上海交通大学
SHANGHAI JIAO TONG UNIVERSITY

Study of residual artificial neural network for PID using the CEPC AHCAL Prototype

Siyuan SONG
Shanghai Jiao Tong University

The 2023 International Workshop on the CEPC from Oct 23-27

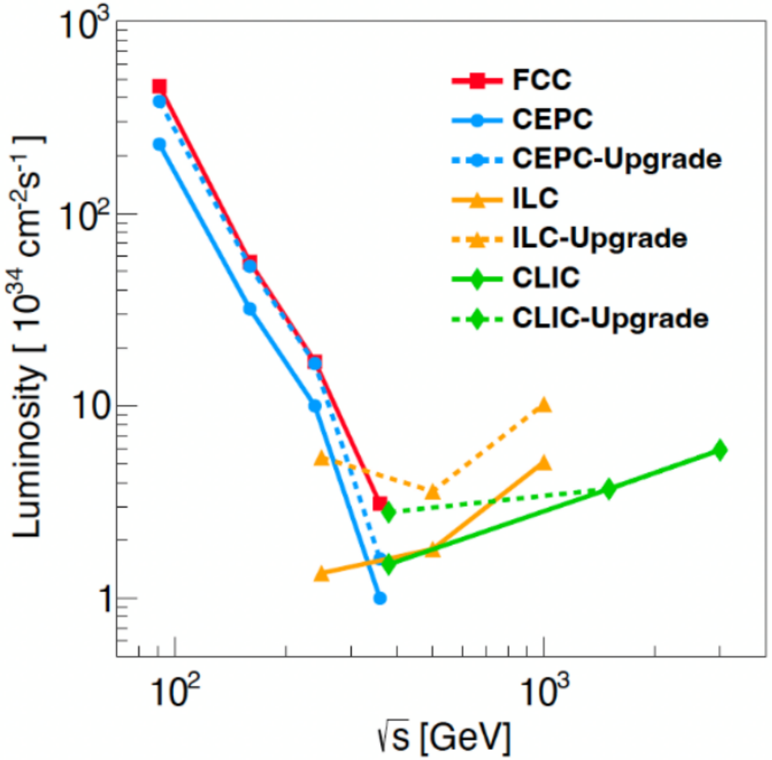
饮 水 思 源 • 爱 国 荣 校

Outline

- **Motivation**
- **MC samples**
- **PID based on BDT**
- **PID based on ANN**
- **PID application on Beam data**

Motivation

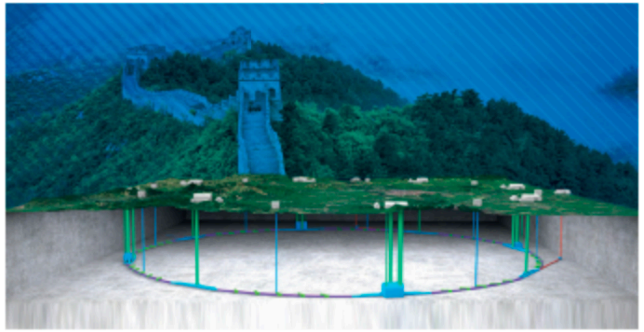
- **Future Higgs/W/Z/Top factories require high jet energy resolution**
  - Particle identification is essential for jet tagging and measurement.
  - Accurate differentiation between hadronic and electromagnetic showers is crucial for determining the energy and type of each particle.



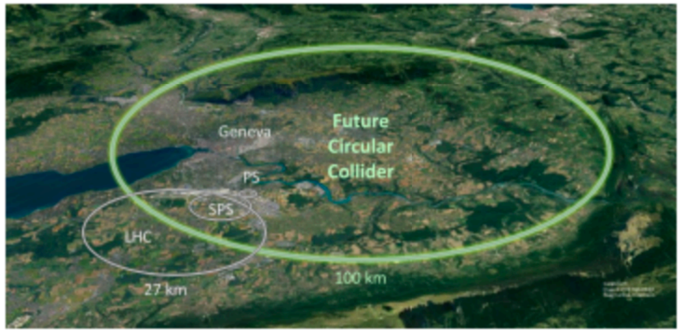
ILC in Japan

CLIC at CERN

CEPC in China

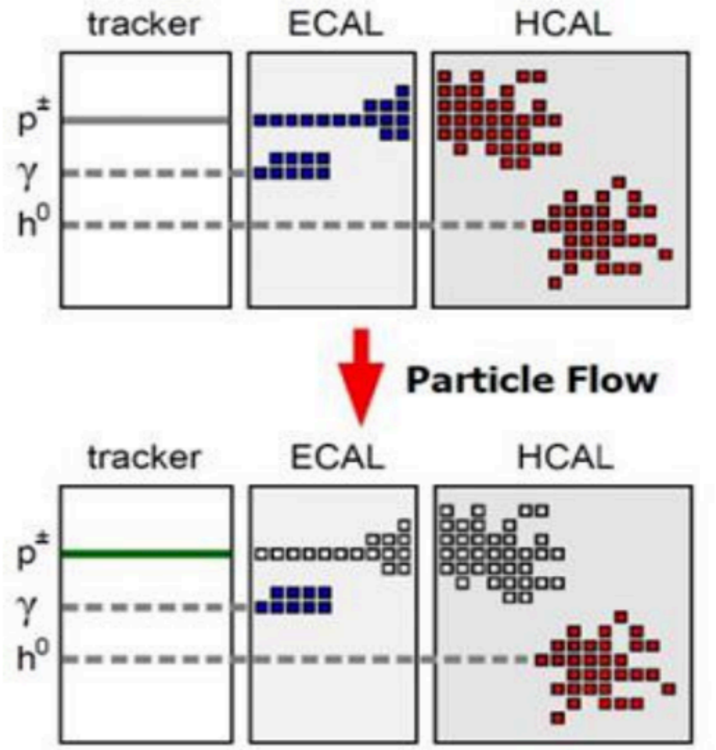


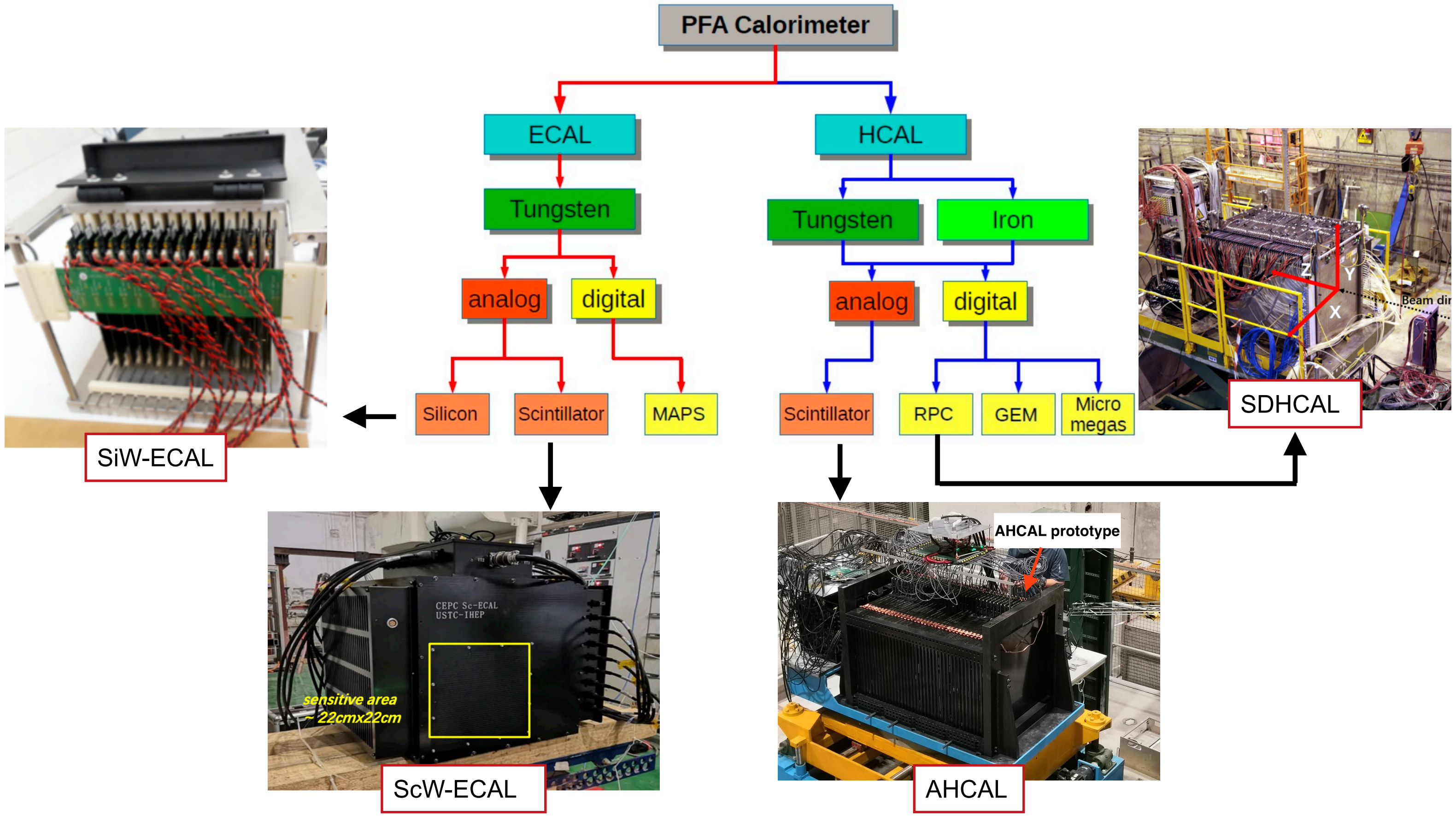
FCC at CERN



- **Exploration for ML PID methods in PFA-oriented calorimeters**

  - A major calroimetry option of future Higgs/W/Z/Top factories:
    - High granularity (imaging) -> A “camera”.

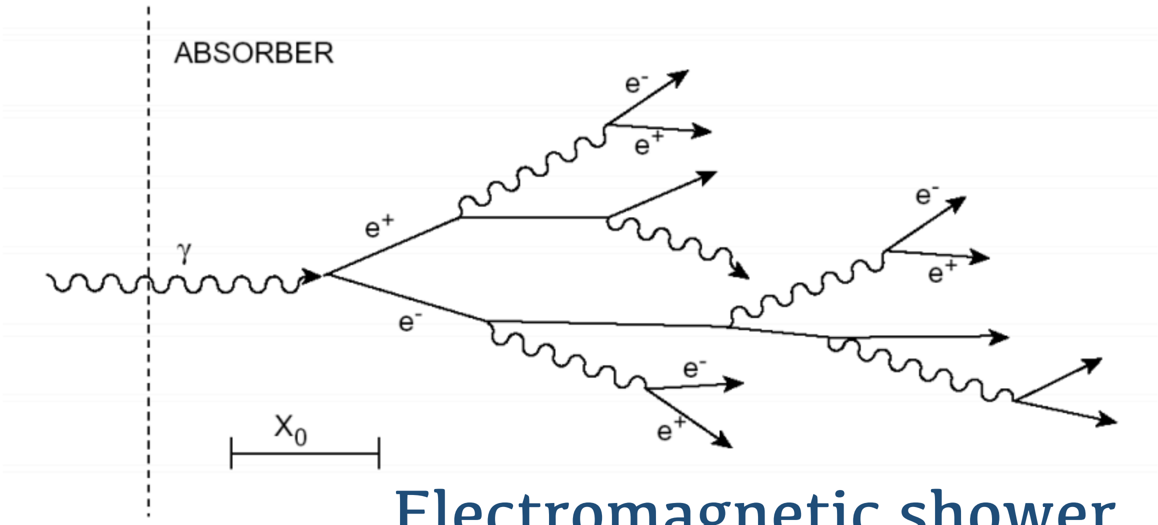




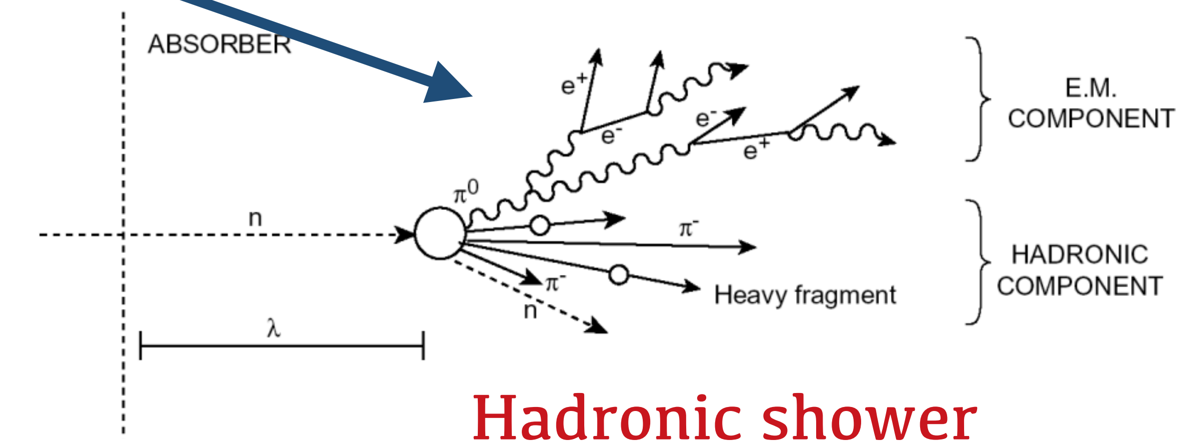
Overview of world-wide development of high-granularity calorimeters

Motivation

- **Challenge:**
  - Hadronic showers may undergo secondary interactions and exhibit complex development patterns.
  - Hadronic showers also contain EM component.
  - Develop effective shower topology variables.



Electromagnetic shower



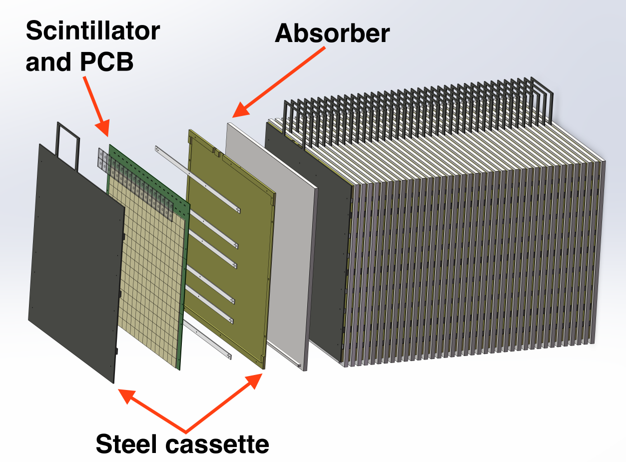
Hadronic shower

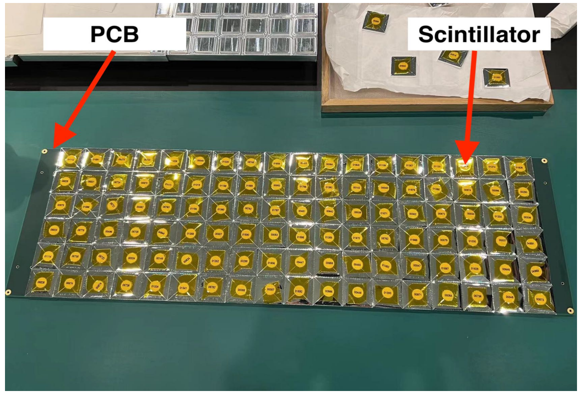
Introduction

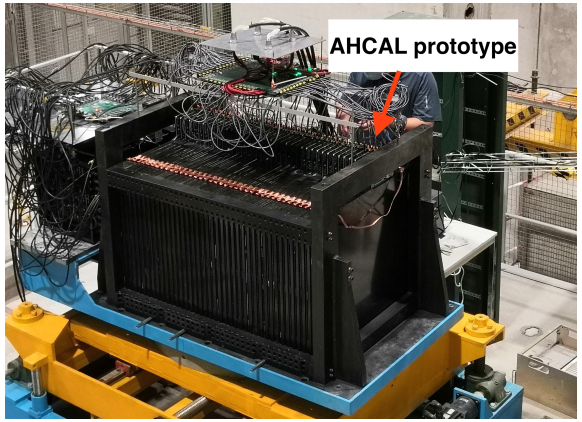
- **CEPC AHCAL prototype parameter**
  - **- Geometry**
    - - 40 sampling layers.
    - - $72\text{cm} \times 72\text{cm}$ in transversal plane.
    - - 120cm in longitudinal direction.
  - **- Absorber**
    - - 2 cm thickness/layer steel.
  - **- Sensitive cells**
    - - $40\text{mm} \times 40\text{mm} \times 3\text{mm}$ scintillator tile coupled with SiPM (SiPM-on-tile).
    - - $18 \times 18 \times 40$ array.

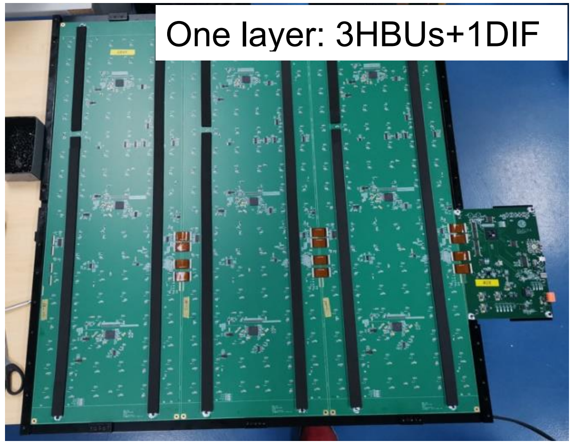
Spatial distribution of hits viewed as a 3D image

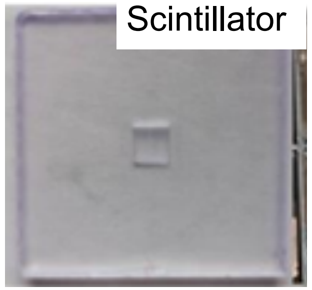
Journal of Instrumentation, 2021, 16(03): P03001.
Journal of Instrumentation, 2022, 17(11): P11034.





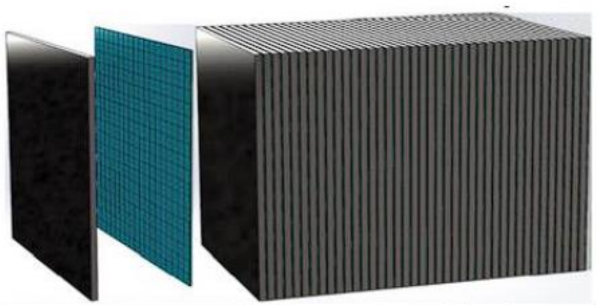


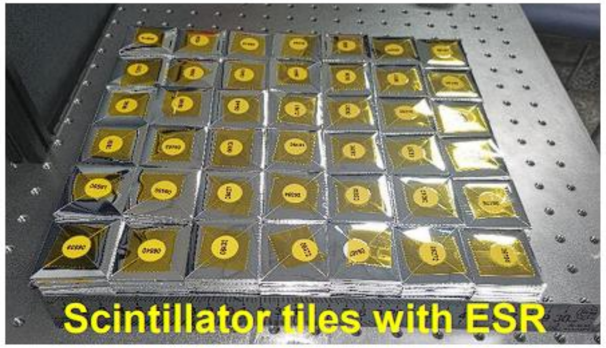


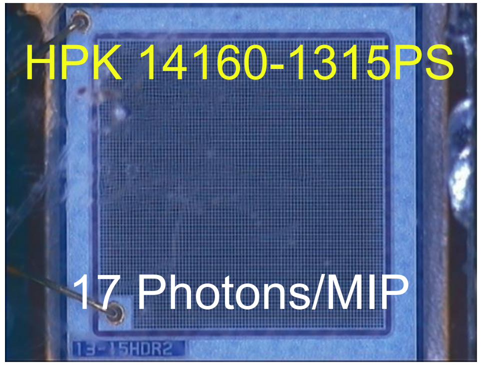


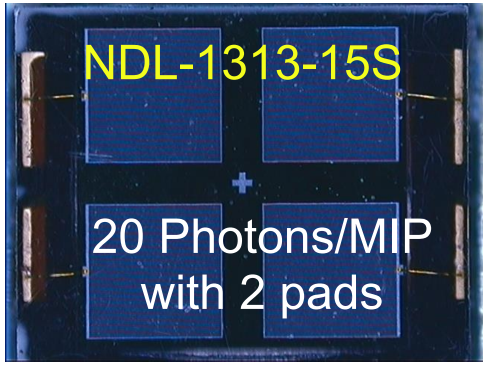
• Simulation set up

- the Geant4 11.1.1 Toolkit with the QGSP_BERT physics list was employed.
- $2\ mm$ plastic scintillator + $0.25\ mm \times 2$ ESR + $20\ mm$ Steel.
- **Digitization:**
  - Photon statistics: Poisson distribution concerning #detected photons (light output).
  - SiPM saturation : $response = \#pixel \times e^{-\frac{photon}{\#pixel}}$.
  - ADC error: assume 0.02%, very low.
  - Energy cut: 0.5 MIP.
- **SiPM:**
  - S14160-1315PS for first 38 layers.
  - EQR15 22-1313D-S for last 2 layers.





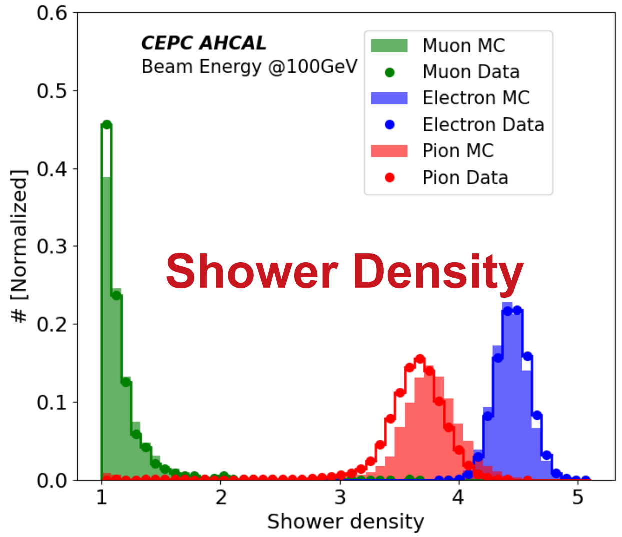


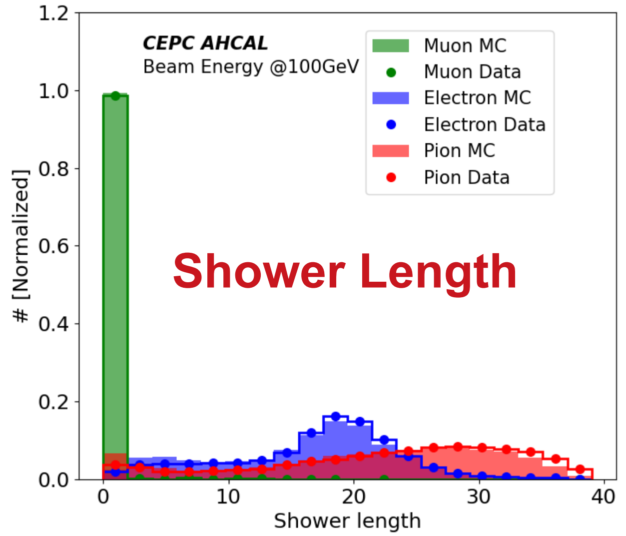


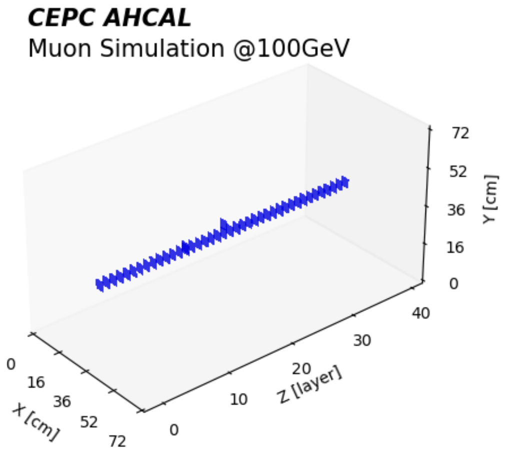
https://indico.cern.ch/event/847884/contributions/4831207/

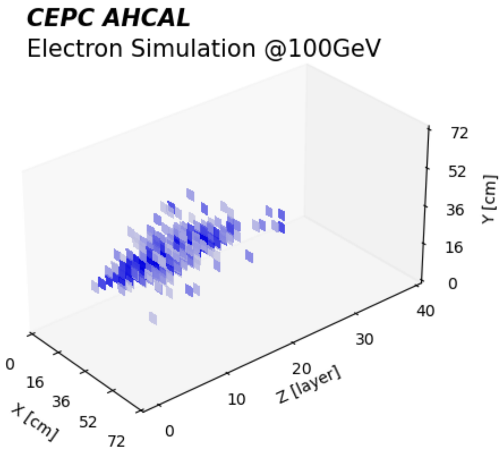
GEANT4 Simulation

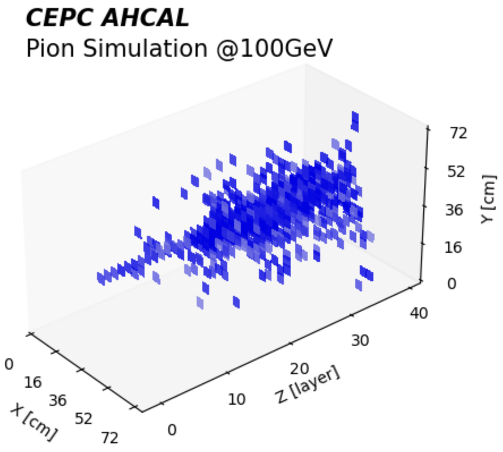
- **Generally, MC and Data are close in shower profile.**
  - Data come from beam test at SPS-H2, CERN.
  - Several shower topology variables are reconstructed:
    - **Shower density**: Mean hits number in a 3×3 cell.
    - **Shower length**： Distance between the start of the shower and the layer with maximum RMS of hit transverse coordinates.











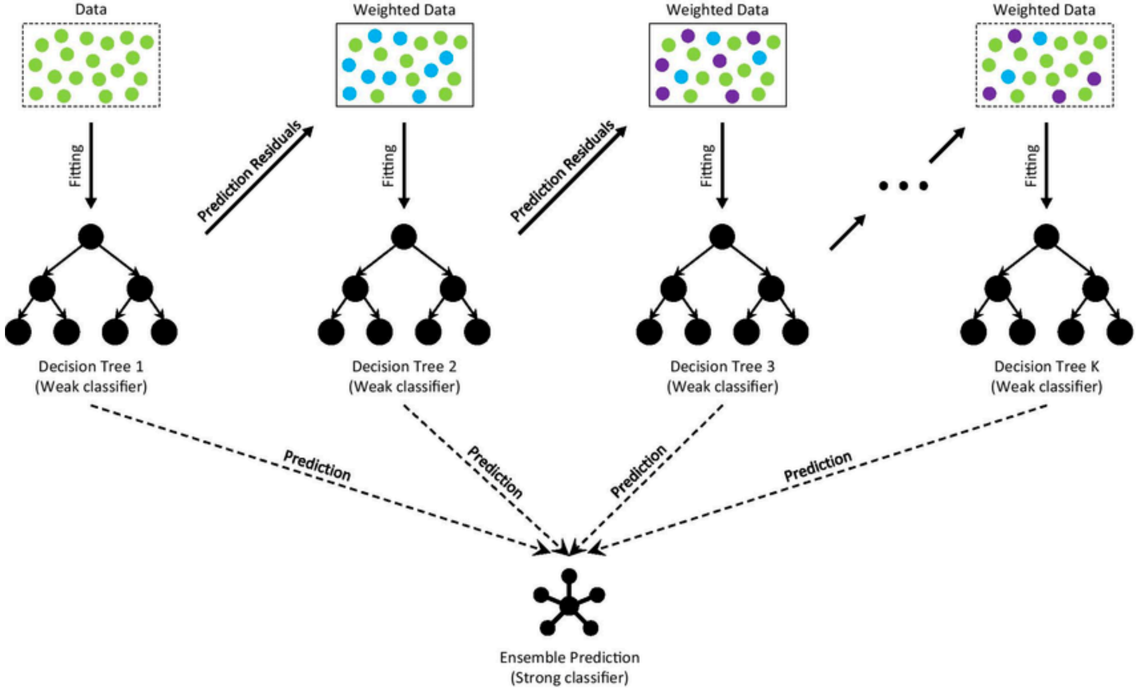
Monte Carlo Samples

- **Monte Carlo Samples**
  - To study the separation power in HAD showers and EM showers.
  - Training set : Validation set : Test set = 5: 1 : 4 .

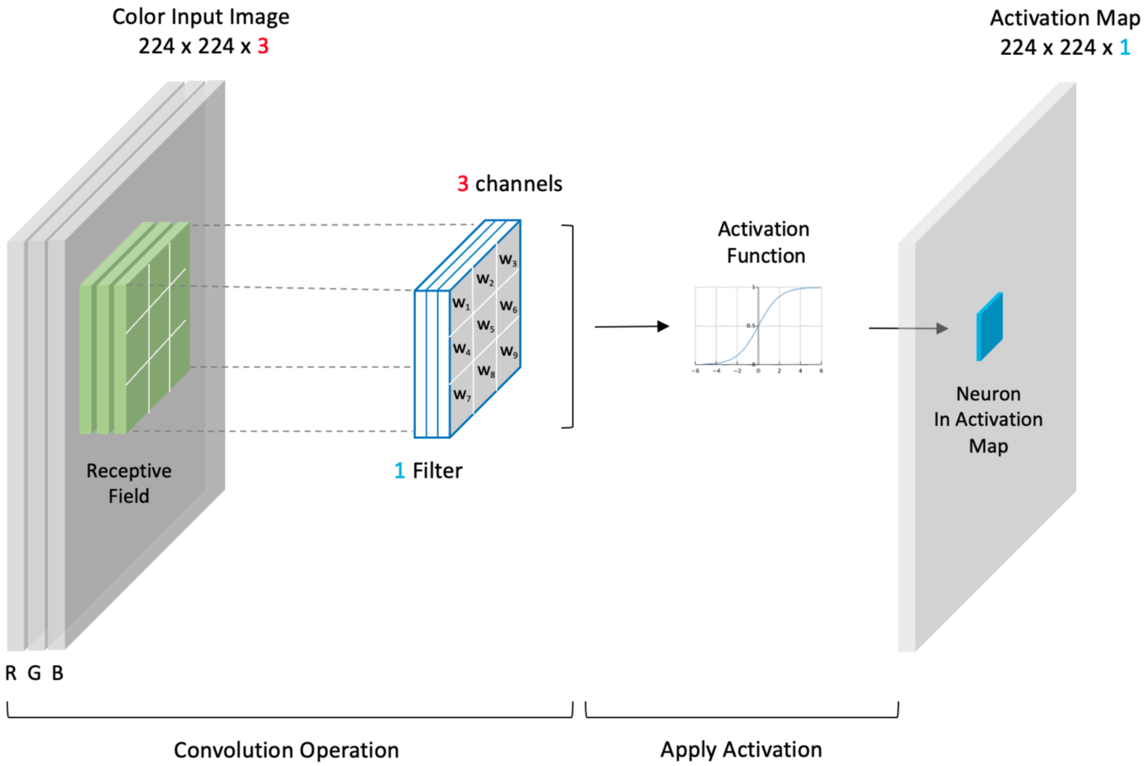
Energy	5GeV	10GeV	30GeV	50 GeV	60GeV	80GeV	100GeV	120GeV
Electron	100k	100k	100k	100k	100k	100k	100k	100k
Pion-	100k	100k	100k	100k	100k	100k	100k	100k

Two machine learning methods

- **Boosted decision tree (BDT)**
  - Need pre-reconstructed input.
- **Cell-based artificial neural networks (ANN)**
  - Treat spatial distribution of hits as images.

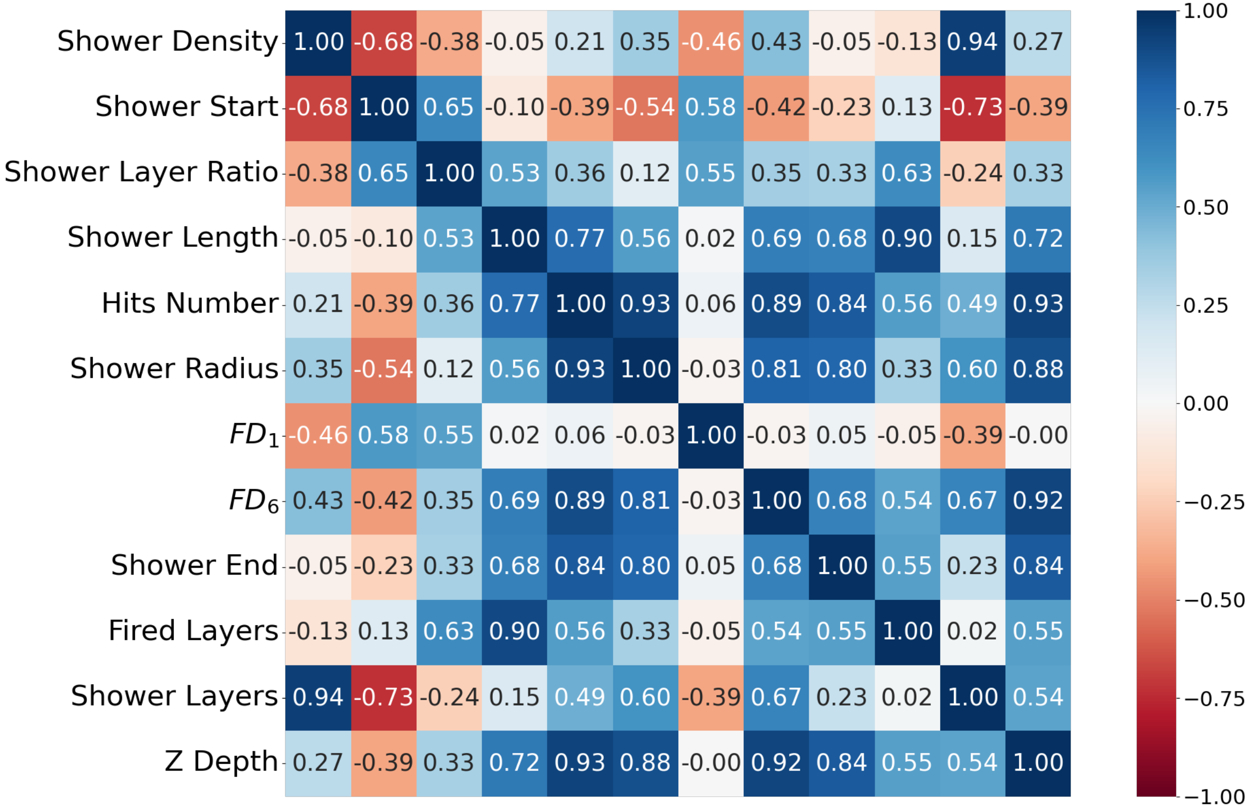


Boost decision tree



Neural networks

- **Apply Extreme Gradient Boosting (XGBoost).**
  - 12 variables are reconstructed.
  - Z depth, Shower radius, and Shower layers are top 3 variables.

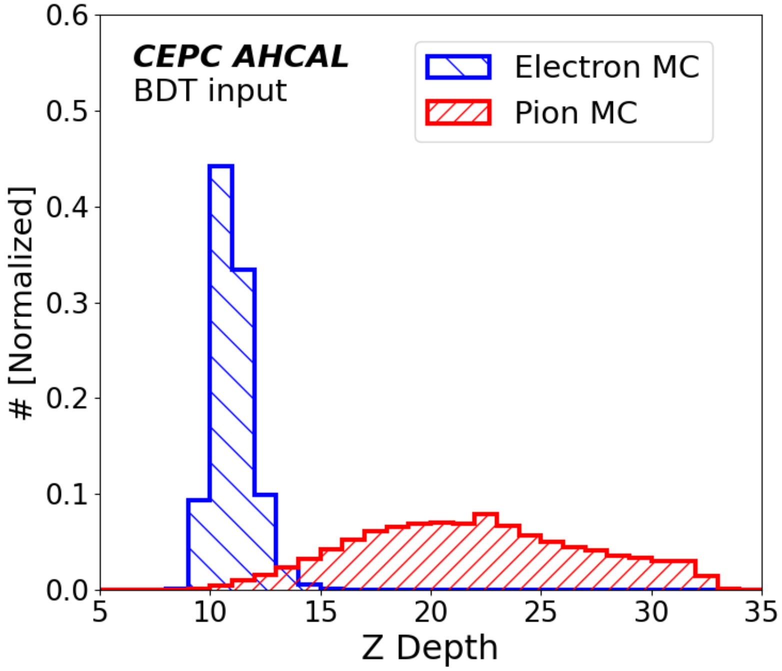


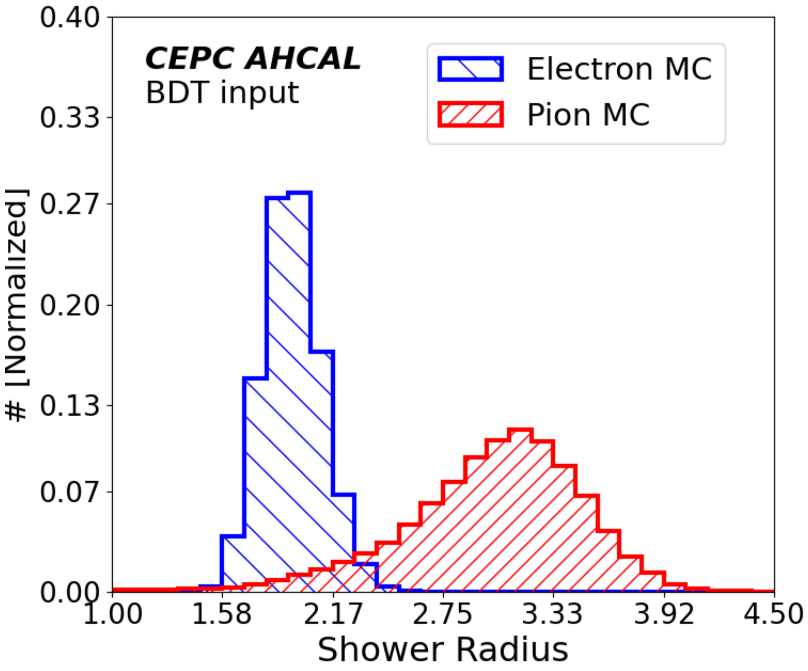
Correlation matrix

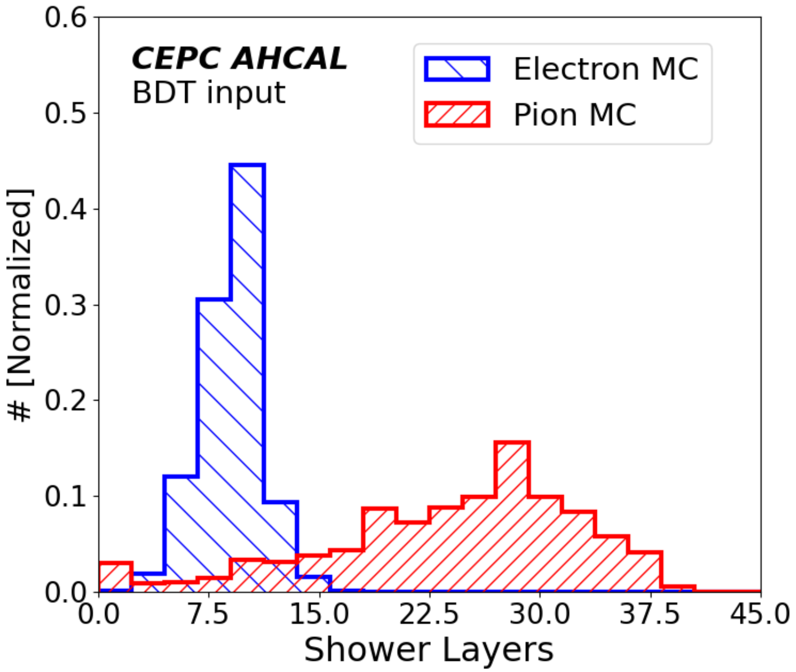
Rank: Variable	Variable weight
1: Z depth	0.532
2: Shower radius	0.186
3: Shower layers	0.073
4: Fired layers	0.065
5: Shower density	0.370
6: Shower start	0.026
7: Shower layer ratio	0.022
8: FD_1	0.018
9: Hits number	0.013
10: FD_6	0.012
11: Shower end	0.009
12: Shower length	0.006

BDT input

- **Top 3 variables in separating between EM showers and HAD showers.**
  - **Z depth**: The RMS of the z-axis coordinates.
  - **Shower Radius**: The RMS of the distance with respect to the z-axis.
  - **Shower layers**: The number of layers in which the RMS of positions in the x-y plane exceeds 4 cm.





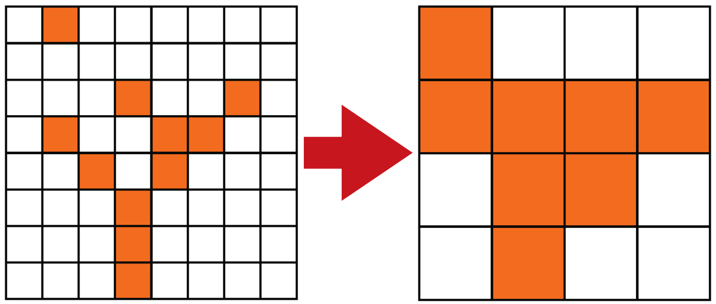


BDT input

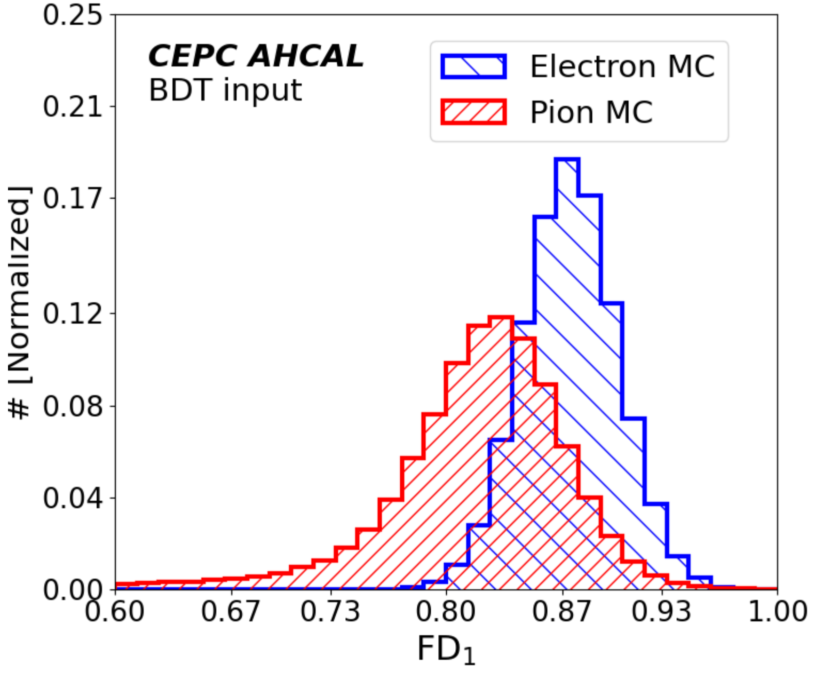
- **Fractal dimension**: $\mathrm{FD}_\beta = \left\langle \frac{\log(\mathrm{R}_{\alpha,\beta})}{\log(\alpha)} \right\rangle + 1$, where $\mathrm{R}_{\alpha,\beta} = \mathrm{N}_\beta / \mathrm{N}_\alpha$.
- **Hits number**: The number of hits.

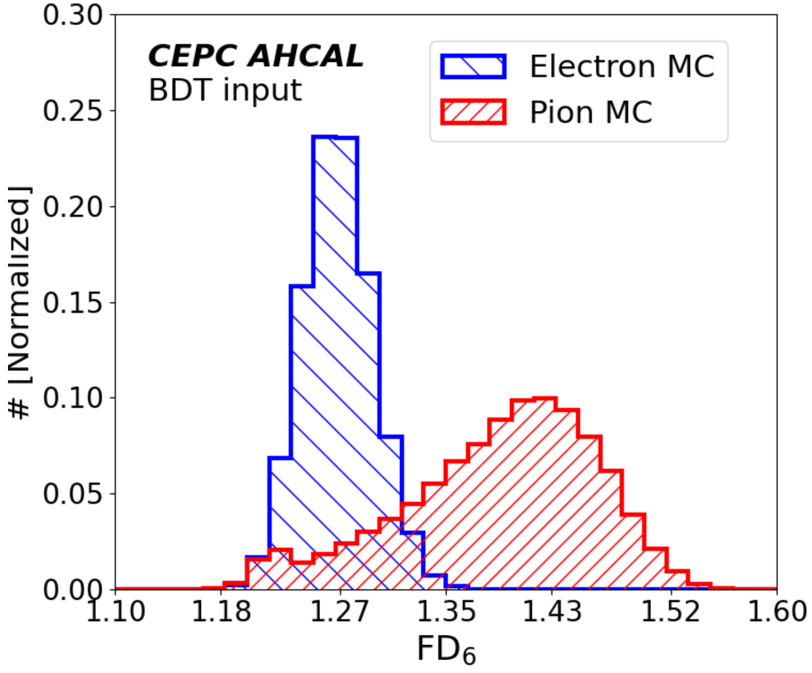
FD Ref: PhysRevLett.112.012001

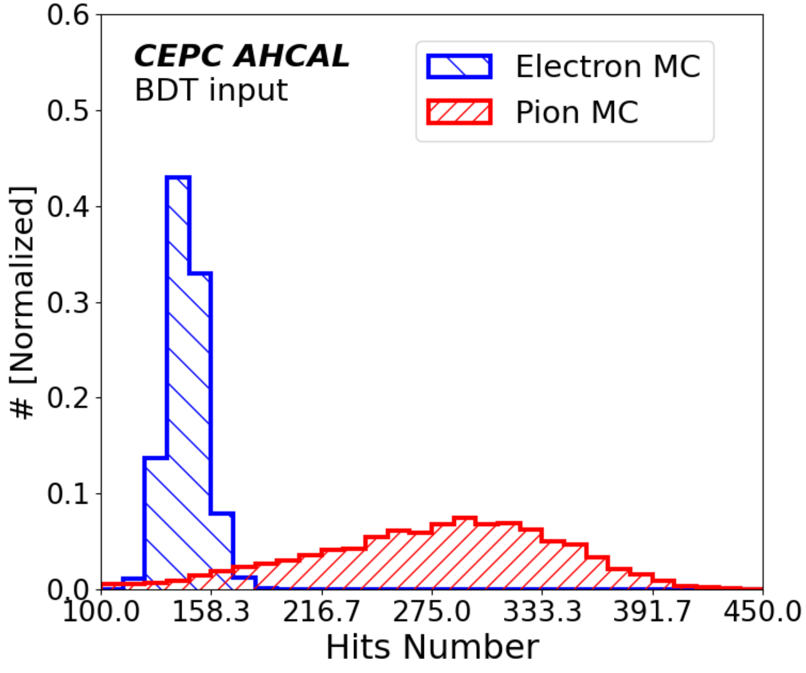
- N_α: number of hits scaled by α.



Take $\alpha = 2$ as an example



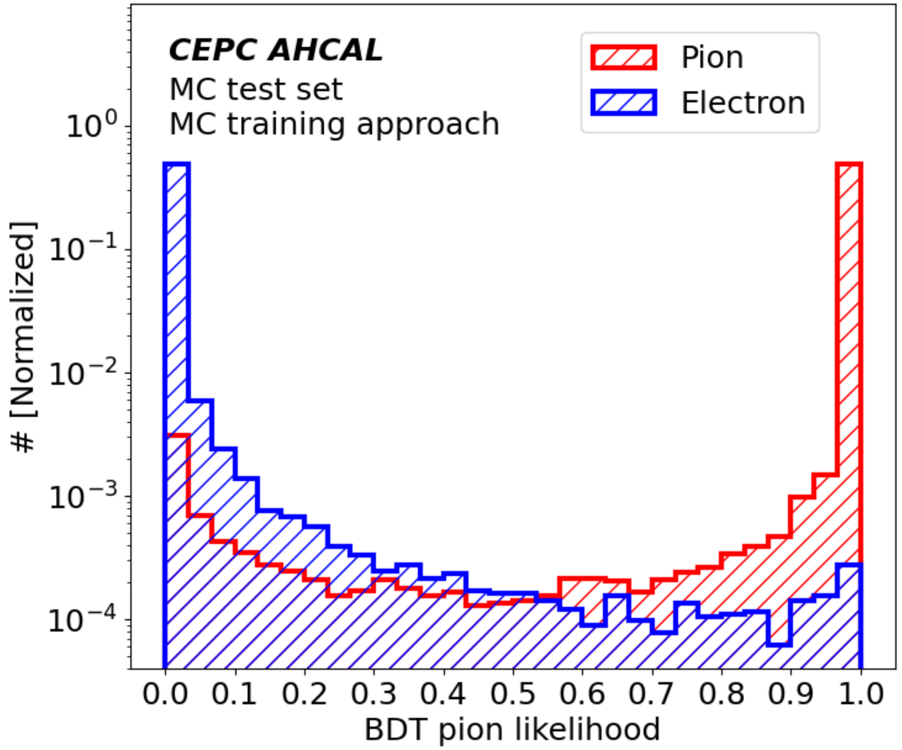


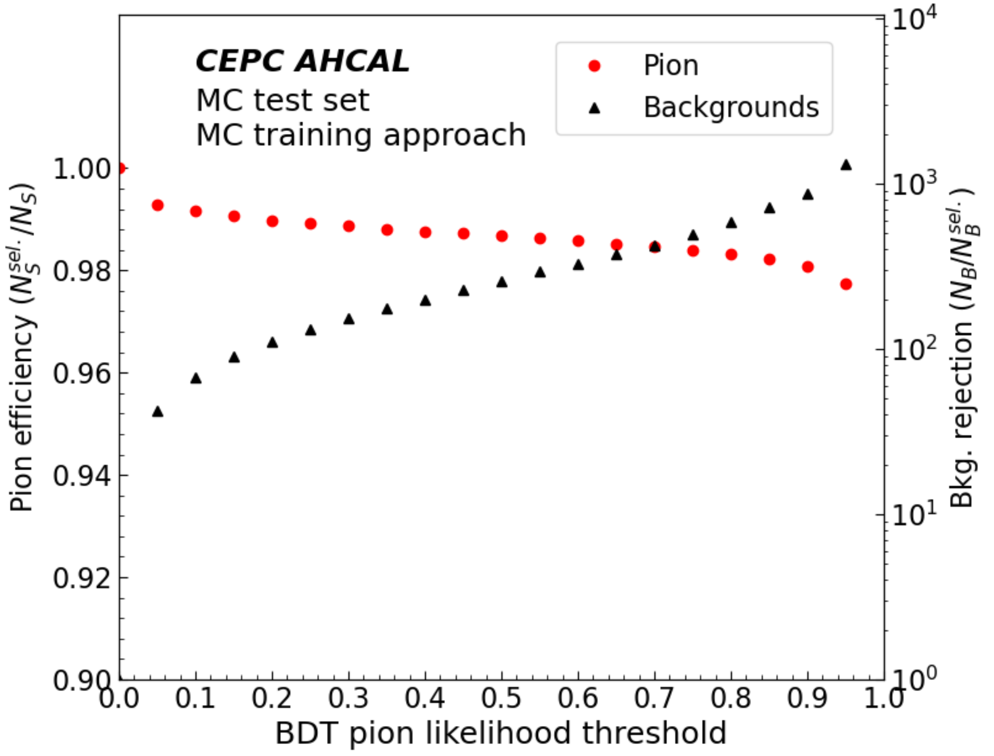


PID based on BDT

- **BDT classifier performance**

Pion efficiency	95%	96%	97%	98%	99%
Electron rejection (1/e efficiency)	16012.7	8734.2	3843.0	970.5	105.3

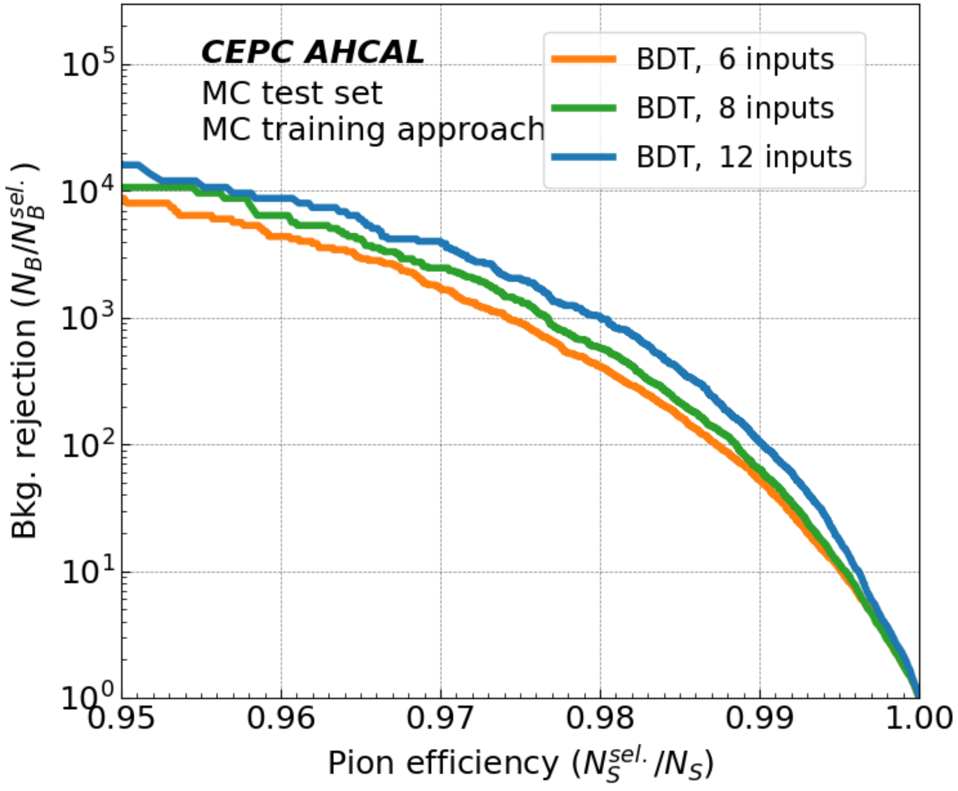




PID based on BDT

- **We observe dependence of BDT performance on input variables**
  - Remove Shower End, Shower Layers, Fired Layers, and Z Depth to build BDT with 8 inputs.
  - Further remove FD_1 and FD_6 to build BDT with 6 inputs.

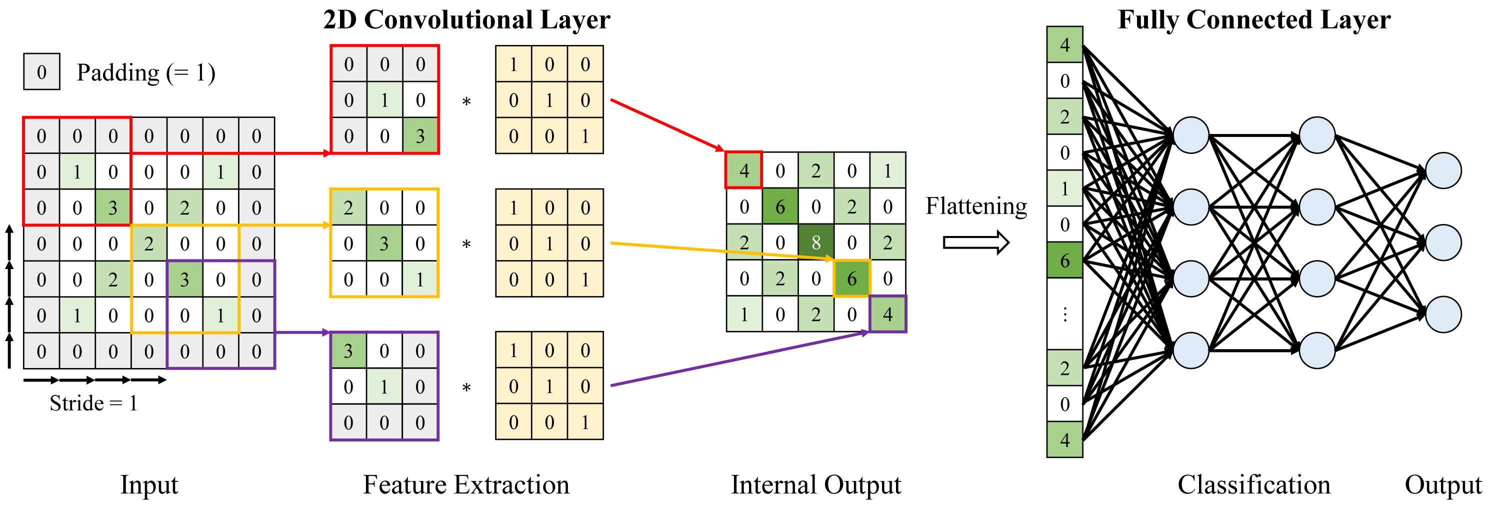
- Feature engineering can be sometimes tricky.



- **BDT optimization is still on going.**

PID based on ANN

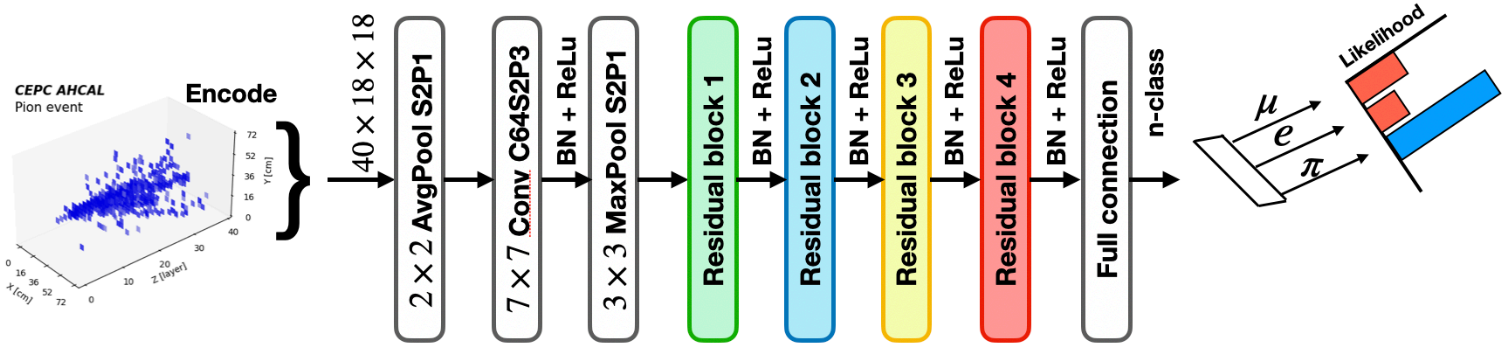
- **Cell-based Artificial Neural Networks (ANN) make full use of high-dimensional input.**
  - Compile layers to extract features.
  - **Input**: Variable stands for events (Spatial distribution of hits).
  - **Output**: the likelihood of each particle type candidate.
  - After iteration, the mapping: **Input**->**Output** close to truth.

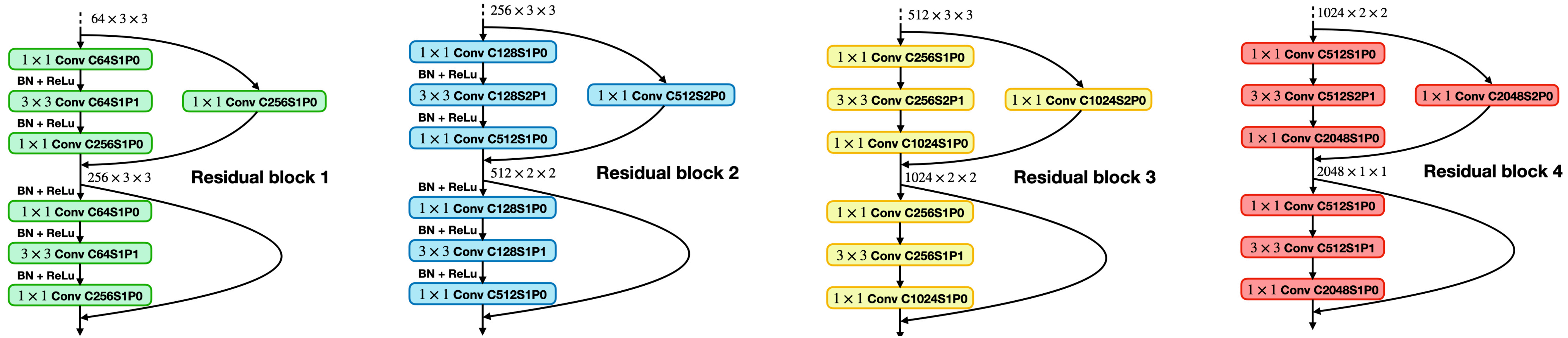


Particle Identification

• ANN-based PID: Taking the advantage of ResNet

- Input: energy depositions in AHCAL (Input tensor size: $40 \times 18 \times 18$) .
- Output: likelihood of each particle type candidate.

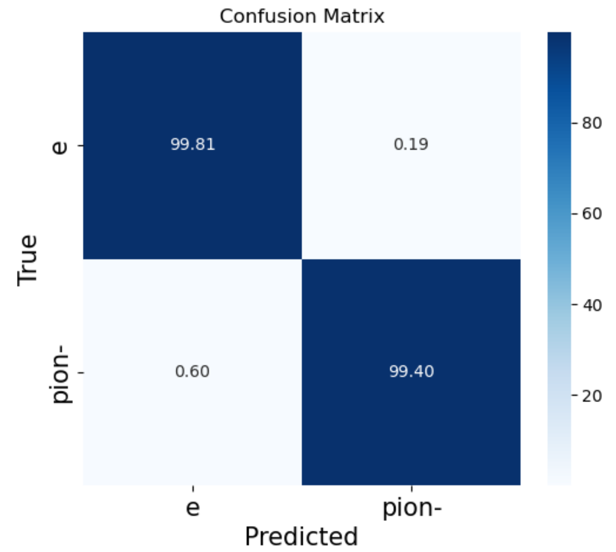




ResNet Ref: He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

Particle Identification

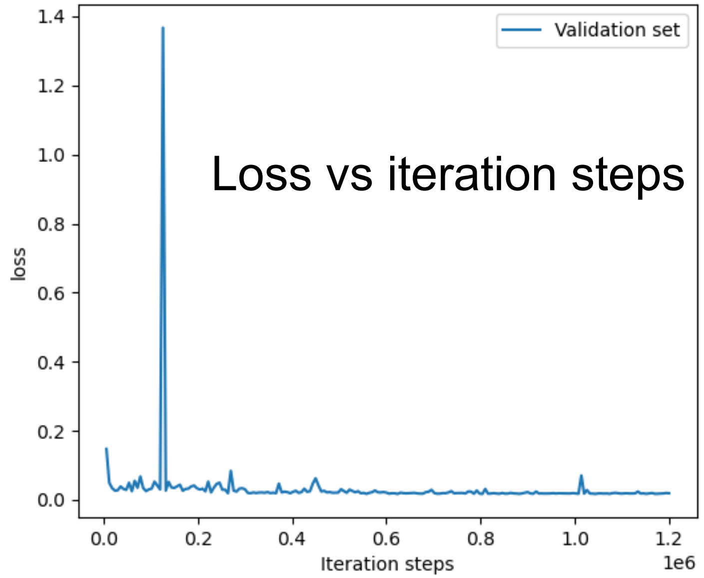
- **Executed on an NVIDIA V100 NVLink GPU.**
- **Loss: Cross Entropy.**
- **Hyper-parameter**
  - **Batch size= 64**
  - **lr=0.0001**
  - **epoch = 200**



Algorithm 1 Artificial Neural Network.

Require: The batch size m, the epoch number n, initial learning rate lr, initial net parameters θ_0.

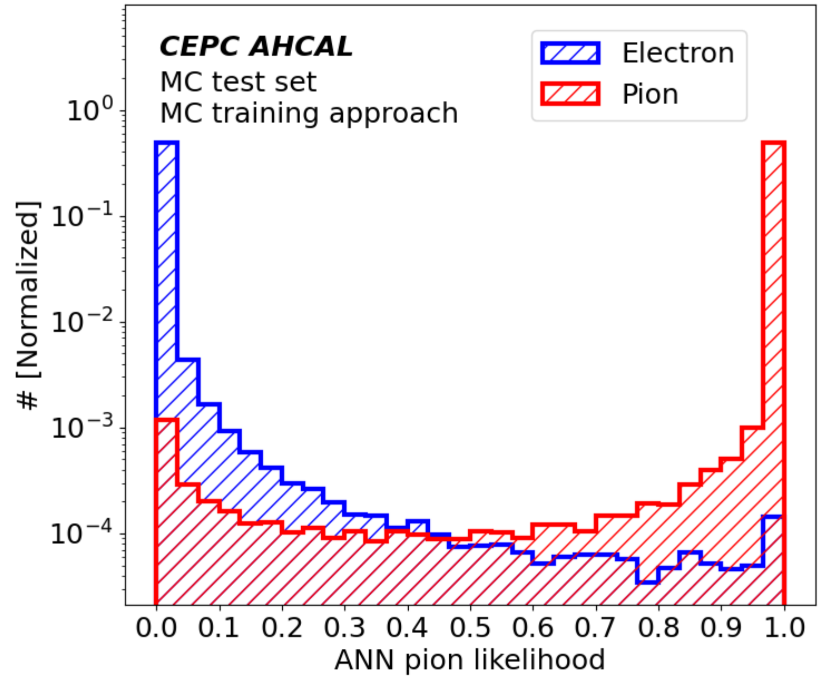
Assign corresponding label y to data x.
for $t = 1, \cdots, k$ iteration steps **do**
    for $i = 1, \cdots, m$ **do**
        $\hat{y} \leftarrow \mathrm{Net}(x, \theta)$
        $\mathrm{Loss}(y_i, \hat{y}_i)_\theta^{(i)} \leftarrow (-\log(\hat{y}_i))$
    end for
    $\theta \leftarrow \mathrm{SGD}\left(\nabla_\theta \frac{1}{m} \sum_{i=1}^m \mathrm{Loss}_\theta^{(i)}\right)$
end for

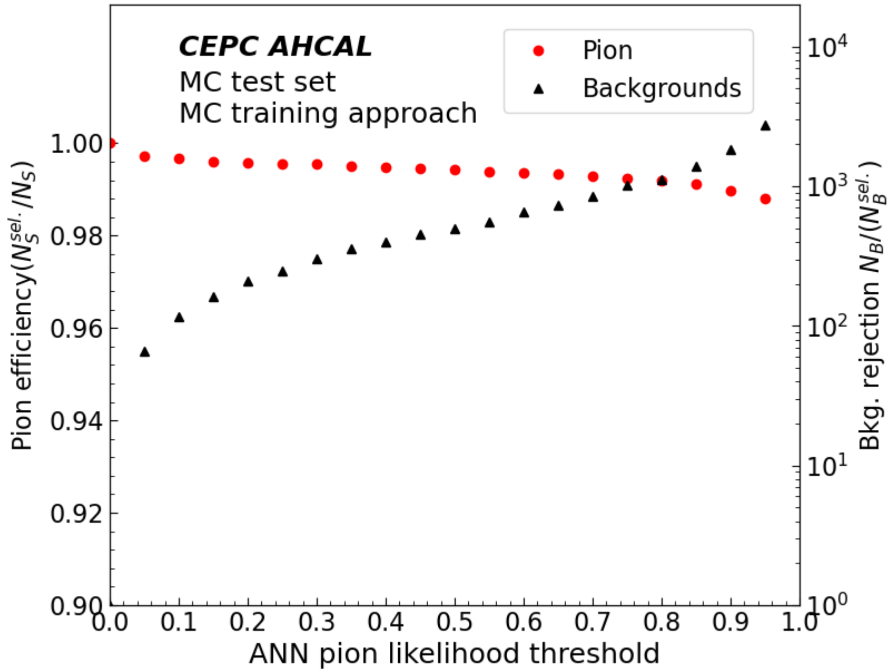


PID based on ANN

- **ANN classifier performance**

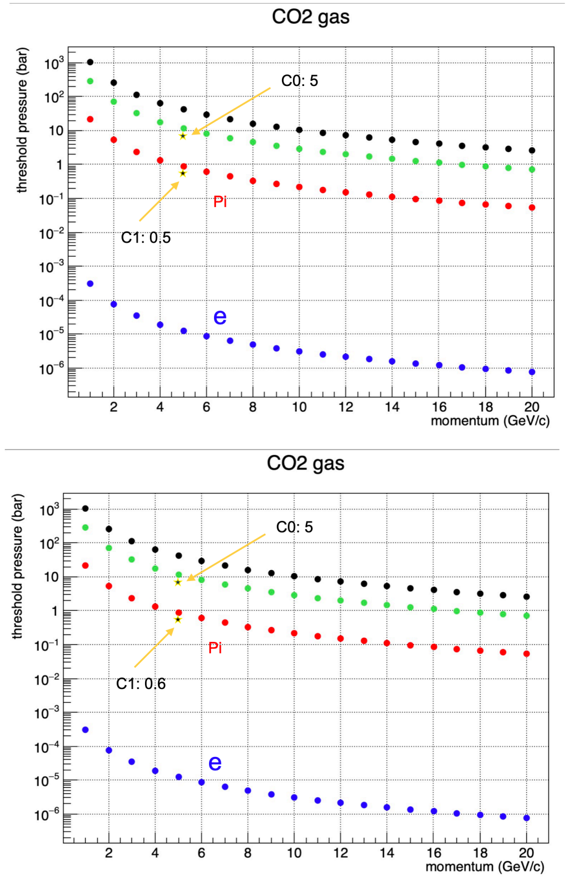
Pion efficiency	95%	96%	97%	98%	99%
Electron rejection (ANN)	**56012.7**	**48010.9**	**19769.2**	**9083.1**	**2154.3**
Electron rejection (BDT)	16012.7	8734.2	3843.0	970.5	105.3
Improvement	199.8%	249.8%	393.8%	752.1%	1945.0%



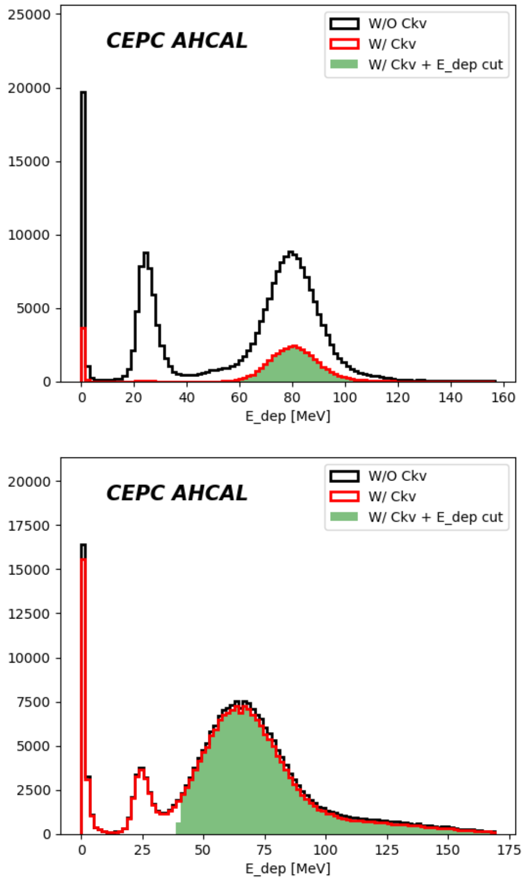


Particle Identification

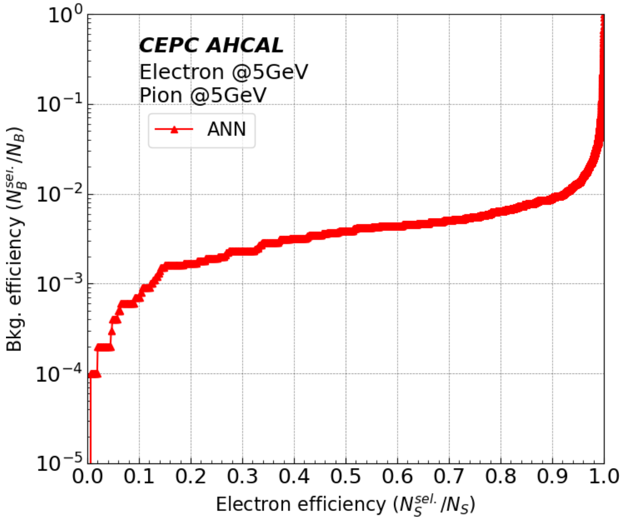
- **ANN PID cross-check using Cherenkov detectors**
  - Two CO_2 Cherenkov detectors are available at PS (<15 GeV).
  - 20,000 Electron and 20, 000 Pion samples are selected as truth.

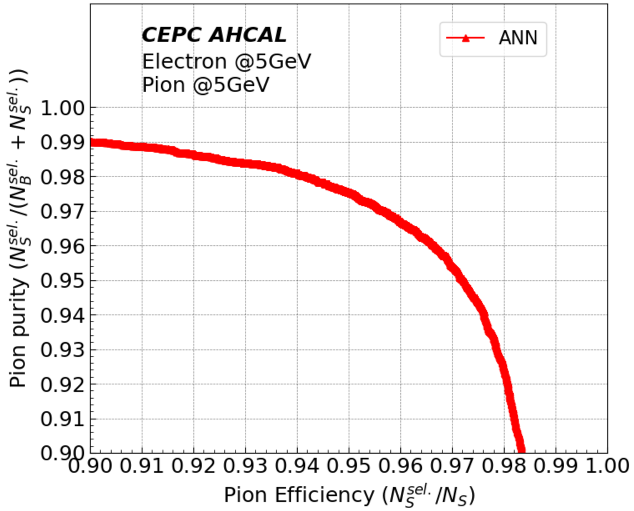


Cherenkov detector setting strategy.



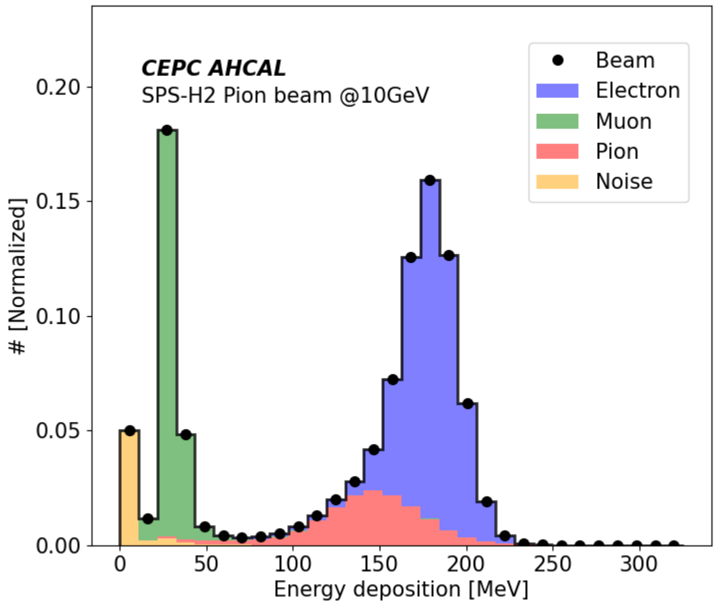
Samples after selection.



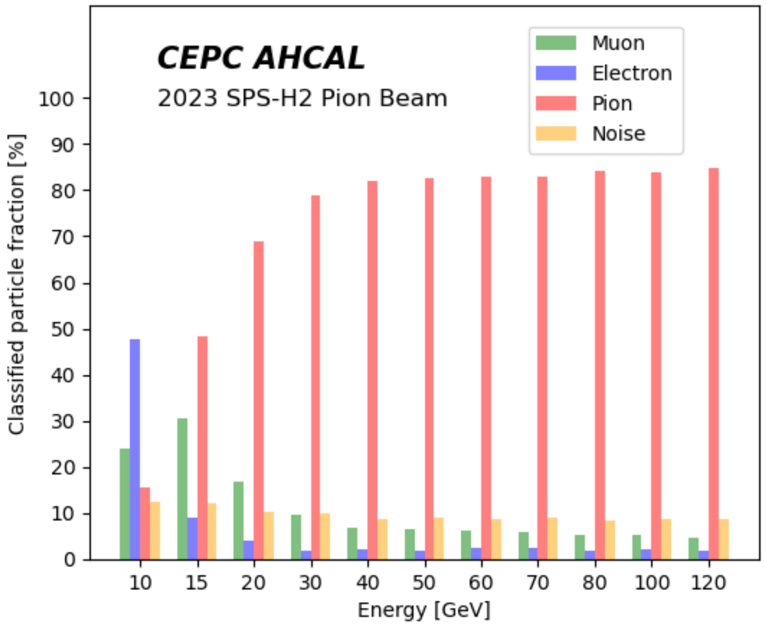


- Achieve 90% Pion efficiency and 99% Pion purity at the same time.

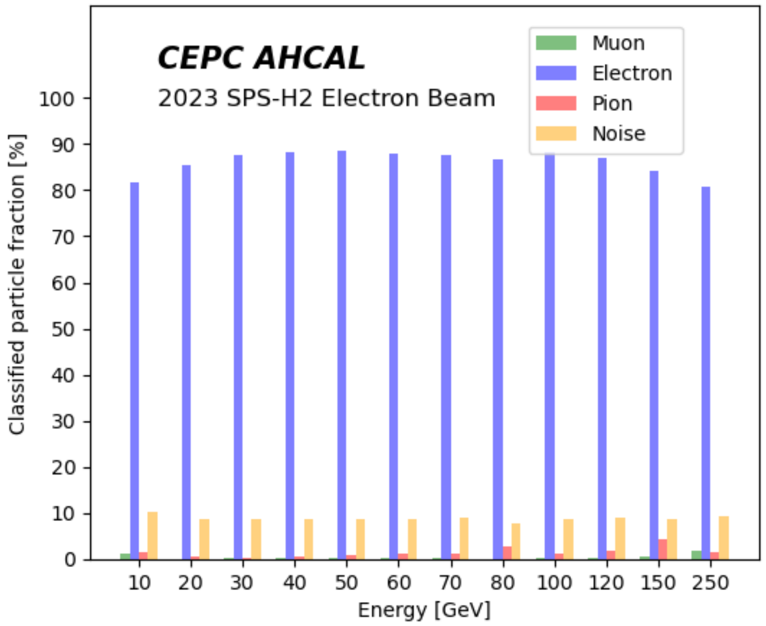
- **Beam composition is given by ANN classifier**
  - Trained on pre-selected data also tagged by Cherenkov detector.
  - Pion beam purity: around 80% when beam energy is over 30 GeV.
  - Electron beam purity: over 80% at each energy point.



Beam composition in Energy deposition spectrum (10 GeV)

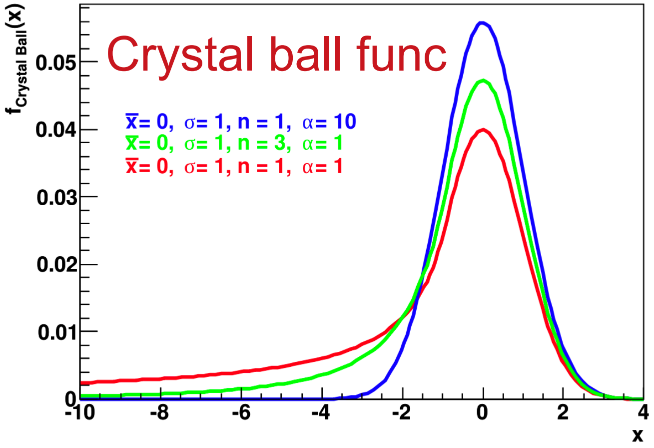


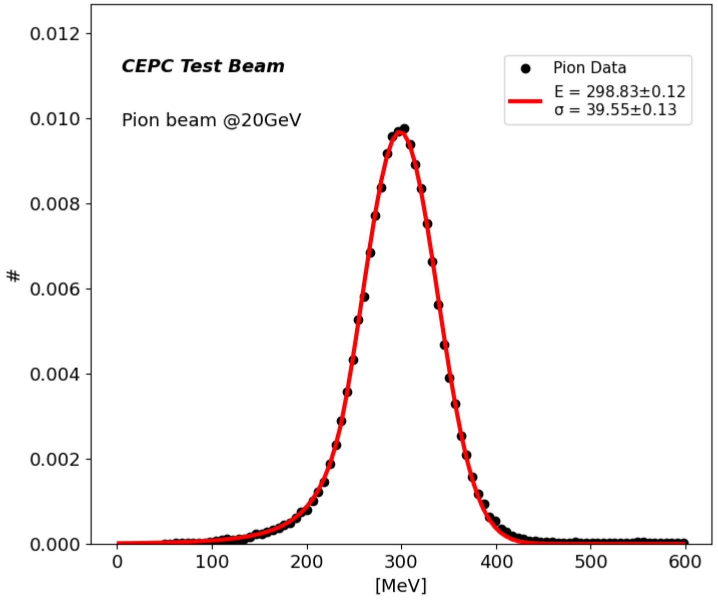
Pion Beam composition

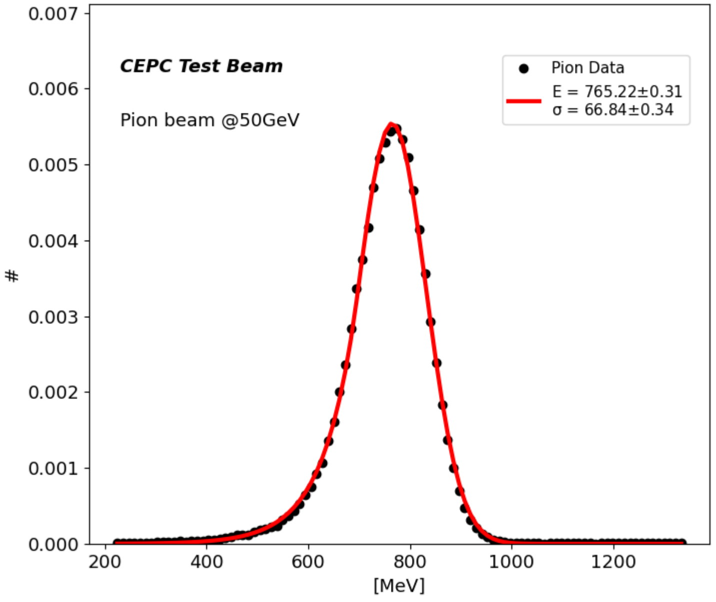


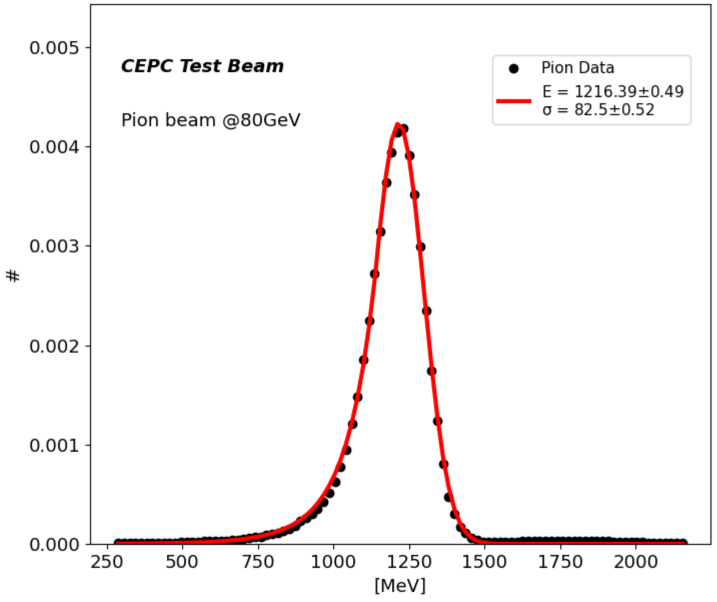
Electron Beam composition

- **Purified beam data fitting**
  - ANN classier is first used for purifying the Pion beam.
  - Crystal ball function is then used for fitting purified Pion Data.









The energy of 20, 50, 80 GeV.

Discussion

- **ANN(ResNet) outperforms BDT in our cases.**
  - ***Automatic feature extraction:*** This allows ANN to make full use of all input information, and potentially uncover hidden patterns in the data that may be missed by BDT, which relies on limited reconstructed features.
  - ***Effective in handling large and high-dimensional inputs:*** ANN is well-equipped to handle high-dimensional data and capture complex patterns within it.
  - ***Non-linearity:*** ANN can model complex non-linear relationships in data more effectively than BDT.
- **More validation is still on going.**

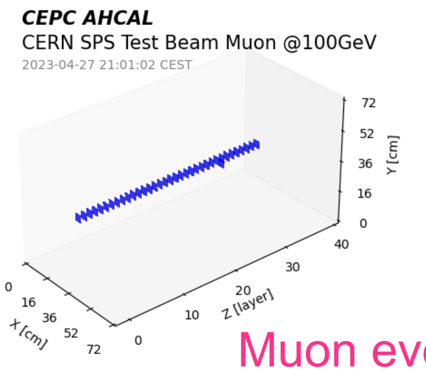
Thank you

感谢聆听

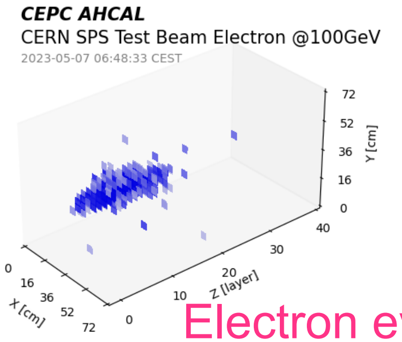
饮 水 思 源 爱 国 荣 校

Beam Test at CERN in 2022 & 2023

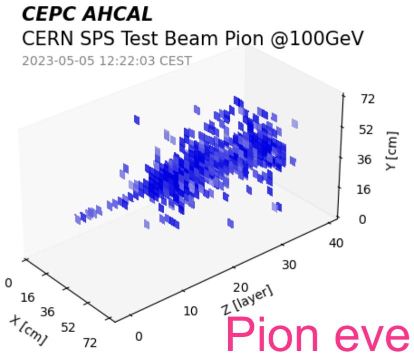
- **SPS-H8: Oct 19 - Nov 2, 2022:**
  - μ^+: 160 GeV (for calibration)
  - π^+: 10, 20, 30, 40, 50, 60, 70 , 80, 90, 100, 120 GeV (**~1M** events each point)
  - e^+: 10, 20, 30, 40, 50, 60, 70 , 80, 90, 100, 120 GeV (**~0.3M** events each point)
- **SPS-H2: Apr 26 - May 10, 2023:**
  - μ^-: 100 GeV (for calibration)
  - π^-: 10, 15, 20, 30, 40, 50, 60, 70 , 80, 100, 120, 350 GeV (**~1M** events each point)
  - e^-: 10, 20, 30, 40, 50, 60, 70 , 80, 100, 120, 150, 250 GeV (**~0.3M** events each point)
- **PS-T9: Apr 17 - May 31, 2023:**
  - μ^-: 10 GeV (for calibration)
  - π^-: 1, 2, 3, 4, 5, 6, 7, 8, 10, 12, 15 GeV (**~0.5M** events each point)
  - e^-: 1, 2, 3, 4, 5 GeV (**~50K** events each point)



Muon event



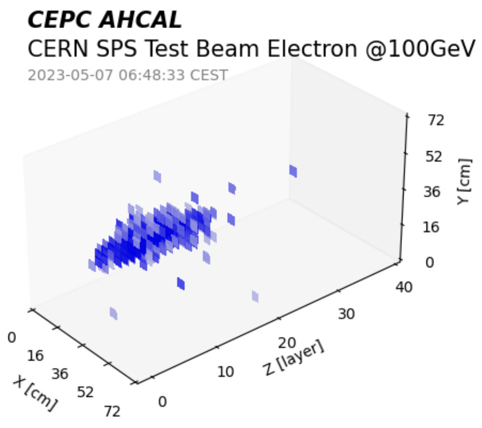
Electron event

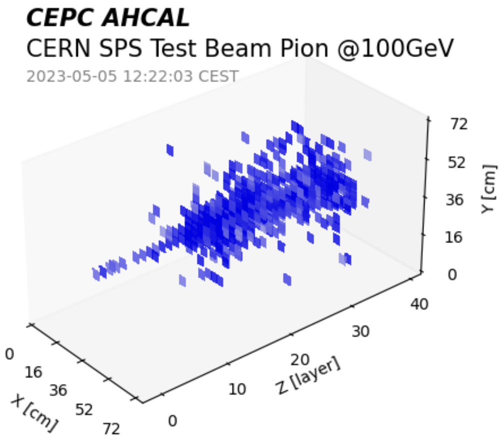


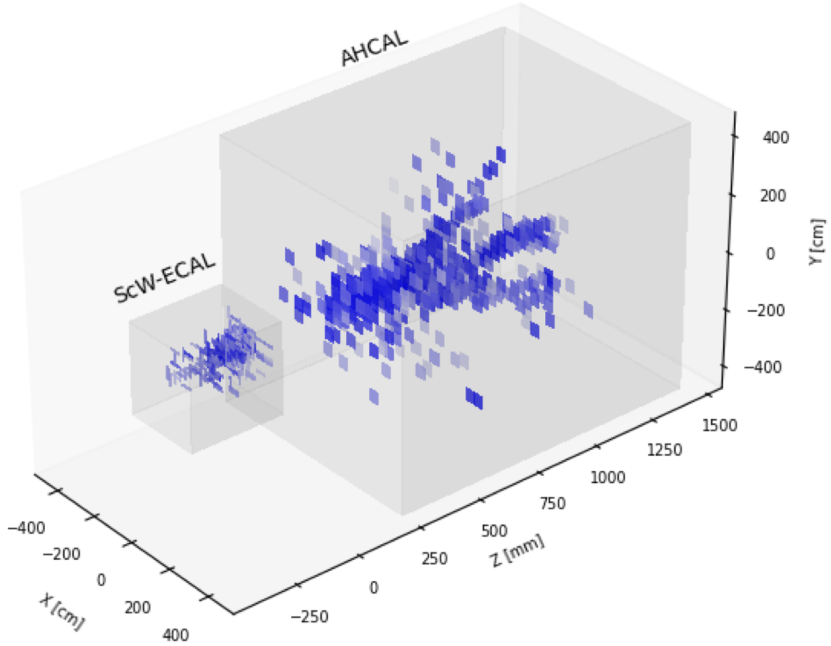
Pion event

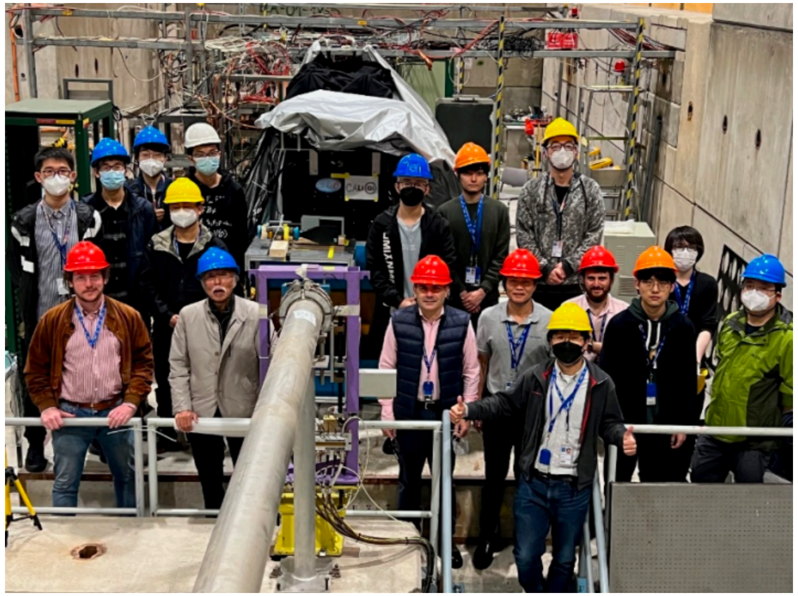
Beam Test at CERN in 2022 & 2023

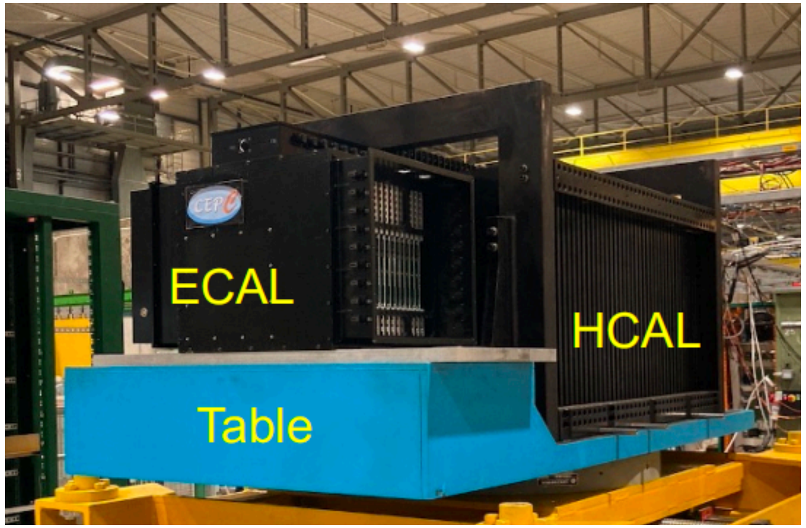
- **Successful beam test for AHCAL and ScW-ECAL was conducted.**
  - Beam test at CERN SPS & PS in 2022, 2023.
  - Beam: Muon, Electron, Pion, Proton (1-350GeV).



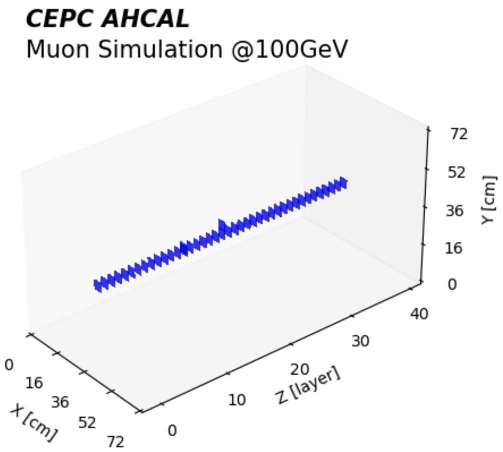


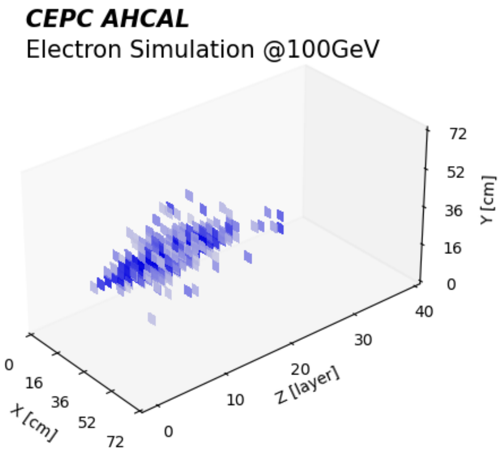


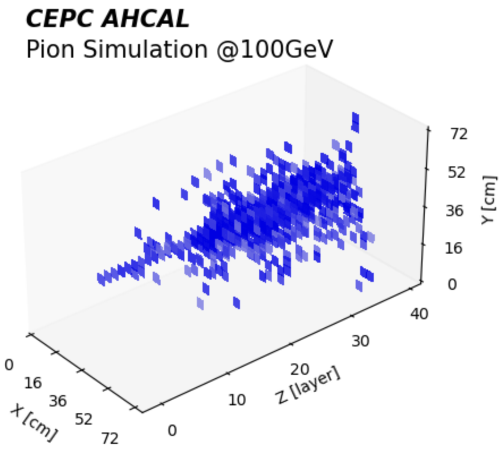




Monte Carlo Samples



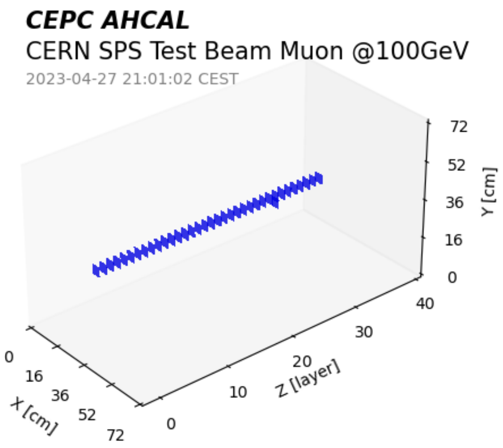


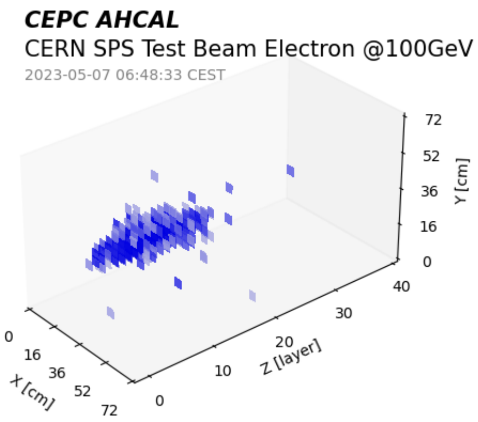


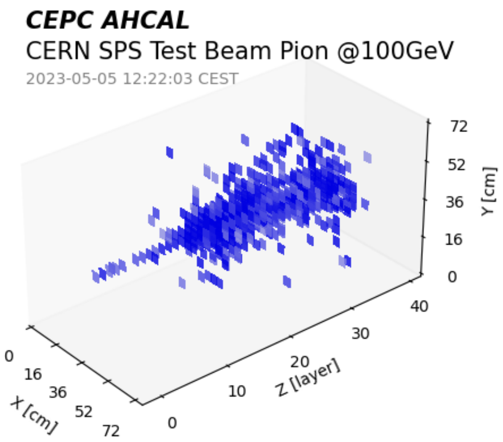
Event Display

- AHCAL only.
- Shower topology of the same particle type is similar between MC and Data.
- Shower type:
  - Muon: Non-showering track.
  - Electron: Electromagnetic shower.
  - Pion: Hadronic shower.

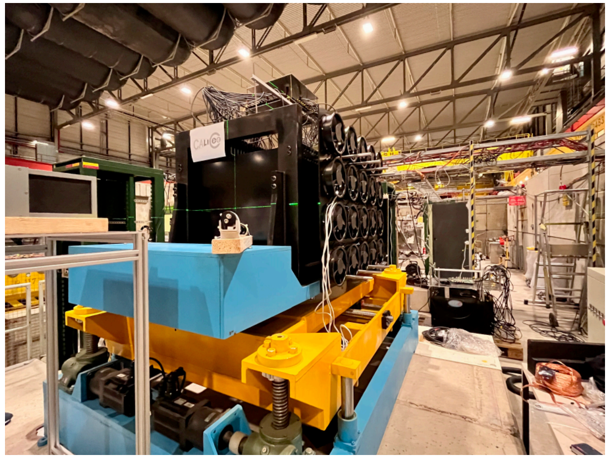
Test Beam Samples

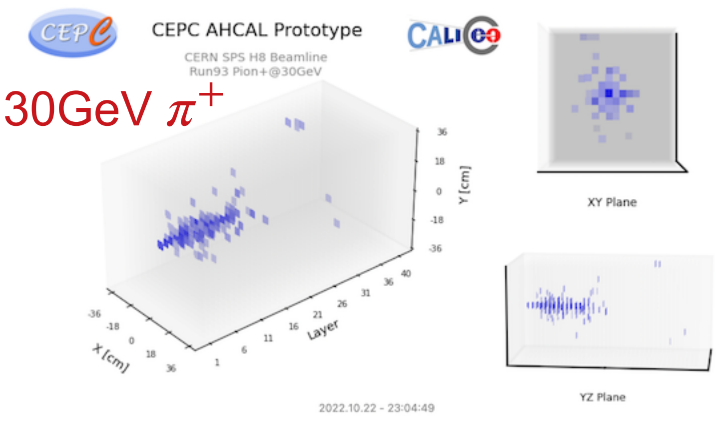


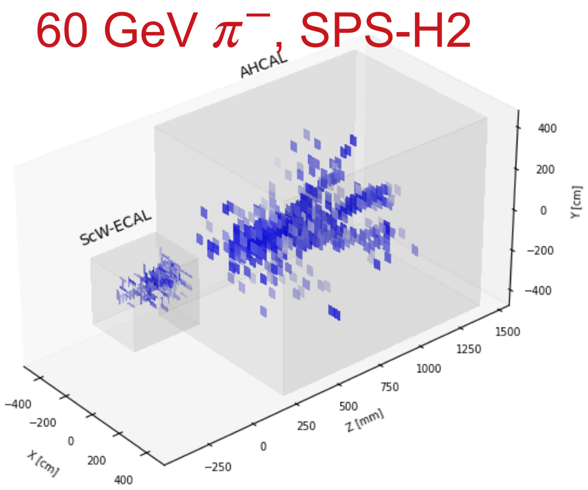


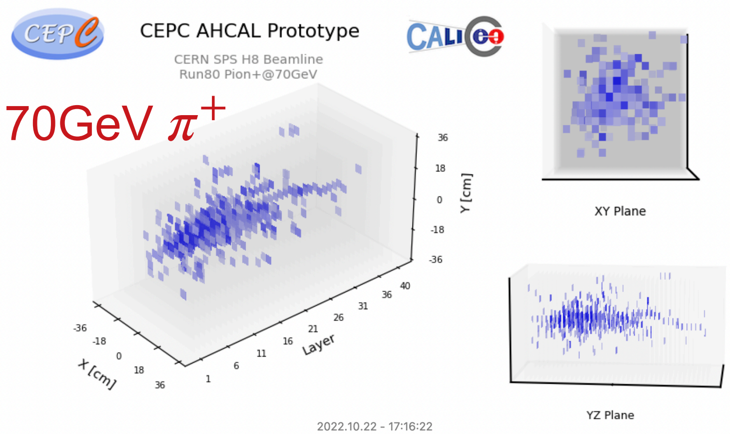


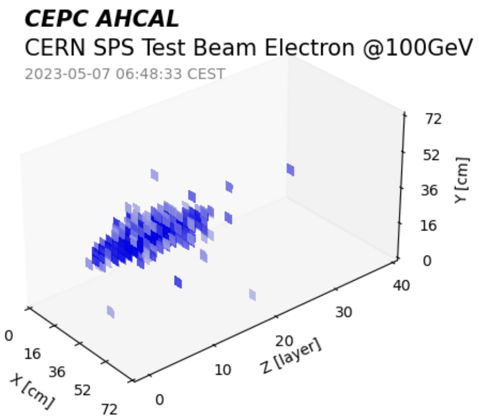
Beam Test at CERN in 2022 & 2023

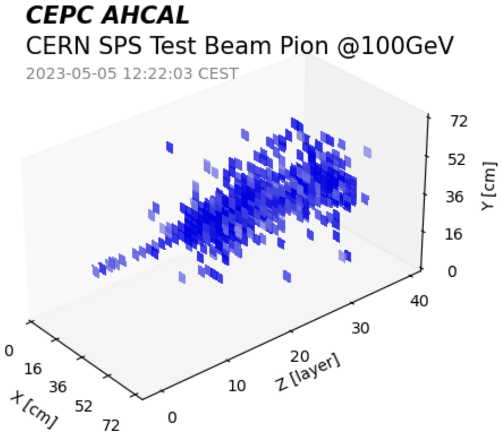












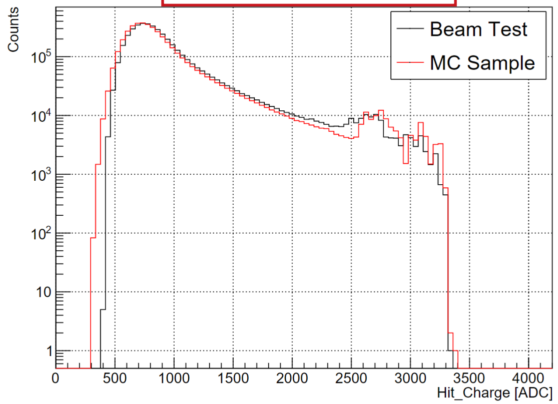
- First beam test at CERN SPS-H8: Oct-Nov, 2022.
- Beam: Muon, Electron, Pion (10-160GeV).
- Notice beam contamination issue.

- Beam test at CERN SPS-H2 & PS-T9: Apr-May, 2023.
- Beam: Muon, Electron, Pion, Proton (1-350GeV).
- Beam purity is better than 2022's.
- Available Cherenkov detectors (effective: E_{Beam} <30GeV).

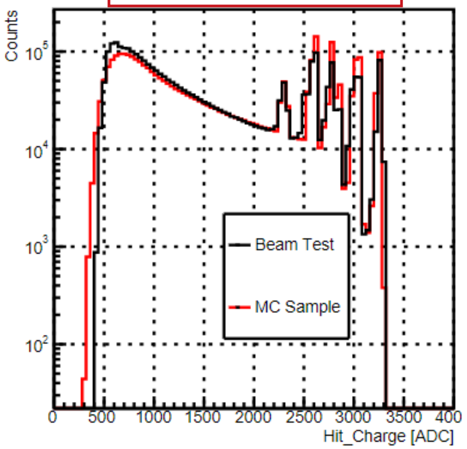
GEANT4 Simulation

- **Validation on High/Low Gain - 0.5 MIP energy threshold**
  - Generally fit in High Gain; Need optimization in Low gain saturation correction.

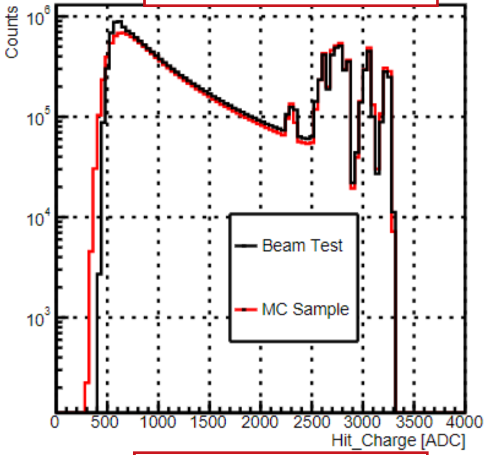
HG 100GeV μ^-



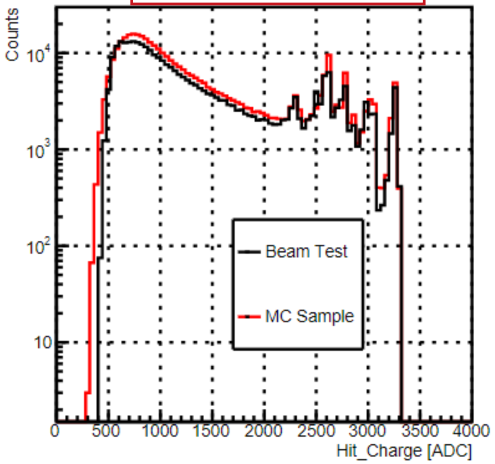
HG 10GeV e^-



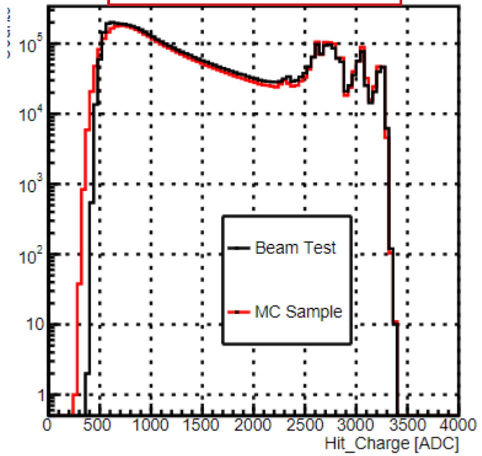
HG 80GeV e^-



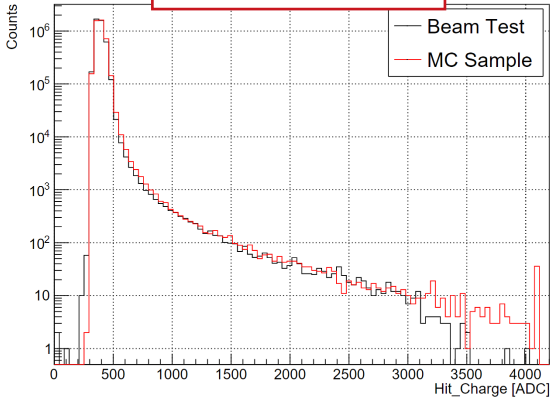
HG 10GeV π^-



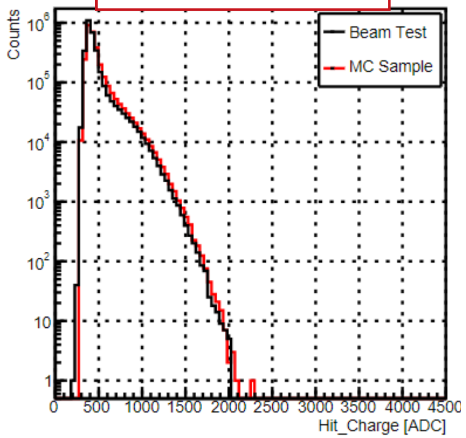
HG 80GeV π^-



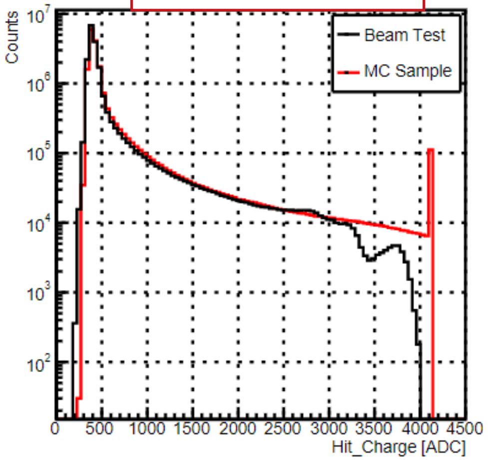
LG 100GeV μ^-



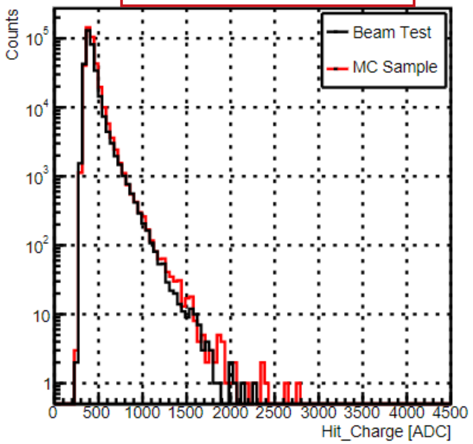
LG 10GeV e^-



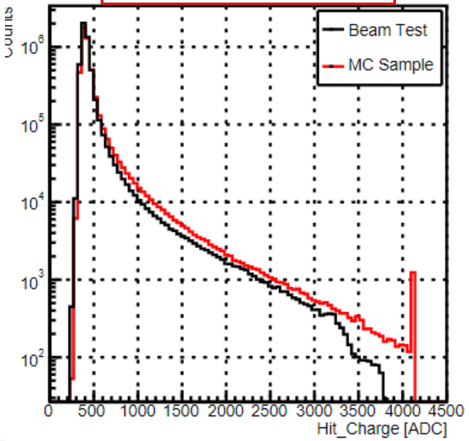
LG 80GeV e^-



LG 10GeV π^-

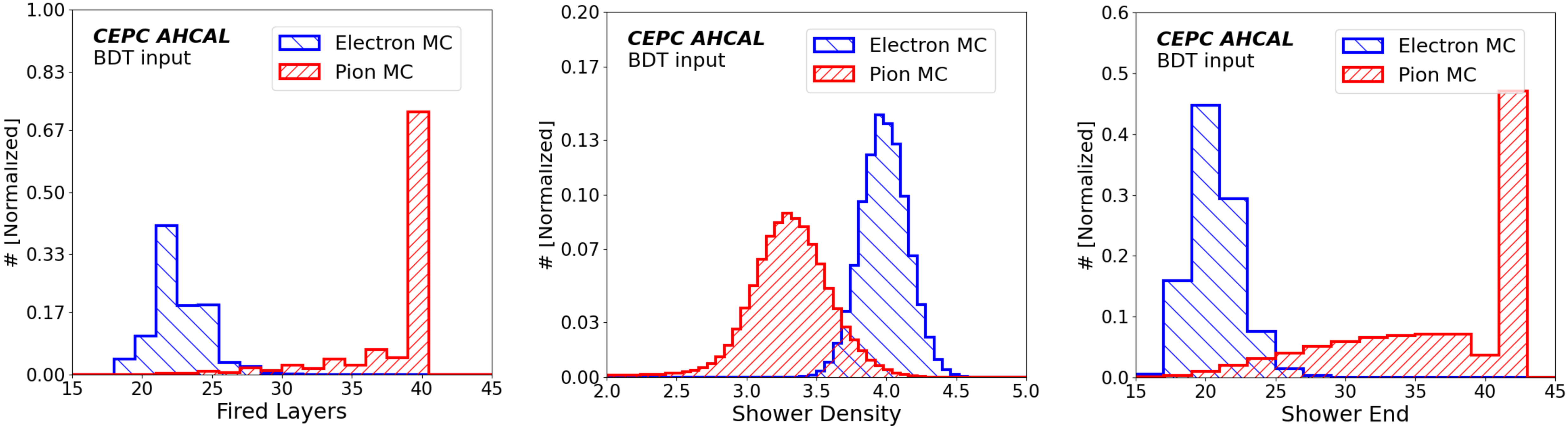


LG 80GeV π^-



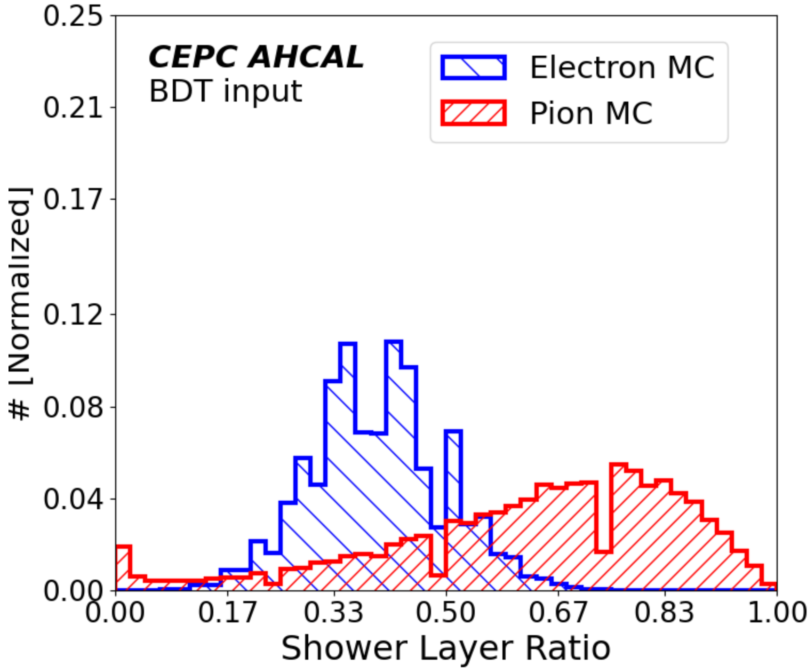
BDT input

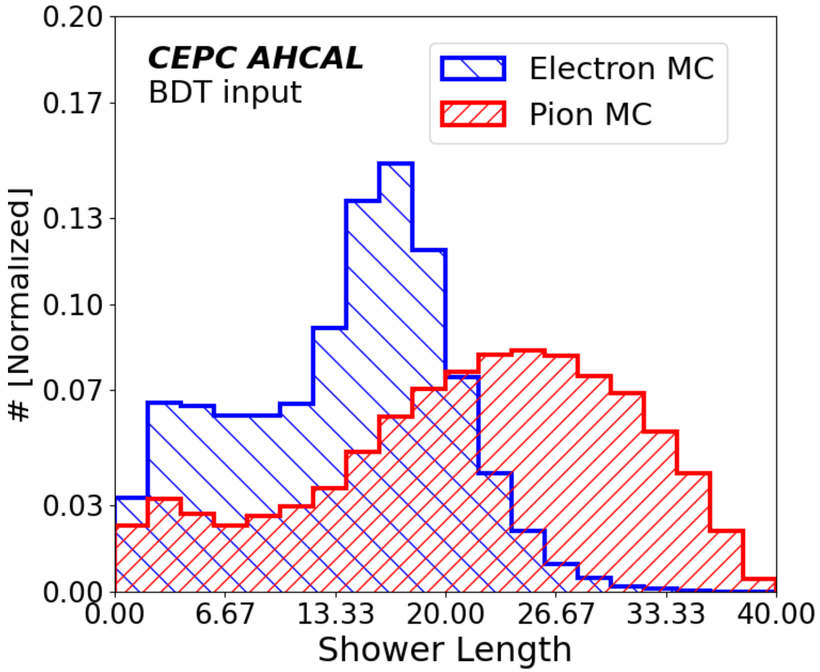
- **Fired layers**: The number of layer with hits.
- **Shower density**: The average number of neighboring hits around one hit, including the hit itself, in a 3×3 cell area in a given event is calculated.
- **Shower end**: After the shower starts, the first layer of two consecutive layers without 2 hits. If no shower is formed, it is set to 42.

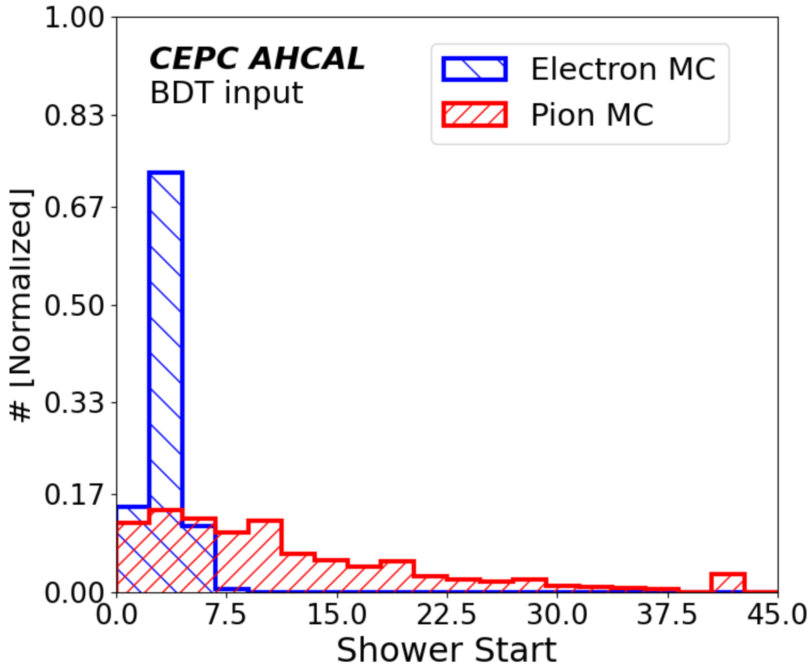


BDT input

- **Shower layer ratio**: Ratio of shower layers over fired layers.
- **Shower length**: The distance between the start of the shower and the layer where the maximum Root Mean Square (RMS) of hit transverse coordinates with respect to the z-axis occurs.
- **Shower start**: The first layer of the first three consecutive layers with at least 5 hits. For events without showers, the shower start layer is set to 42.

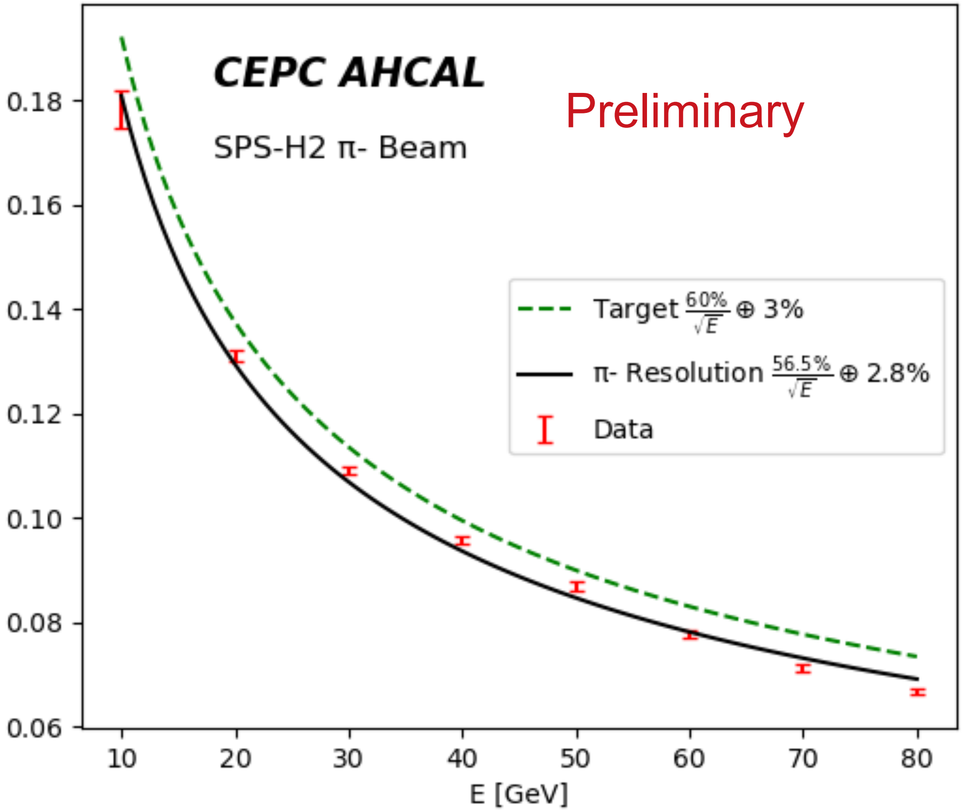


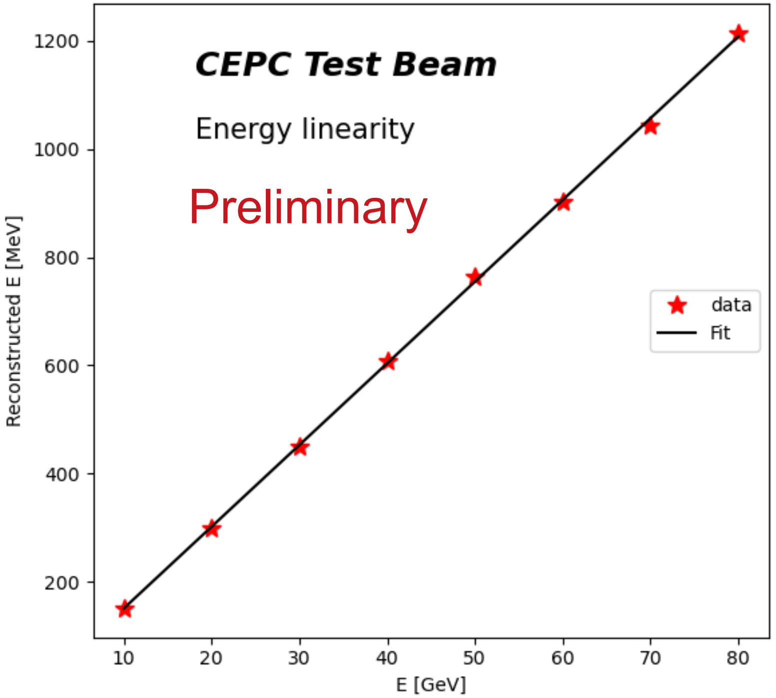


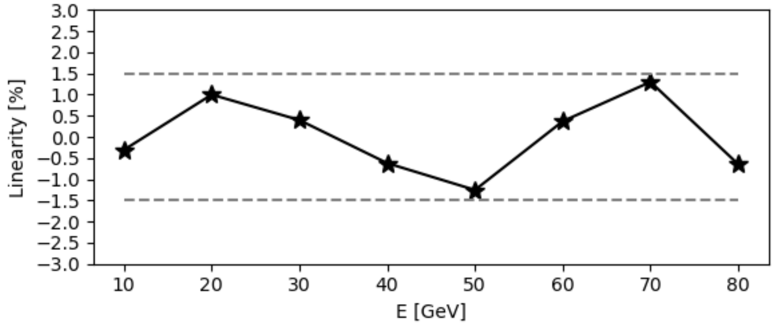


AHCAL Prototype Performance

- The Energy resolution is $\frac{56.5\,\%}{\sqrt{E}} \oplus 2.8\,\%$.
- The Energy linearity is $\pm 1.5\,\%$.







PID based on BDT

- **Apply Extreme Gradient Boosting (XGBoost)**
  - 12 variables are reconstructed (Signal: π, Bkg.: e , μ)
- MC samples to build $\mathrm{BDT}_{\mathrm{MC}-12}$,Data samples to build $\mathrm{BDT}_{\mathrm{Data}-12}$

MC samples

Shower density	1	-0.88	-0.65	0.0084	0.63	0.61	0.7	0.7	0.35	-0.25	0.98	0.51
Shower start	-0.88	1	0.79	0.0071	-0.69	-0.66	-0.81	-0.69	-0.36	0.32	-0.9	-0.39
Shower layer ratio	-0.65	0.79	1	0.28	-0.16	-0.28	-0.42	-0.14	-0.11	0.56	-0.6	-0.15
Shower length	0.0084	0.0071	0.28	1	0.27	-0.074	0.18	0.35	-0.19	0.67	0.02	0.49
Hits number	0.63	-0.69	-0.16	0.27	1	0.86	0.82	0.92	0.6	0.1	0.73	0.5
Shower radius	0.61	-0.66	-0.28	-0.074	0.86	1	0.71	0.73	0.83	-0.016	0.71	0.49
FD_1	0.7	-0.81	-0.42	0.18	0.82	0.71	1	0.79	0.43	-0.09	0.76	0.42
FD_6	0.7	-0.69	-0.14	0.35	0.92	0.73	0.79	1	0.43	0.13	0.77	0.51
Shower end	0.35	-0.36	-0.11	-0.19	0.6	0.83	0.43	0.43	1	0.13	0.46	0.48
Fired layers	-0.25	0.32	0.56	0.67	0.1	-0.016	-0.09	0.13	0.13	1	-0.21	0.46
Shower layers	0.98	-0.9	-0.6	0.02	0.73	0.71	0.76	0.77	0.46	-0.21	1	0.54
Z width	0.51	-0.39	-0.15	0.49	0.5	0.49	0.42	0.51	0.48	0.46	0.54	1

Correlation matrix

Data samples

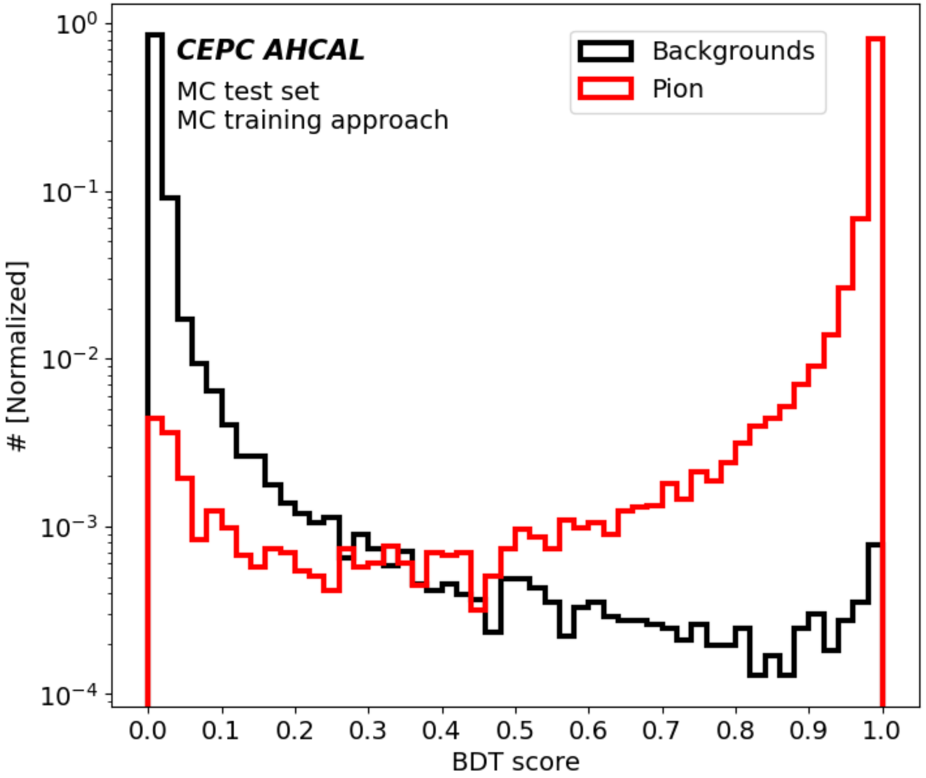
Shower density	1	-0.89	-0.65	-0.0012	0.62	0.64	0.7	0.68	0.36	-0.26	0.98	0.52
Shower start	-0.89	1	0.77	0.0073	-0.7	-0.7	-0.81	-0.69	-0.39	0.33	-0.9	-0.42
Shower layer ratio	-0.65	0.77	1	0.29	-0.17	-0.32	-0.42	-0.12	-0.12	0.57	-0.59	-0.16
Shower length	-0.0012	0.0073	0.29	1	0.27	-0.085	0.18	0.35	-0.17	0.67	0.016	0.49
Hits number	0.62	-0.7	-0.17	0.27	1	0.87	0.82	0.9	0.63	0.076	0.74	0.52
Shower radius	0.64	-0.7	-0.32	-0.085	0.87	1	0.72	0.72	0.83	-0.072	0.75	0.51
FD_1	0.7	-0.81	-0.42	0.18	0.82	0.72	1	0.78	0.45	-0.11	0.77	0.44
FD_6	0.68	-0.69	-0.12	0.35	0.9	0.72	0.78	1	0.45	0.12	0.77	0.52
Shower end	0.36	-0.39	-0.12	-0.17	0.63	0.83	0.45	0.45	1	0.11	0.49	0.5
Fired layers	-0.26	0.33	0.57	0.67	0.076	-0.072	-0.11	0.12	0.11	1	-0.22	0.43
Shower layers	0.98	-0.9	-0.59	0.016	0.74	0.75	0.77	0.77	0.49	-0.22	1	0.57
Z width	0.52	-0.42	-0.16	0.49	0.52	0.51	0.44	0.52	0.5	0.43	0.57	1

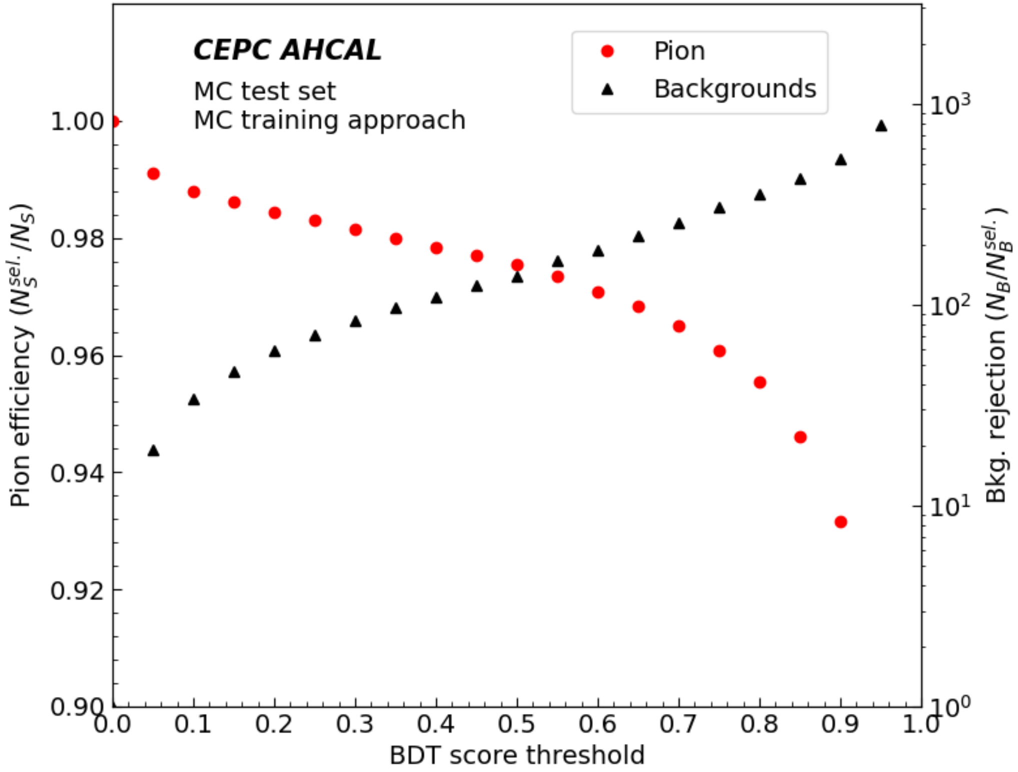
Correlation matrix

PID based on BDT

- **MC training approach**

Pion efficiency	90%	95%	99%
Bkg. rejection (1/Bkg. efficiency)	1701.2	617.4	29.6

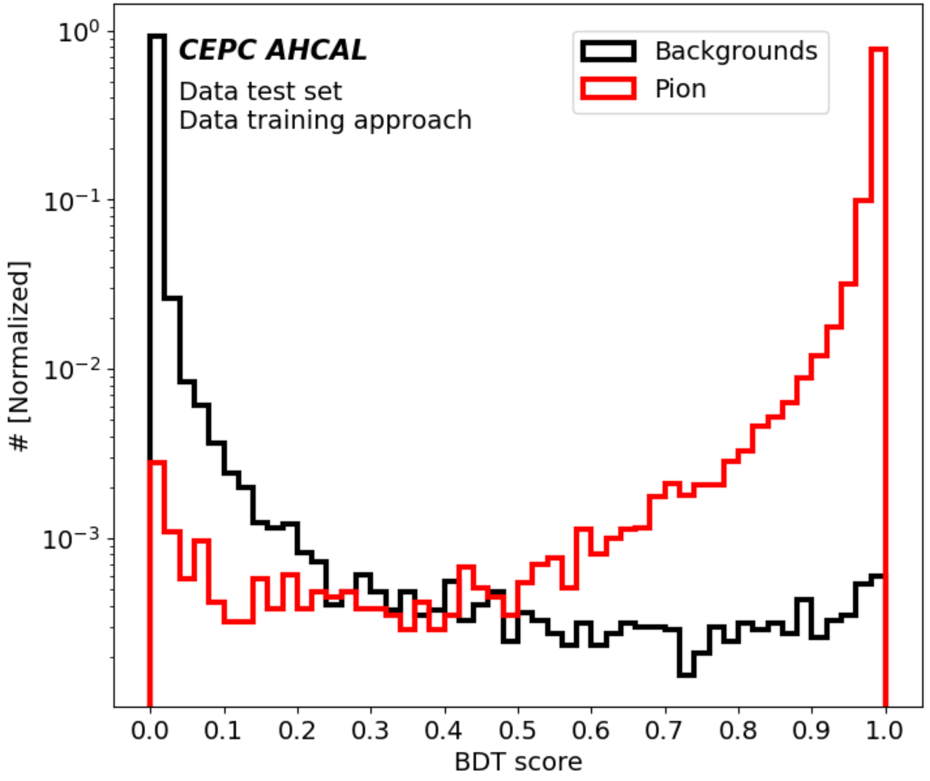


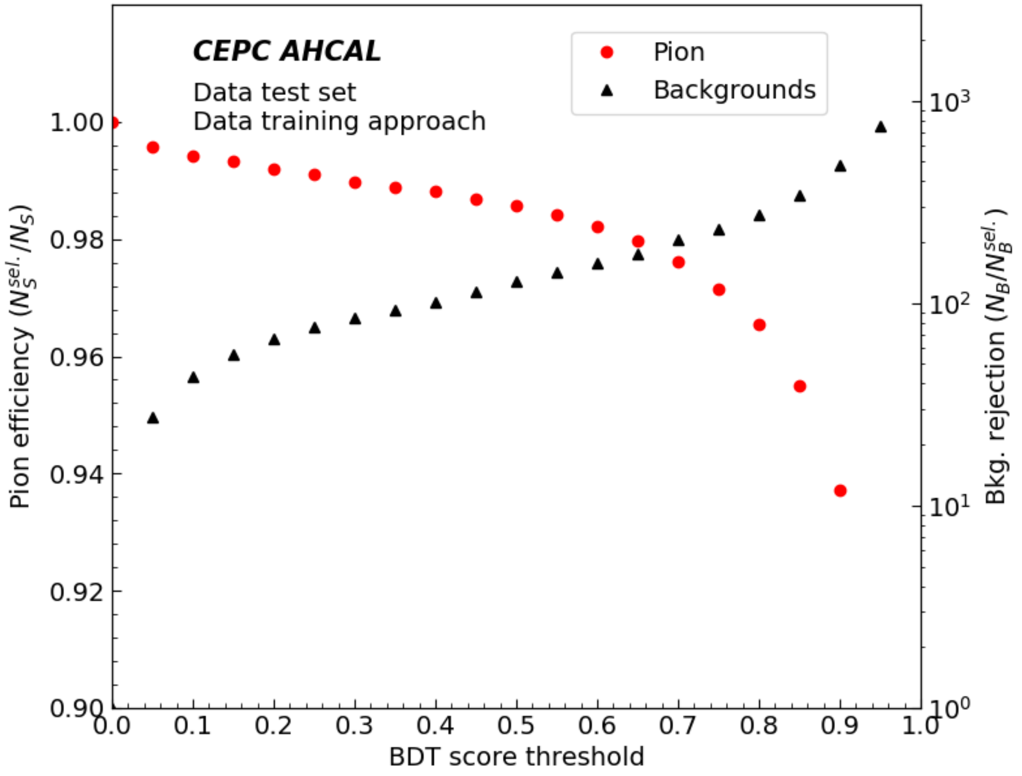


PID based on BDT

- **Data training approach**

Pion efficiency	90%	95%	99%
Bkg. rejection (1/Bkg. efficiency)	1448.5	691.6	143.0

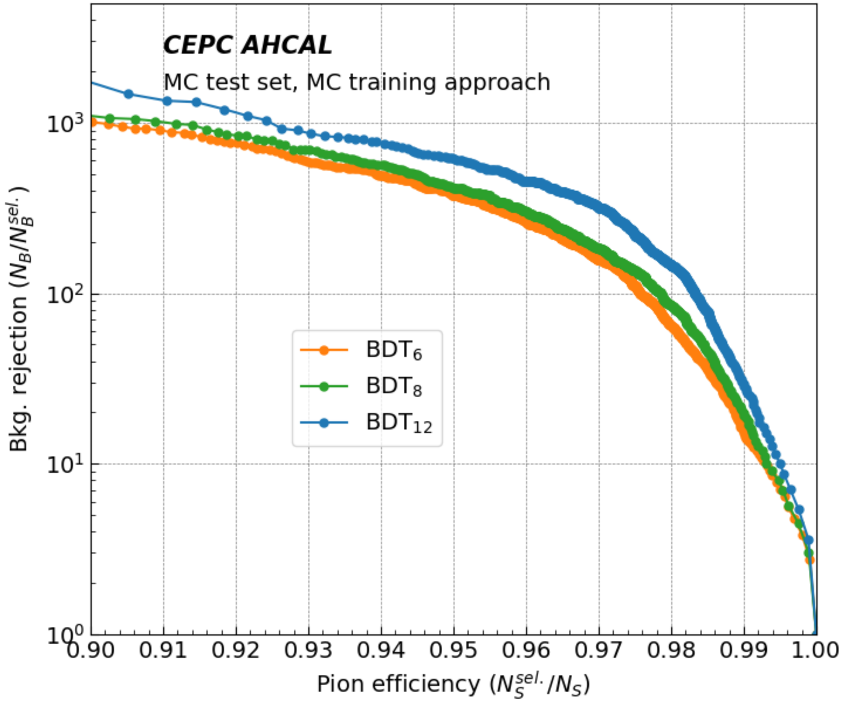


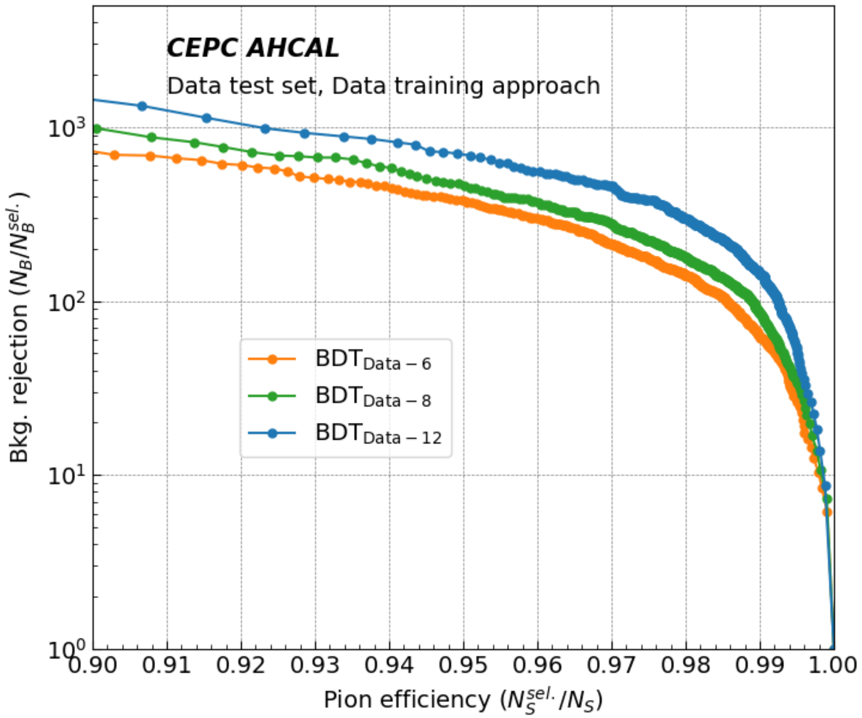


PID based on BDT

- **We observe dependence of BDT performance on input variables**
  - Remove Shower End, Shower Layers, Fired Layers, and Z Width to build BDT with 8 inputs.
  - Further remove FD_1 and FD_6 to build BDT with 6 inputs.

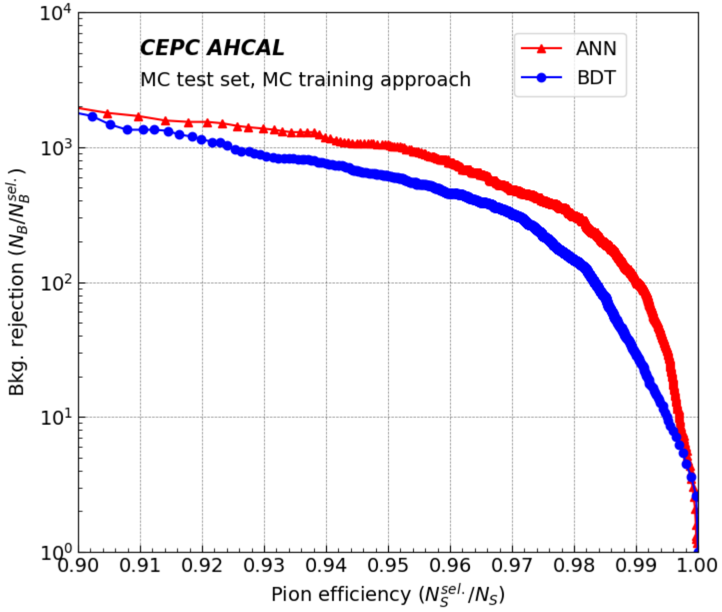
- **Feature engineering matters in BDT**
  - BDT optimization is on going.

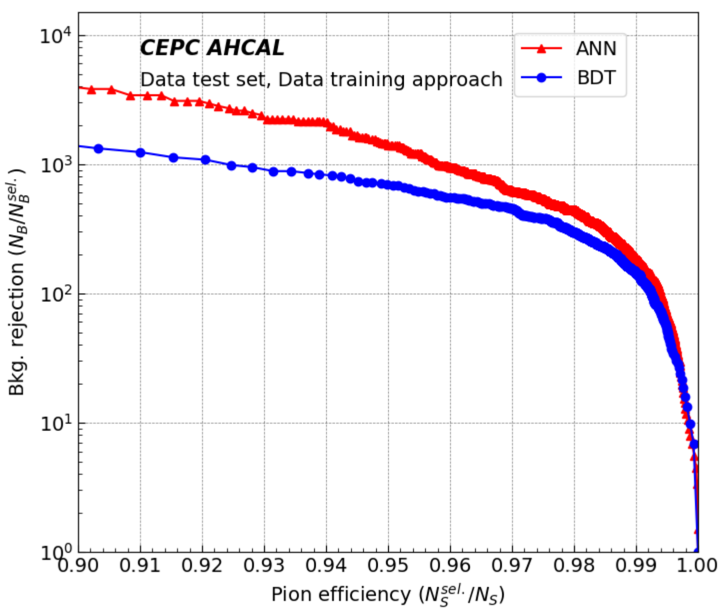




Comparison between ANN and BDT

- **We observe obvious improvement in terms of Background rejection when tested on two sets of samples.**

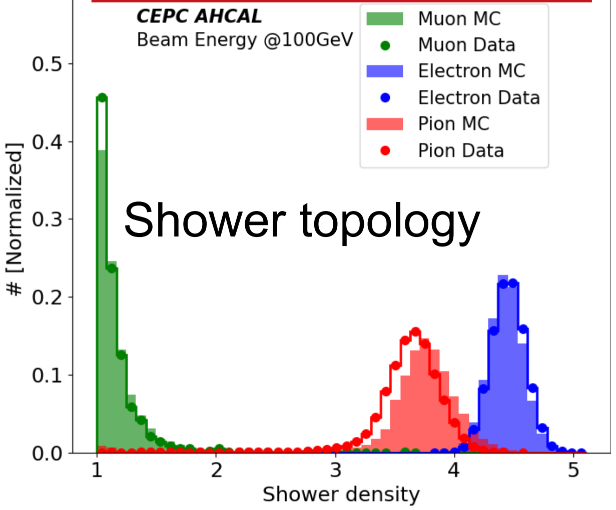




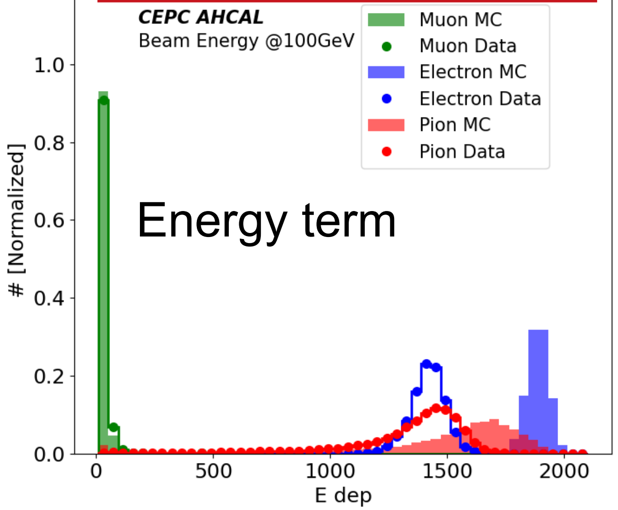
Pion efficiency	90%		95%		99%	
	MC	Data	MC	Data	MC	Data
BDT bkg. rejection	1701.2	1448.5	617.4	691.6	29.6	143.0
ANN bkg. rejection	**2015.7**	**3811.2**	**1040.3**	**1408.5**	**103.9**	**187.8**
Improvement	18.49%	163.12%	68.51%	103.65%	251.14%	31.37%

- **Quick view on shower topology variables**
  - Data samples are pre-selected by Cherenkov detectors and rough F.D. cut.
  - MC and Data are close in Shower topology term .
  - Obvious discrepancy in energy term.

Mean hits number in a 3×3 cell.



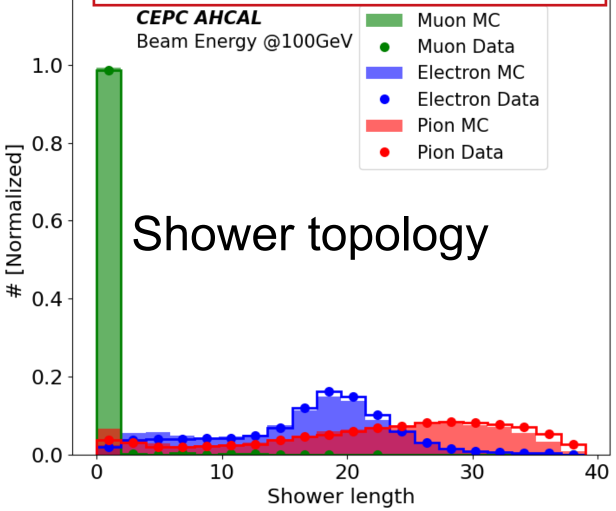
Total energy deposited in AHCAL.



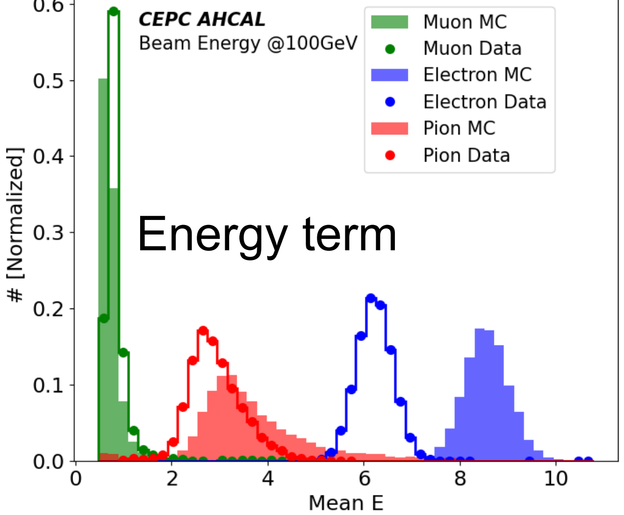
- **Currently**
  - ✓Do Research on PID approach.
    - Separate Data training and MC training.
  - ❖Trained on MC and then applied on Data.
    - Discrepancy between Data and MC.
    - Bridging gap is on going.

Muon data: SPS_2023/100GeV_mu_Run25.
Electron data: SPS_2023/100GeV_ e_run267.
Pion data: SPS/100GeV_pi_run230.

Distance between the start of the shower and the layer with maximum RMS of hit transverse coordinates.



Mean energy deposited in the cell.



PID based on BDT

- **Apply Extreme Gradient Boosting (XGBoost)**
  - 12 variables are reconstructed (Signal: π, Bkg.: e , μ)
- MC samples to build $\mathrm{BDT}_{\mathrm{MC-12}}$,Data samples to build $\mathrm{BDT}_{\mathrm{Data-12}}$

MC samples

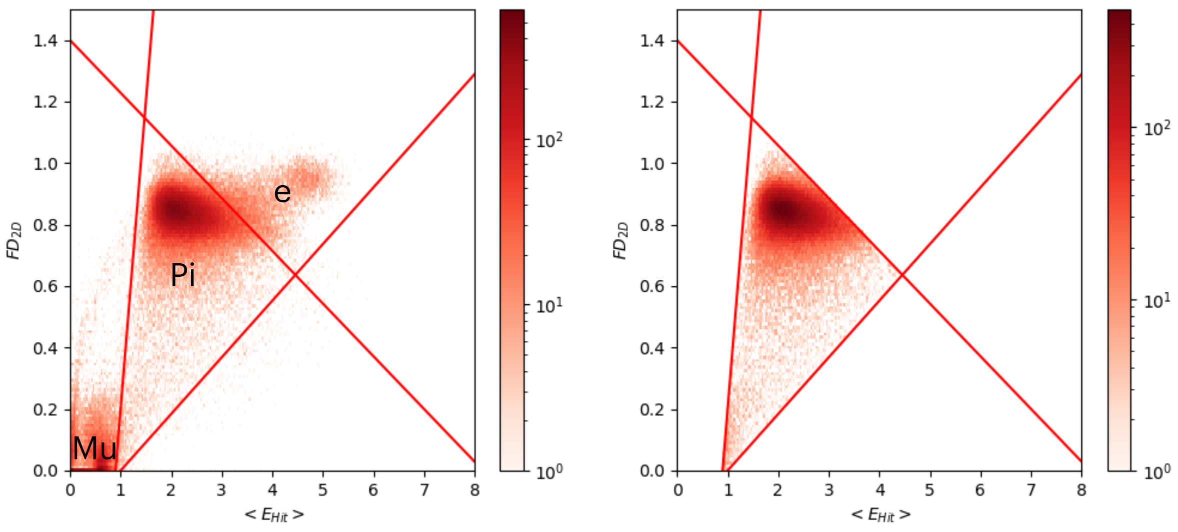
Rank: Variable	Variable weight
1: Shower radius	0.377
2: Shower layers	0.232
3: Hits number	0.088
4: Fired layers	0.083
5: Shower start	0.080
6: Shower density	0.049
7: Z width	0.034
8: FD_6	0.017
9: FD_1	0.015
10: Shower layer ratio	0.014
11: Shower end	0.006
12: Shower length	0.006

Data samples

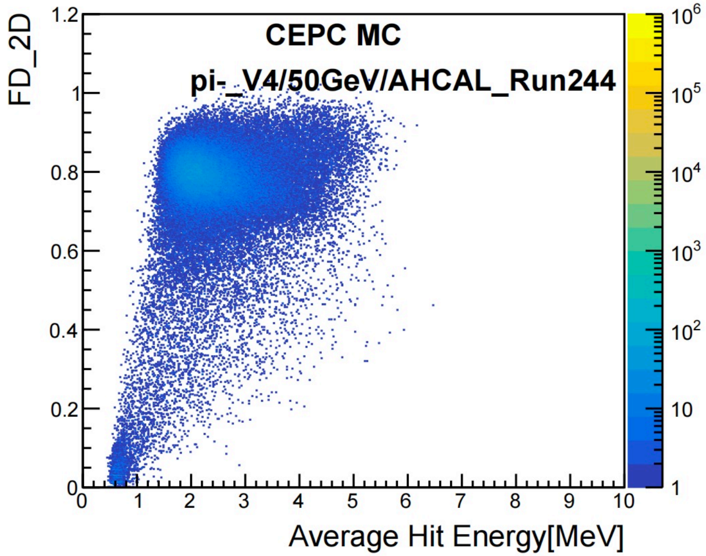
Rank: Variable	Variable weight
1: Shower radius	0.379
2: Shower layers	0.228
3: Hits number	0.133
4: Shower density	0.058
5: Fired layers	0.058
6: Z width	0.042
7: Shower start	0.039
8: FD_6	0.019
9: FD_1	0.016
10: Shower layer ratio	0.010
11: Shower length	0.010
12: Shower end	0.008

Data pre-selection

- Collect pion samples in 20pion run files.
- Cut approach is guided by MC.



FD cut

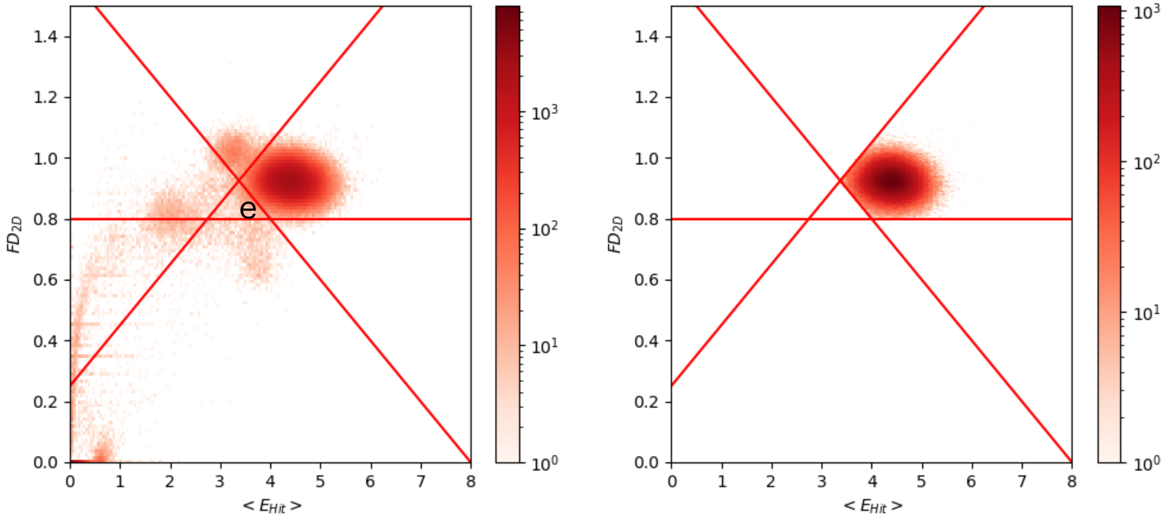


MC

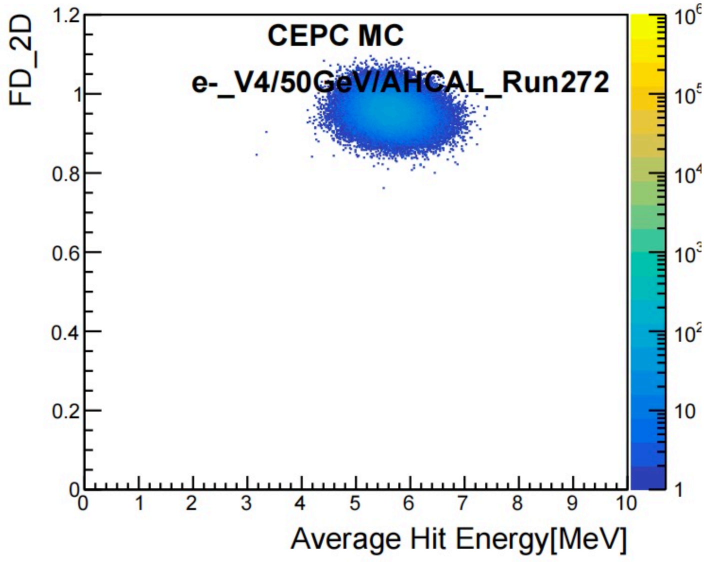
For data collected in 2023, SPS and PS

Data pre-selection

- Collect e samples in e run files.
- Cut approach is guided by MC.



FD cut



MC

For data collected in 2023, SPS and PS