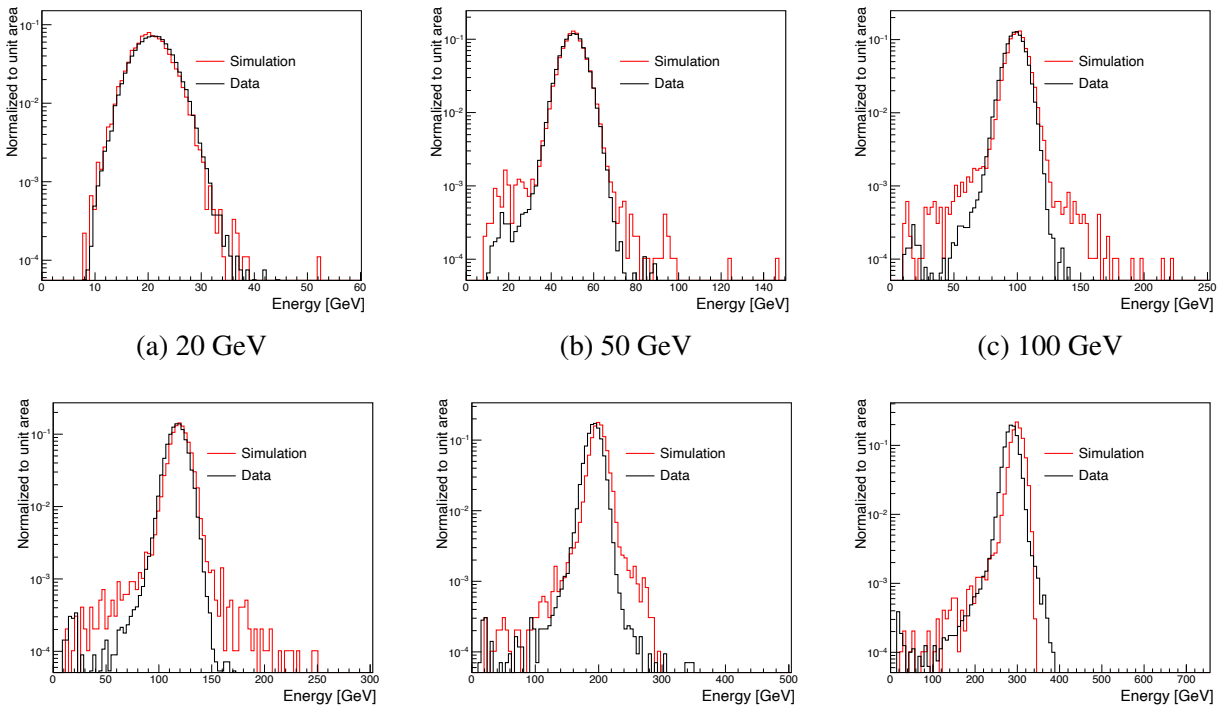
(d) 120 GeV (e) 200 GeV (f) 300 GeV

Figure 4.13: Predicted energy distribution comparison of the validation dataset for three different models: DRN-Ratio (blue curve), DRN-TrueE (magenta curve) and χ^2 Method (black curve).



(a) Resolution (b) Response

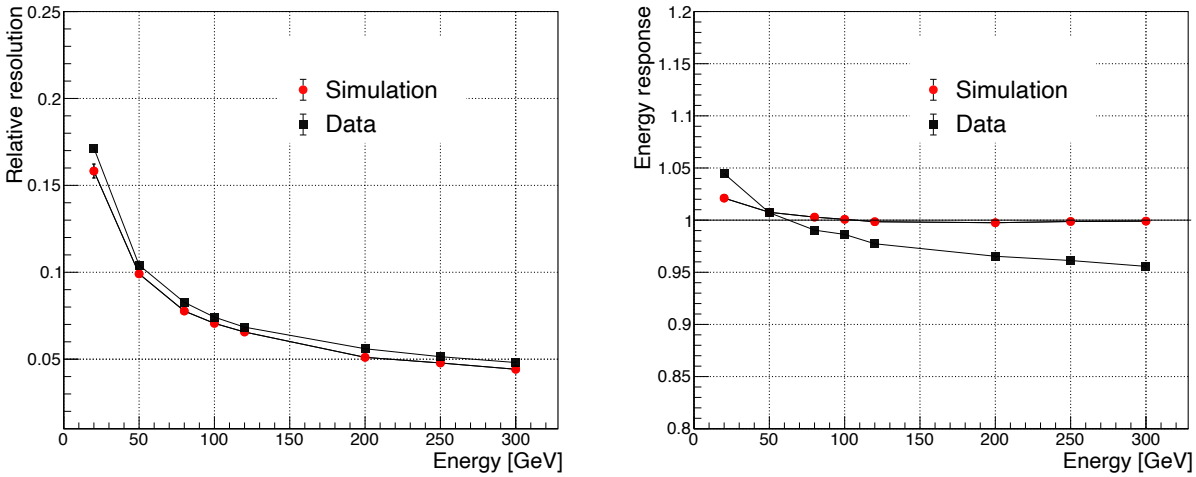
Figure 4.14: The comparison of resolution (a) and response (b) of the validation dataset for three different models: DRN-Ratio (blue curve), DRN-TrueE (magenta curve) and χ^2 Method (black curve).



(a) 20 GeV (b) 50 GeV (c) 100 GeV

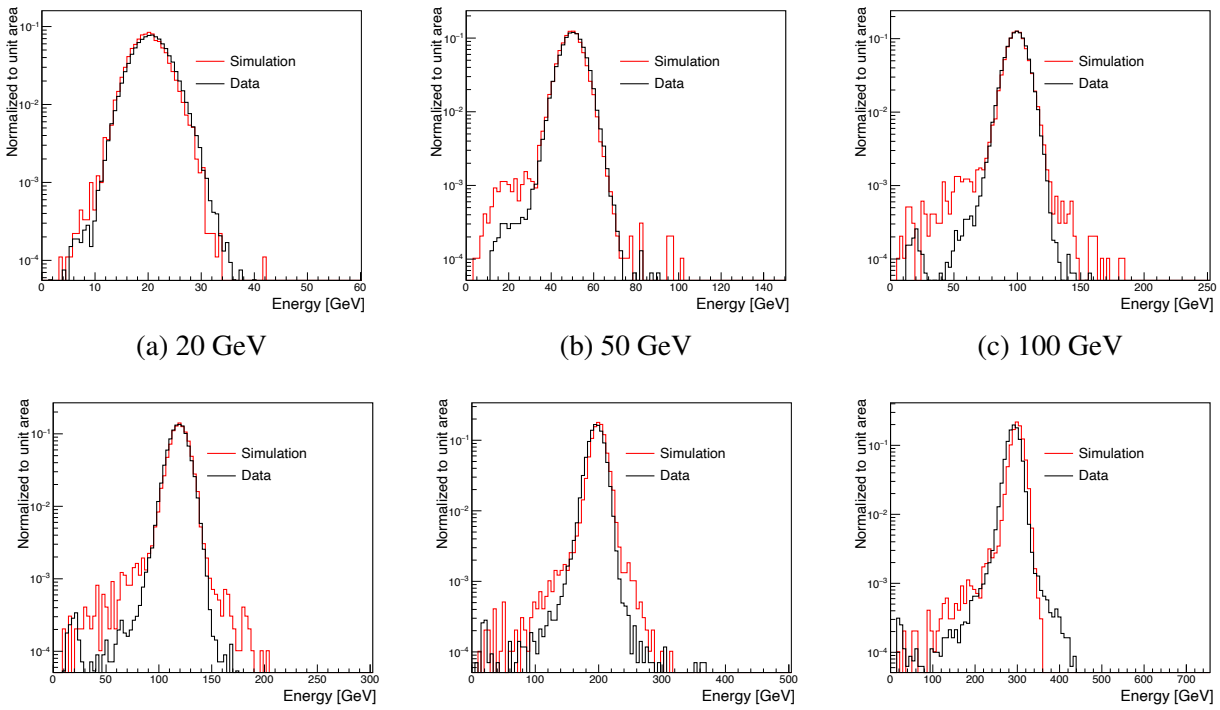
(d) 120 GeV (e) 200 GeV (f) 300 GeV

Figure 4.15: Predicted energy distribution comparison of the validation dataset (red curve) and test beam data (black curve) for the DRN-TrueE model.



(a) Resolution (b) Response

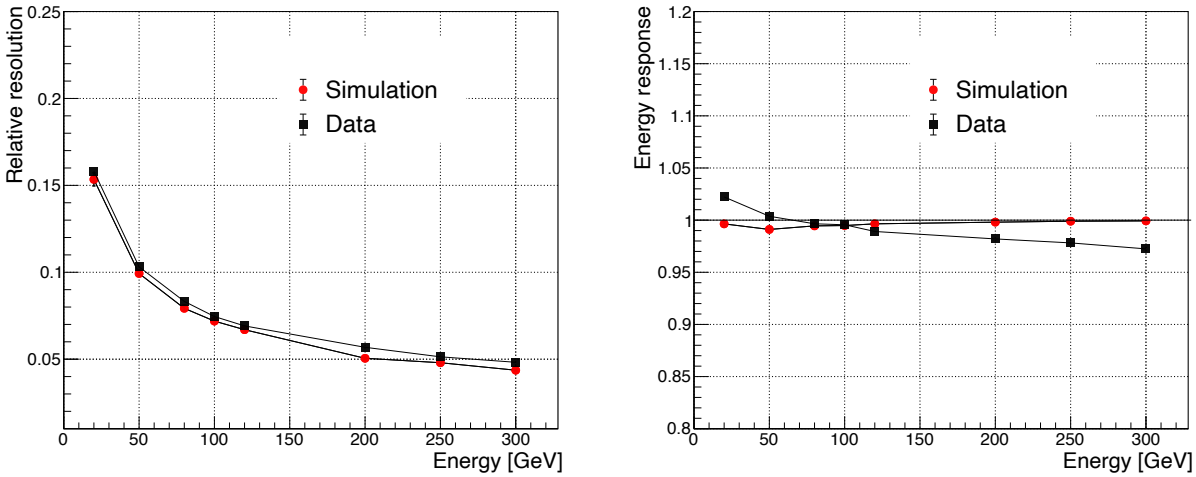
Figure 4.16: Comparison of resolution (a) and response (b) of the validation dataset (red) and test beam data (black) for the DRN-TrueE model.



(a) 20 GeV (b) 50 GeV (c) 100 GeV

(d) 120 GeV (e) 200 GeV (f) 300 GeV

Figure 4.17: Predicted energy distribution comparison of the validation dataset (red curve) and test beam data (black curve) for the DRN-Ratio model.



(a) Resolution (b) Response

Figure 4.18: Comparison of resolution (a) and response (b) of the validation dataset (red) and test beam data (black) for the DRN-Ratio model.

4.6 Resolving Edge Effects

Figure 4.19 shows peaks at the extreme ends of the predicted energy distribution (beyond 20–300 GeV). The higher energy peak has significantly reduced for the DRN-Ratio model compared to the DRN-TrueE model. In Figure 4.14, we see that the resolution and response of the DRN-TrueE model deviate more compared to the DRN-Ratio model for energy greater than 320 GeV. Changing the target from E_{true} to $\frac{E_{true}}{E_{raw}}$ has significantly reduced the edge effects for energy greater than 320 GeV. As mentioned earlier, the sharp edges are smoothened out in the $\frac{E_{true}}{E_{raw}}$ target distribution and thus, help in reducing the edge effects. Also, the DRN-Ratio model has a better overall response for both simulation and test beam data than the DRN-TrueE model. For these reasons, the DRN-Ratio model is chosen for any further analysis in the rest of the thesis.

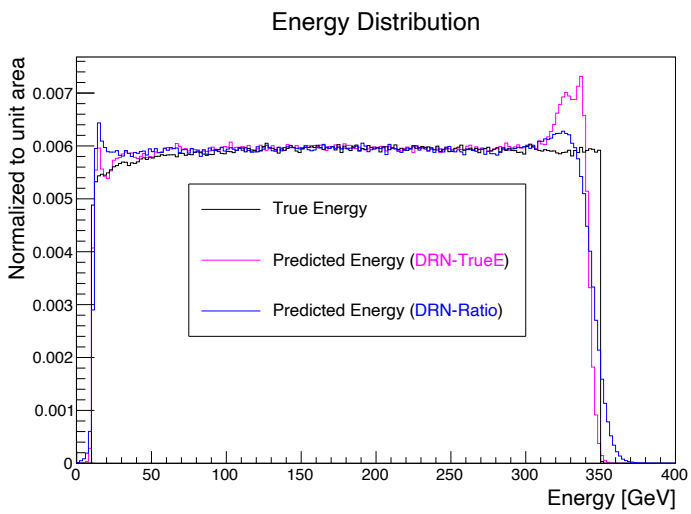
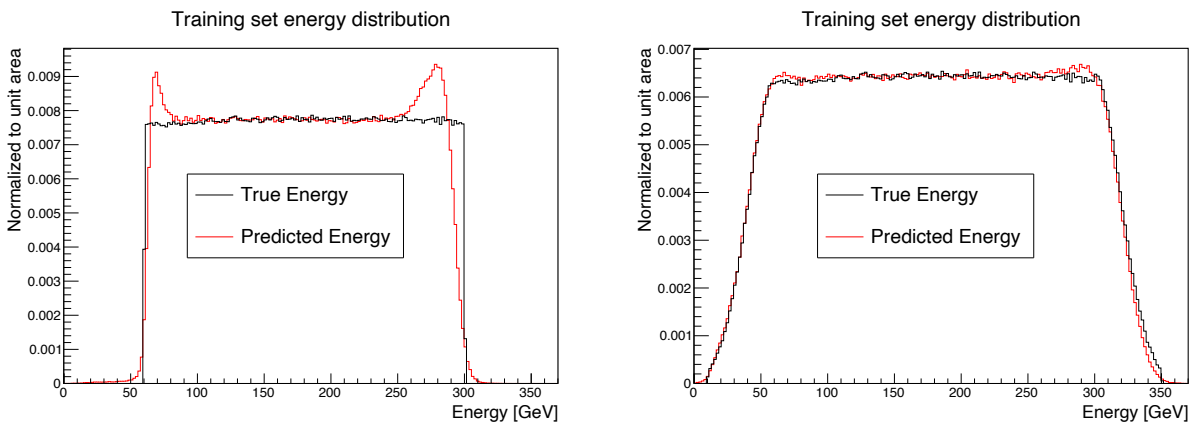


Figure 4.19: Pion energy distribution: true energy (black curve), predicted energy for the DRN-TrueE model (magenta curve) and predicted energy for the DRN-Ratio model.

Since we train the model with pion energies ranging from 10–350 GeV, it is always safer for the model not to predict energy beyond 10–350 GeV. This is also concluded by training the DRN-Ratio model with pion energies ranging from 60–300 GeV [DRN-Flat60to300], which gives a similar predicted energy distribution as shown in Figure 4.20a.

To resolve the edge effect, we train a DRN model with a gaussian edge in pion energy distribution [DRN-GaussEdge]. The pion events with energy between 60–300 GeV are untouched. Beyond 60 and 300 GeV, we select pion events for training according to a gaussian distribution (σ=20 GeV). There is no transformations of distributions involved, we limit the number of events for training according to a gaussian distribution.

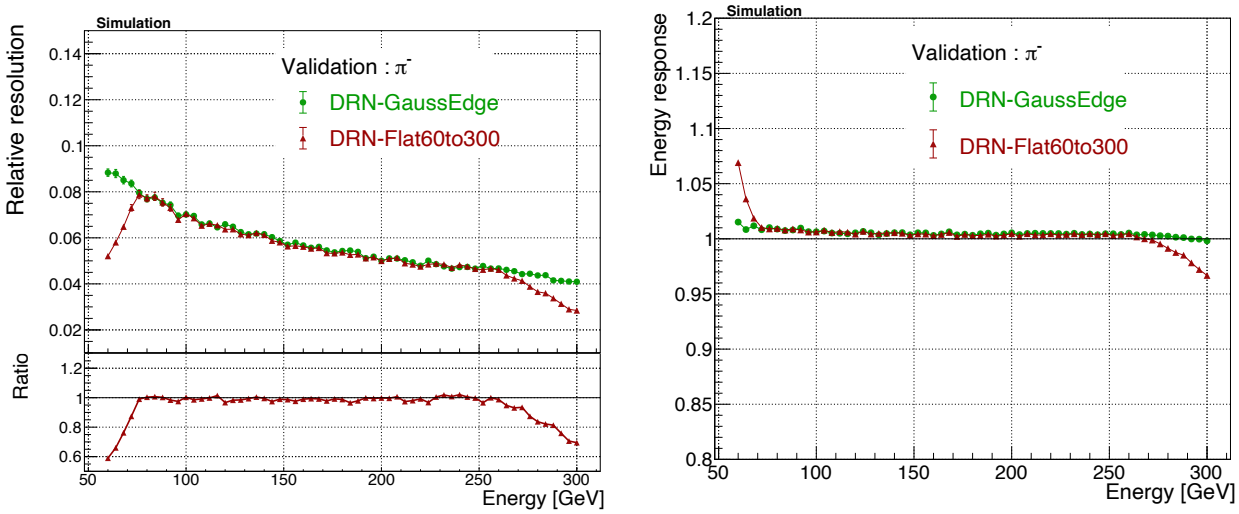


(a) DRN-Flat60to300

(b) DRN-GaussEdge

Figure 4.20: True (black curve) and predicted (red curve) energy distribution (training set) for a model trained with a guassian edge in pion energy distribution.

Using a gaussian edge for training has significantly reduced the edge effects. The predicted energy distribution follows the true energy distribution as shown in Figure 4.20b. Compared to the model DRN-Flat60to300, the response of the model DRN-GaussEdge is linear (Figure 4.21b), and the resolution curve is smooth (Figure 4.21a) for the pion energy ranging from 60–300 GeV. This concludes that the edge effects can be reduced by extending the energy range of pion samples according to a gaussian distribution. The time and the computational power required to simulate a pion shower increases with an increase in pion energy. So extending the energy range according to a gaussian distribution is better than a flat distribution at higher energies.



(a) Resolution

(b) Response

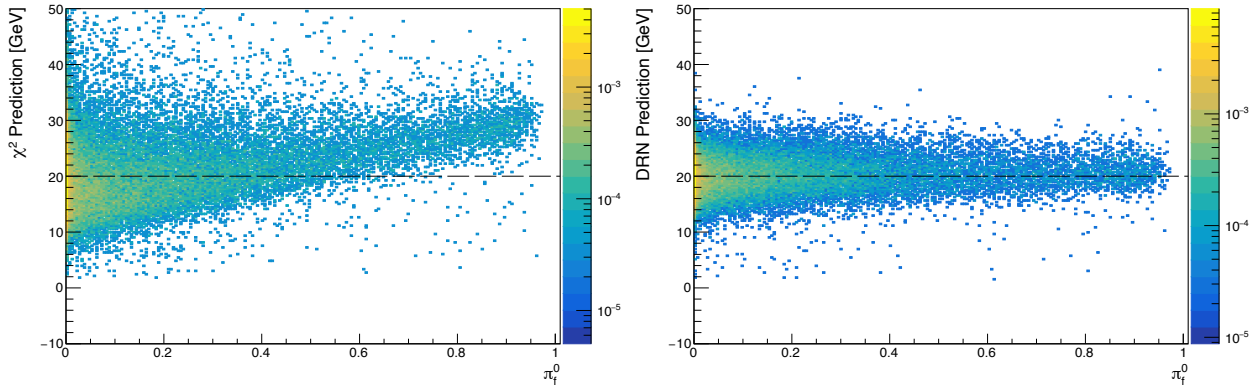
Figure 4.21: Comparison of resolution (a) and response (b) of the models DRN-GaussEdge (green curve) and DRN-Flat60to300 (brown curve).

4.7 DRN performance as a function of π_f^0

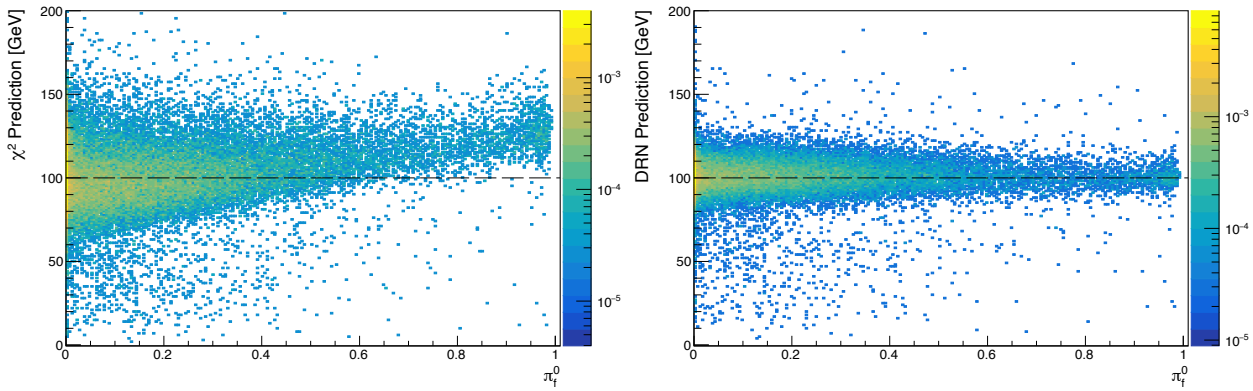
To study the effect of π^0 component on the energy reconstruction of hadron showers, we see how the predicted energy changes with the π_f^0. Here, π_f^0 is the fraction of energy carried by π^0 created in the first hadronic interaction of pion and the subscript f stands for 'first'. This information is obtained by saving the information of the secondary particles created in the first hadronic interaction using the GEANT4 simulation. Refer to section 3.4.1 for more details.

Figure 4.22 shows the predicted energy of χ^2 method and the DRN model (DRN-Ratio) as a function of π_f^0. The predicted energy by the χ^2 method increases with an increase in the π_f^0. This is expected as the χ^2 weights are obtained as an average over many events falling into a specific energy bin and do not account for shower-to-shower fluctuation in π_f^0. In comparison, the predicted energy of the DRN model is not correlated with the π_f^0.

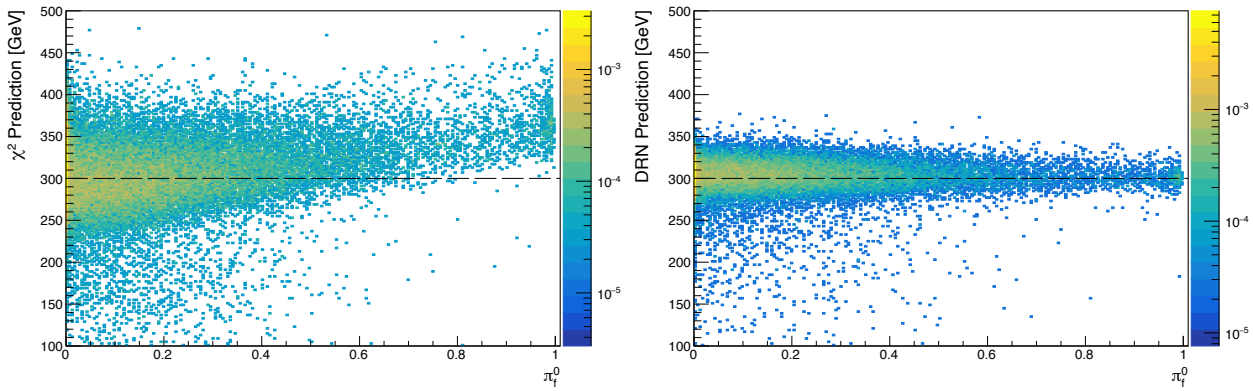
To further study how the predcited energy varies with π_f^0, we divide π_f^0 into different bins as shown in Figure 4.23. The peak of the predicted energy distribution of χ^2 method in different bins of π_f^0, shifts to the right with an increase in π_f^0 as shown in Figure 4.24 (left plots) and this worsens the overall resolution for a particular beam energy. In comparison, the peak of the predicted energy distribution of the DRN model in different bins of π_f^0 remains at a constant value as shown in Figure 4.24 (right plot), which concludes that the DRN model is effectively learning the shower-to-shower fluctuation in π_f^0.



(a) 20 GeV (b) 20 GeV



(c) 100 GeV (d) 100 GeV



(e) 300 GeV (f) 300 GeV

Figure 4.22: Predicted energy as a function of π_f^0: χ^2 prediction (left) and DRN prediction (right).

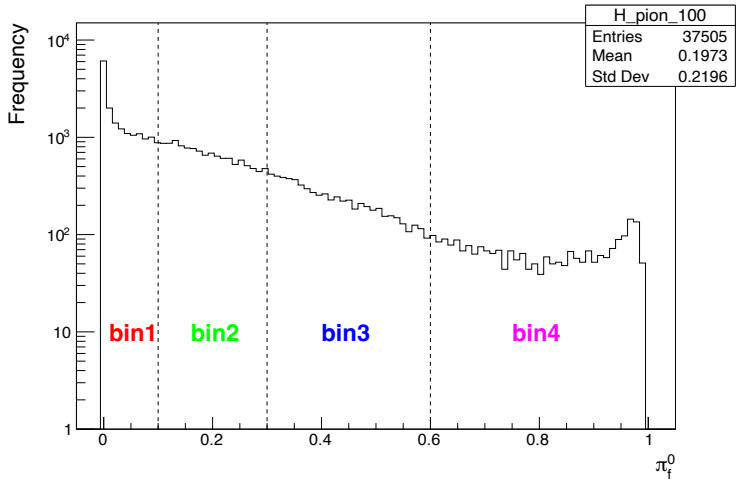
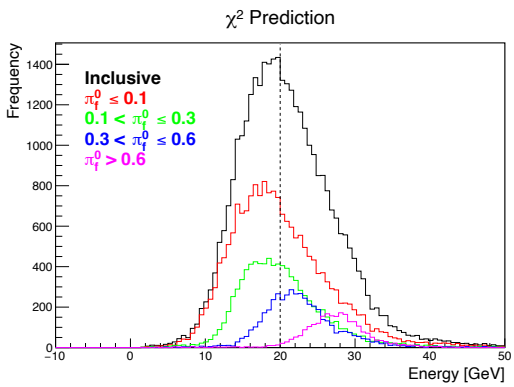
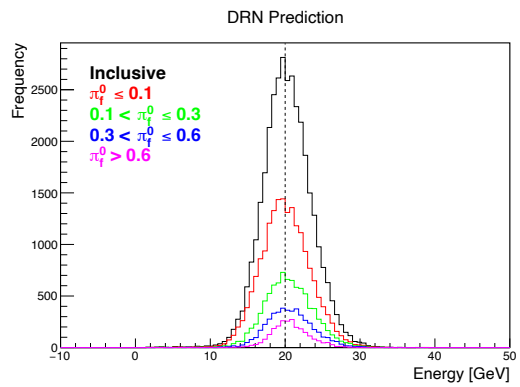


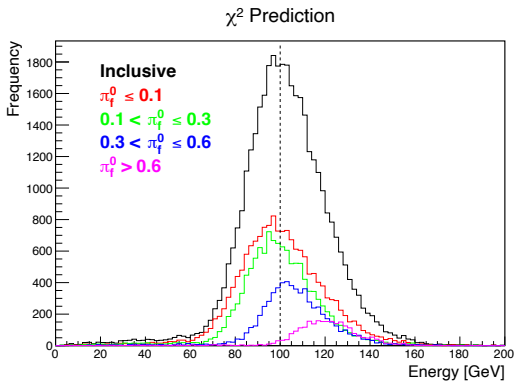
Figure 4.23: Distribution of π_f^0 for 100 GeV pion and the different bins used for analysis.



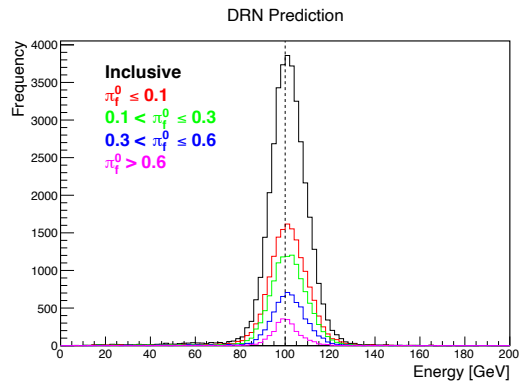
(a) 20 GeV



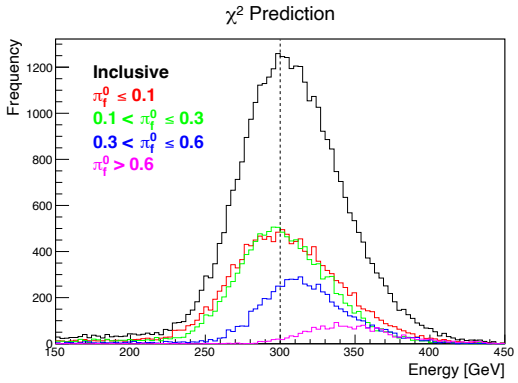
(b) 20 GeV



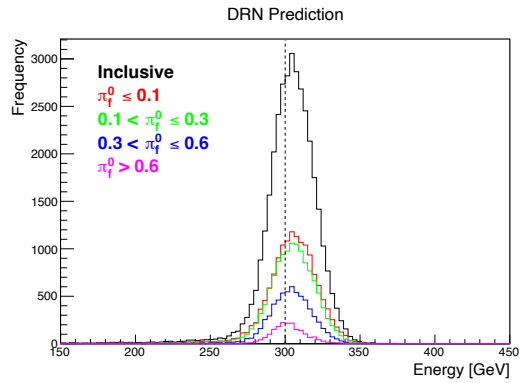
(c) 100 GeV



(d) 100 GeV



(e) 300 GeV



(f) 300 GeV

Figure 4.24: Predicted energy distribution for different bins of π_f^0 for χ^2 (left) and DRN (right).

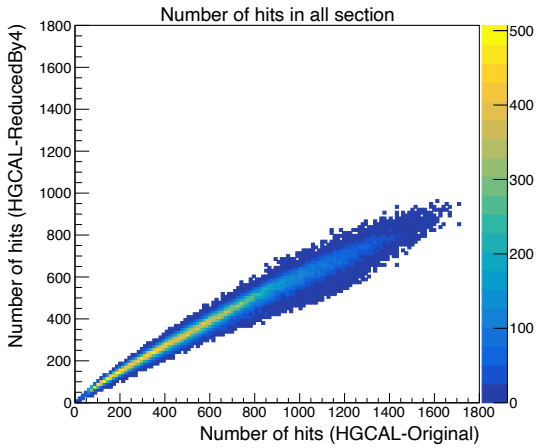
4.8 Performance of the DRN when providing reduced information as inputs

The following studies were additionally done to check the performance of the DRN model when we provide reduced information as inputs. The following sections describe how the information is reduced and the performance of the DRN model in terms of resolution and response.

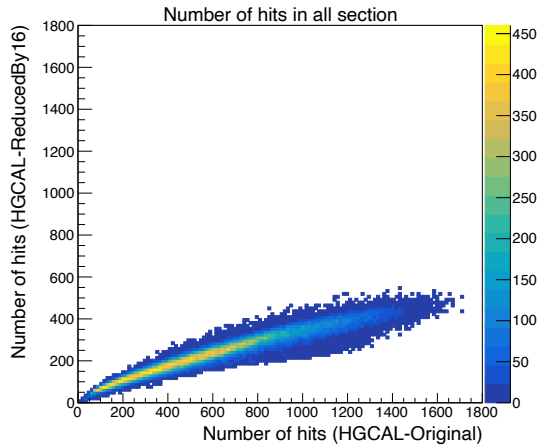
4.8.1 Performance of DRN when the rechit level information is combined along the transverse direction

We combine the adjacent cells in each sensitive layer of the HGCAL test beam setup to reduce the rechit level information along the transverse direction. This is done artificially by defining new x-y coordinates of the cells as the average of the x-y position of the combined adjacent cells in the original configuration. The rechit energy in the reduced configuration is defined as the sum of the rechit energy of the combined cells, because of which the total rechit energy sum in each layer for a pion event will be the same as the original configuration. There is no simulation of new pion events involved. After the new cell x-y positions are defined, we calculate the rechits for the same pion events in different reduced configurations of the HGCAL prototype. Specifically, we use two different configurations, i.e., combining the rechit level information from 4 (HGCAL-ReducedBy4) and 16 (HGCAL-ReducedBy16) adjacent cells for each sensitive layer. Figure 4.25 shows the number of hits recorded for different pion events in the reduced configuration versus the original configuration of the HGCAL prototype. The number of hits in the reduced configuration is not exactly reduced by a factor of 4 or 16 as not all the cells of the detector record a signal in a single pion event.

The DRN was trained for three different configurations (HGCAL-Original, HGCAL-ReducedBy4 and HGCAL-ReducedBy16) using 850k flat energy pion simulation samples (80% for training). The inputs for the models are the rechit energy and rechit x, y and z coordinates. All three models were trained for 200 epochs and the epoch with minimum validation loss was used for energy prediction. Figure 4.26 compares the predicted energy distribution of the DRN model trained on three different configurations for the training dataset (plots for the validation dataset are in the appendix). The predicted energy distribution closely overlaps for all three models. Figure 4.27 shows the performance in terms of resolution and response for the training dataset. The resolution and response of the validation dataset fluctuate because of the low statistics in each energy bin and



(a) HGCAL-ReducedBy4 versus HGCAL-Original

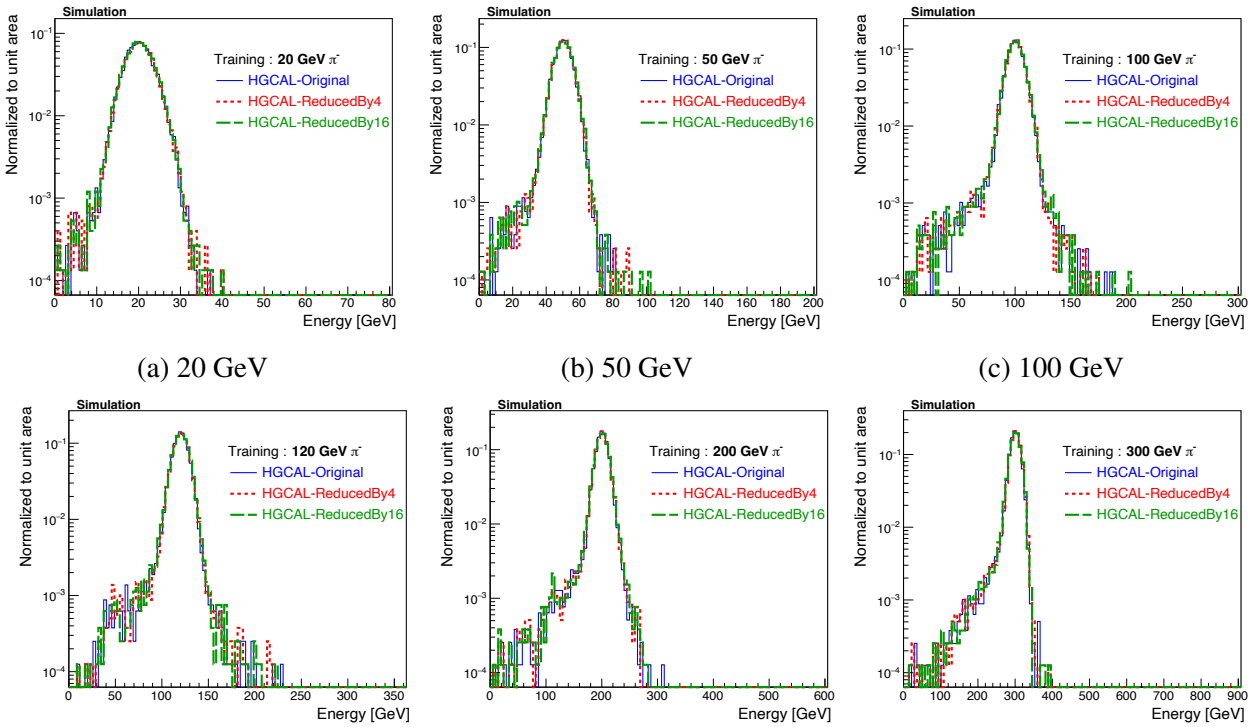


(b) HGCAL-ReducedBy16 versus HGCAL-Original

Figure 4.25: The number of hits recorded for different pion events in the reduced configuration versus the original configuration of the HGCAL prototype

the plots can be found in the appendix. The bottom panel of Figure 4.27a plots the ratio of resolution values of the HGCAL-ReducedBy4 (red curve) and HGCAL-ReducedBy16 (green curve) with the HGCAL-Original model. The performance of the HGCAL-ReducedBy4 is similar to the HGCAL-Original, but there is a slight degradation in the resolution of the HGCAL-ReducedBy16 (within 10%).

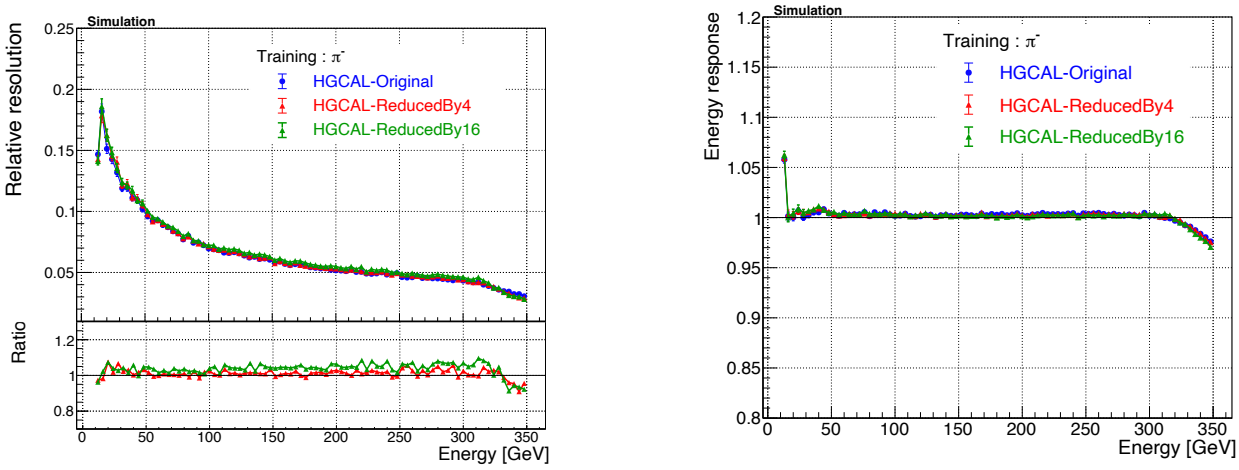
To check further how reducing the rechit level information in the transverse direction affects the DRN performance, we provide the layerwise rechit energy sum and z value of the layer as input to the DRN model (HGCAL-Layerwise_Ez). So there are only two input features to the model, i.e., the layerwise rechit energy sum and the corresponding z value of the sensitive layer. The HGCAL-Layerwise_Ez model was also trained using 850k flat energy pion simulation samples (80% for training) for 200 epochs. Figure 4.28 compares the predicted energy distribution of the HGCAL-Original and HGCAL-Layerwise_Ez for the training dataset. We see that the peak of the predicted energy distribution of the model HGCAL-Layerwise_Ez is broader compared to the HGCAL-Original model. Figure 4.29 shows the performance in terms of resolution and response for the training dataset. The bottom panel of Figure 4.29a plots the ratio of resolution values of the HGCAL-Layerwise_Ez with the HGCAL-Original model. There is degradation in the resolution of the HGCAL-Layerwise_Ez by about 30–60%. This is expected as the transverse features (rechit x and rechit y) are not available for the DRN model to learn the details of the transverse profile of the pion showers at each depth, which also gives information about π^0 content in the shower.



(a) 20 GeV (b) 50 GeV (c) 100 GeV

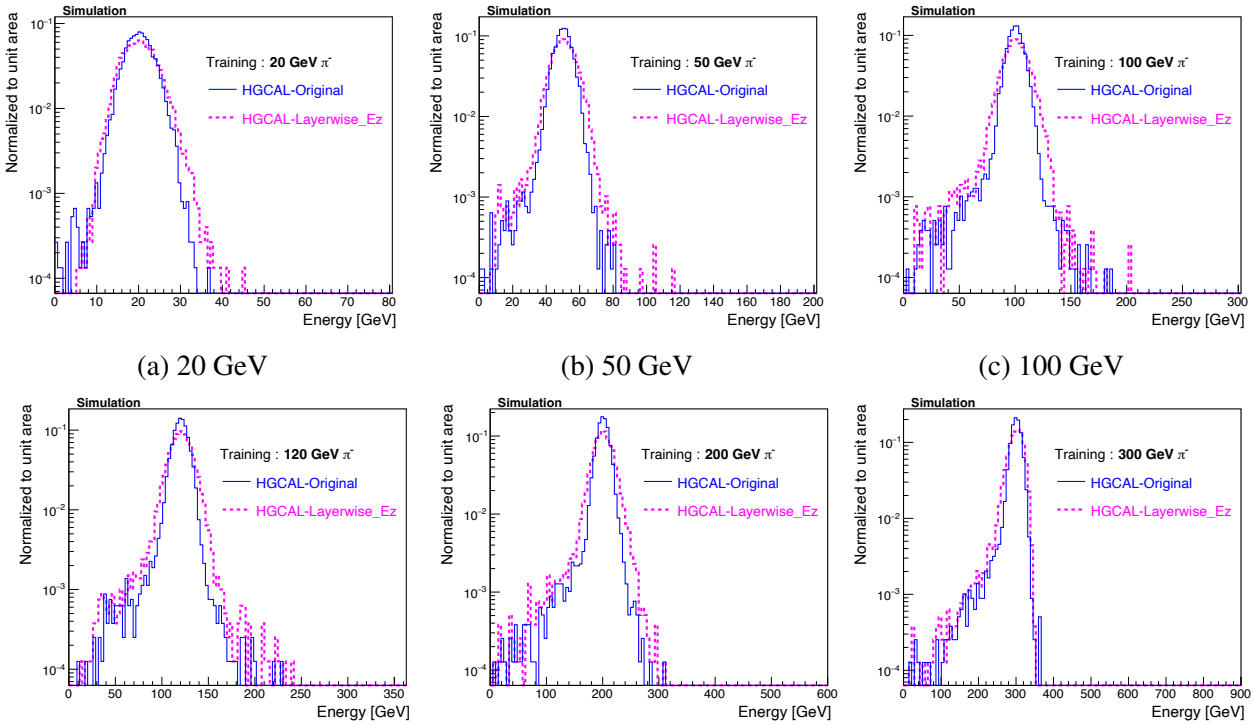
(d) 120 GeV (e) 200 GeV (f) 300 GeV

Figure 4.26: Predicted energy distribution comparison of the training dataset for three different configurations: HGCAL-Original (blue curve), HGCAL-ReducedBy4 (red curve) and HGCAL-ReducedBy16 (green curve).



(a) Resolution (b) Response

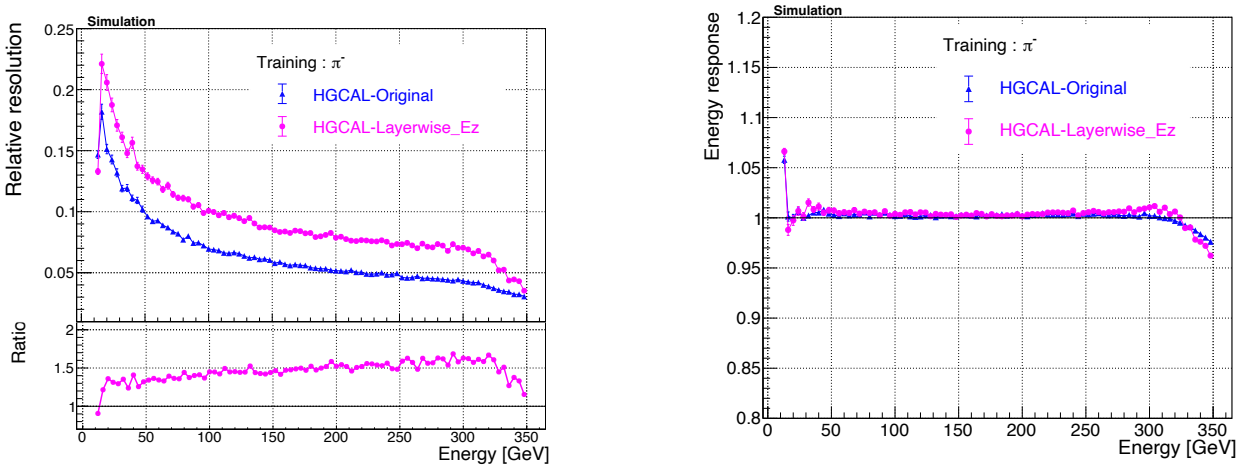
Figure 4.27: The comparison of resolution (a) and response (b) of the training dataset for three different configurations: HGCAL-Original (blue curve), HGCAL-ReducedBy4 (red curve) and HGCAL-ReducedBy16 (green curve).



(a) 20 GeV (b) 50 GeV (c) 100 GeV

(d) 120 GeV (e) 200 GeV (f) 300 GeV

Figure 4.28: Predicted energy distribution comparison of the training dataset for three different configurations: HGCAL-Original (blue curve), HGCAL-Layerwise_Ez (magenta curve).



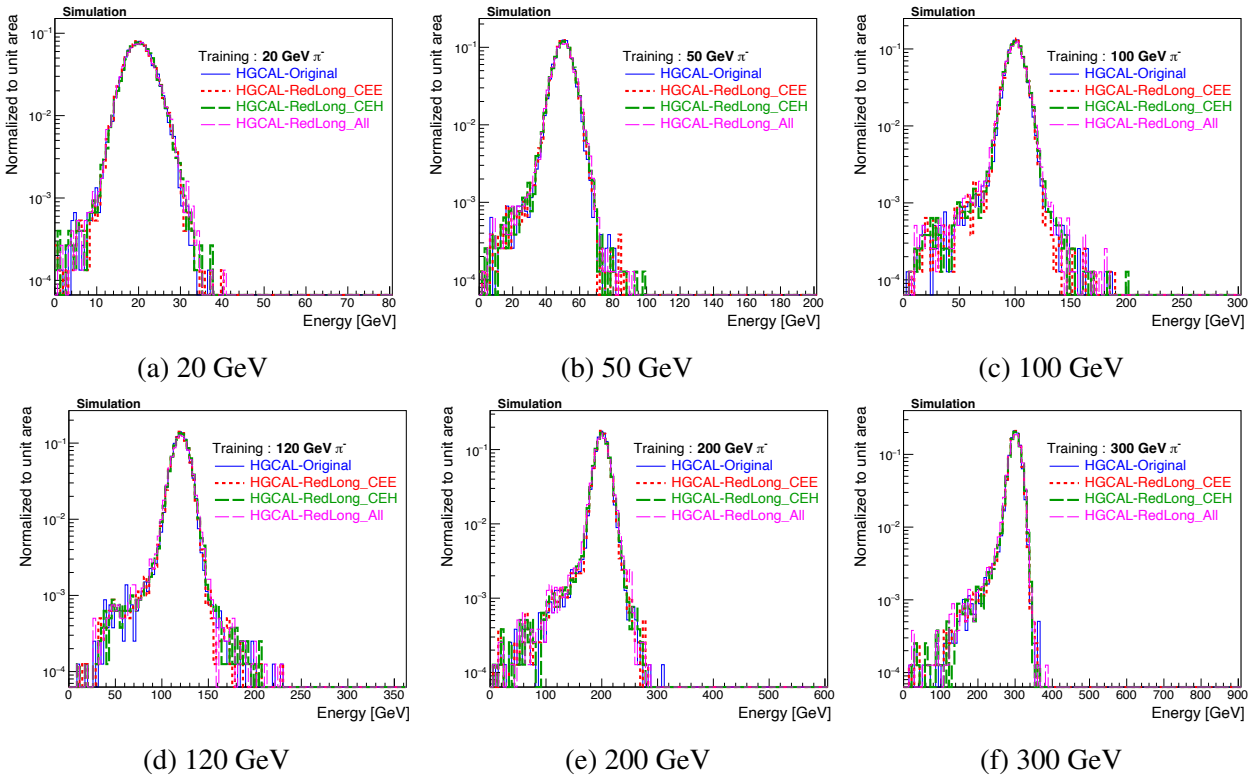
(a) Resolution (b) Response

Figure 4.29: The comparison of resolution (a) and response (b) of the training dataset for three different configurations: HGCAL-Original (blue curve), HGCAL-Layerwise_Ez (magenta curve).

4.8.2 Performance of DRN when the rechit level information is combined along the longitudinal direction

To reduce the rechit level information along the longitudinal direction in different sections, we integrate the energy deposited in each sensitive layer along the z-axis. There will be one sensitive layer per section and the rechit energy of the sensitive cells with the same (x,y) position is added for each section separately. The z value of the single sensitive layer is chosen as the z value of the first sensitive layer of each section. We train the DRN model with three different configurations, i.e., the energy deposited in each sensitive layer is integrated along the z-axis for only the CE-E section (HGCAL-RedLong_CEE), for only the CE-H section (HGCAL-RedLong_CEH) and for all three sections (HGCAL-RedLong_All). The inputs for the models are the rechit energy and rechit x, y and z coordinates. All the models were trained using 850k flat energy pion simulation samples (80% for training) for 200 epochs and the epoch with minimum validation loss was used for energy prediction.

Figure 4.30 compares the predicted energy distribution of the DRN model trained on different configurations for the training dataset (plots for the validation dataset are in the appendix). Figure 4.31 shows the performance in terms of resolution and response for the training dataset (plots for the validation dataset are in the appendix). The bottom panel of Figure 4.31a plots the ratio of resolution values of the HGCAL-RedLong_CEE (red curve), HGCAL-RedLong_CEH (green curve) and HGCAL-RedLong_All (magenta curve) with the HGCAL-Original model. The response of the DRN trained on all configuration closely match. The resolution of the HGCAL-RedLong_CEE model closely matches with the HGCAL-Original, but the resolution of the HGCAL-RedLong_CEH model is degraded (within 10%). Integrating the energy deposited in each sensitive layer along the z-axis in the CE-H section degrades the resolution more than in the CE-E section. This is mainly because the CE-H section covers a total depth of about 3.8 λ_{int}, whereas the CE-E section covers a total depth of about 1.4 λ_{int}. In the HGCAL-RedLong_CEH configuration, there is only one sensitive layer for 3.8 λ_{int} in the CE-H section and the performance of the DRN degrades due to less information available on the longitudinal development of the pion shower. The resolution is degraded even further (within 20%) for the HGCAL-RedLong_All model.

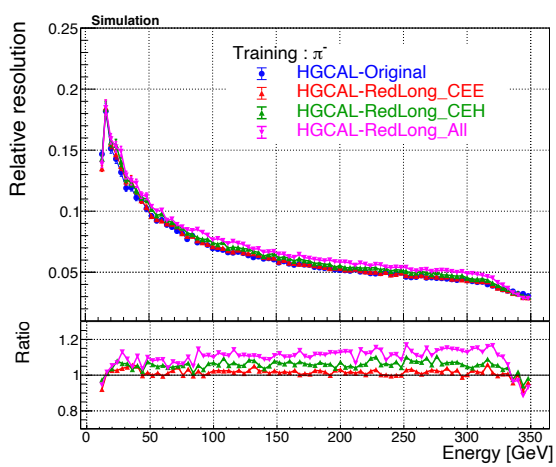


(a) 20 GeV (b) 50 GeV (c) 100 GeV

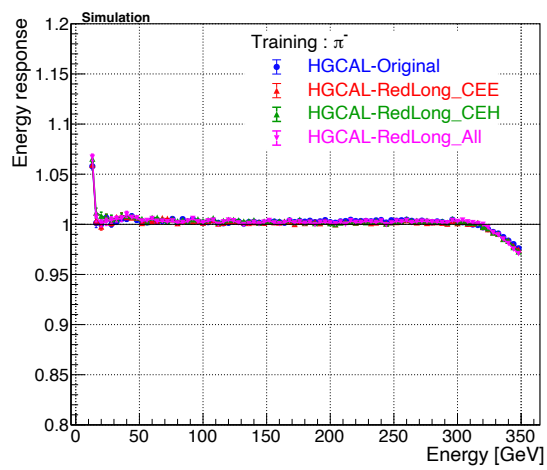
(d) 120 GeV (e) 200 GeV (f) 300 GeV

Figure 4.30: Predicted energy distribution comparison of the training dataset for four different configurations: HGCAL-Original (blue curve), HGCAL-RedLong_CEE (red curve), HGCAL-RedLong_CEH (green curve) and HGCAL-RedLong_All (magenta curve).

The above studies on providing reduced information for the DRN model show that the distribution of rechit energy in both longitudinal and transverse directions is essential for the DRN model to learn about the π^0 component of the pion shower (which has a different shape than the hadronic component) and get the best performance in terms of energy resolution. The degradation in pion energy resolution is the maximum in the HGCAL-Layerwise_Ez model (about 30–60%), where we have completely reduced the rechit energy distribution in the transverse direction and only provide one energy value per layer as input to the DRN model. In the HGCAL-ReducedBy16 model, the amount of information provided to the model is reduced, but seems to be sufficient for the model to learn the event-by-event variations with degradation in resolution by 10%. Similar is the case in the HGCAL-RedLong_All model, where the distribution of rechit energy in the transverse direction is still preserved and degradation in energy resolution is within 20%.



(a) Resolution



(b) Response

Figure 4.31: The comparison of resolution (a) and response (b) of the training dataset for four different configurations: HGCAL-Original (blue curve), HGCAL-RedLong_CEE (red curve), HGCAL-RedLong_CEH (green curve) and HGCAL-RedLong_All (magenta curve).

Chapter 5

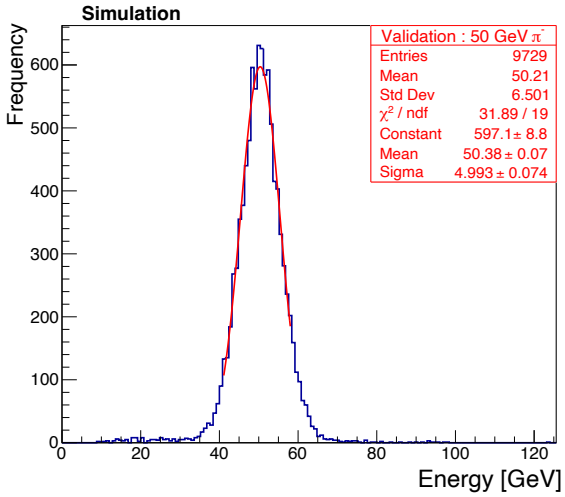
Summary

In this thesis, we started by studying the interactions of particles with matter, electromagnetic and hadronic showers, calorimetry and detector simulation using the GEANT4 toolkit. In Chapter 3, we discussed the details of the proposed CMS HGCAL detector and the HGCAL test beam setup, along with detector calibration techniques. Using the GEANT4 simulation, we also studied the kinematics of various particles produced in the first hadronic interaction of the pion and the variation in the fraction of energy carried by the π^0 produced in the pion showers.

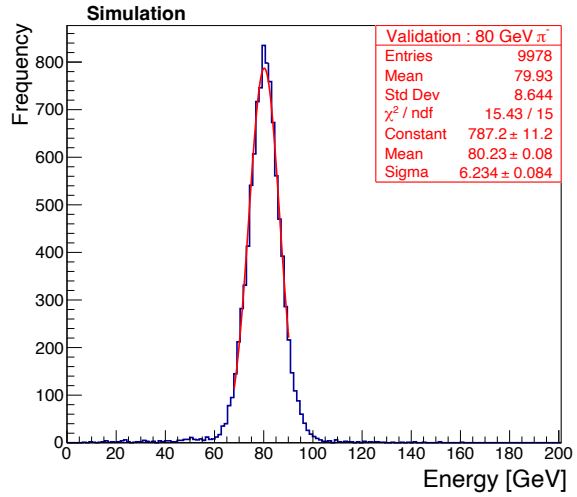
In chapter 4 we discussed the details of the semiparametric dynamic reduction network (DRN), which is a graph neural network-based model and checked its performance on energy regression of charged pions. The semiparametric DRN model efficiently uses the rechit information of the HGCAL-TB setup and reconstructs the pion energy better than the χ^2 method. Compared to the χ^2 method, the pion energy resolution of the semiparametric DRN model is improved by a factor of 2. Unlike χ^2 method, the semiparametric DRN model is able to learn the shower-to-shower fluctuation π^0_f. We also saw that the edge effects could be reduced by extending the energy range of the pion samples according to a gaussian distribution for training the model. When the reduced information was provided to the semiparametric DRN model by combining 16 adjacent cells of each layer, we saw degradation (within 10%) in the pion energy resolution. The pion energy resolution is degraded by about 30–60% when we provide layerwise rechit energy sum and corresponding z value of layer as input. Also, we saw degradation in the pion energy resolution when reduced information was provided by integrating the energy deposited along the longitudinal direction. The resolution degradation was about 10% (20%) when the energy deposited was integrated along the longitudinal direction only for the CE-H section (for all sections).

Appendix A

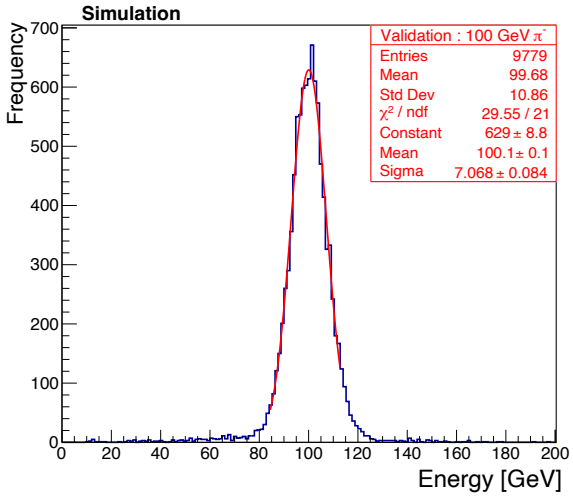
Plots and graph



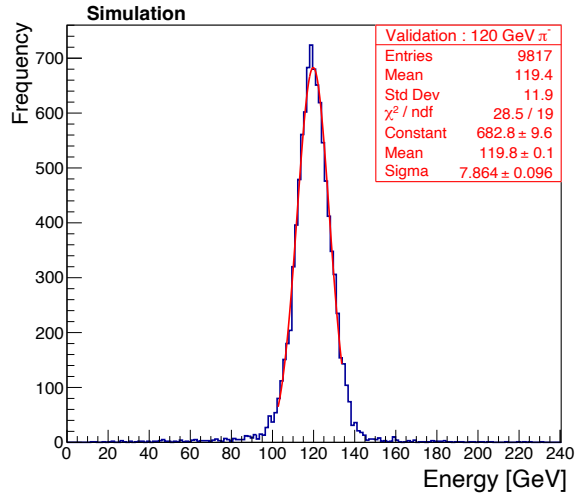
(a) 50 GeV



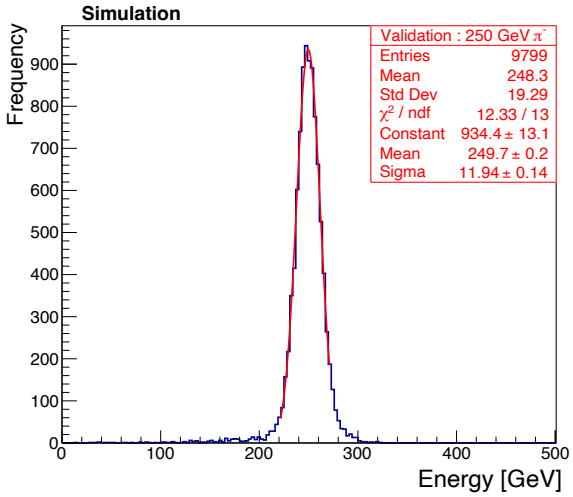
(b) 80 GeV



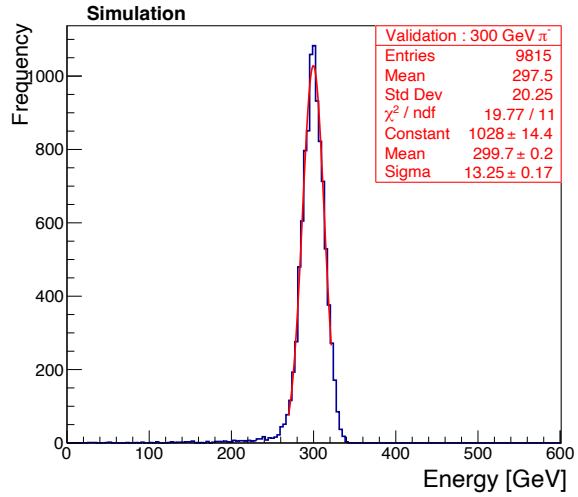
(c) 100 GeV



(d) 120 GeV

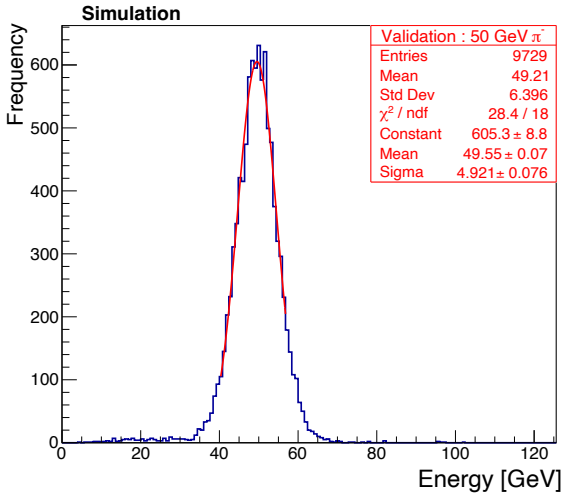


(e) 250 GeV

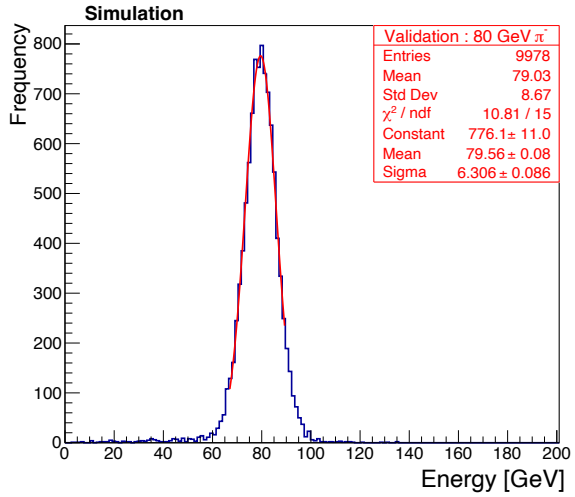


(f) 300 GeV

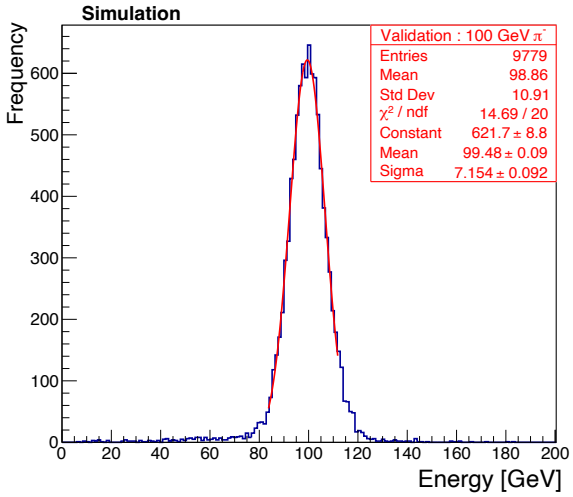
Figure A.1: The predicted energy distributions and the fitted gaussian function is shown for the prediction made by the DRN-TrueE model on the validation dataset.



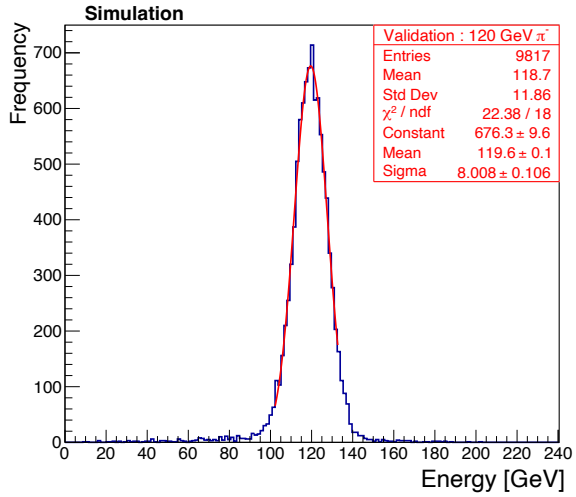
(a) 50 GeV



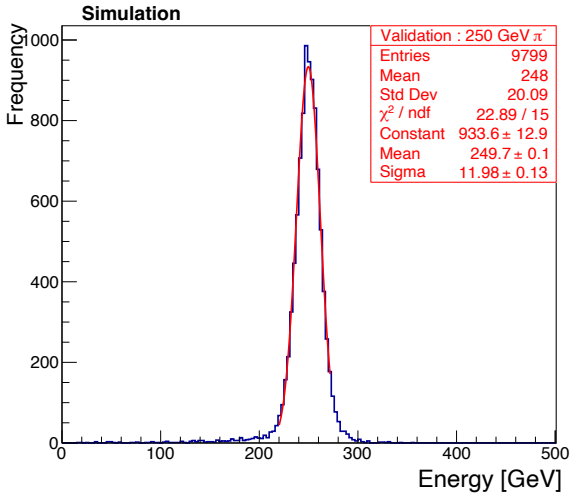
(b) 80 GeV



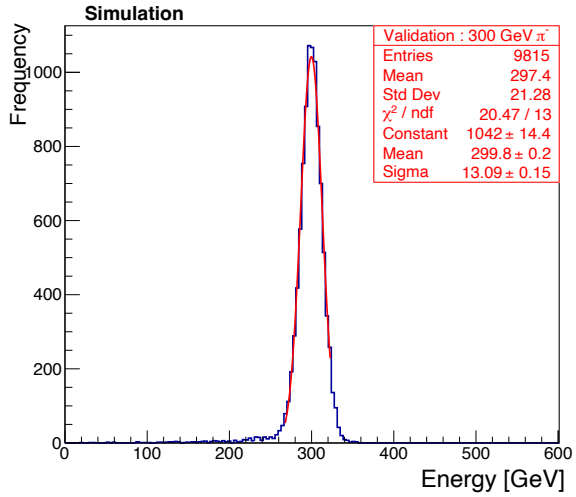
(c) 100 GeV



(d) 120 GeV

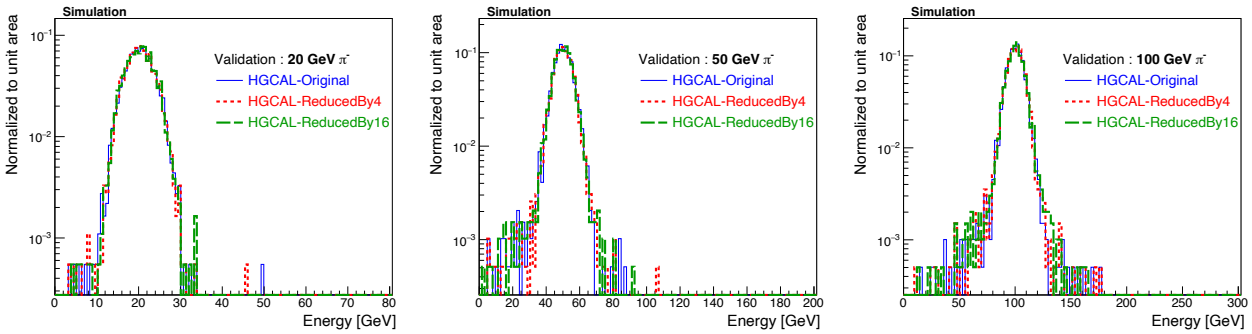


(e) 250 GeV



(f) 300 GeV

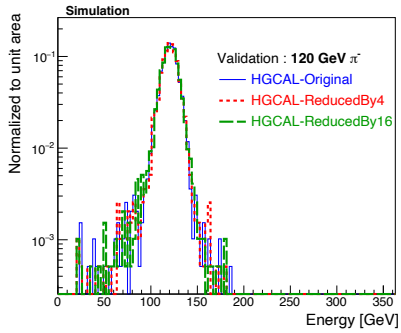
Figure A.2: The predicted energy distributions and the fitted gaussian function is shown for the prediction made by the DRN-Ratio model on the validation dataset.



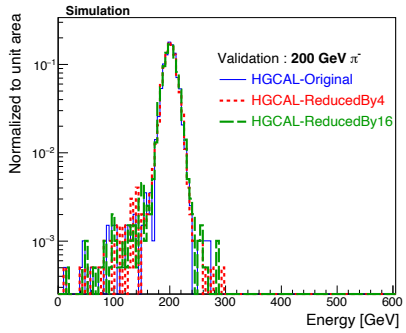
(a) 20 GeV

(b) 50 GeV

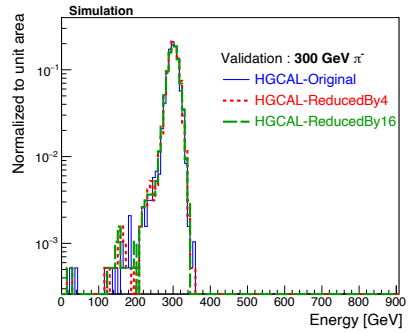
(c) 100 GeV



(d) 120 GeV

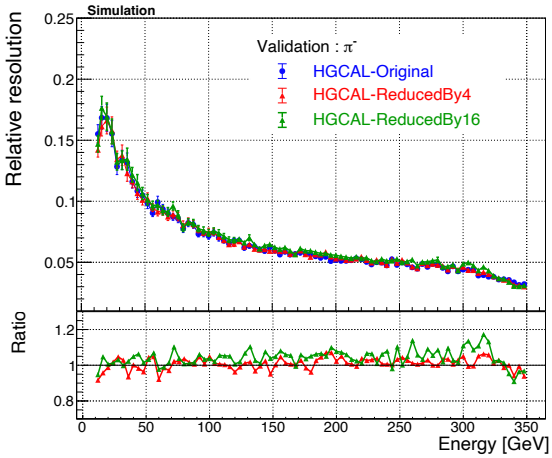


(e) 200 GeV

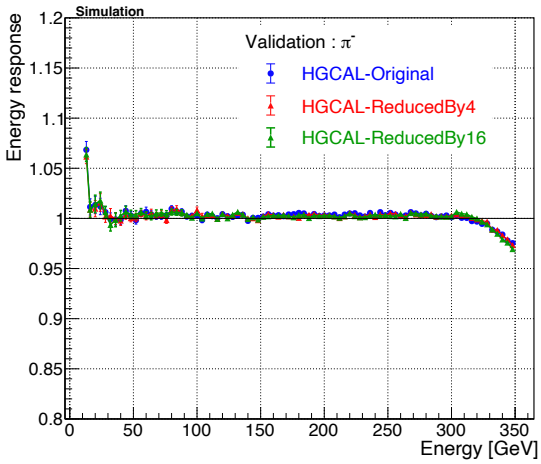


(f) 300 GeV

Figure A.3: Predicted energy distribution comparison of the validation dataset for three different configurations: original (blue curve), reduced by a factor of 4 (red curve) and 16 (green curve).

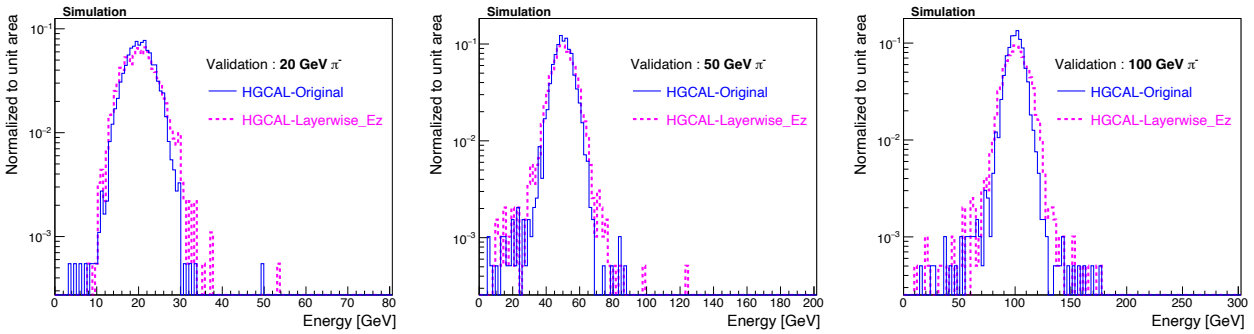


(a) Resolution

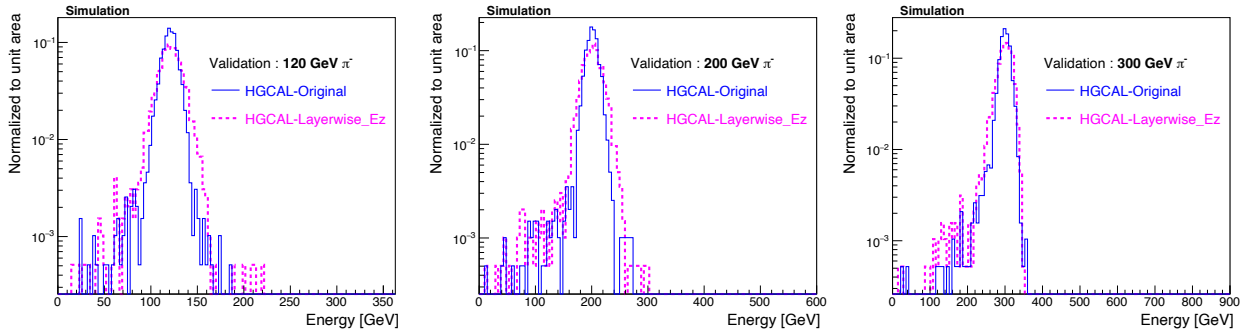


(b) Response

Figure A.4: The comparison of resolution (a) and response (b) of the validation dataset for three different configurations: original (blue curve), reduced by a factor of 4 (red curve) and 16 (green curve).

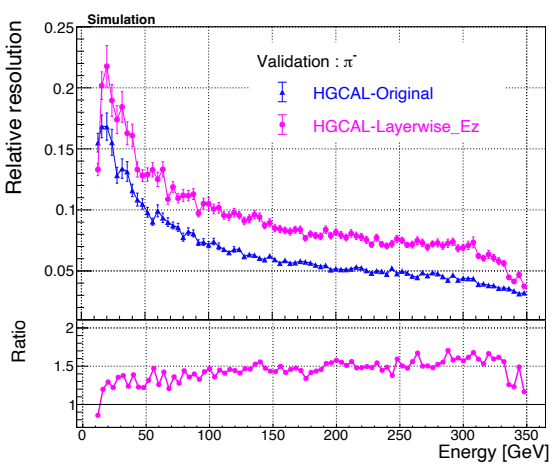


(a) 20 GeV (b) 50 GeV (c) 100 GeV

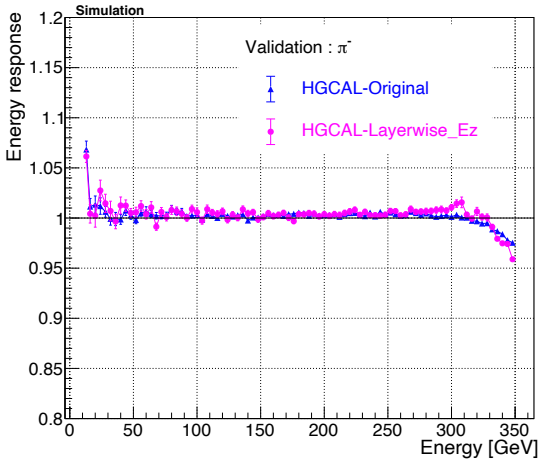


(d) 120 GeV (e) 200 GeV (f) 300 GeV

Figure A.5: Predicted energy distribution comparison of the validation dataset for three different configurations: HGCAL-Original (blue curve), HGCAL-Layerwise_Ez (magenta curve).

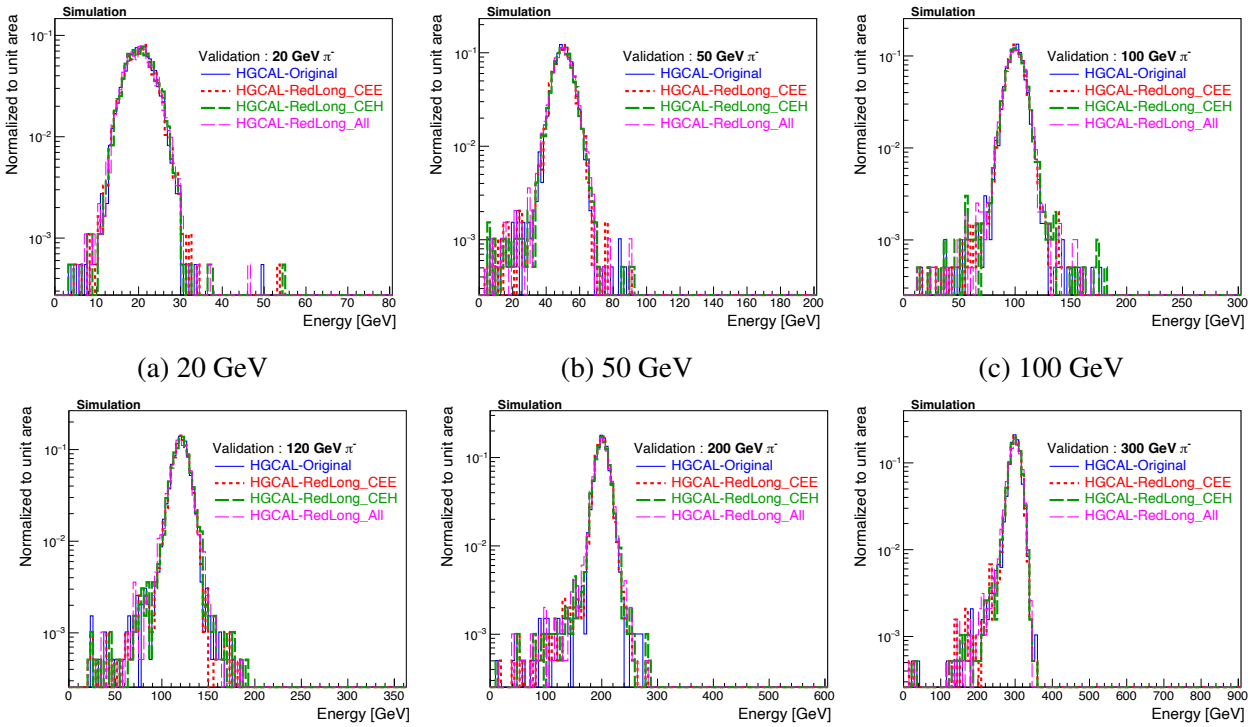


(a) Resolution



(b) Response

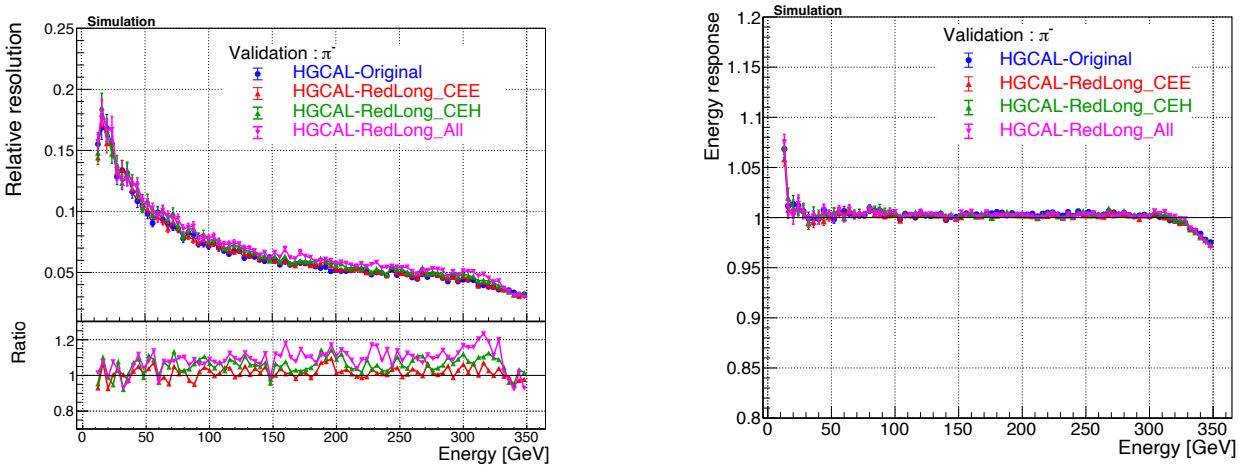
Figure A.6: The comparison of resolution (a) and response (b) of the validation dataset for three different configurations: HGCAL-Original (blue curve), HGCAL-Layerwise_Ez (magenta curve).



(a) 20 GeV (b) 50 GeV (c) 100 GeV

(d) 120 GeV (e) 200 GeV (f) 300 GeV

Figure A.7: Predicted energy distribution comparison of the validation dataset for four different configurations: HGCAL-Original (blue curve), HGCAL-RedLong_CEE (red curve), HGCAL-RedLong_CEH (green curve) and HGCAL-RedLong_All (magenta curve).



(a) Resolution (b) Response

Figure A.8: The comparison of resolution (a) and response (b) of the validation dataset for four different configurations: HGCAL-Original (blue curve), HGCAL-RedLong_CEE (red curve), HGCAL-RedLong_CEH (green curve) and HGCAL-RedLong_All (magenta curve).

Bibliography

[1] Review of Particle Physics. *Progress of Theoretical and Experimental Physics*, 2020(8), 08 2020. ISSN 2050-3911. doi: 10.1093/ptep/ptaa104. URL https://doi.org/10.1093/ptep/ptaa104. 083C01.

[2] Michele Livan and Richard Wigmans. *Calorimetry for Collider Physics, an Introduction*. Springer International Publishing, Cham, 2019. ISBN 978-3-030-23653-3. doi: 10.1007/978-3-030-23653-3_2. URL https://doi.org/10.1007/978-3-030-23653-3_2.

[3] Tejinder S Virdee. *Experimental Techniques*. Technical report, CERN, 1999.

[4] The Phase-2 Upgrade of the CMS Endcap Calorimeter. Technical report, CERN, Geneva, 2017. URL https://cds.cern.ch/record/2293646.

[5] Construction and commissioning of CMS CE prototype silicon modules. *JINST*, 16(04): T04002, 2021. doi: 10.1088/1748-0221/16/04/T04002. URL https://cds.cern.ch/record/2748054. 37 pages, submitted to JINST.

[6] CMS Collaboration and CALICE Collaboration. Performance of the CMS High Granularity Calorimeter prototype to charged pion beams of 20−300 GeV/c, 2022. URL https://arxiv.org/abs/2211.04740.

[7] Shubham Pandey. Performance of High Granularity Calorimeter prototypes for the CMS HL-LHC upgrade in beam test experiments at CERN, 2022. URL http://cds.cern.ch/record/2810369.

[8] CS231n Convolutional Neural Networks for Visual Recognition — cs231n.github.io. https://cs231n.github.io/neural-networks-1/#layers.

[9] Aurélien Géron. *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*. O'Reilly Media, Inc., second edition, 2019.

[10] Lyndon R Evans. The Large Hadron Collider Project. 1997. URL https://cds.cern.ch/record/313675.

[11] The CMS experiment at the CERN LHC. *Journal of Instrumentation*, 3(08):S08004, aug 2008. doi: 10.1088/1748-0221/3/08/S08004. URL https://dx.doi.org/10.1088/1748-0221/3/08/S08004.

[12] The ATLAS Experiment at the CERN Large Hadron Collider. *JINST*, 3:S08003, 2008. doi: 10.1088/1748-0221/3/08/S08003. URL http://cds.cern.ch/record/1129811. Also published by CERN Geneva in 2010.

[13] William R. Leo. *Techniques for Nuclear and Particle Physics Experiments: A How-to Approach*. Springer Berlin Heidelberg, Berlin, Heidelberg, 1994. ISBN 978-3-642-57920-2. doi: 10.1007/978-3-642-57920-2_7. URL https://doi.org/10.1007/978-3-642-57920-2_7.

[14] Geant4—a simulation toolkit. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 506(3):250–303, 2003. ISSN 0168-9002. doi: https://doi.org/10.1016/S0168-9002(03)01368-8. URL https://www.sciencedirect.com/science/article/pii/S0168900203013688.

[15] CMS Collaboration. Introduction to GEANT4. URL https://cds.cern.ch/record/491492/files/p107.pdf.

[16] GEANT4 Collaboration. *GEANT4 Physics Reference Manual*, 2017.

[17] GEANT4 Collaboration. *GEANT4 Book For Application Developers*, 2017.

[18] GEANT4 Collaboration. GEANT4 simulation of Si-based sampling calorimeter. URL https://github.com/pfs/PFCal/tree/master/PFCalEE.

[19] CMS Collaboration. Investigations on the impact of re-optimization of the HGCAL longitudinal structure on hadron energy resolution, CMS Internal Note, 15 June 2021.

[20] GEANT4 Collaboration. Saving information regarding secondary particles produced in pion showers, 2023. URL https://github.com/nitishkumarkv/PFCal/blob/master/PFCalEE/src/SteppingAction.cc#L36-L121.

[21] Lindsey Gray, Thomas Klijnsma, and Shamik Ghosh. A dynamic reduction network for point clouds. *CoRR*, abs/2003.08013, 2020. URL https://arxiv.org/abs/2003.08013.

[22] Ilya Loshchilov and Frank Hutter. Fixing weight decay regularization in adam. *CoRR*, abs/1711.05101, 2017. URL http://arxiv.org/abs/1711.05101.

[23] Simon Rothman. Calibrating Electrons and Photons in the CMS ECAL using Graph Neural Networks. Technical report, CERN-CMS-CR-2021-156, 2021. URL https://cds.cern.ch/record/2799575.

[24] π^- energy reconstruction in HGCAL Beam Test prototype detector using Graph Neural Networks, CERN-CMS-DP-2022-022, 2022. URL http://cds.cern.ch/record/2815404.

[25] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E. Sarma, Michael M. Bronstein, and Justin M. Solomon. Dynamic graph CNN for learning on point clouds. *CoRR*, abs/1801.07829, 2018. URL http://arxiv.org/abs/1801.07829.

[26] Inderjit S. Dhillon, Yuqiang Guan, and Brian Kulis. Weighted graph cuts without eigenvectors a multilevel approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29 (11):1944–1957, 2007. doi: 10.1109/TPAMI.2007.1115.

[27] Response of a CMS HGCAL silicon-pad electromagnetic calorimeter prototype to 20–300 GeV positrons. *Journal of Instrumentation*, 17(05):P05022, may 2022. doi: 10.1088/1748-0221/17/05/P05022. URL https://dx.doi.org/10.1088/1748-0221/17/05/P05022.