Experimental best practices from CMS and introducing benchmark datasets for the workshop

Rajdeep M Chatterjee, Avik Das
TIFR

Machine Learning For High Energy Physics V4, TIFR-Mumbai

Credits to several colleagues and collaborators who have shared content and code

The benchmark datasets for the ML-Workshop

1) Electron energy reconstruction with a HGCAL prototype : https://zenodo.org/records/7504164

Electron Energy Regression in High-Granularity Calorimeter Prototype

Bhargav Joshi[1] ; Alpana Alpana[2]

Show affiliations

The dataset consists of simulations of calibrated reconstructed hits produced by a positron passing through the HGCAL test beam prototype. For the simulations, Monte Carlo method is used to produce the positrons with energy ranging from 20 to 350 GeV. The dataset contains the coordinates of the calibrated reconstructed hits in the prototype along with the calibrated energy in units of MIP. The HDF5 files can be extracted from the gzip files.

Files

Files (16.8 GB)

Name	Size	Download all
hgcal_electron_data_0001.h5.gz md5:c4c5dbbd3025dc917b9386921236f527	2.8 GB	Download
hgcal_electron_data_large.h5.gz md5:3515f983cd0f154cfc22e68c3d1e73c3	14.0 GB	Download

The benchmark datasets for the ML-Workshop

2) Pion energy reconstruction with a HGCAL prototype :

https://zenodo.org/records/7864471

Pion Energy Regression in High-Granularity Calorimeter Prototype

Alpana Alpana[1] ; Bhargav Joshi[2] ; Shubham Pandey[2]

Show affiliations

The dataset consists of simulations of calibrated reconstructed hits produced by a pion passing through the HGCAL test beam prototype. For the simulations, Monte Carlo method is used to produce the pions with energy ranging from 10 to as high as 500 GeV. The dataset contains the coordinates of the calibrated reconstructed hits in the prototype along with the calibrated energy in units of MIP. The HDF5 files can be extracted from the gzip files.

Files

Files (20.5 GB)

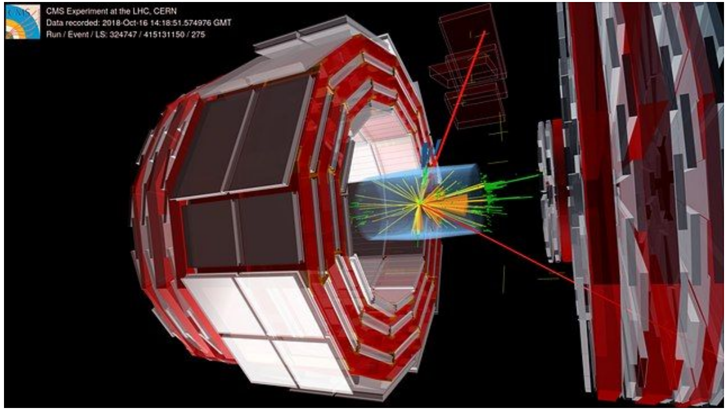
Name	Size	Download all
home/rusack/shared/hdf5/hgcal_pion/zenodo/hgcal_pions_combinedHgc_Ahc_1.h5.gz md5:23a8d2d21c1bb5f3e396b35ffdc034f1	4.1 GB	Download

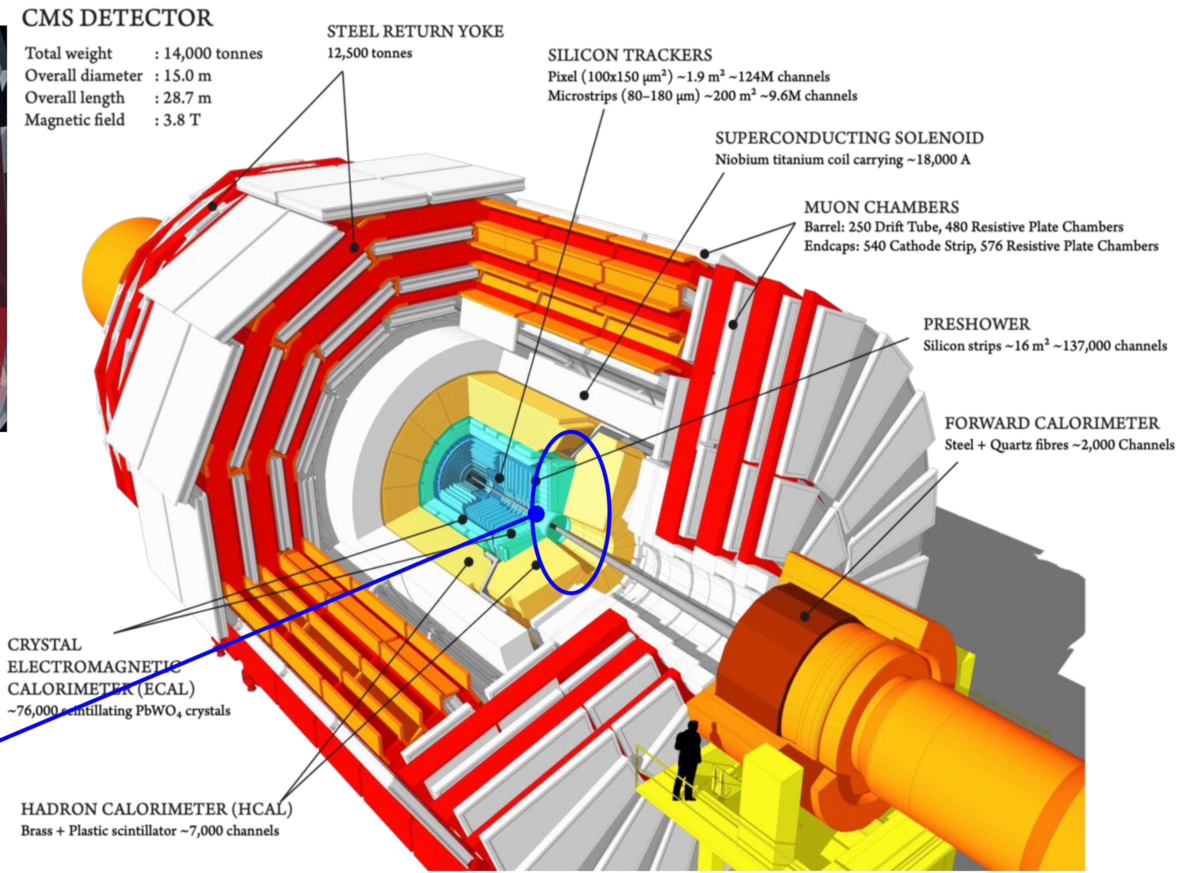
Outline

- Introduction to the CMS High Granularity Calorimeter (HGCAL)

- The HGCAL beam test prototype and benchmark results

Compact muon solenoid (CMS) at LHC





- Designed to work well till the end of Run 3.
- Various sub-detector need to be upgrade for HL–LHC.
- Today's talk focuses on endcap calorimeters.

HL-LHC Requirements for the CMS Detector

High radiation exposure

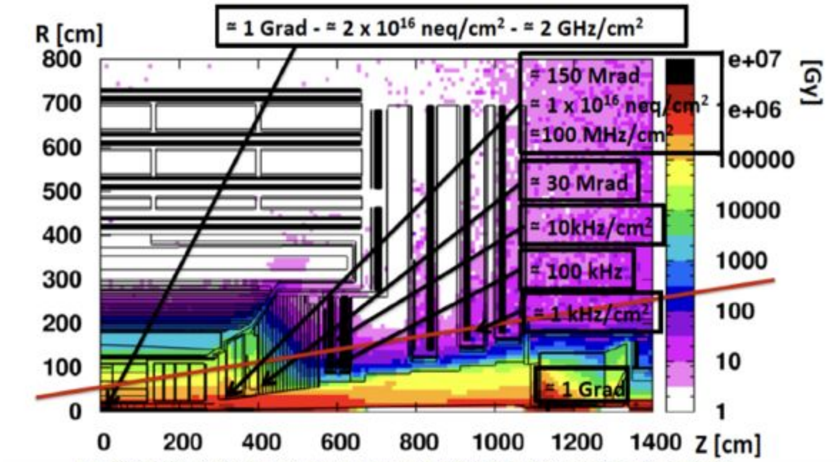
- ➢ Cumulative radiation exposure: radiation levels up to 2×10^{16} neq /cm^2
- ➢ Requires radiation hard detectors and electronics.

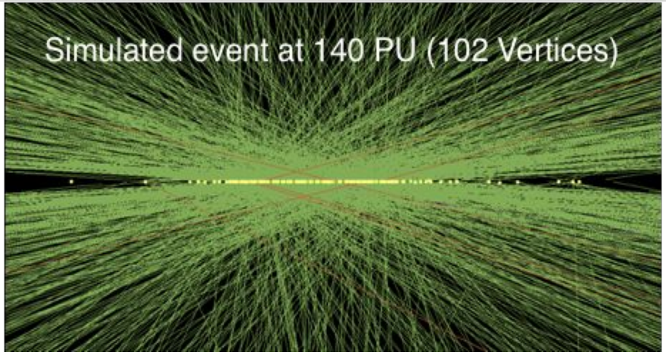
High instantaneous luminosity

- ★ High pileup (number of overlapping hard interactions per bunch crossing): <μ>=140-200
- ★ More particles per event
  - ○ Higher detector occupancy
- ★ Require higher granularity of detectors

Large data volumes to be processed

- ★ Require new trigger strategy and readout system upgrade.





Upgrades

- Full replacement - tracker, endcap calorimeters, electronics system of barrel calorimeters & of Muon system.
- Adding new technology - additional GEM detectors for Muon chambers, MIP timing detectors.

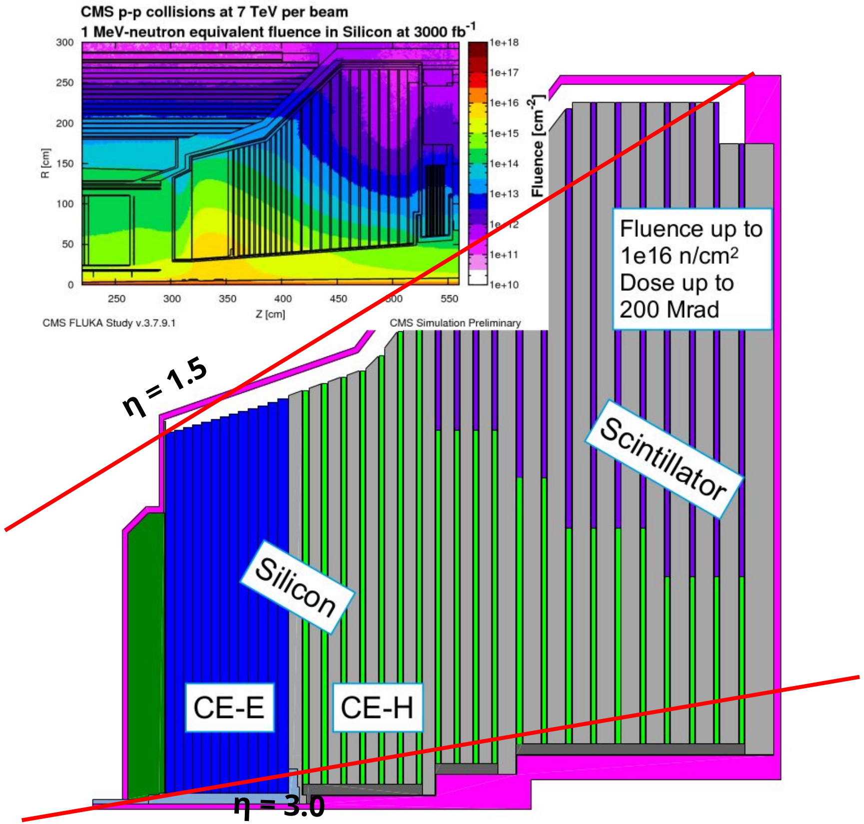
HGCAL - CMS endcap calorimeter for HL-LHC

Calorimeters are important for Jet energy measurement, E/gamma measurements, measurement of METs.

Salient features:

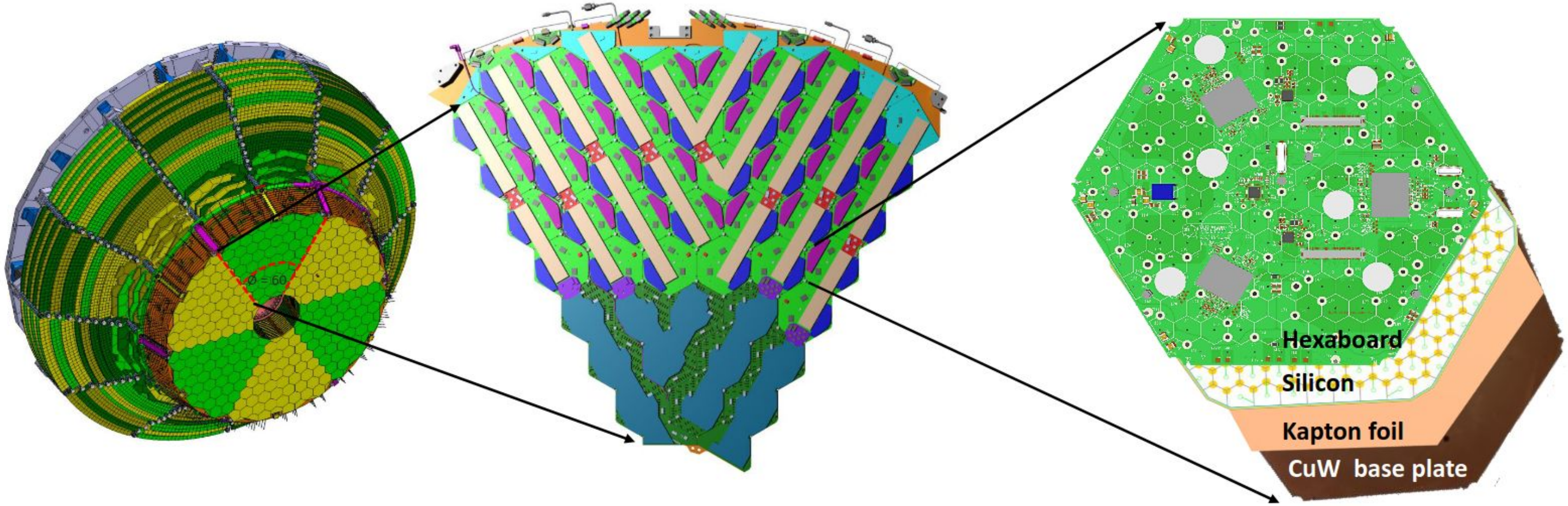
- **A sampling calorimeter with 5D measurements-**
  - Energy, position & time
- Si in high radiation regions and scintillators otherwise.
- **Electromagnetic section (CE-E):**
  - 26 sampling layers to maintain current EM resolution
  - Silicon as active & Cu/CuW, Pb as absorbers
- **Hadronic section (CE-H):**
  - 21 sampling layers
  - Silicon and Scint as active & steel as the absorber

- **A total of ~6M silicon channels & ~240k scintillator channels to be read out .**
  - Currently in endcap: ~14k ECAL crystals & ~2.3k channels in HCAL.
- Whole system is to be maintained at -35°.



Building blocks of HGCAL

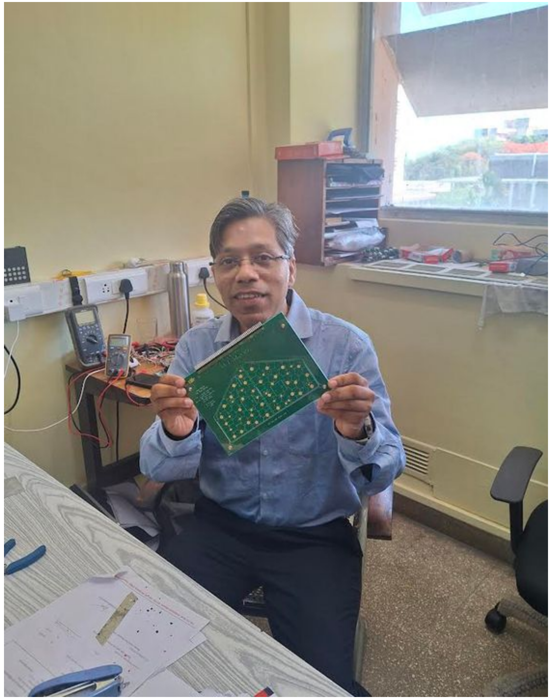
Silicon modules are tiled together to form a cassettes of HGCAL

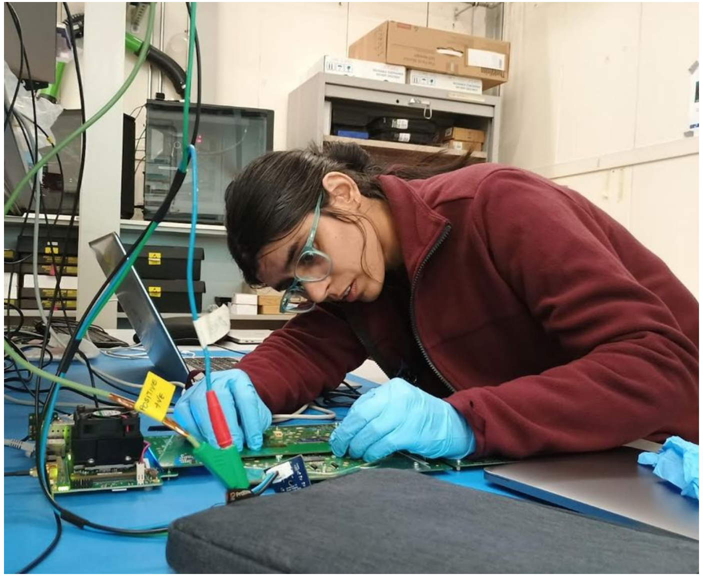


- ➔ Constructed in 60° (30°) cassettes in CE-E (CE-H) - each having silicon modules (Silicon modules + Scintillator tiles) along with hexaboards.

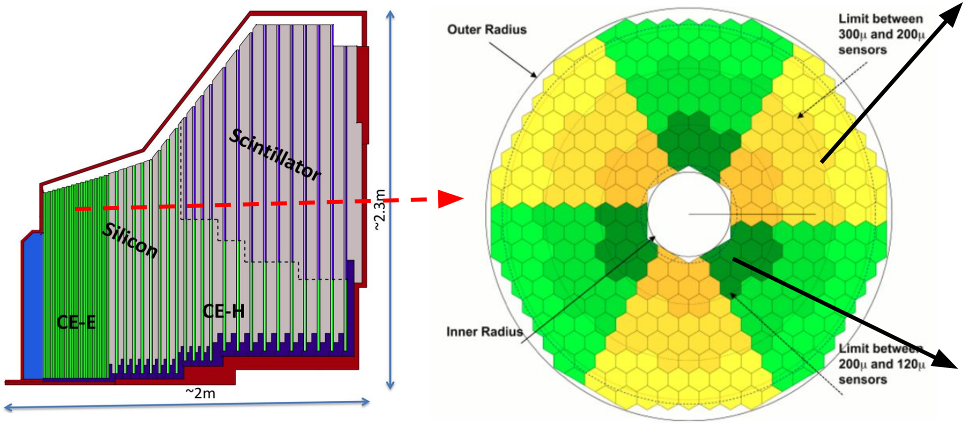
- ➔ Work going on different aspects of the detector:
  - ◆ Design, production of absorber & active elements.
  - ◆ Production of 26k silicon modules & 3.7k scintillator modules
  - ◆ Data acquisition system to handle over 6M channels, services, connectivity.

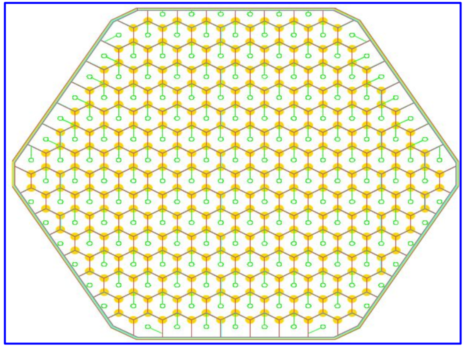
Large fraction of production of hexaboards boards in India

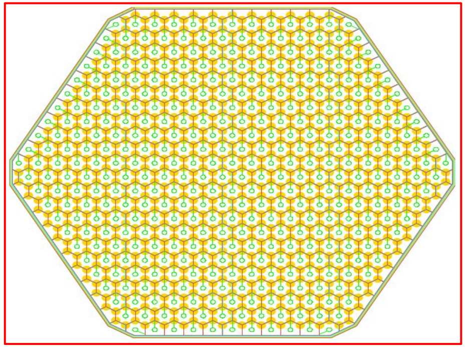




Challenges in handling large data rates







➔ Low Density Modules (LDM)
 ○ 192 channels with cell size = 1.18 cm²

➔ High Density Modules (HDM)
 ○ 432 channels with cell size = 0.54 cm²

➔ Scintillator tiles size ~4-30 cm²

- Sensors with thickness of 120, 200, 300 μm in very high radiation & less radiation region respectively.
- At the L1 accept rate of 750 kHz, the data will be read out at up to ~O(10) Tb/s.
  - Total latency at CMS - L1 trigger for HL-LHC ~12.5 μs.
  - ~1M trigger channels at L1 to be read out & make decision for HGCAL in ~5 μs.

Need to have high-bandwidth low-latency specialized computer hardware for off-detector data processing.

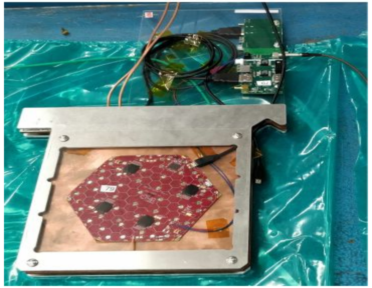
System testing & physics performance

- ➔ Build a large prototype testing
- ➔ Readout electronics
- ➔ Performance to EM showers
- ➔ Performance to Hadronic showers
- ➔ GEANT4 simulation validation

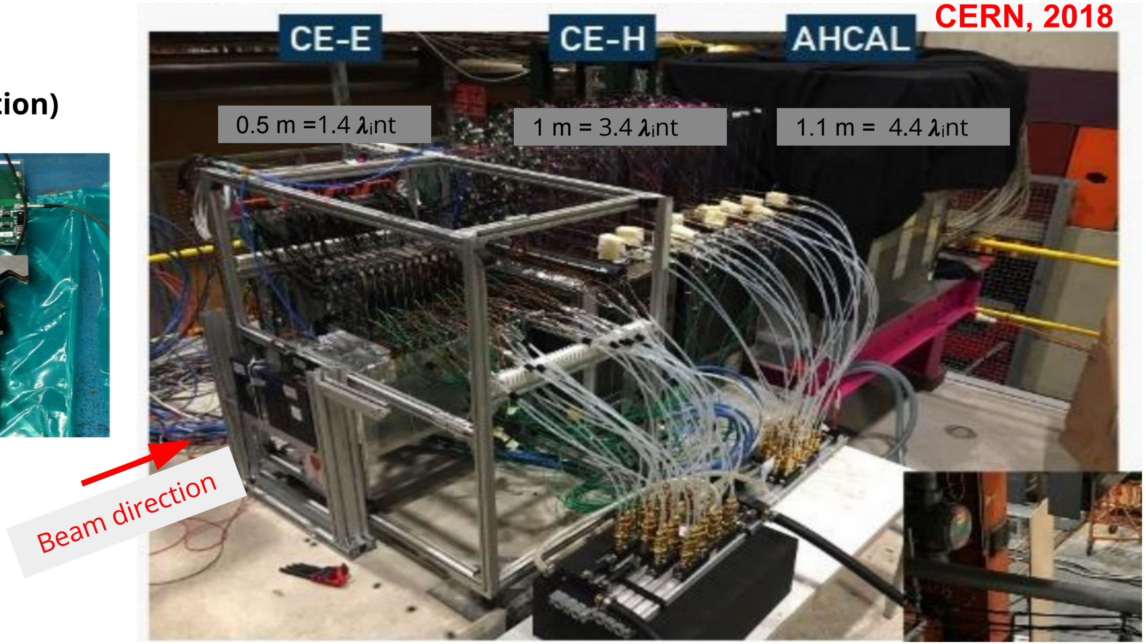
Beam Test Experiments with HGCAL Detector Prototype

A prototype HGCAL detector setup comprised of Si based electromagnetic (CE-E) and a hadronic (CE-H) sections followed by scintillator-tile based AHCAL was exposed to e^+ and π^- beams of energies ranging from 20 – 300 GeV.

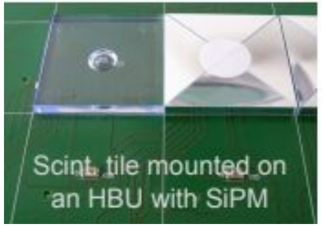
CE-E (EM section)



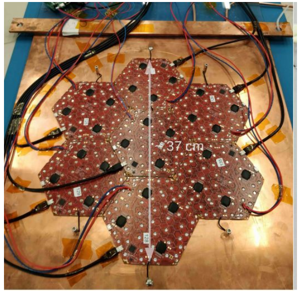
H2 beamline at SPS, CERN



CALICE AHCAL



CE-H



Detector Setup and Its Simulation

CE-E

- Si sensors (1.1 cm^2) + Cu/CuW & Pb absorbers
- 28 sampling layers
  - one module per layer
- Lateral dimensions (x,y) = (15 cm x 15 cm)

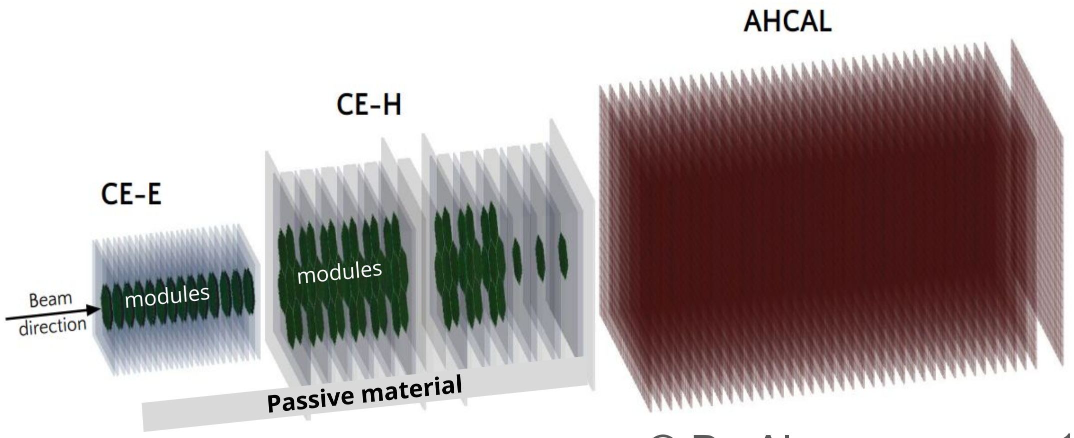
CE-H

- Si sensors (1.1 cm^2) + Cu/CuW & Steel absorbers
- 12 sampling layers
  - 7 modules per layer, arranged in a daisy structure
  - last three layers have only one module
- Lateral dimensions (x,y) = (37 cm x 37 cm)

CALICE AHCAL

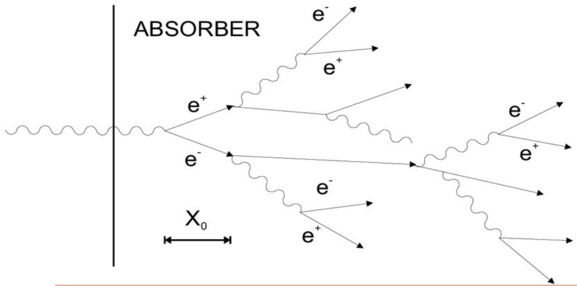
- Scintillators on SiPMs (3 × 3 × 0.3 cm^3) + Steel absorbers
- 39 sampling layers
- Lateral dimensions (x,y) = (72 cm x 72 cm)

- A total of 12k & 22k readout channels in Si HGCAL prototype & Scint. AHCAL prototype.
- The detector setup, along with absorbers and electronics board material, is simulated using GEANT4.10.4.p03.

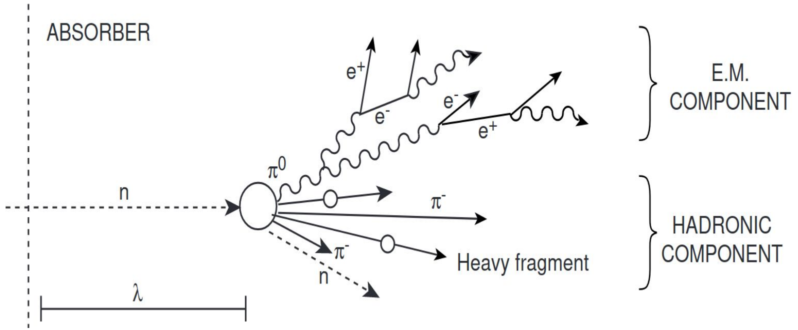


Electromagnetic & hadronic showers

- High energy $e^{\pm}$ & photons produce EM showers of secondary $e^{\pm}$, photons.
  - Response of calorimetry is linear with energy.
  - Large particle multiplicity in EM showers.
- Shower scaling variable - radiation length (X_0), Moliere radius (ρ_m)



For lead (Pb) : X_0 = 0.56 cm, ρ_m = 1.6 cm,
For steel : X_0 = ~1.75 cm, ρ_m = ~1.71 cm,

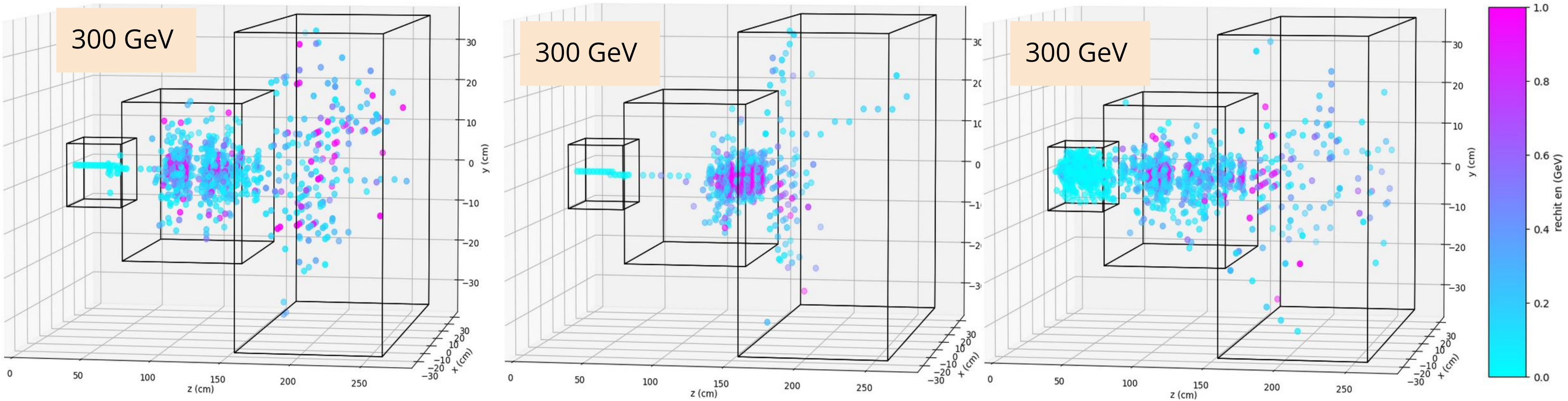


Shower scaling variable: interaction length (λ^{nuc}_{int})
For steel: = ~17 cm

- ➢ **Hadronic component:** energy lost in breaking up of nuclei → no corresponding signal in calorimetry (Invisible energy).
  - Response is less than 1.
- ➢ Shower development of **EM component** is as the EM showers
  - Fraction of energy carried increase with incident hadrons energy.
  - Results into non-linearity of response.

Event displays of pion showers in HGCAL beam test prototype

- ➢ Rechits (reconstructed cell hits) are plotted according to their (x,y,z) position.
- ➢ Colour represents the relative energy of these rechits .

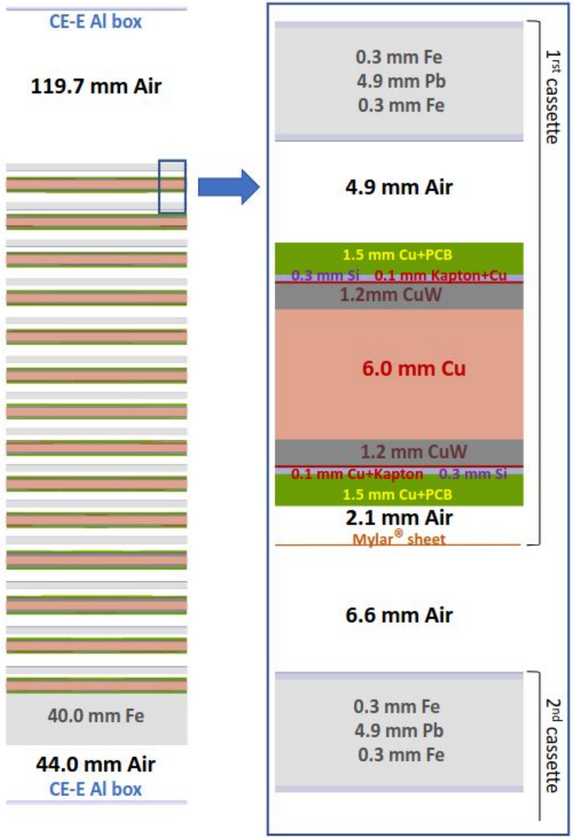


- ➢ Transverse and longitudinal leakage → contribute to fluctuations in pion showers.
  - ○ Mostly due to single modules used in CE-E and last three layers of CE-H.

Energy reconstruction of positron @HGCAL-TB

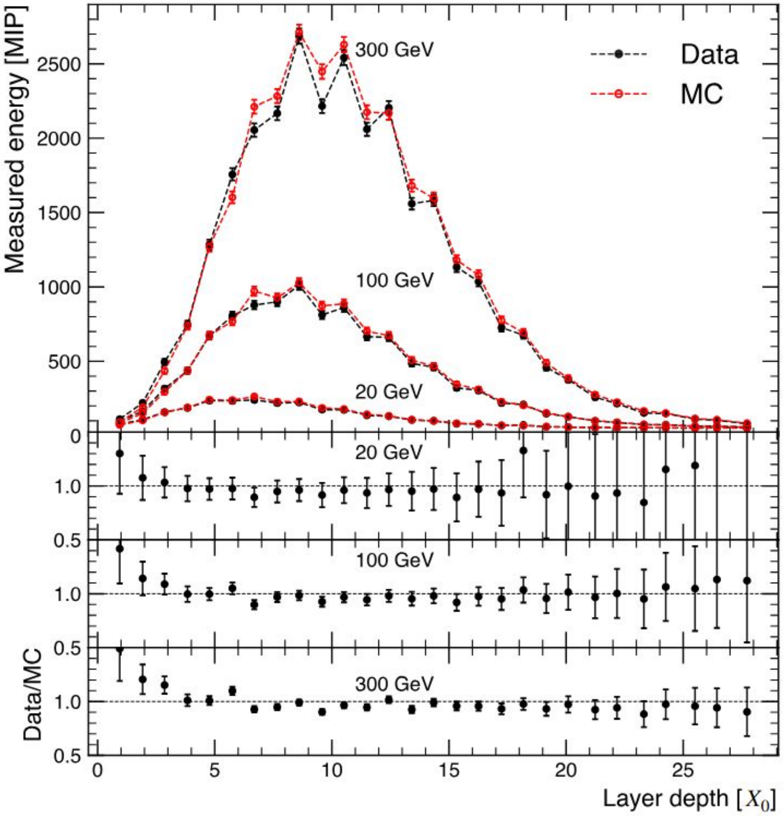
1. Using classical method (interpolation)

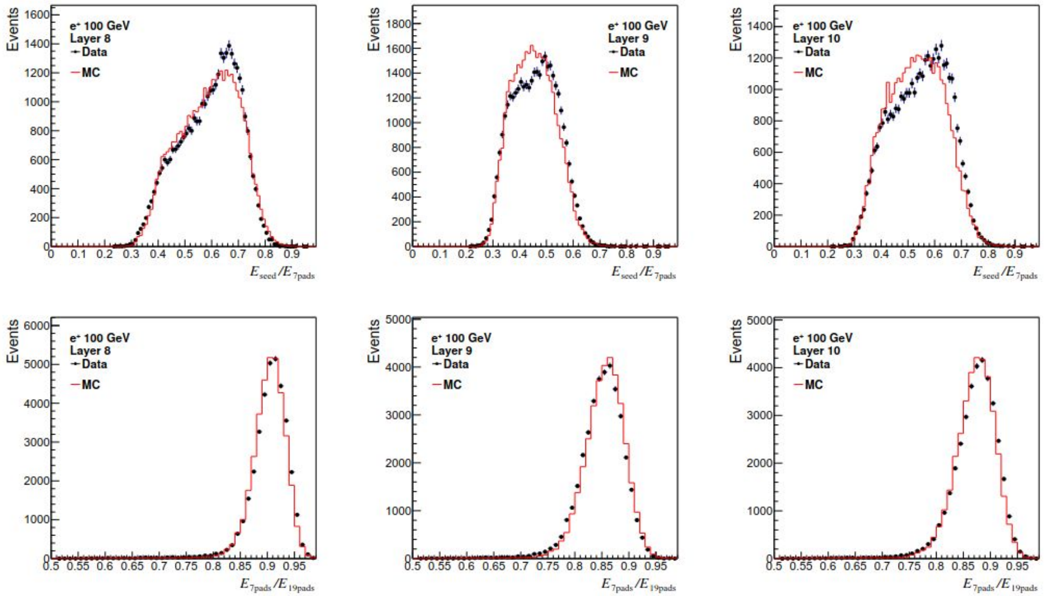
The EM-section of the prototype



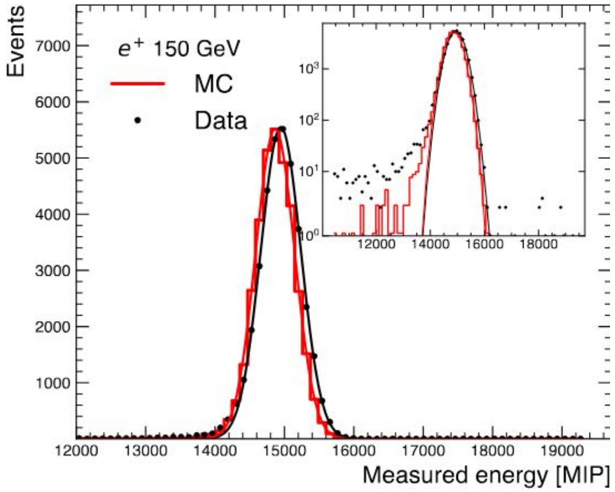
Cassette /Layer h/i		Energy Loss ΔE_i^{Si} [MeV]	Depth z_i $[X_0]$	Absorber Thickness Δz_i $[X_0]$	Energy Loss $\Delta E_i^{\mathrm{Abs}}$ [MeV]	Average $(\Delta E_i^{\mathrm{Abs}} + \Delta E_{i+1}^{\mathrm{Abs}})/2$ [MeV]
1	1	0.085	1.00	1.00	10.2	11.29
	2	0.085	1.98	0.98	12.3	9.85
2	3	0.085	2.92	0.94	7.4	9.85
	4	0.085	3,90	0.98	12.3	9.85
⋮	⋮	⋮	⋮	⋮	⋮	⋮
10	19	0.085	18.23	0.94	7.4	9.85
	20	0.085	19.21	0.98	12.3	9.85
11	21	0.085	20.15	0.94	7.4	11.36
	22	0.085	21.30	1.15	15.4	11.36
12	23	0.085	22.23	0.94	7.4	11.36
	24	0.085	23.38	1.15	15.4	11.36
13	25	0.085	24.31	0.94	7.4	9.85
	26	0.085	25.29	0.98	12.3	9.85
14	27	0.057	26.23	0.94	7.4	9.85
	28	0.057	27.21	0.98	12.3	6.17

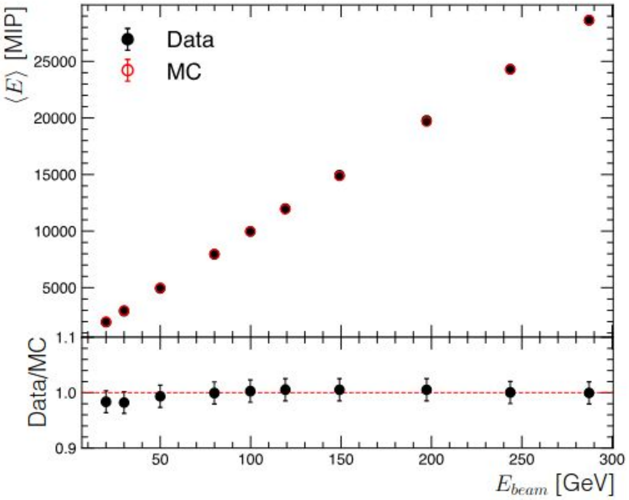
EM Shower Shape and Profile

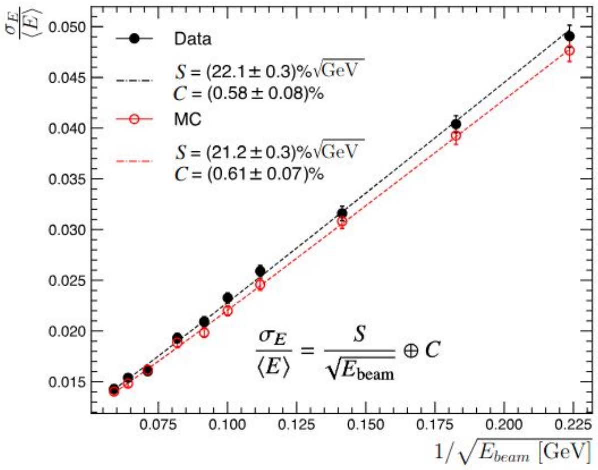




Calorimetric response







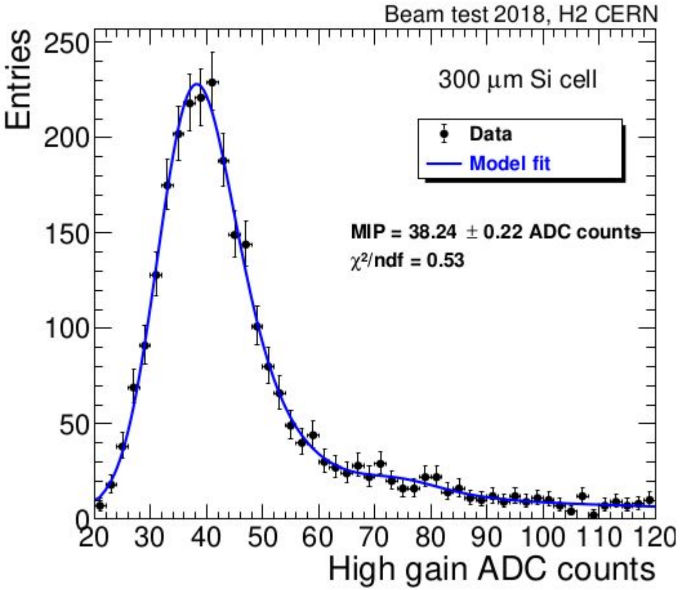
Energy reconstruction of charged pions @HGCAL-TB

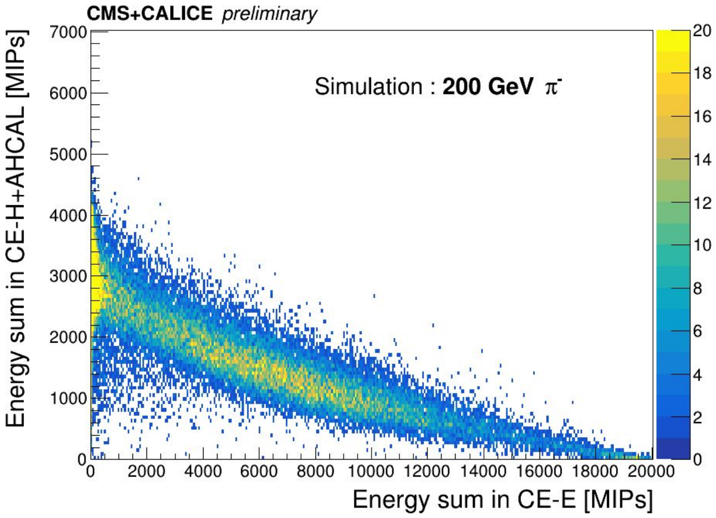
1. Using classical method (χ^2 method)
2. Using DRN (a type of graph neural network)

Energy scale for pions

S. Pandey et.al

- ➢ The digitized signal measured for muons is used to set the scale for a MIP.
- ➢ Energy of particles is presented in the units of number of MIPs.





- ➢ Set the energy scales in GeV :
  - 100 MIPs per GeV in CE-E using 50 GeV e+ beams
  - 12 MIPs per GeV in CE-H+AHCAL using 50 GeV π^-
- ➢ Energy reconstructed as "Detector level calibration" or "fixed weights".

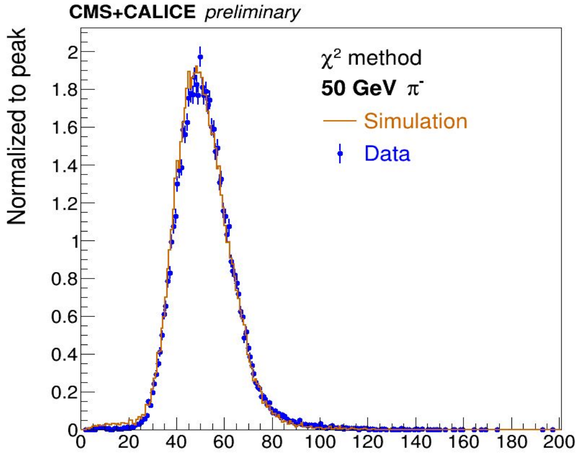
Energy reconstruction of pions - conventional ways

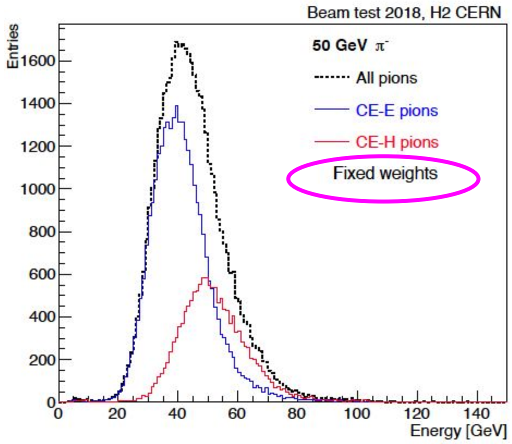
S. Pandey et.al

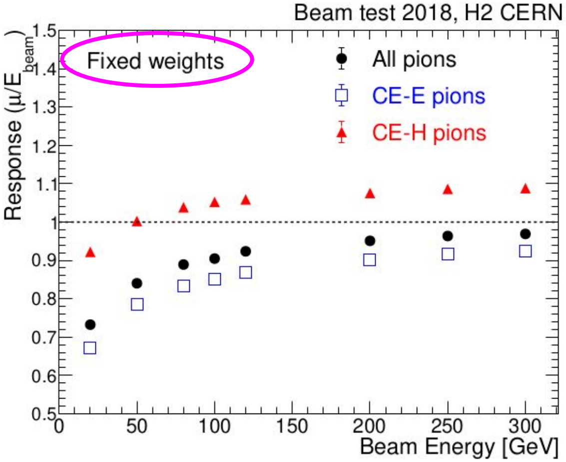
- χ^2-method : energies deposited in CE-E, CE-H and AHCAL are combined using energy dependent weight factors extracted after minimizing an estimator defined as,

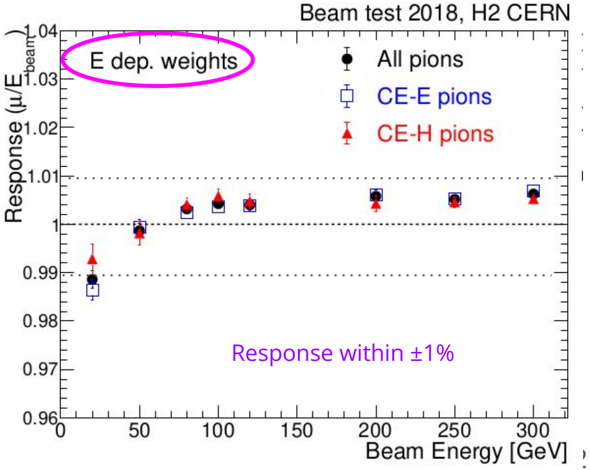
$$\chi^2 = \sum_{pions} \frac{(E_{beam} - E_{corr})^2}{\sigma^2(E_{fix})}$$

- Doesn't compensate for event-by-event fluctuations of showers.
- Used as a published reference.

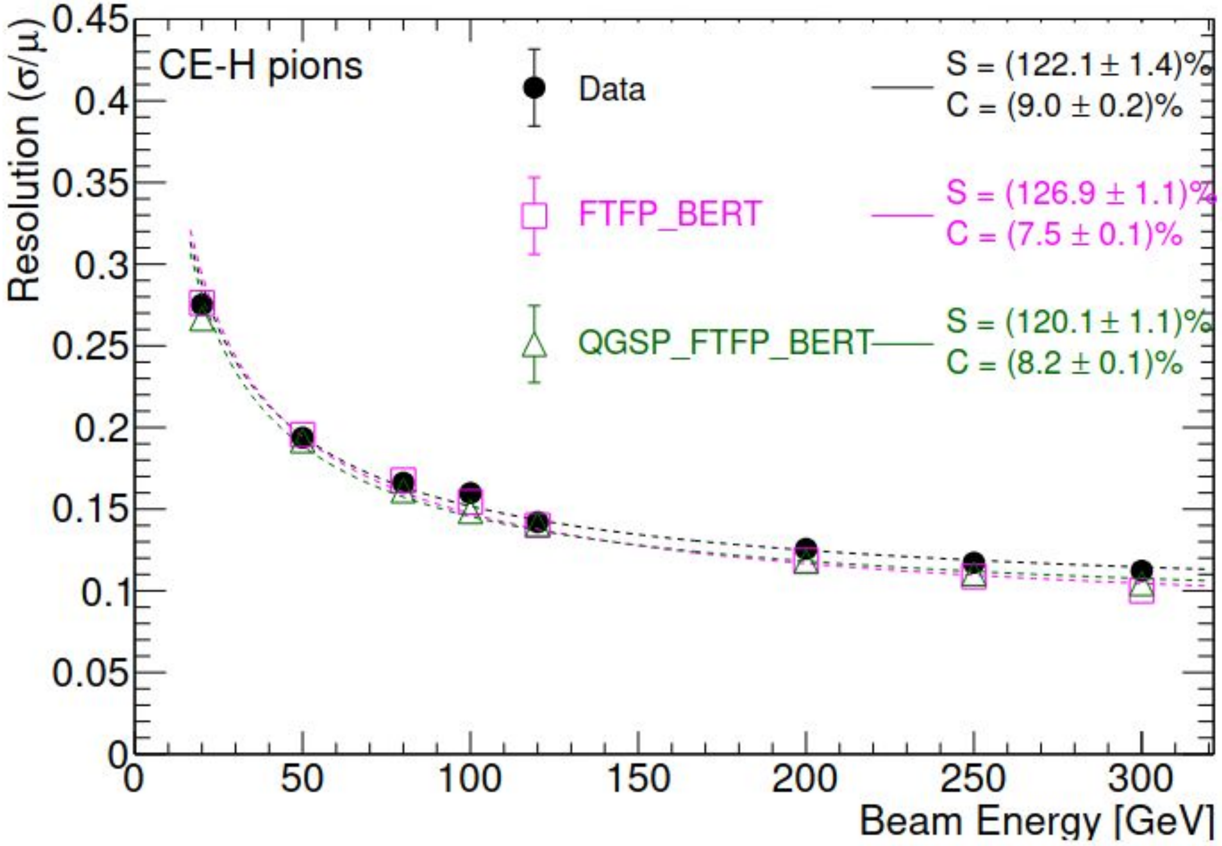








Benchmark relative energy resolution for charged pions



TUTORIAL

Open either dataset

```
import numpy as np

import pandas as pd

import h5py

import matplotlib.pyplot as plt

import awkward as ak


train_dataset = h5py.File('hgcal_electron_data_0001.h5', "r")

nhits = np.array(train_dataset['nhits'])

target_energy = np.array(train_dataset['target'])
```

The dynamic reduction network

The Dynamic Reduction Network

Arxiv: 2003.08013v1

https://github.com/lgray/hgcal_ldrd

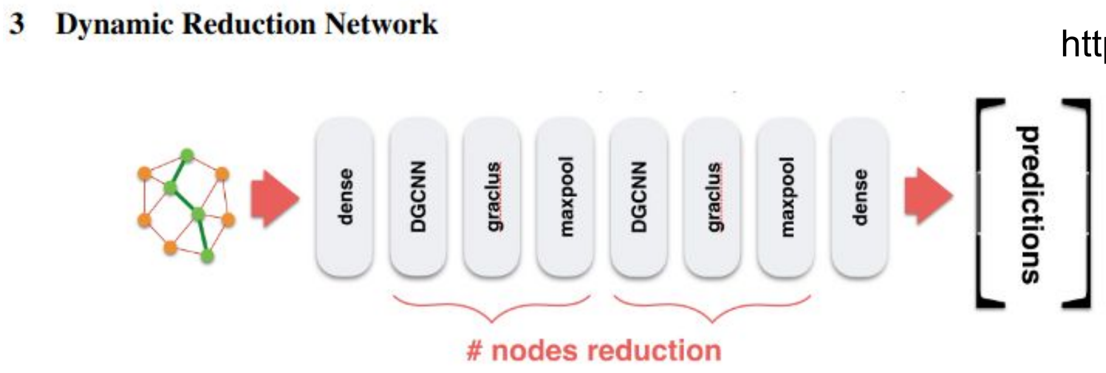
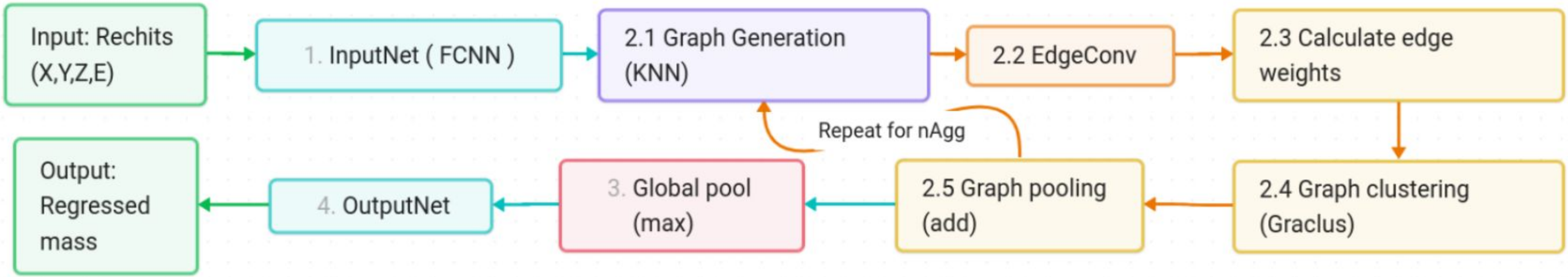


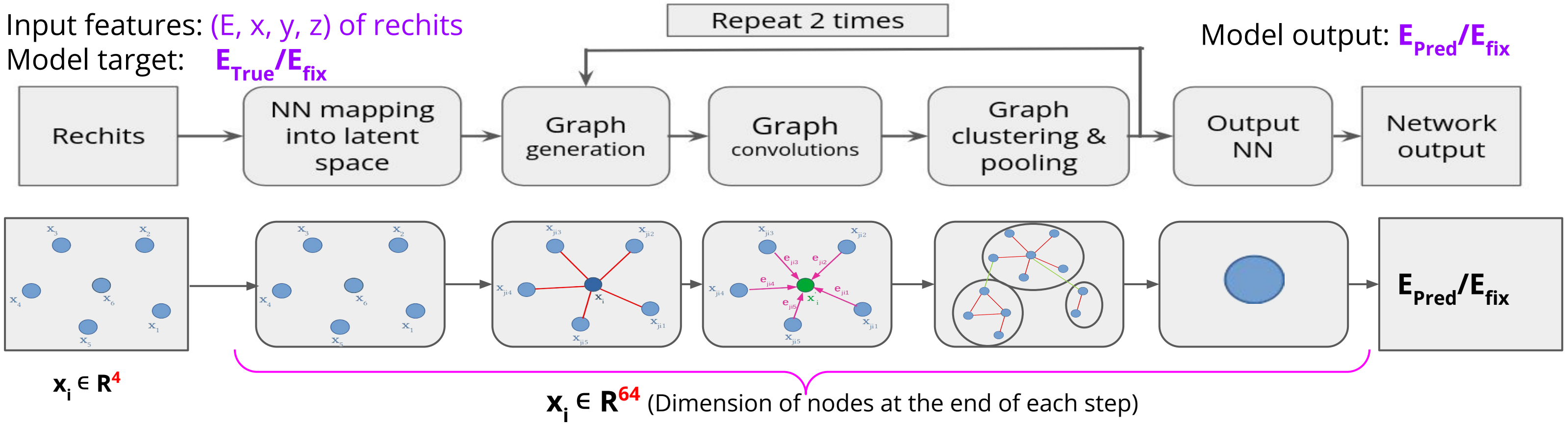
Figure 1: The basic workflow of the Dynamic Reduction Network, details given below.



Dynamic Reduction Network (DRN) for HGCAL

(UMN, IISER, TIFR)

- Input features are mapped onto a higher dimensional latent space & add clustering, pooling step to learn high level information iteratively.

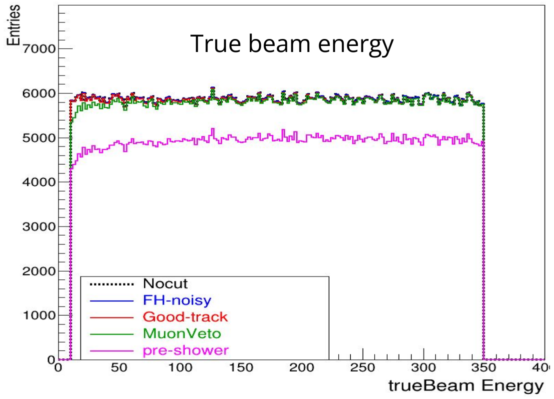


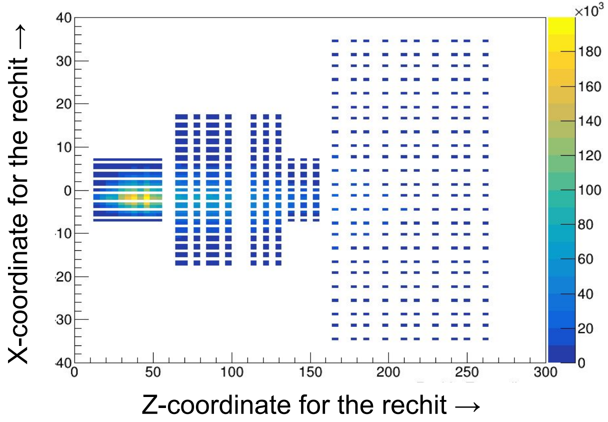
➢ A total of 63k parameters to learn in the model. The most expensive step is the graph convolutions, followed by construction of the nearest neighbors graph.

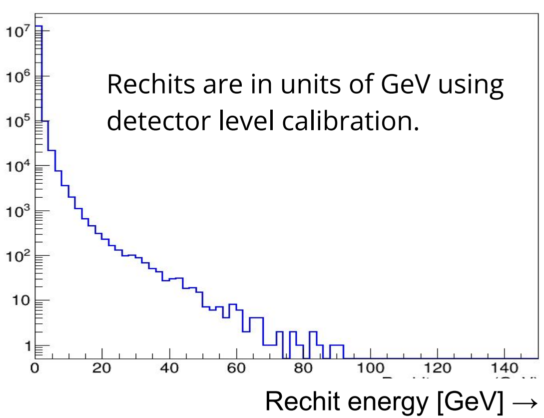
➢ The model is trained on a flat energy sample of 10-350 GeV with a total of 4.1M events simulated using GEANT4.10.4.p03 and FTFP BERT EMN hadronic physics list.

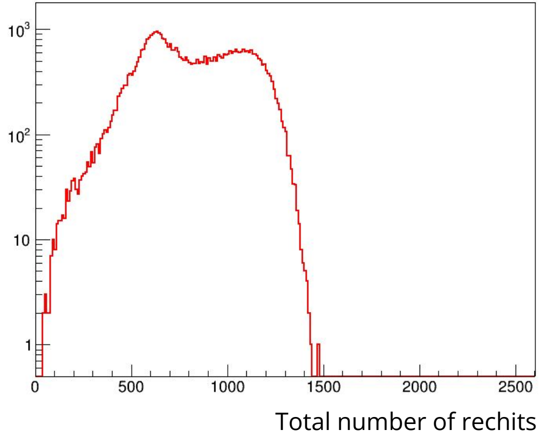
➢ Training time ~2:30 hrs which reduces further upto 30 mins if you train your model parallelly using data parallelism on GPUs.

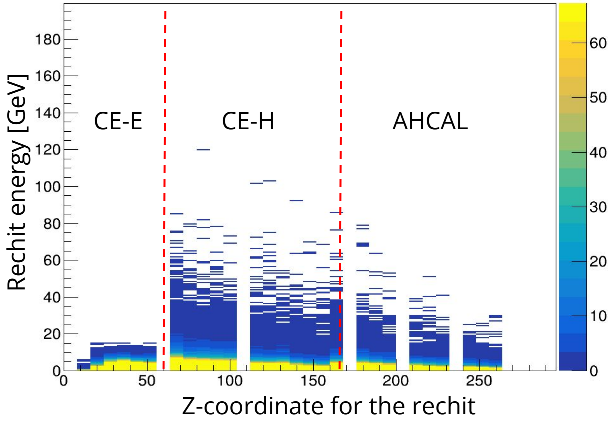
Input features & target used for training the model

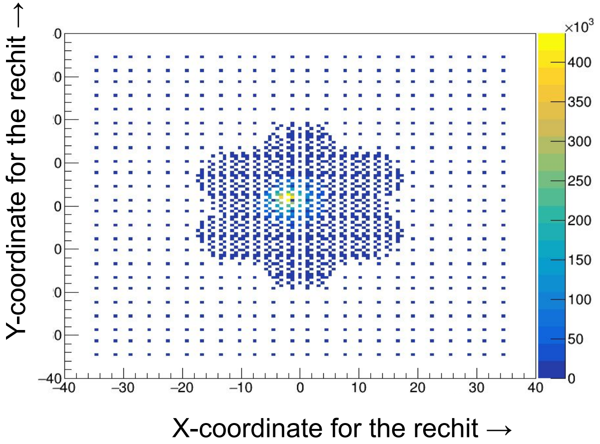






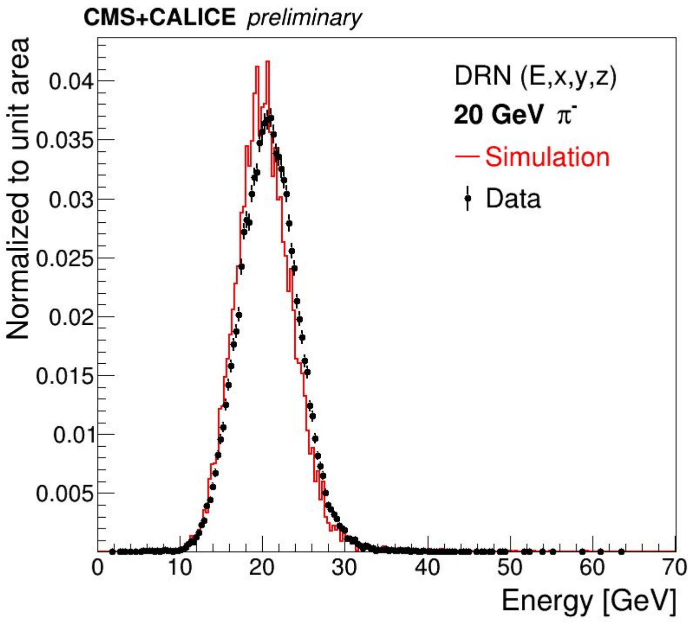


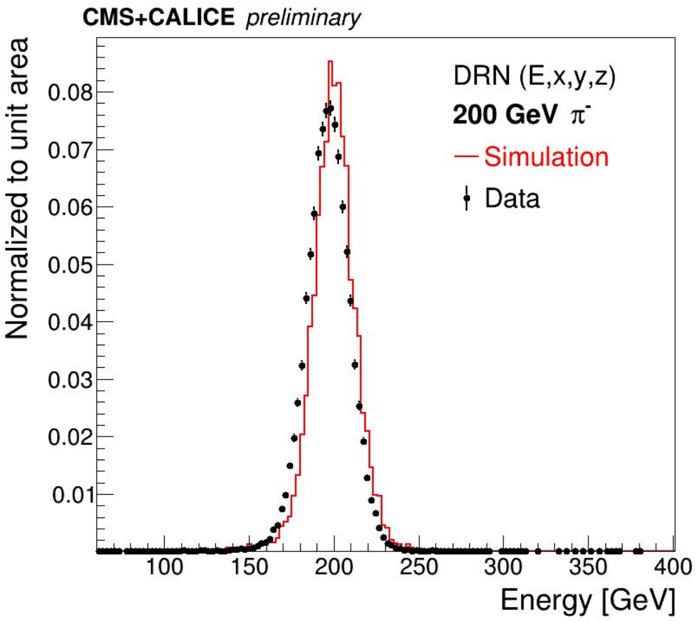




Energy reconstruction of charged pions using DRN

DPS-2022-022

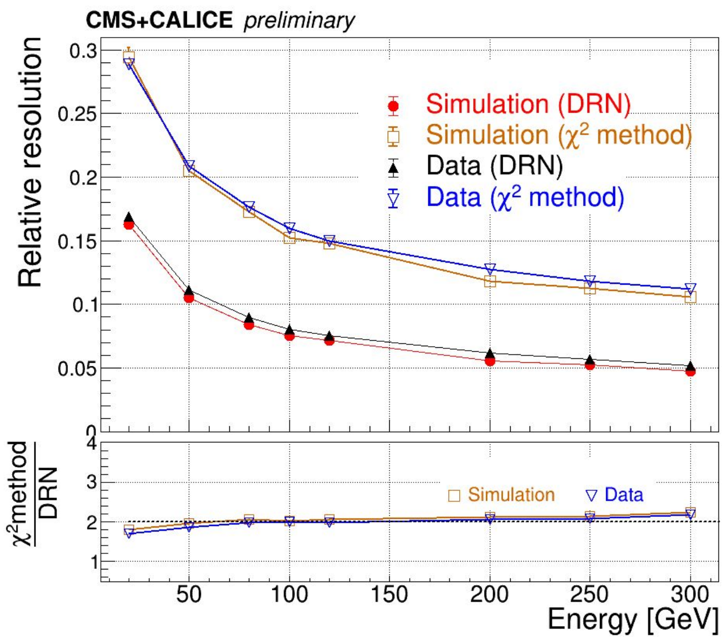


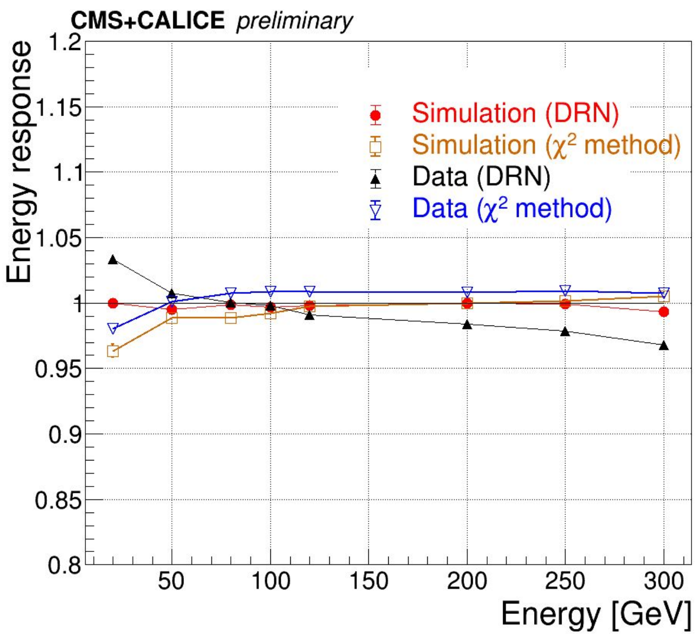


The bulk of distribution predicted by the simulation matches well with data (on average within few %).

Energy resolution & response of charged pions

DPS-2022-022





- ➢ Significant improvement in energy resolution using DRN as compared to simple χ^2 method at all the energies.
- ➢ **Similar performance in real data**

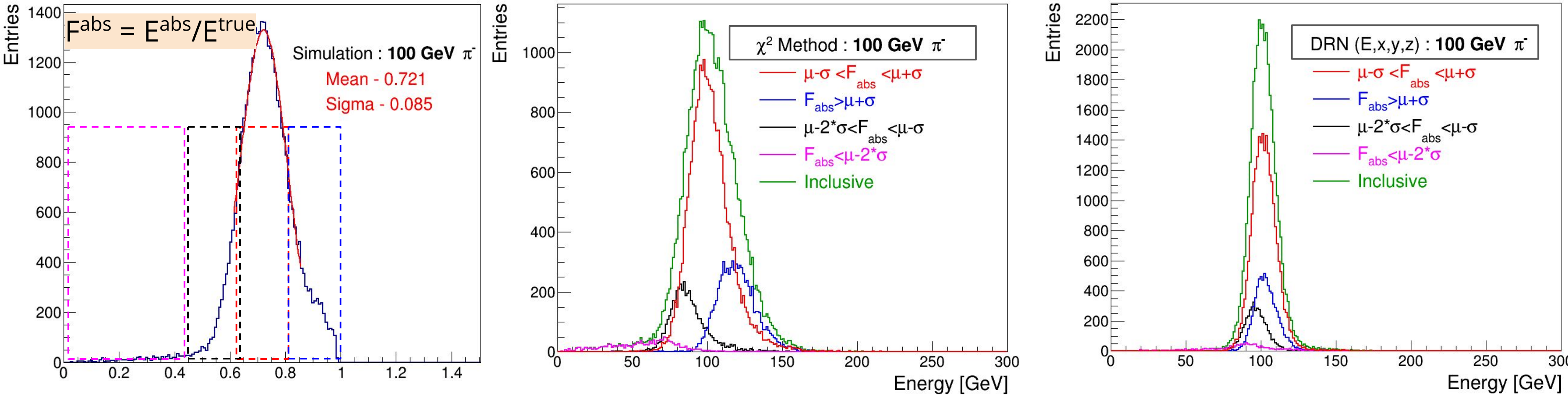
Understanding the gain in resolution using DRN

Opening the black box

- Using GEANT4 truth information
- Contribution of different features used in training.
- Effect of calibration methodology on DRN
- Leakage effects

Energy reconstructed using DRN & X^2 in different absorber bins

For a given true energy, large variation in (visible) energy deposited in absorbers, hence that sampled by the active layers.

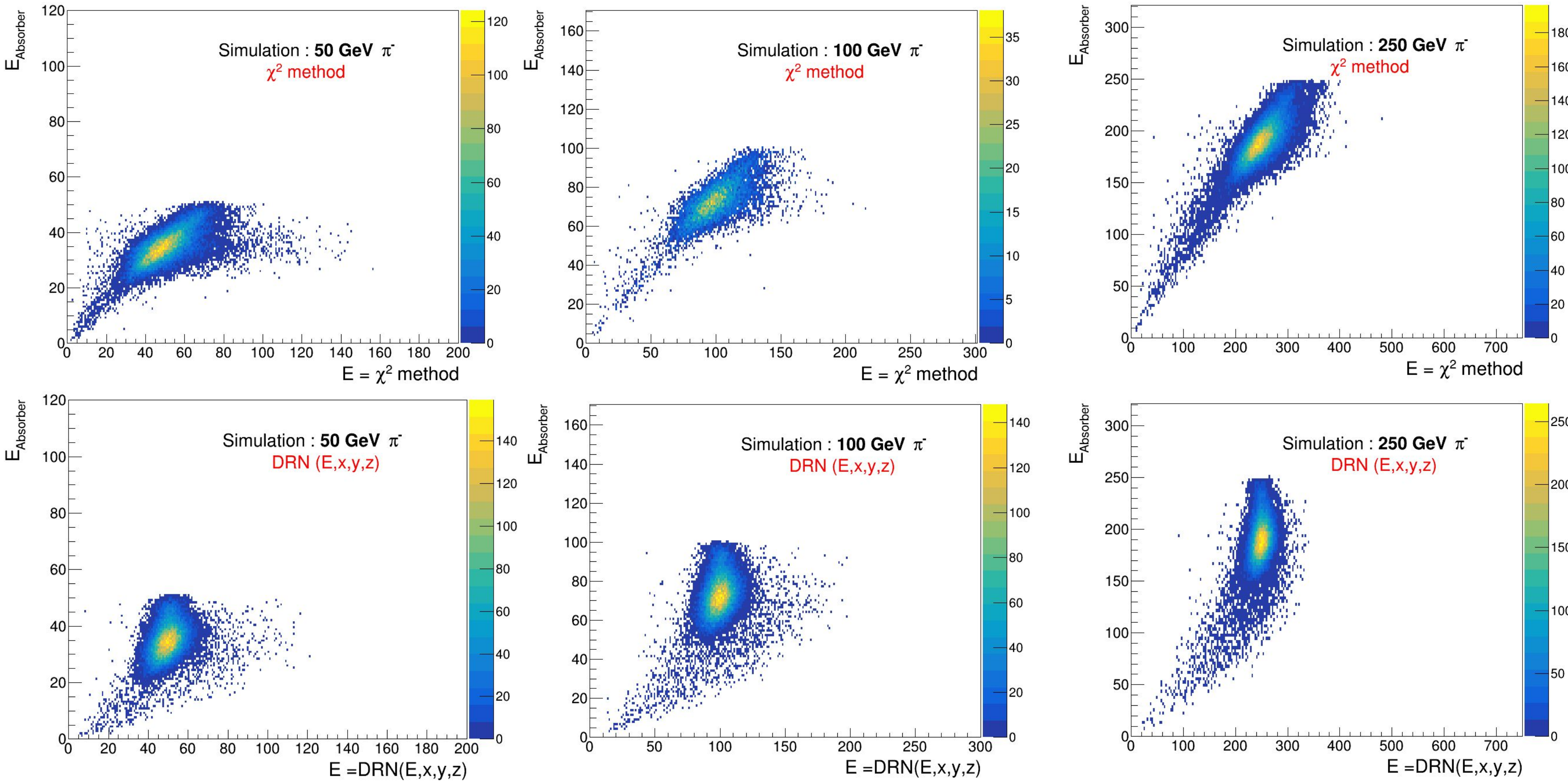


In the bins of energy deposited in absorbers,

➢ DRN response is similar while **X^2** shows large variation resulting in overall degraded resolution.

  ○ Same X^2 weights are applied to all showers of a given pion true energy.

Predicted energy as a function of absorber energy

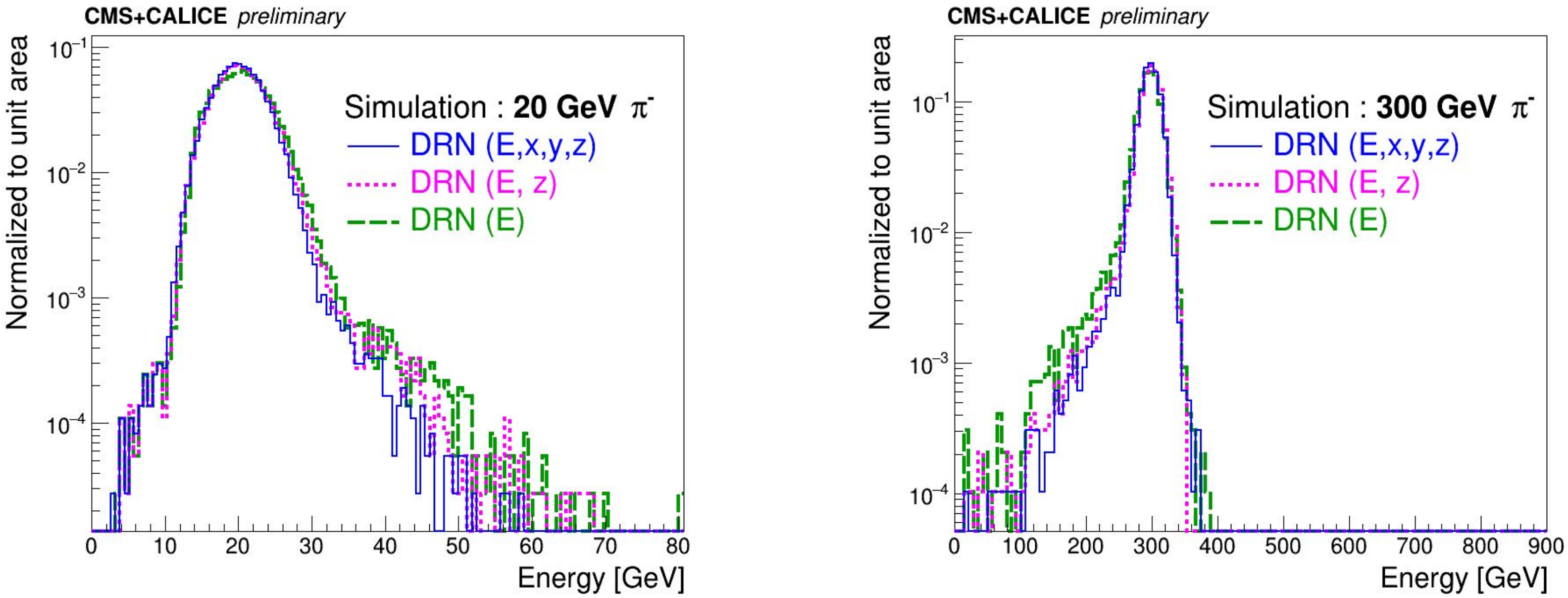


Impact of input features on the DRN learning

Training DRN with different input features

DPS-2022-022

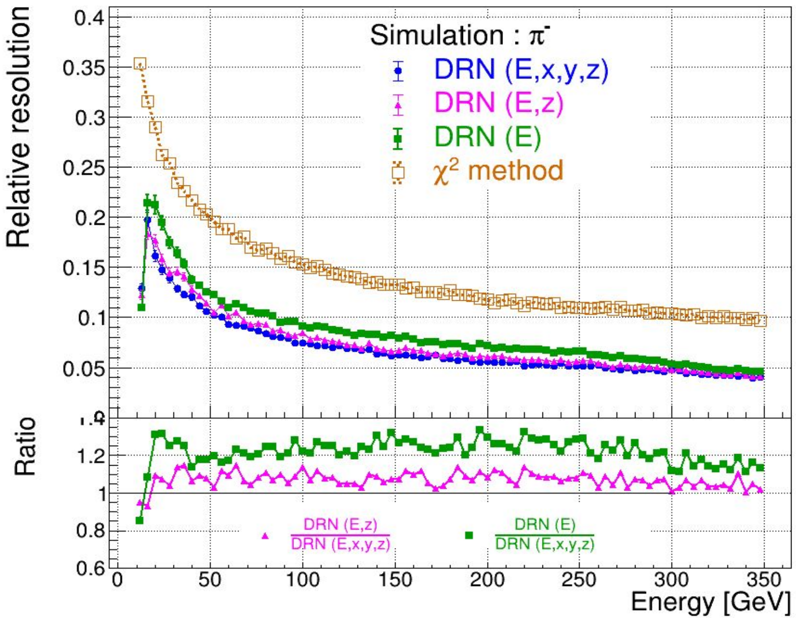
- We considered three combinations of these input information:
  - DRN (E), DRN (E, z), DRN (E, x, y, z)
  - z is the longitudinal coordinate & x-y are the transverse coordinates.

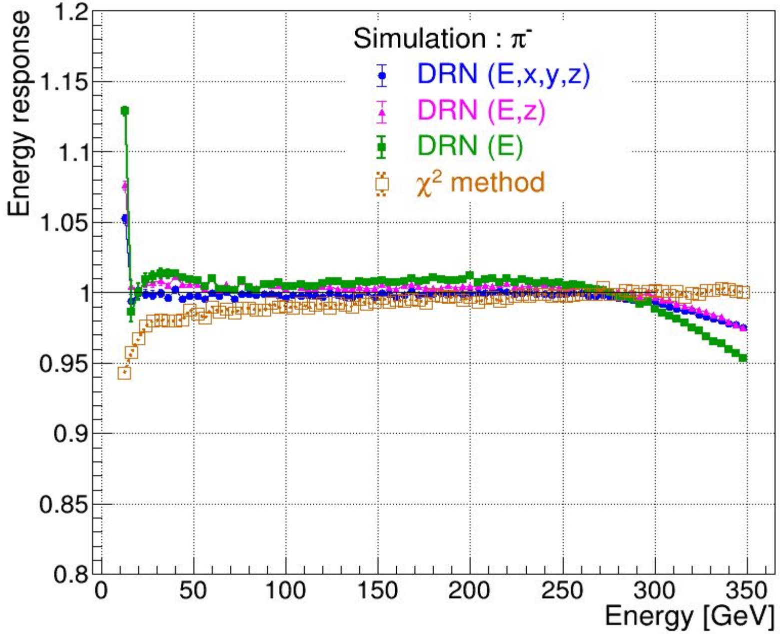


By increasing the input features the resolution improves overall and the tails are reduced.

Improvement in energy resolution with different DRNs

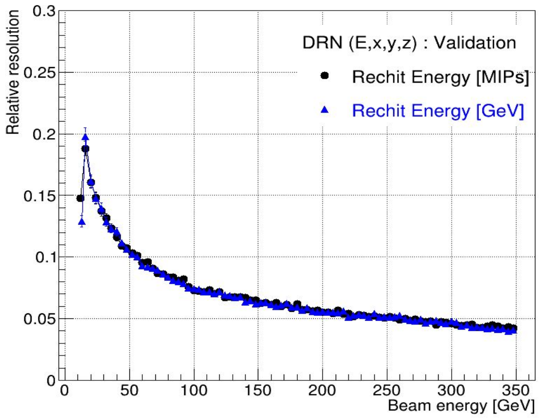
DPS-2022-022

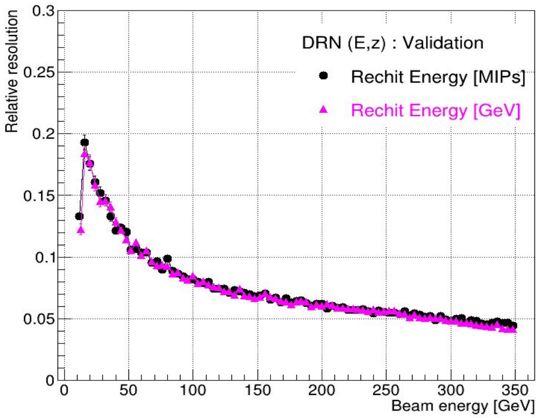




- ★ DRN(E) is effectively learns the event-by-event fluctuations in pions showers.
- ★ Further more improvement in resolution (20-30%) by providing spatial information
  - ○ Learns about transverse leakage along with longitudinal leakage.

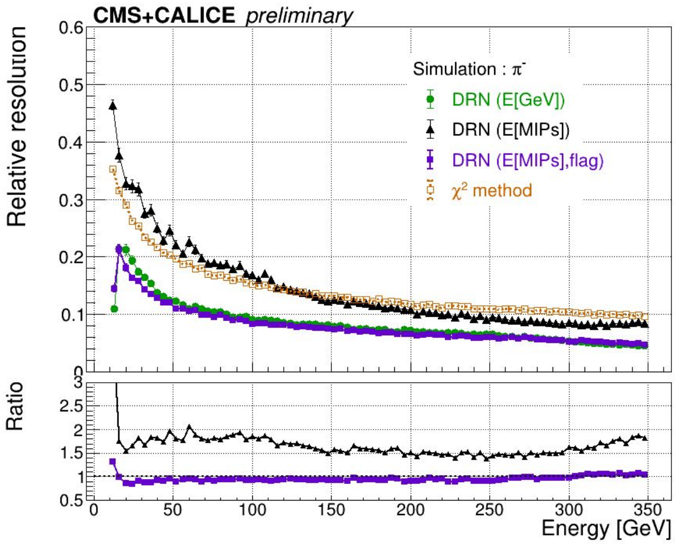
Impact of detector level calibrations on DRN's performance





➔ Providing spatial information of rechits along with energy makes model independent of the energy calibrations used (Efix or MIPs).

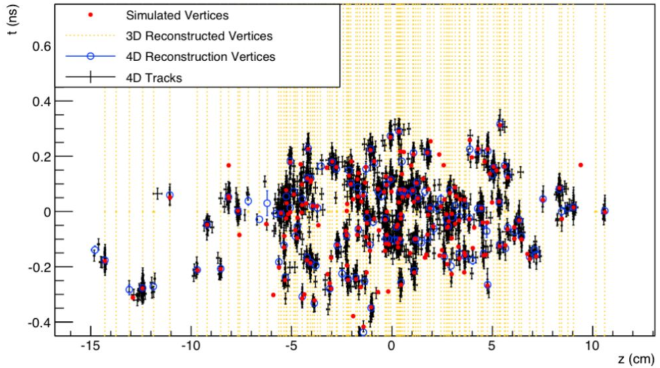
Model does learn about the rechit's positions across the detector with the help of three fixed weights used for MIP → GeV conversion.



BACKUP

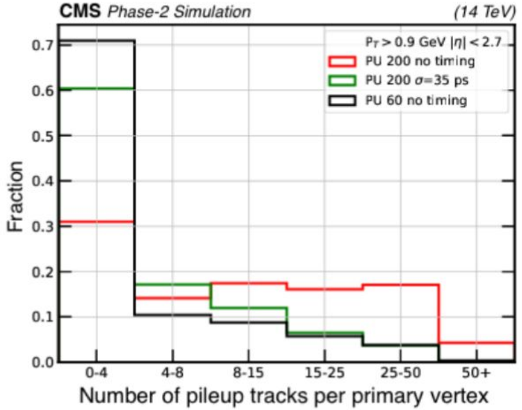
Precision timing as a tool to mitigate pileup

- Key idea: Under HL-LHC conditions timing information can be used to "separate" pileup vertices that otherwise appear to be "merged" in 3D space.
- This separation improves with the time resolution: ~30 ps resolution is already effective.
- The upgraded Barrel ECAL will provide precision timing information for high energy photon showers.
- The new Endcap calorimeter will provide precision timing information for both high energy photons and hadrons.
- The MIP Timing Detector (MTD) will be used at the HL-LHC for precision timing of all charged particles.



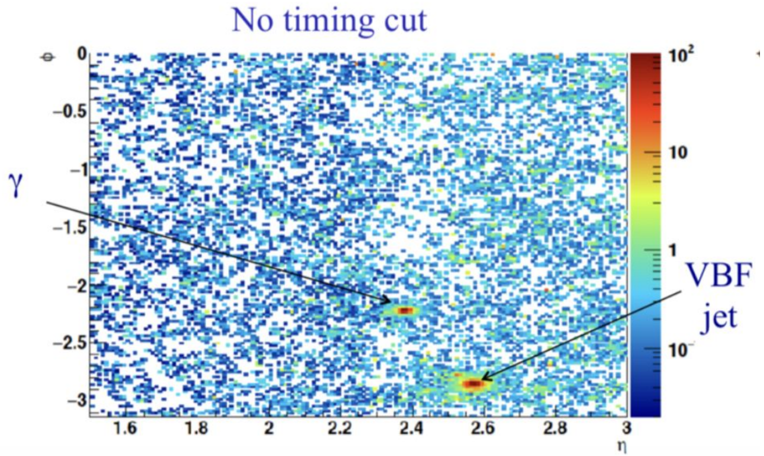
Collision vertices within a bunch crossing at 200 PU.

Pileup tracks incorrectly associated to the interaction vertex



Impact of 5D measurements on physics performance

- Although CMS uses silicon based preshower detector with 2 planes. However, this is the first silicon based calorimeter of such a scale.
- Measurements in 5D can be done with HGCAL:
  - Energy measurement:
    - Dynamic range: 0.2 fC - 10 pC (corresponds to an energy range of ~ TeV)
  - Spatial granularity (x, y and z): Cell sizes from 0.5 to 30 cm^2
  - Timing measurement: **O(25 ps)** per channel energy above O(10) MIPs Essential to mitigate pileup at HL-LHC



VBF H→γγ event with 1 photon and 1 jet in the same quadrant

Shown for cells with energy > 3.5 MIP

Timing cut can reduce most of the PU in the event