上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

SJTU

Machine-learning-based Particle Identification in using CEPC AHCAL Prototype

Siyuan SONG, Jiyuan CHEN

Advisor: Prof. Haijun YANG

2023-09-06

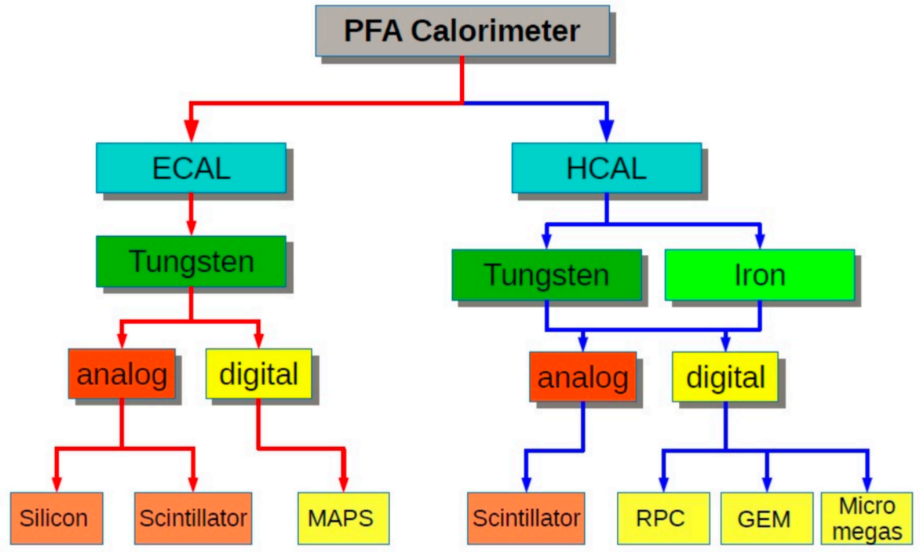
饮 水 思 源 • 爱 国 荣 校

Outline

1. Introduction.
2. Monte Carlo Samples & Test Beam Samples.
3. PID based on BDT.
4. PID based on ANN.
5. Purity of 2022 AHCAL Test Beam.
6. Summary.

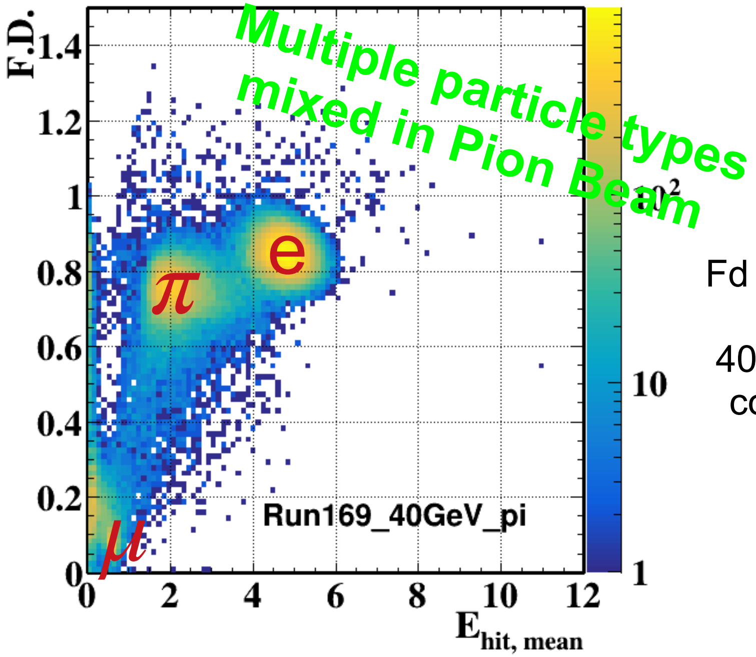
Introduction

- **Exploration for multiple PID methods in high-granularity calorimeters**
  - Several PFA oriented high-granularity calorimeters have been developed.



Overview of world-wide development of high-granularity calorimeters

- **CEPC AHCAL prototype test beam data require PID**
  - Contamination in test beam data collected in 2022 at CERN SPS-H8.
  - Main task is purifying Pion beams.

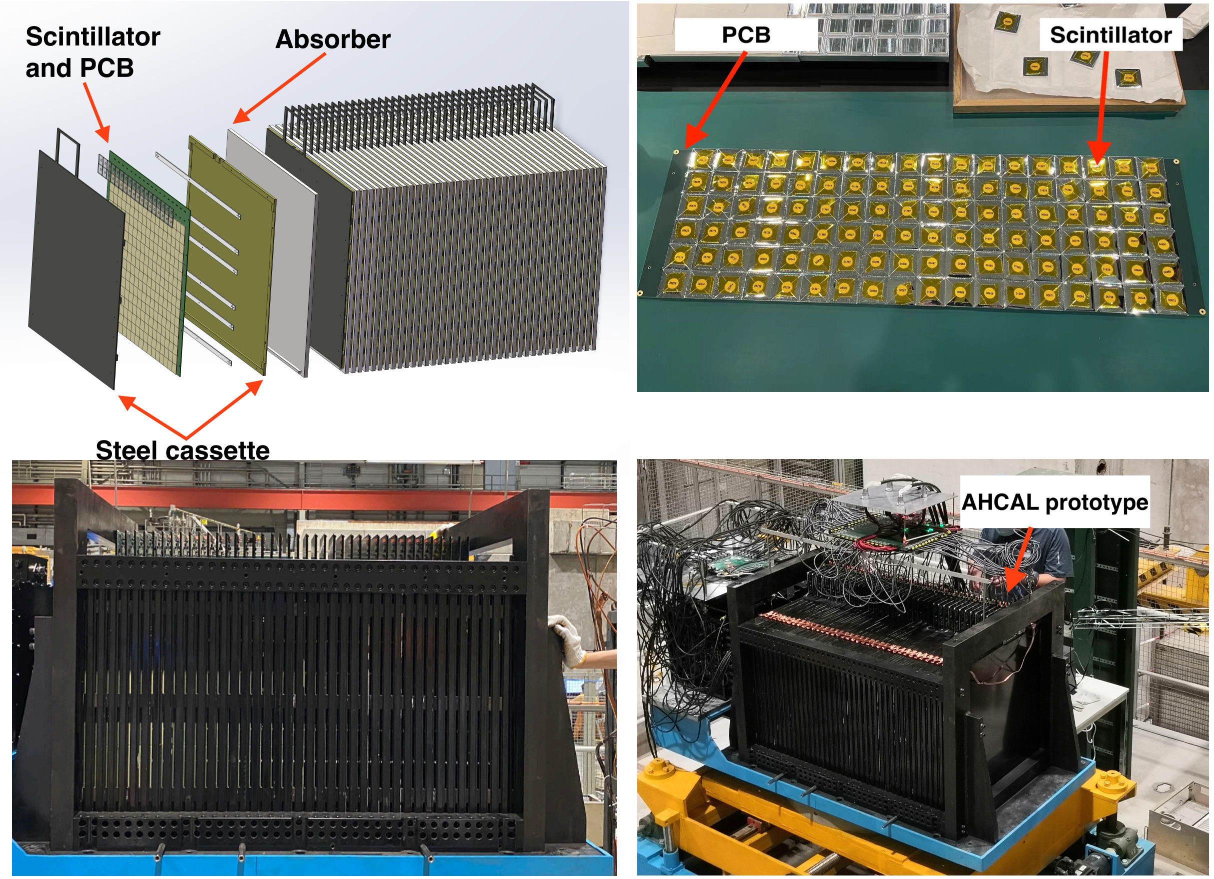


Fd vs mean hit energy

40 GeV Pion Run file collected in 2022 at CERN SPS-H8

Introduction

- **Utilize High-granularity CEPC AHCAL prototype in this PID research**



- **CEPC AHCAL prototype parameter**

- Geometry

- 40 sampling layers.
- 72cm $\times$ 72cm in transversal plane.
- 120cm in longitudinal direction.

- Absorber

- 2 cm thickness/layer steel.

- Sensitive cells

- 40mm $\times$ 40mm $\times$ 3mm scintillator tile coupled with SiPM.

- Electronics readout channels
  - 12960 ($18 \times 18 \times 40$).

Monte Carlo Samples & Test Beam Samples

- **Monte Carlo Samples:** Employ Geant4 11.1.1 Toolkit with the $\mathrm{QGSP}_{\mathrm{BERT}}$ physics list.

Energy point	5 GeV		10 GeV		30 GeV		50 GeV		60 GeV		80 GeV		100 GeV		120 GeV	
	#	Source	#	Source	#	Source	#	Source	#	Source	#	Source	#	Source	#	Source
Muon	10k	MC	10k	MC	10k	MC	10k	MC	10k	MC	10k	MC	10k	MC	10k	MC
Electron	10k	MC	10k	MC	10k	MC	10k	MC	10k	MC	10k	MC	10k	MC	10k	MC
Pion	10k	MC	10k	MC	10k	MC	10k	MC	10k	MC	10k	MC	10k	MC	10k	MC

- **Test Beam Samples:** Pre-processed purer 2023 CERN SPS-H2 & PS-T9 test beam data.

Energy point	5 GeV		10 GeV		30 GeV		50 GeV		60 GeV		80 GeV		100 GeV		120 GeV	
	#	Source	#	Source	#	Source	#	Source	#	Source	#	Source	#	Source	#	Source
Muon	-	-	40k	Data	-	-	-	-	-	-	-	-	40k	Data	-	-
Electron	10k	Data	10k	Data	10k	Data	10k	Data	10k	Data	10k	Data	10k	Data	10k	Data
Pion	10k	Data	10k	Data	10k	Data	10k	Data	10k	Data	10k	Data	10k	Data	10k	Data

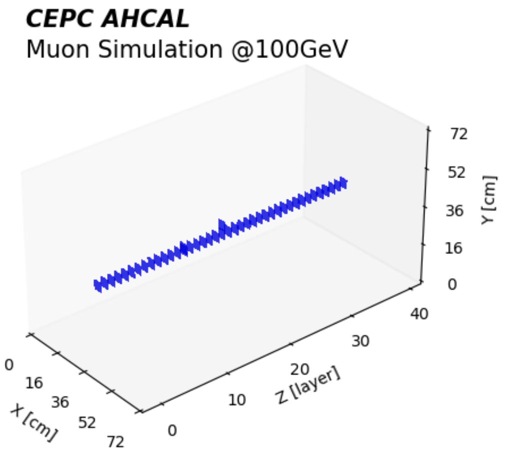
<u>Each sample set is split to a Train set and a Test set in a ration of 3:2.</u>
<u>Each sample set would be utilized to build classifiers.</u>

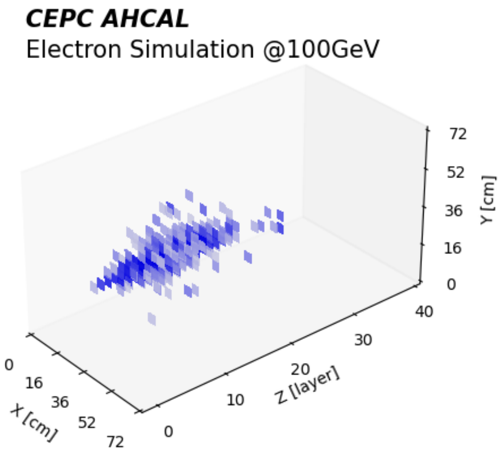
Monte Carlo Samples & Test Beam Samples

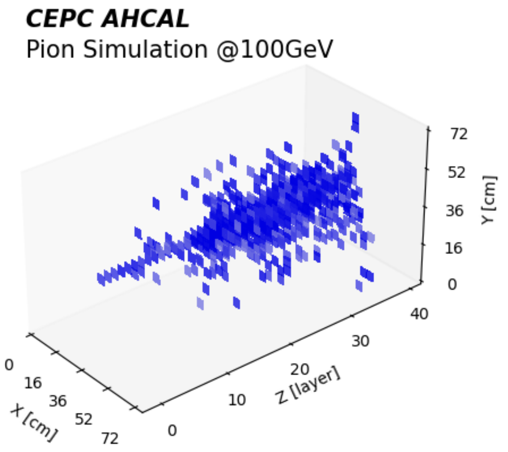
Event Display

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

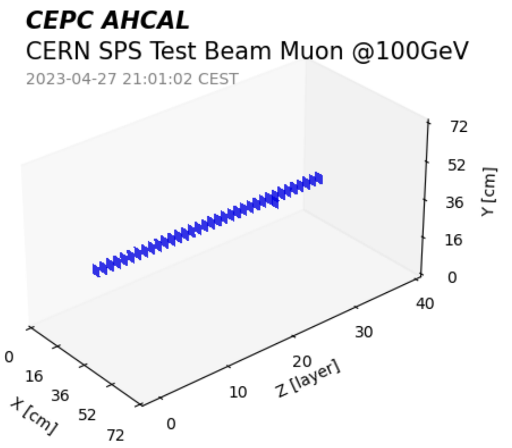
Monte Carlo Samples

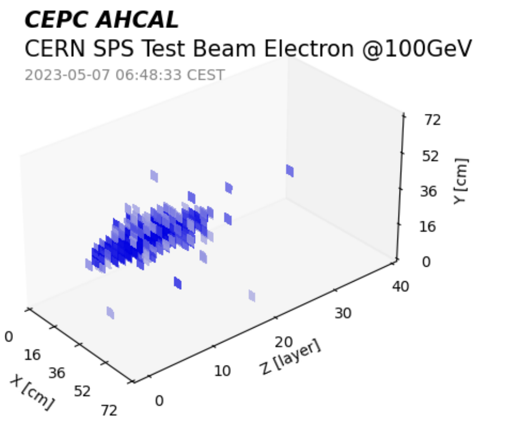


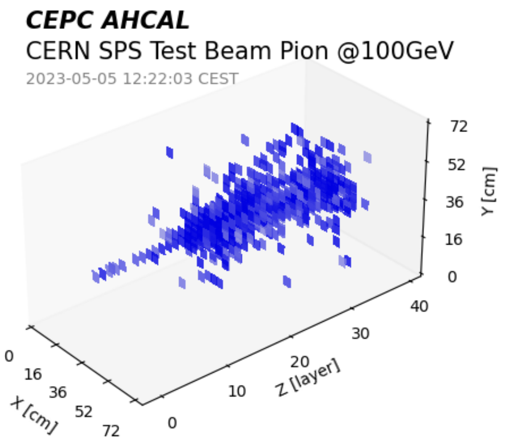




Test Beam Samples







PID based on BDT

- **Correlation Matrix of 12 input variables.**

MC samples

	Shower density	Shower start	Shower layer ratio	Shower length	Hits number	Shower radius	FD_1	FD_6	Shower end	Fired layers	Shower layers	Z width
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

Data samples

	Shower density	Shower start	Shower layer ratio	Shower length	Hits number	Shower radius	FD_1	FD_6	Shower end	Fired layers	Shower layers	Z width
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

PID based on BDT

- **Apply XGBoost**
- **Variable Ranking in Pion identification**
  - Shower radius,
  - Shower layers ,
  - Hits number are important.
  - (Signal: π, Background: e & μ)
- MC samples to build $\mathrm{BDT}_{\mathrm{MC-12}}$
- Data samples to build $\mathrm{BDT}_{\mathrm{Data-12}}$

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

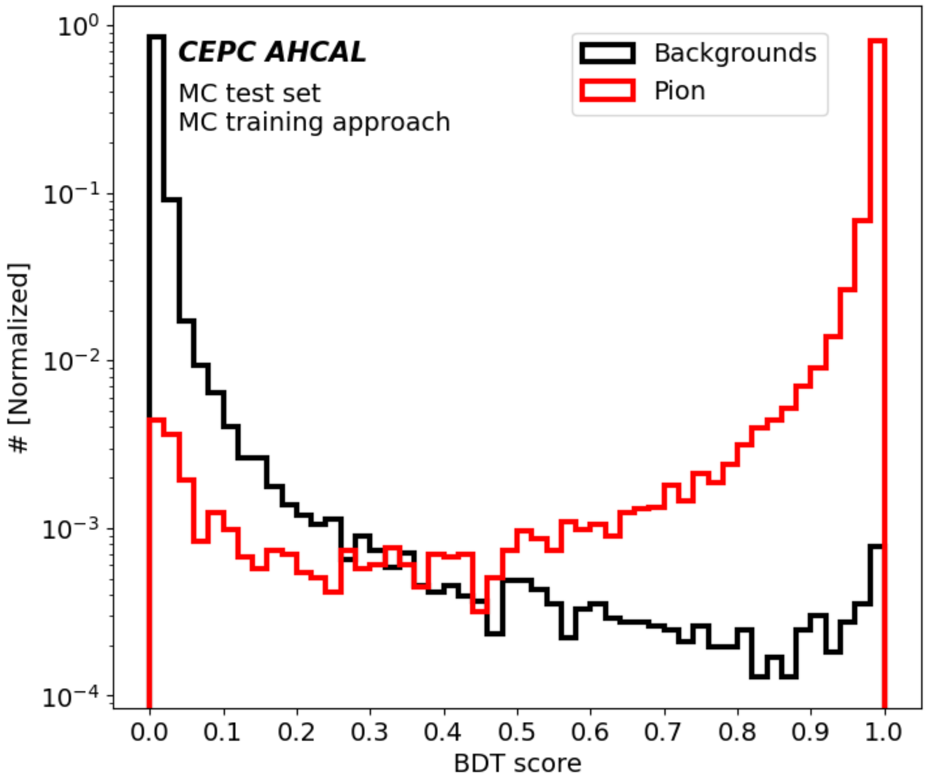
MC samples

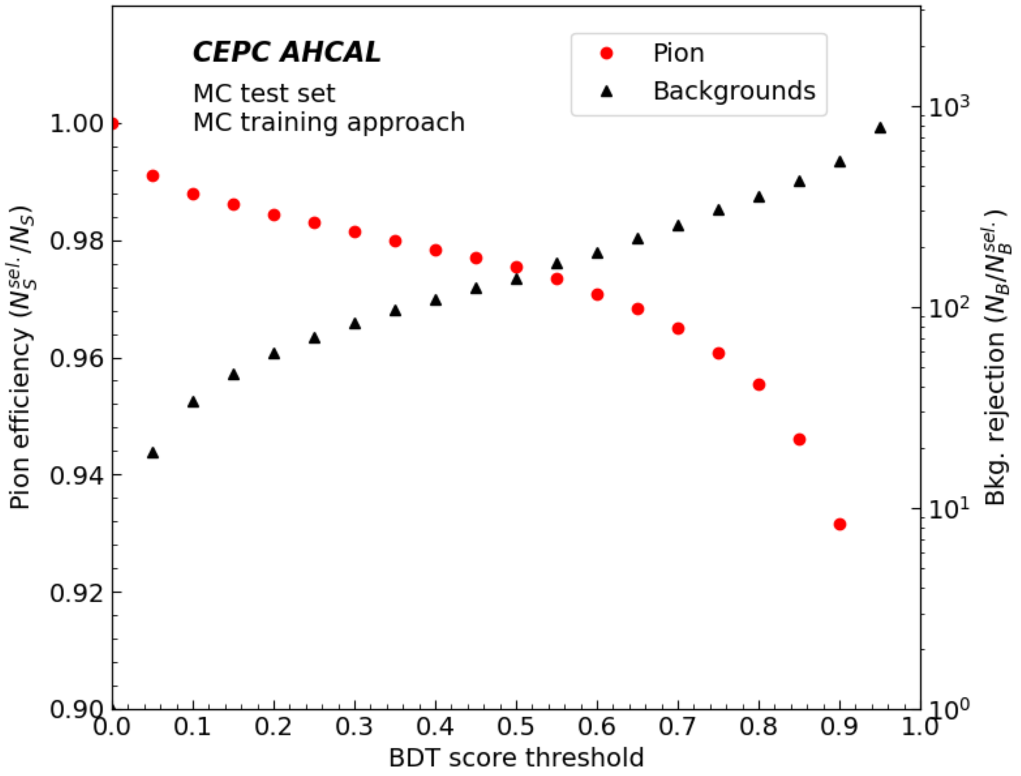
Rank: Variable	Variable weight
1: Shower radius	0.379
2: Shower layers	0.228
3: Hits number	0.133
4: Shower density	0.058
5: Fired layers	0.058
6: Z width	0.042
7: Shower start	0.039
8: FD_6	0.019
9: FD_1	0.016
10: Shower layer ratio	0.010
11: Shower length	0.010
12: Shower end	0.008

Data samples

PID based on BDT

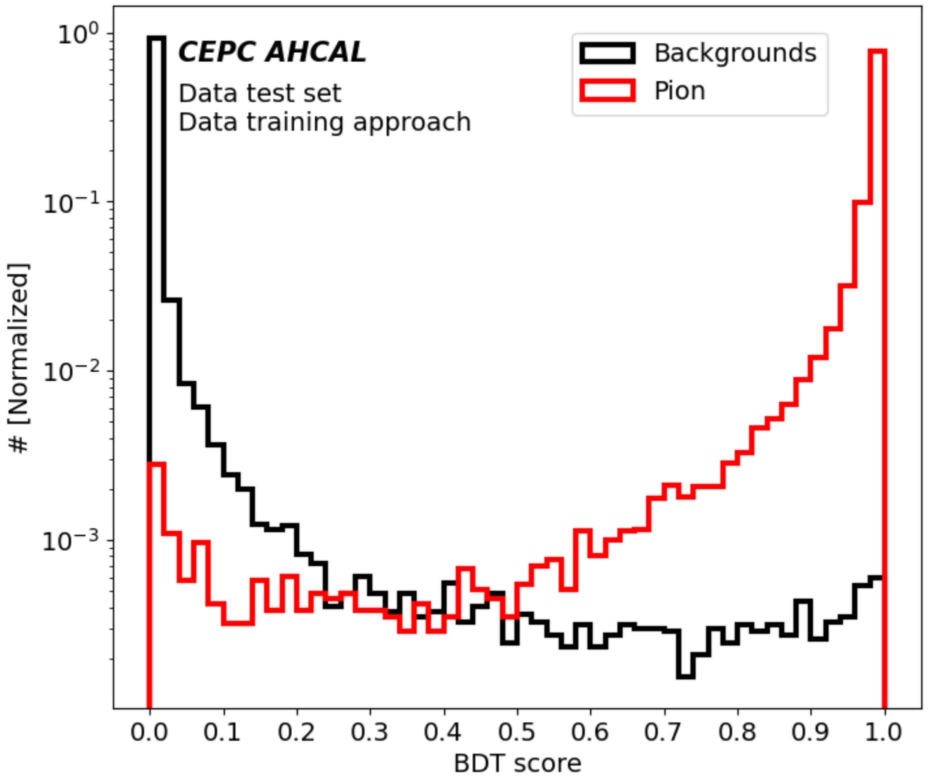
- **MC training approach**
  - At 99% pion signal efficiency, Bkg. Rejection is 29.6 ($N_{Bkg.}/N^{sel.}_{Bkg.}$)

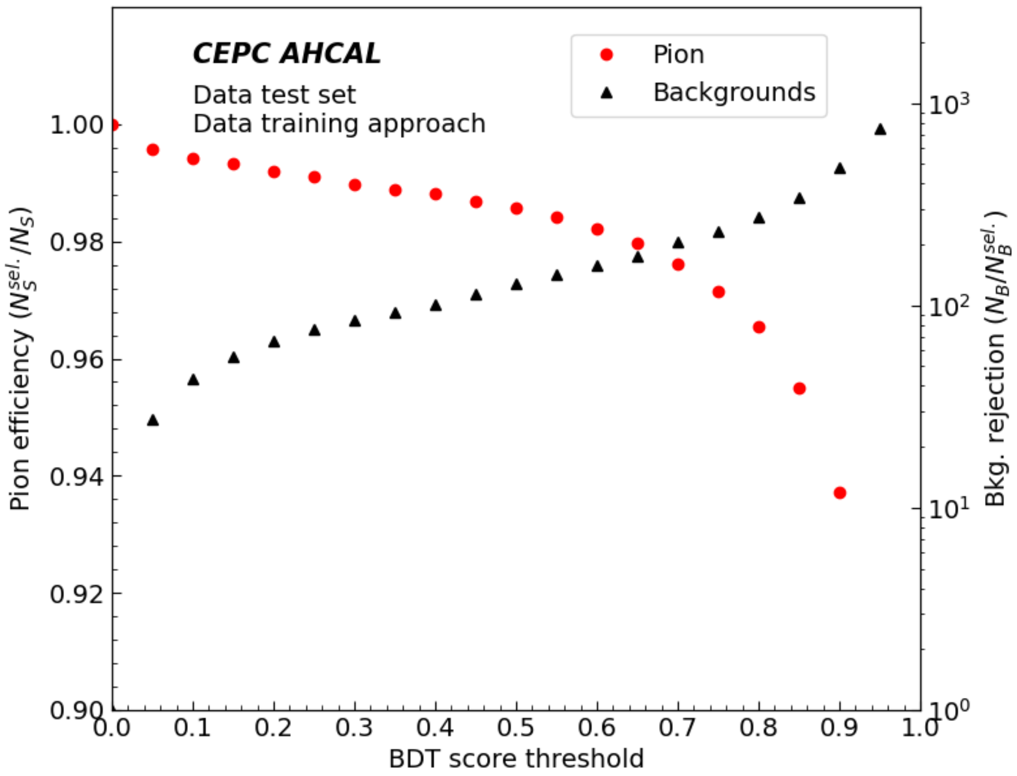




PID based on BDT

- **Data training approach**
  - At 99% pion signal efficiency, Bkg. Rejection is 143.0 ($N_{Bkg.}/N^{sel.}_{Bkg.}$)

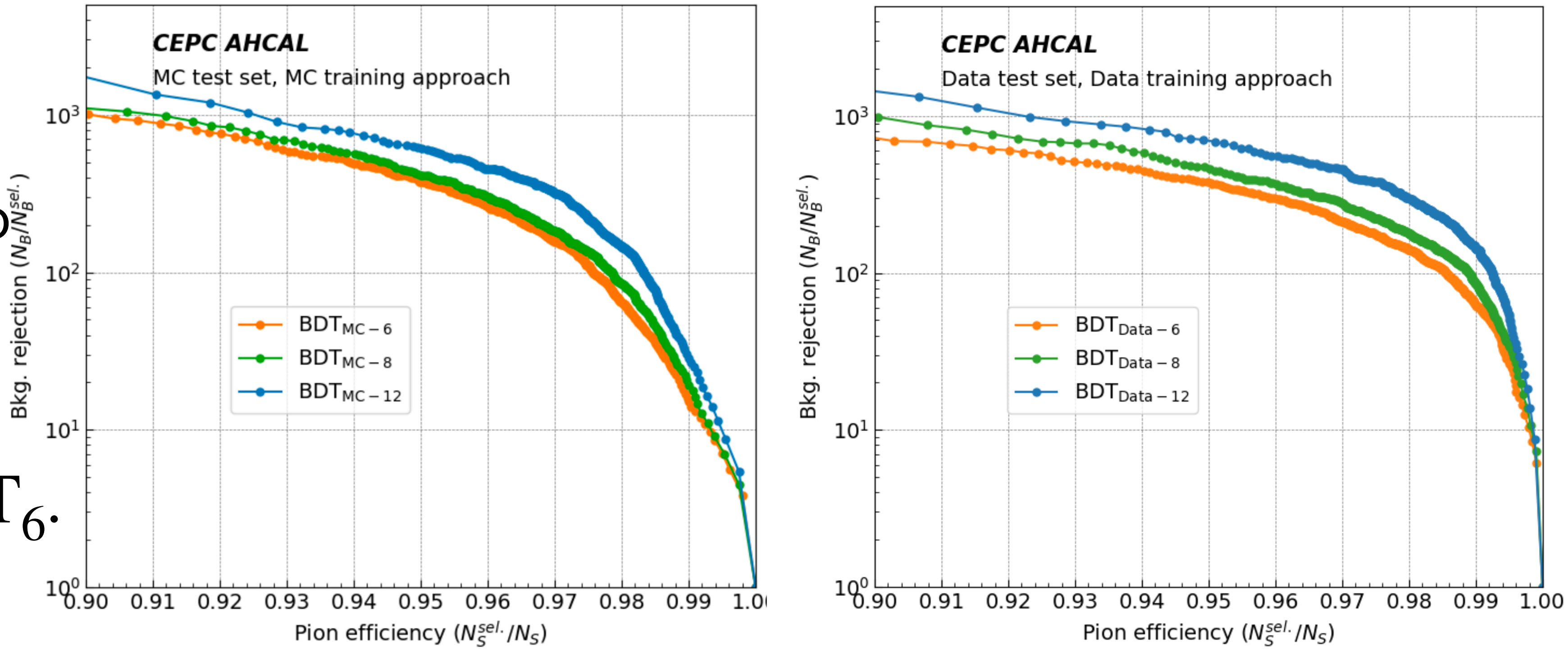




PID based on BDT

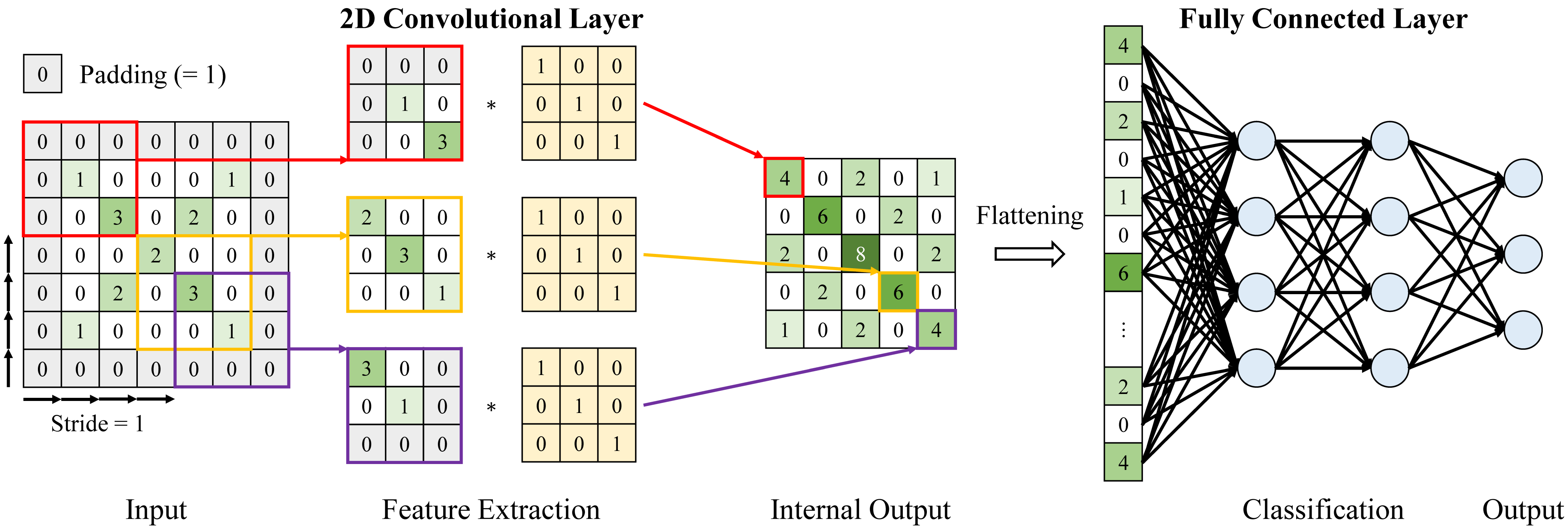
- **Dependence of BDT performance on input variables**
  - Remove Shower End, Shower Layers, Fired Layers, and Z Width to build BDT_8.
  - Further remove FD_1 and FD_6 to build BDT_6.

- **Feature engineering matters in BDT**
  - Increasing variable number can improve BDT



PID based on ANN

- **Cell-based Artificial Neural Networks (ANN) make full use of high-dimensional input ($18 \times 18 \times 40$).**
  - Compile layers to extract features.
  - Output is the probability of each particle type candidate.

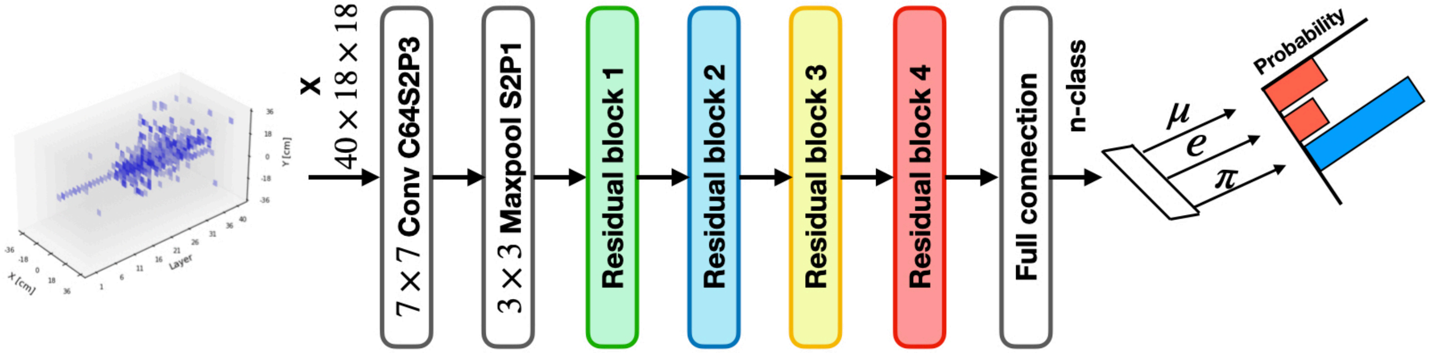


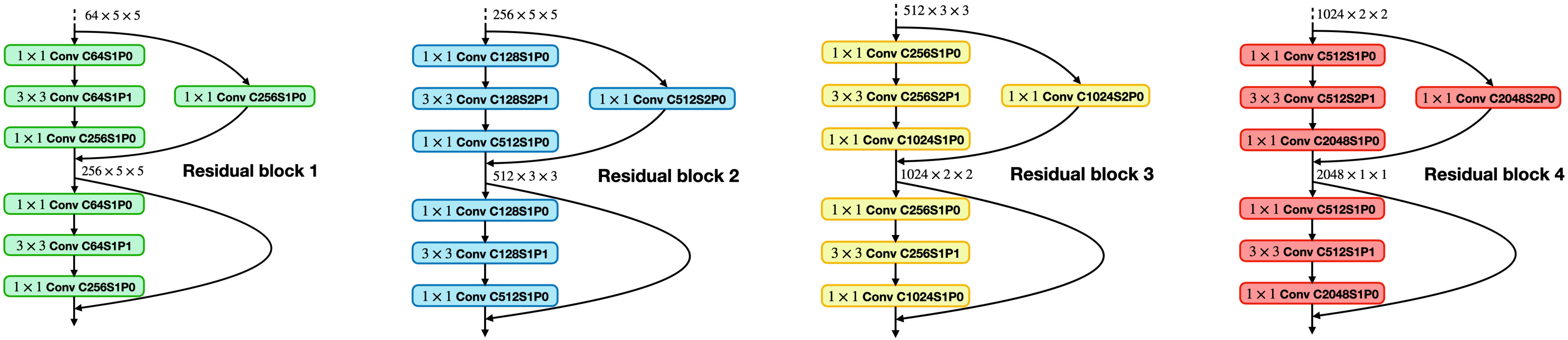
PID based on ANN

- **Architecture: take the advantage of the Residual Block**

Input: energy deposits in AHCAL $(18 \times 18 \times 40)$.

output: probability of each particle type candidate.

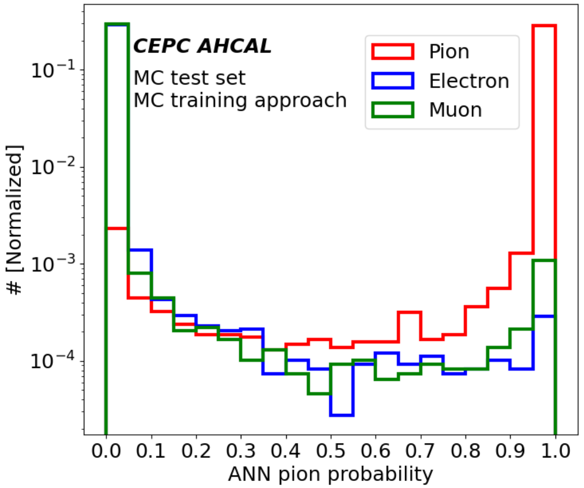


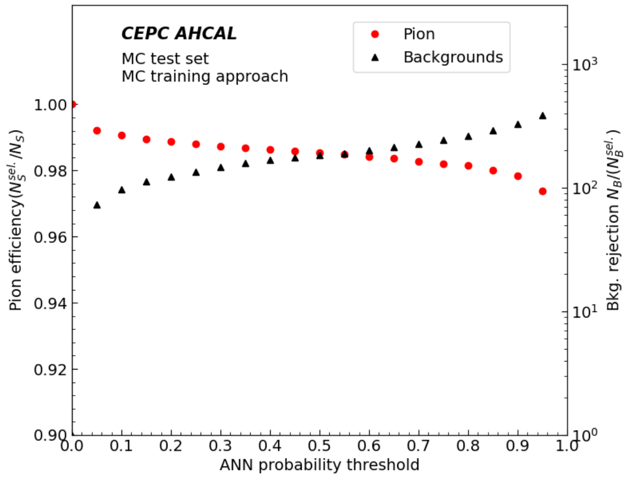


He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

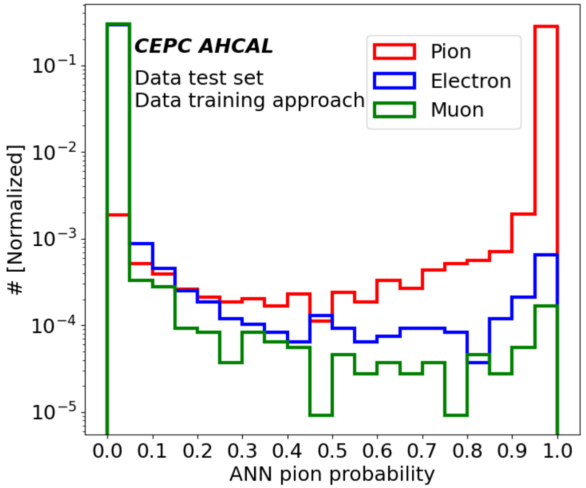
PID based on ANN

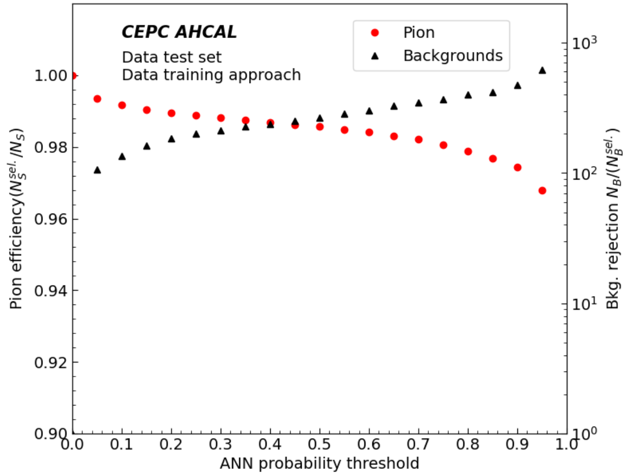
- $\mathrm{ANN_{MC}}$ **is trained on MC samples.**
  - At 99% pion efficiency, background rejection is 103.9, pion purity is 98.0%





- $\mathrm{ANN_{Data}}$ **is trained on Test Beam samples.**
  - At 99% pion efficiency, background rejection is 187.8, pion purity is 98.9%

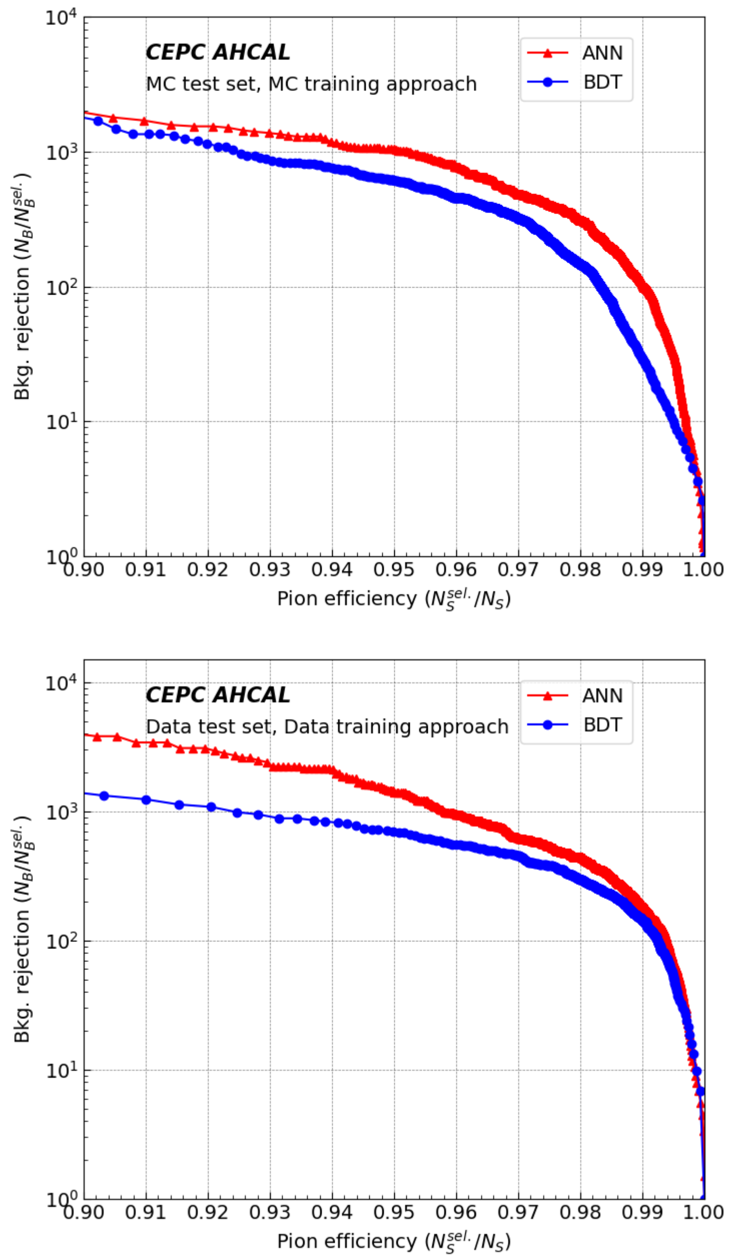




Comparison between ANN and BDT

- **We observe better performance of ANN in terms of Background rejection and Pion purity.**

	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%**
BDT pion purity	0.998	0.998	0.996	0.996	0.923	0.983
ANN pion purity	**0.999**	**0.999**	**0.998**	**0.998**	**0.980**	**0.989**
Improvement	**↑ 0.05%**	**↑ 0.13%**	**↑↑ 0.21%**	**↑ 0.22%**	**↑ 6.20%**	**↑ 0.61%**

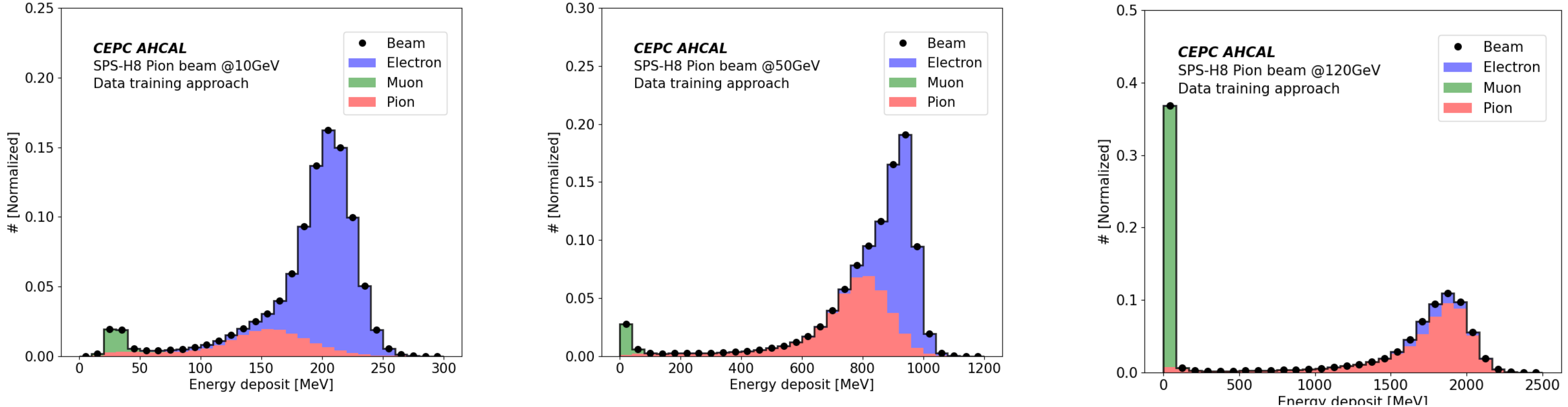


Discussion

- **ANN 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.

2022 SPS-H8 beam composition

Apply the ANN classifier to SPS-H8 beam collected in 2022.



Energy [GeV]	10	20	30	40	50	60	70	80	90	100	120
Muon percentage [%]	3.6	4.3	3.2	3.3	3.1	3.4	3.1	3.7	5.5	28.4	36.3
Electron percentage [%]	78.7	61.7	64.2	58.3	51.7	41.9	41.2	36.0	31.1	20.3	7.9
Pion percentage [%]	**17.7**	**34.0**	**32.6**	**38.5**	**45.3**	**54.8**	**55.7**	**60.3**	**63.4**	**51.3**	**55.8**

Summary

1. BDT and ANN based PID methods are all developed.
2. The preliminary purity of CEPC AHCAL test beam data on the type of incident particles is given.
3. This research promises the application prospect of cell-based ANN classifier in high-granularity calorimeters.

感谢聆听

饮 水 思 源　爱 国 荣 校

Backup

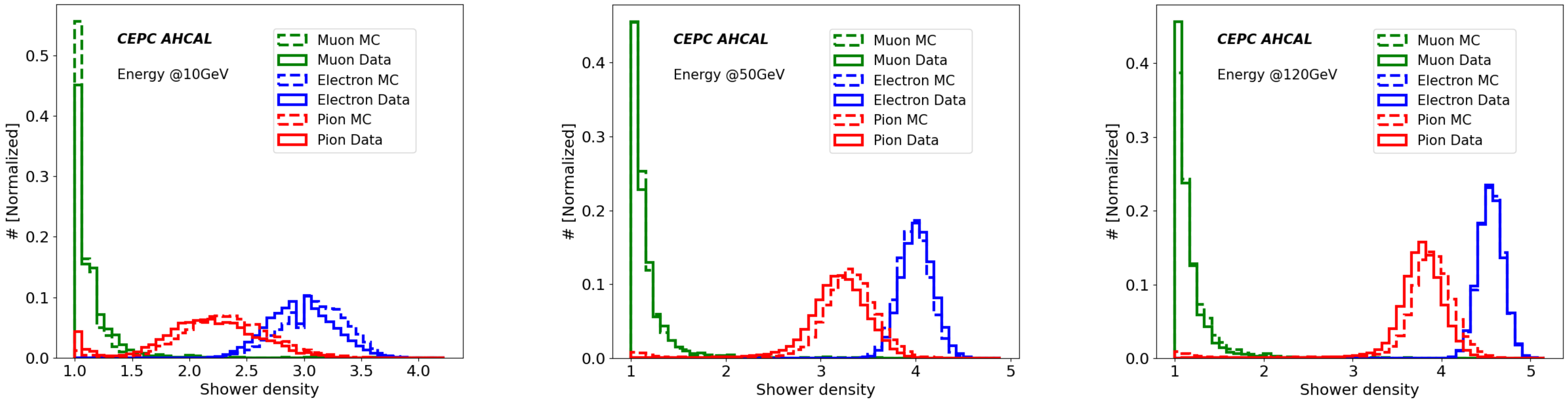
饮 水 思 源　爱 国 荣 校

MC & Data comparison

Shower density

- The average number of the neighbouring hits located in the 3×3 cell around one of the hits including the hit itself.

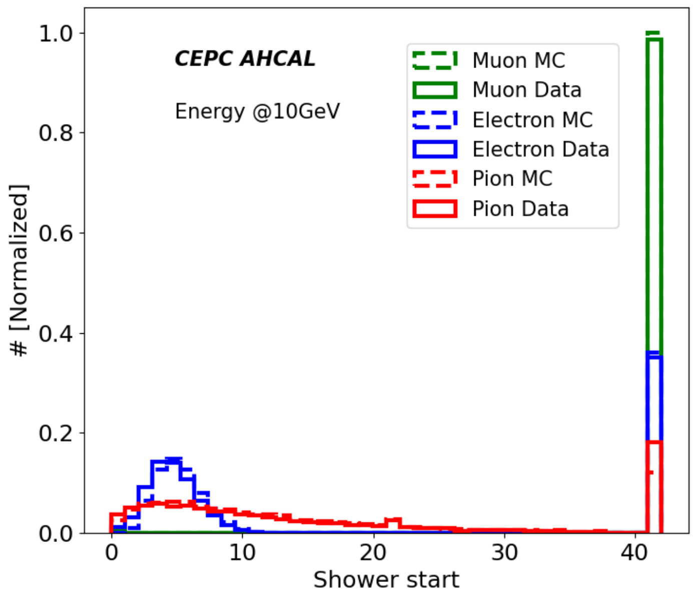
- Data come from 2023's Data, MC samples come from: ***run20230515_AHCAL_Shaping150ns_Window4us***
- The same 0.5 Mip hit energy threshold is pre-applied on both.
- They generally fit.

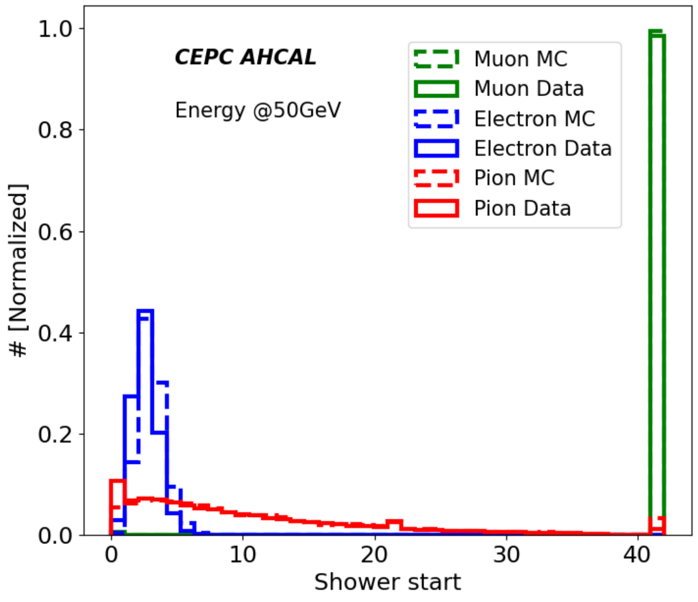


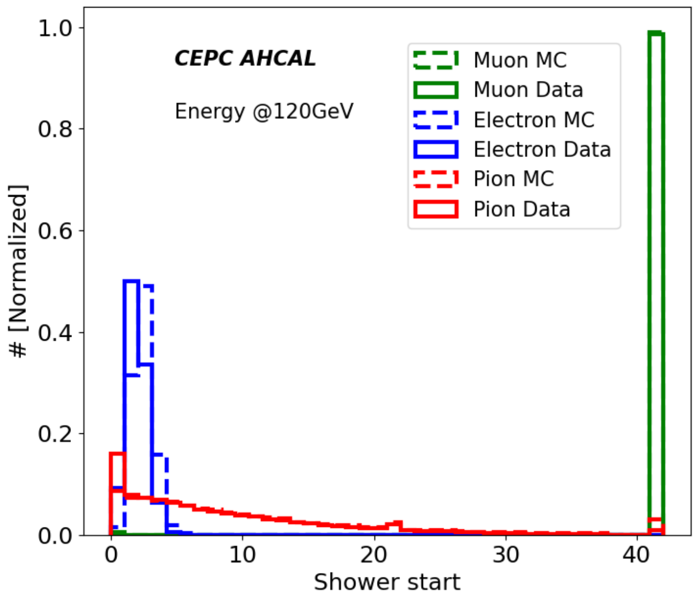
MC & Data comparison

Shower layer start

- The first layer of 3 consecutive layers with at least 5 hits. If no shower, it would be set as 42.
- Data come from 2023's Data, MC samples come from: ***run20230515_AHCAL_Shaping150ns_Window4us***
- The same 0.5 Mip hit energy threshold is pre-applied on both.
- They generally fit.



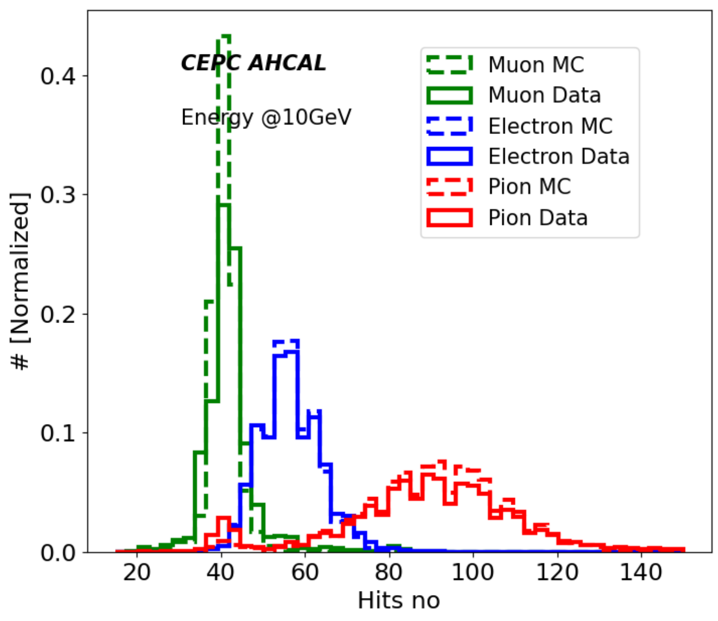


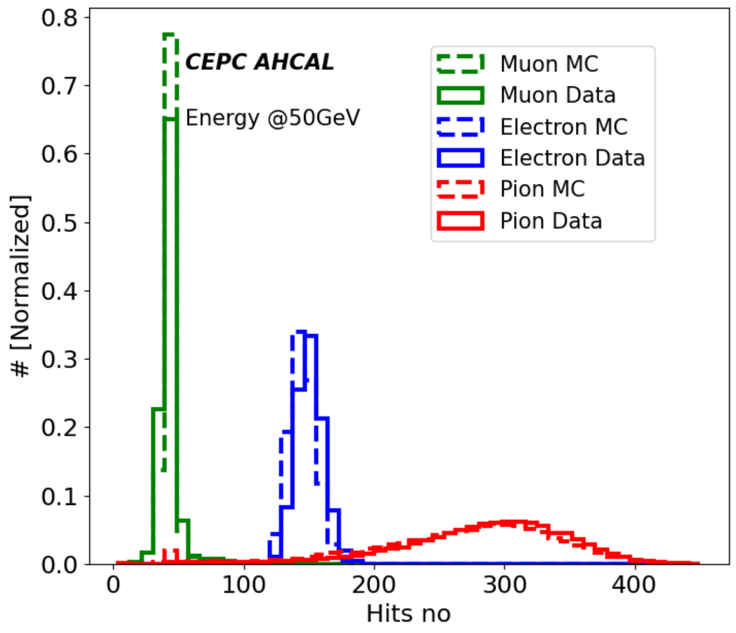


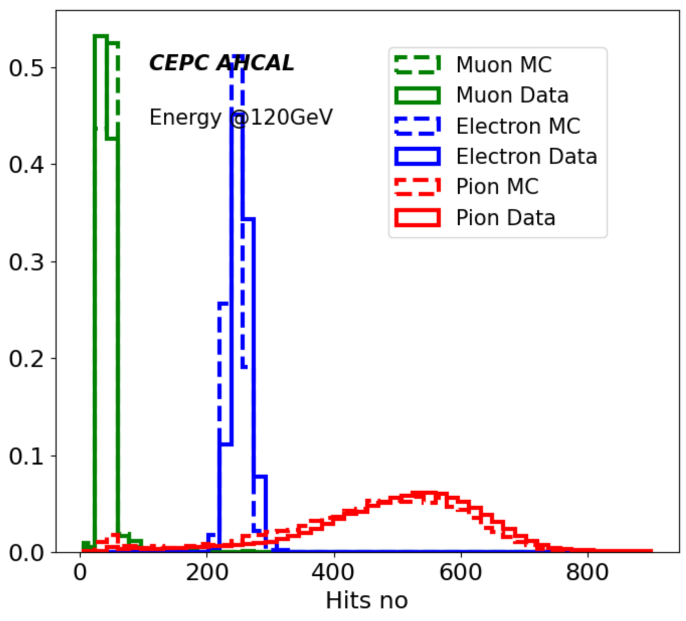
MC & Data comparison

Hits number

- The same 0.5 Mip hit energy threshold is pre-applied on both.
- They generally fit.





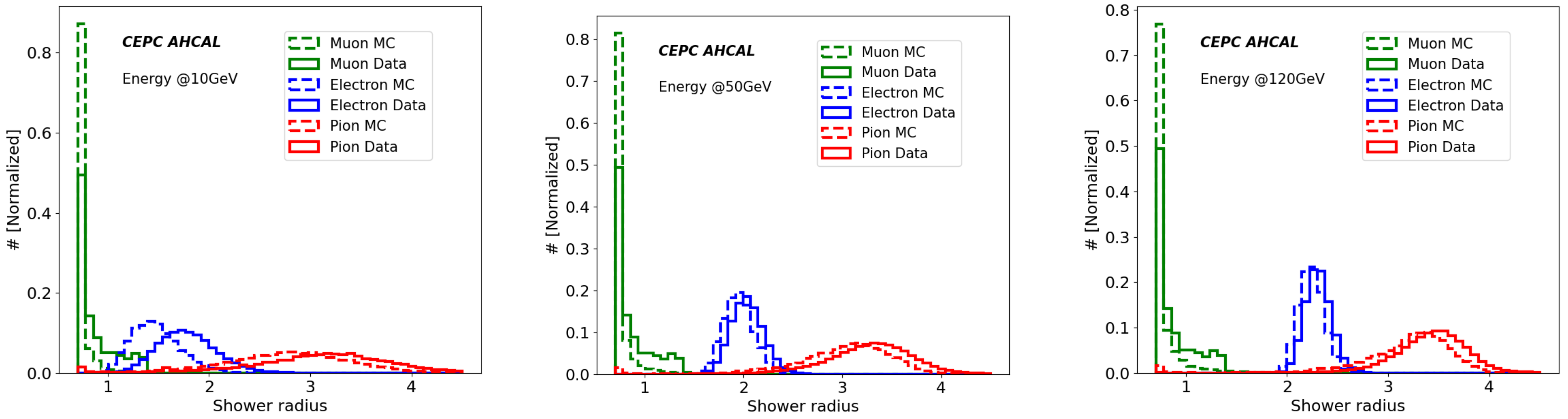


MC & Data comparison

Shower radius

- The average rms of the hits distance to the longitudinal axis of AHCAL (cross the center).

- Data come from 2023's Data, MC samples come from: ***run20230515_AHCAL_Shaping150ns_Window4us***
- The same 0.5 Mip hit energy threshold is pre-applied on both.
- Deviation of peak value is observed.

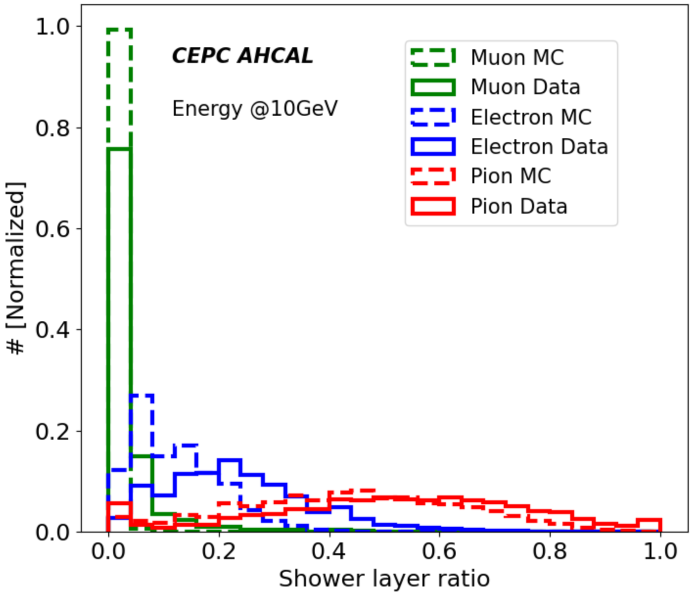


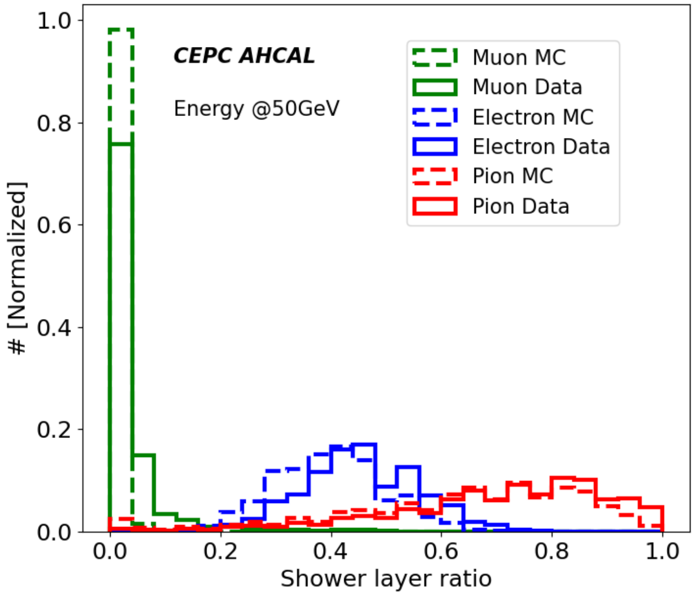
MC & Data comparison

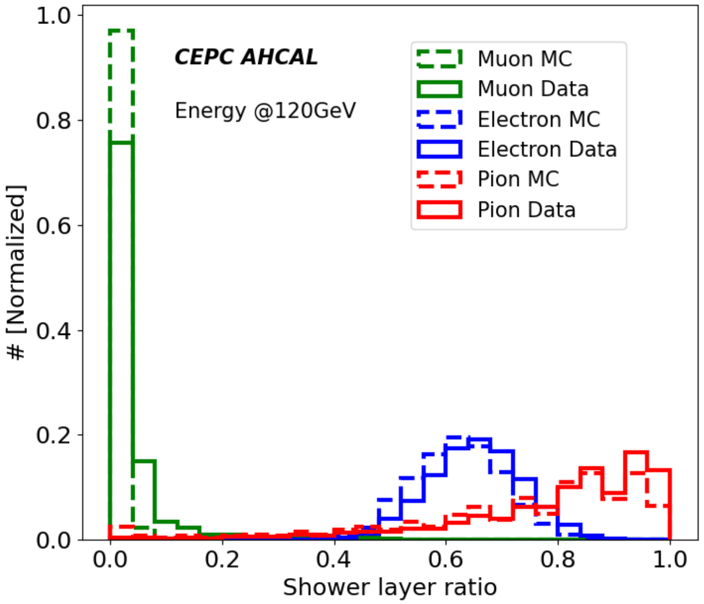
Ratio of shower layers over total hit layers

- The ratio between the number of layers in which the Root Mean Square (RMS) of the hits' position in the x-y plane exceeds 4 cm in both x and y directions and the total number of layers with at least one fired cell.

- Data come from 2023's Data, MC samples come from: ***run20230515_AHCAL_Shaping150ns_Window4us***
- The same 0.5 Mip hit energy threshold is pre-applied on both.
- They generally fit.



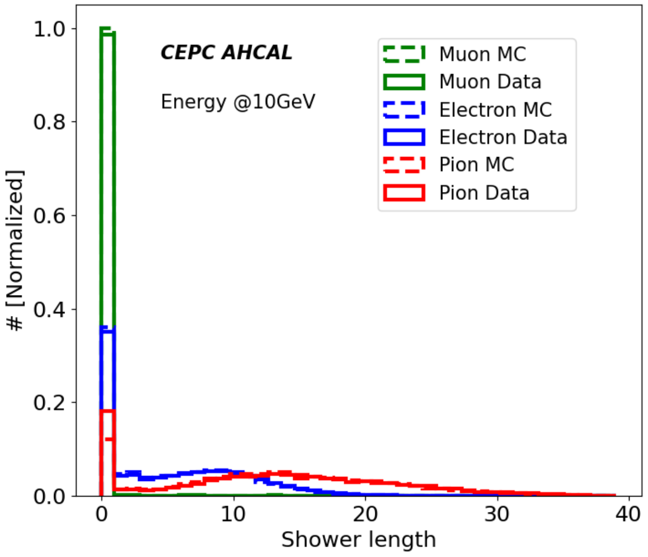


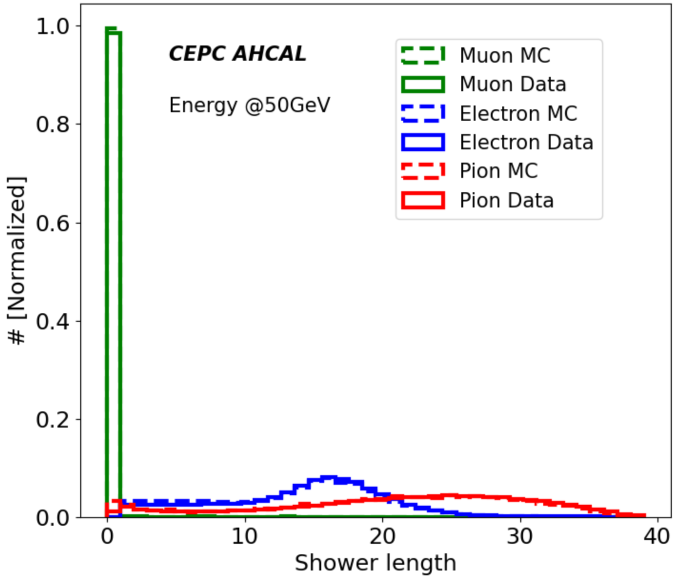


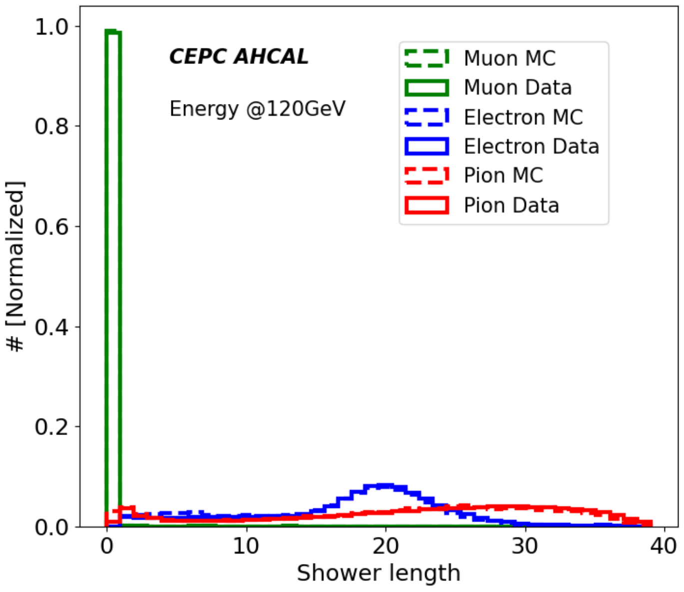
MC & Data comparison

Shower length

- This is the distance between the start of the shower and the layer where the maximum RMS of hit transverse coordinates with respect to the z-axis occurs.
- Data come from 2023's Data, MC samples come from: ***run20230515_AHCAL_Shaping150ns_Window4us***
- The same 0.5 Mip hit energy threshold is pre-applied on both.
- They generally fit.



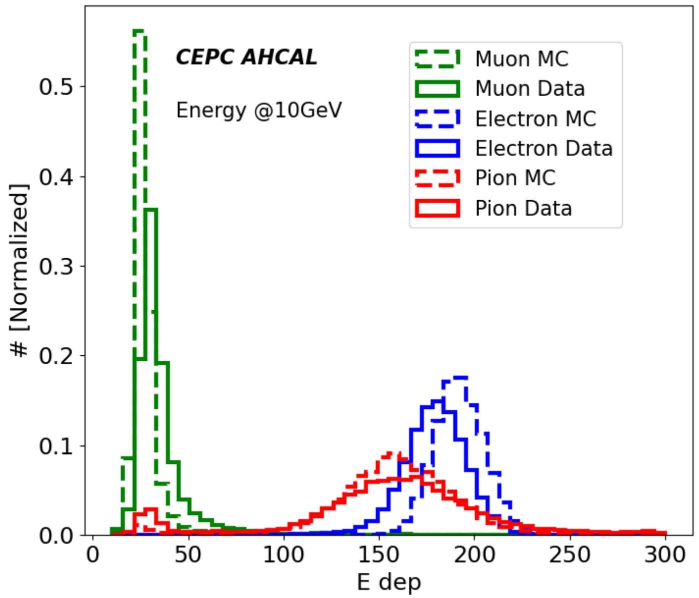


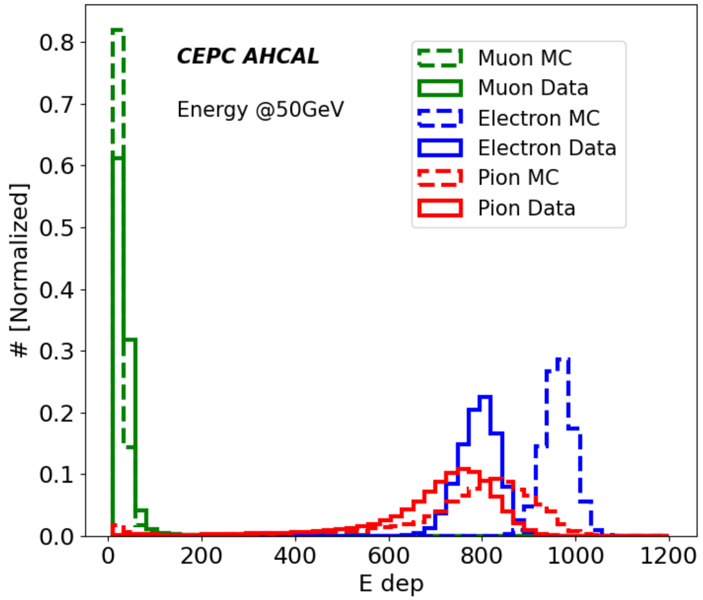


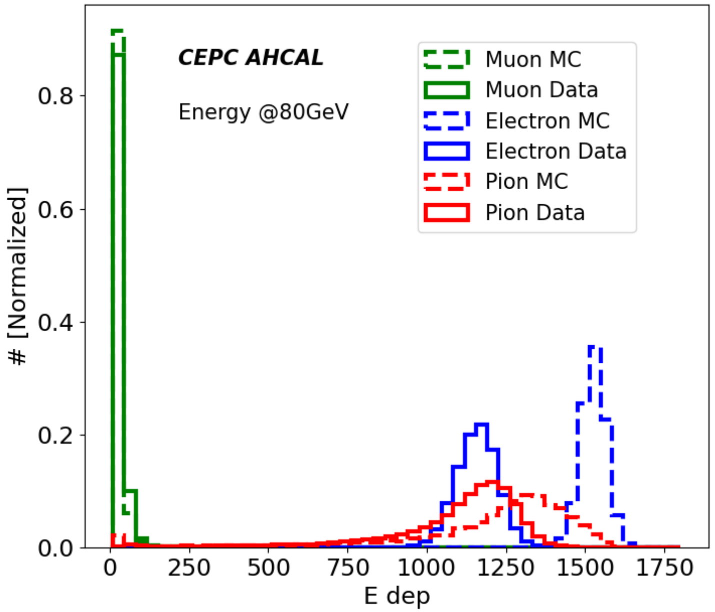
MC & Data comparison

Total energy deposit.

- Data come from 2023's Data, MC samples come from: ***run20230515_AHCAL_Shaping150ns_Window4us***
- The same 0.5 Mip hit energy threshold is pre-applied on both.
- Obvious discrepancy appeared, especially on electron events.

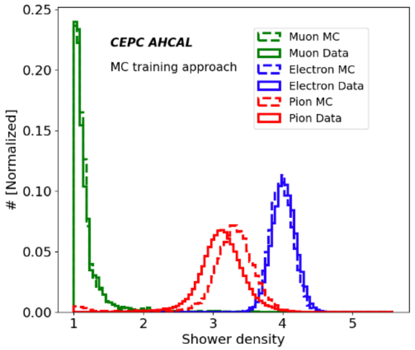




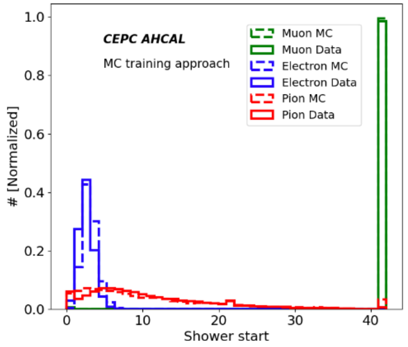


2022 SPS-H8 beam composition

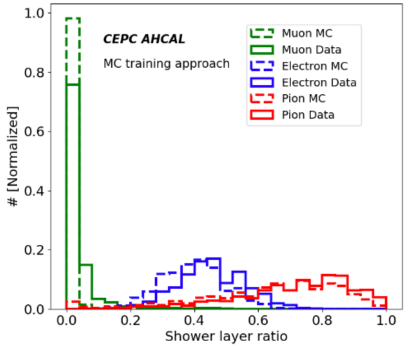
- The Pion samples are classified by using ANN trained on MC data set and are compared with MC.
- Not large disagreement observed.



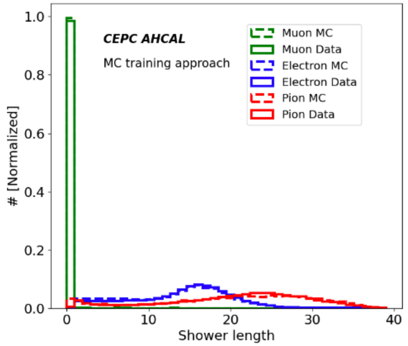
(a) Distribution of the shower density.



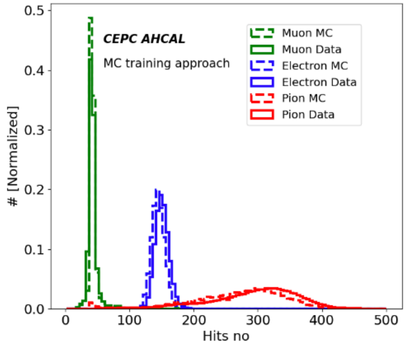
(b) Distribution of the layer that the shower starts.



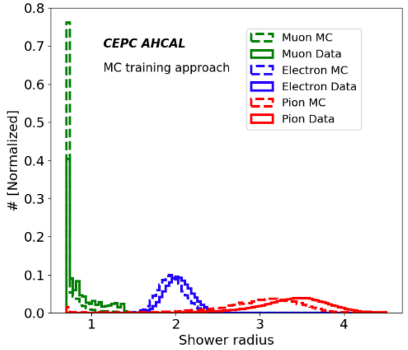
(c) Distribution of the ratio of shower layers over total hit layers.



(d) Distribution of the shower length.



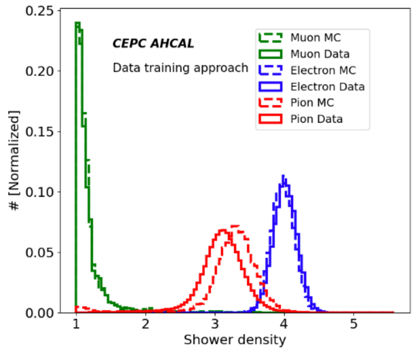
(e) Distribution of the total number of hits.



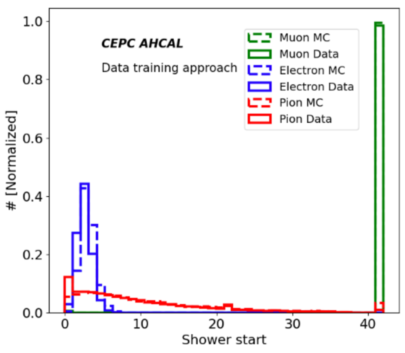
(f) Distribution of the shower radius.

2022 SPS-H8 beam composition

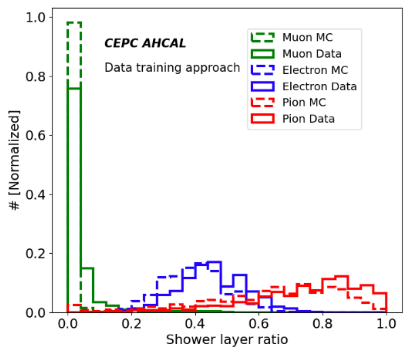
- The Pion samples are classified by using ANN trained on TB data set and are compared with MC.
- Not large disagreement observed.



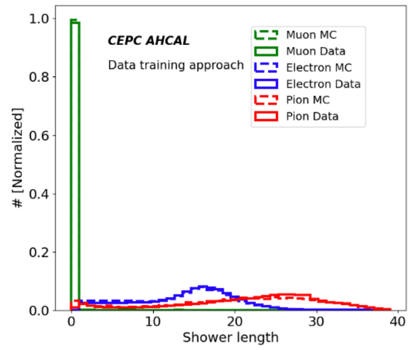
(a) Distribution of the shower density.



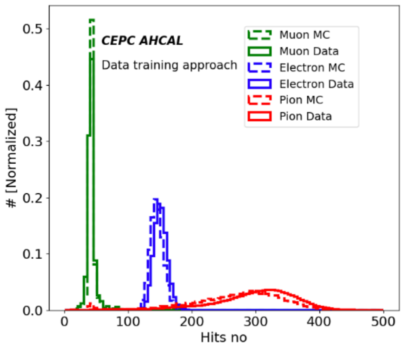
(b) Distribution of the layer that the shower starts.



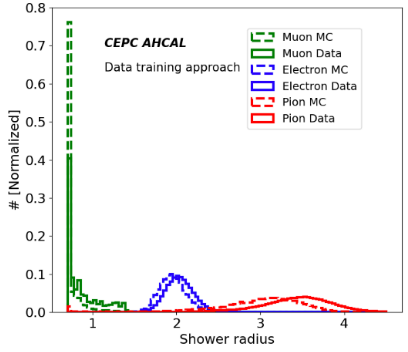
(c) Distribution of the ratio of shower layers over total hit layers.



(d) Distribution of the shower length.



(e) Distribution of the total number of hits.



(f) Distribution of the shower radius.

Two Data sets are prepared due to discrepancy between MC and Data

TB Data set

- Uniformly mixture of particle types.

→ TB train set → Train ANN/BDT (**Data training approach**)

→ TB test set → Evaluate ANN/BDT Performance

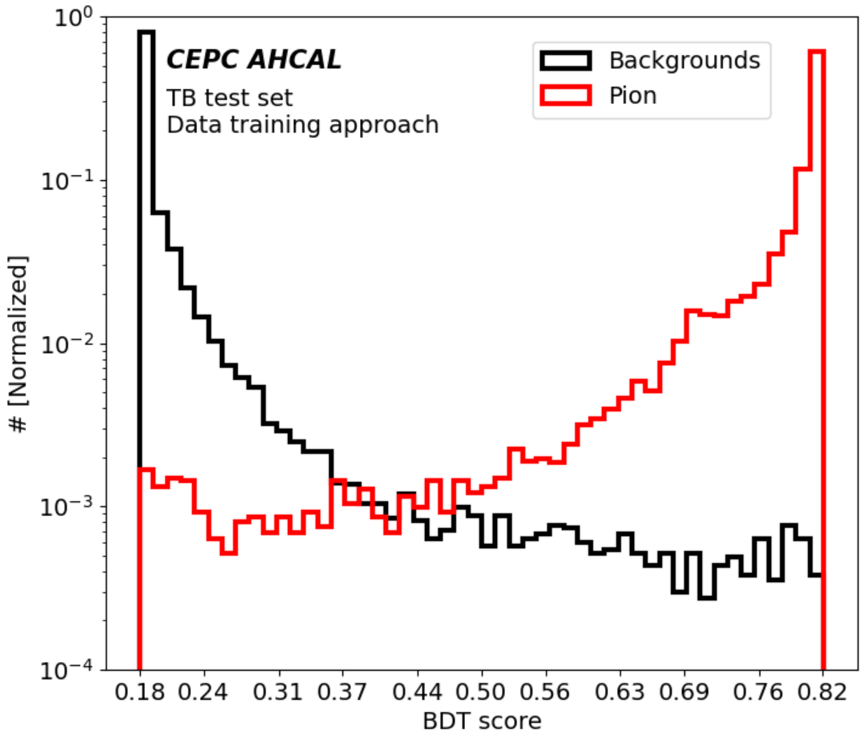
MC Data set

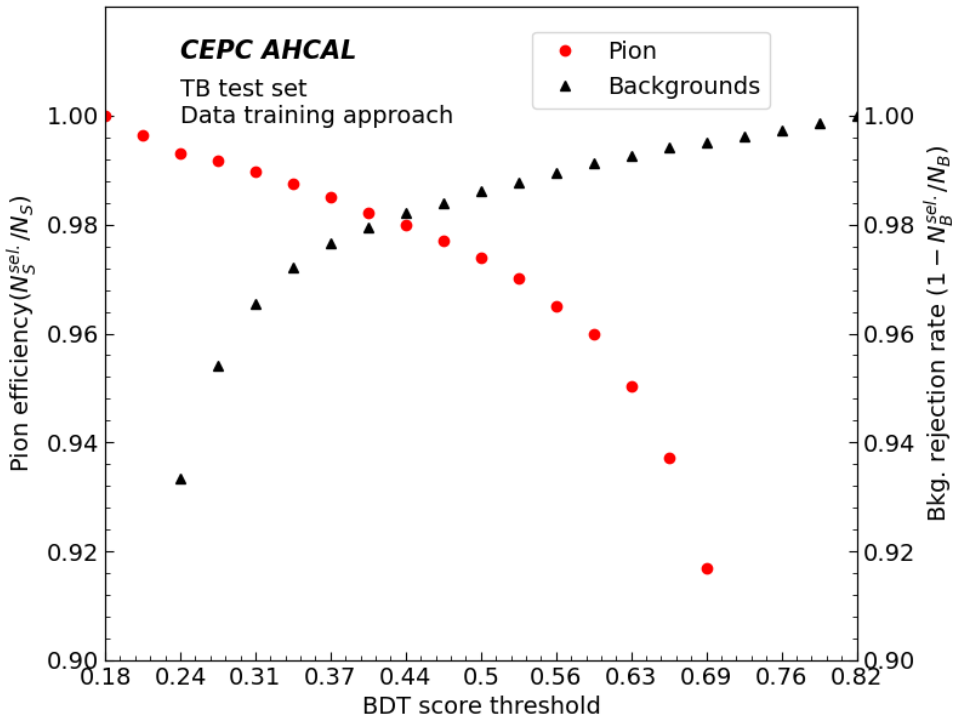
- Uniformly mixture of particle types.

→ MC train set → Train ANN/BDT (**MC training approach**)

→ MC test set → Evaluate ANN/BDT Performance

BDT Output

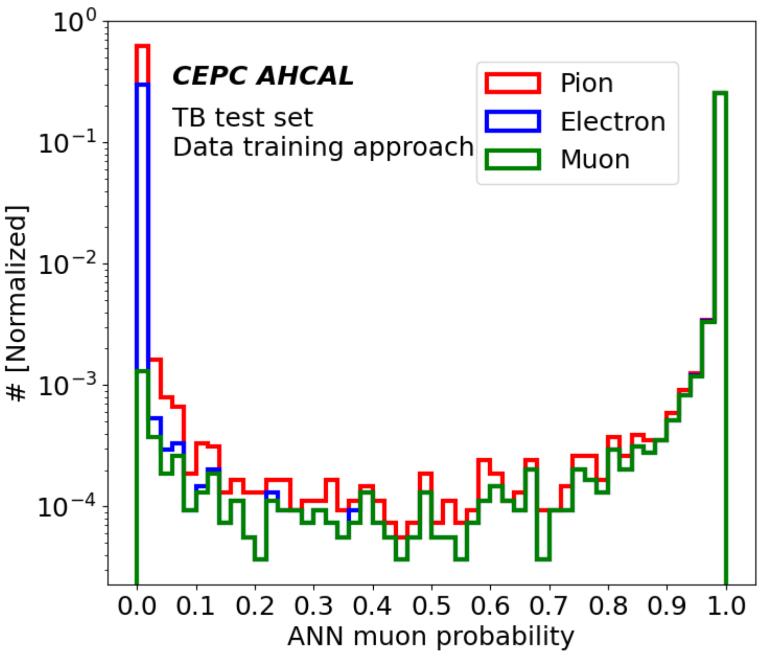


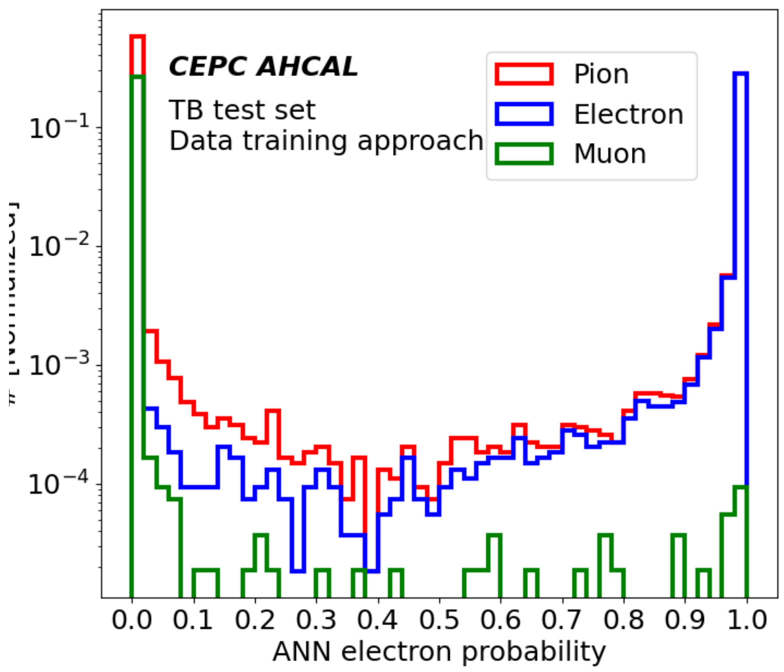


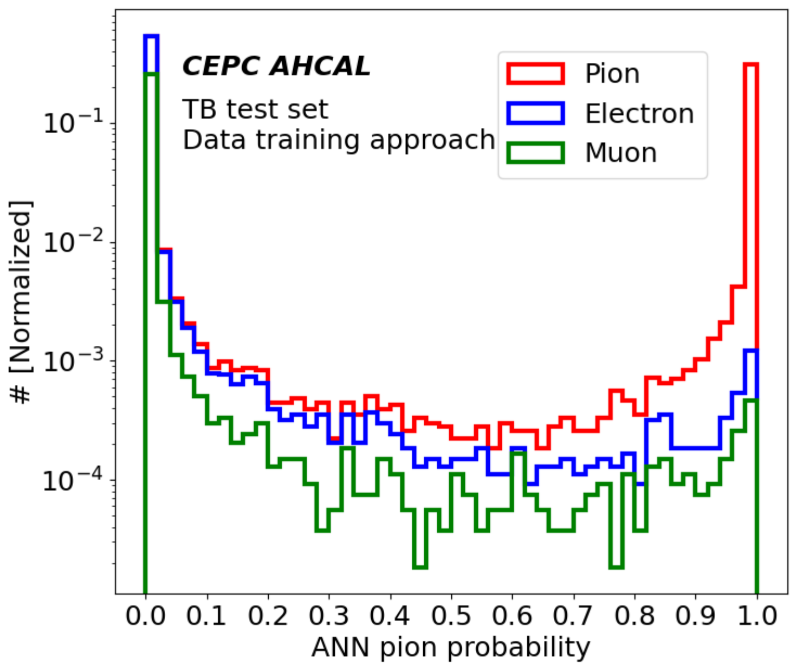
Backgrounds :Signal =2:1

ANN Output

Separation power achieved. e.g. Pions get higher probability classified as pions.



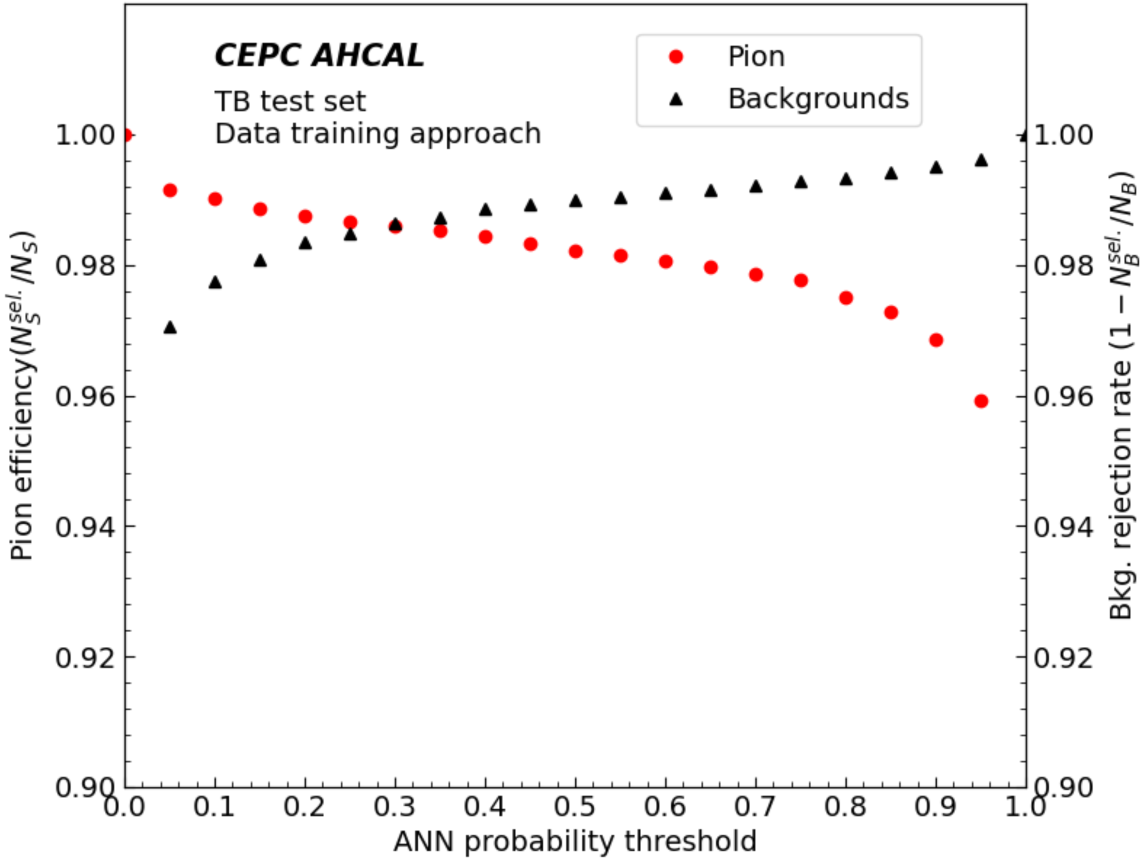




Backgrounds :Signal =2:1

ANN Output

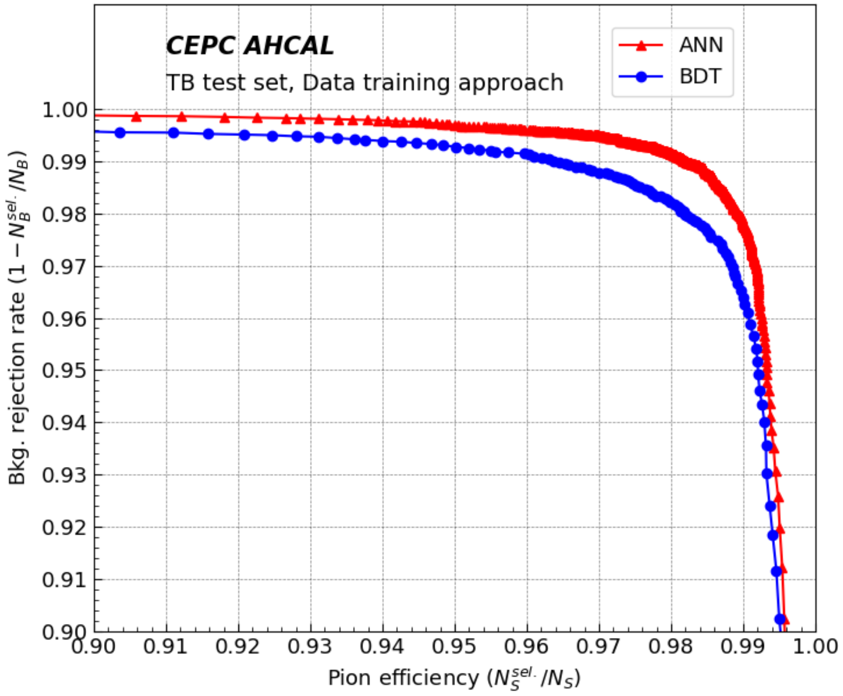
- Separation power is confirmed.
  - Pion efficiency and background rejection rate can be both over 98%

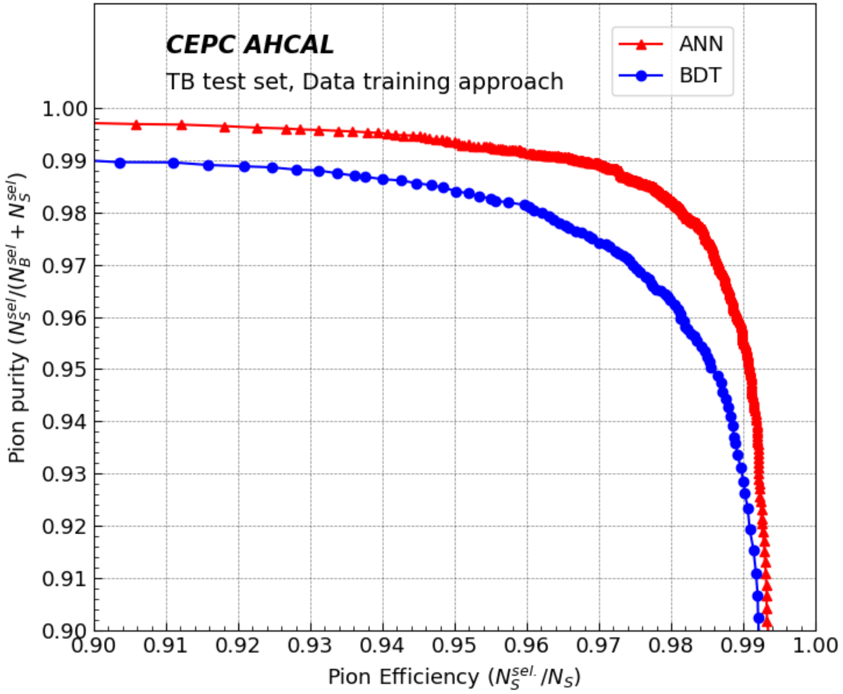


Comparison with ANN PID method

Pion efficiency	0.90	0.92	0.94	0.96	0.98	0.99
BDT_{TB} bkg. rej. rate	0.996	0.995	0.994	0.991	0.982	0.964
ANN_{TB} bkg. rej. rate	**0.999**	**0.998**	**0.998**	**0.996**	**0.991**	**0.978**
	↑ 0.3%	**↑ 0.3%**	**↑ 0.4%**	**↑ 0.5%**	**↑ 0.9%**	**↑ 1.5%**
BDT_{TB} pion purity	0.99	0.989	0.986	0.981	0.963	0.929
ANN_{TB} pion purity	**0.997**	**0.996**	**0.995**	**0.991**	**0.982**	**0.956**
	↑ 0.7%	**↑ 0.7 %**	**↑ 0.9%**	**↑ 1.0%**	**↑ 2.0%**	**↑ 2.9%**

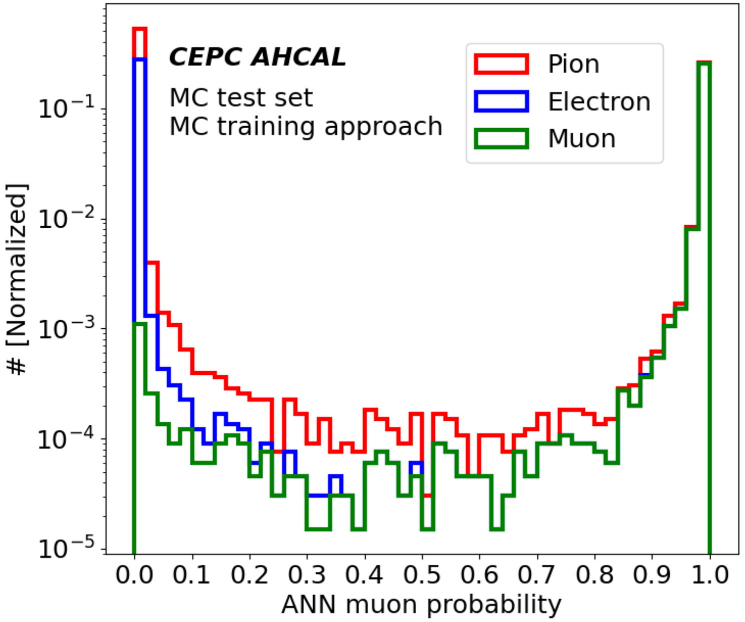
TABLE V. The background rejection rate and the pion purity of the BDT_{TB} and the ANN_{TB}. The results highlight the improvement of the ANN_{TB} compared to the BDT_{TB}.

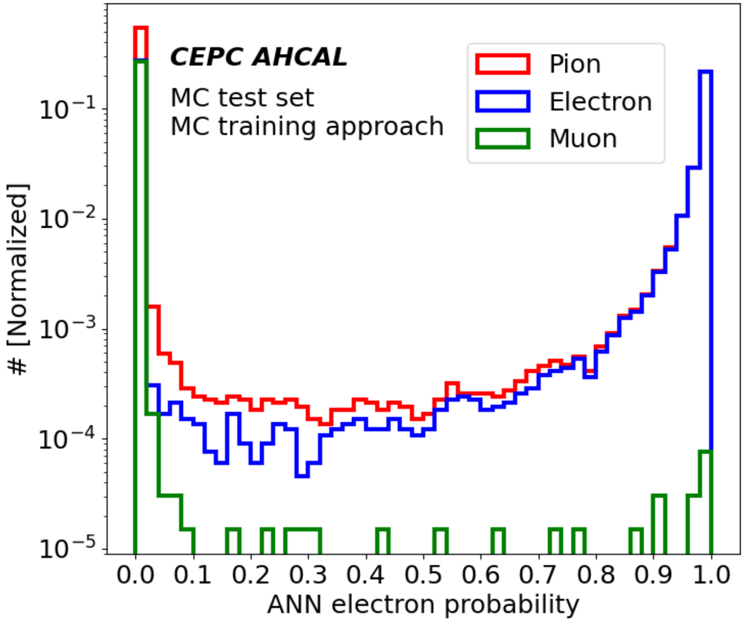


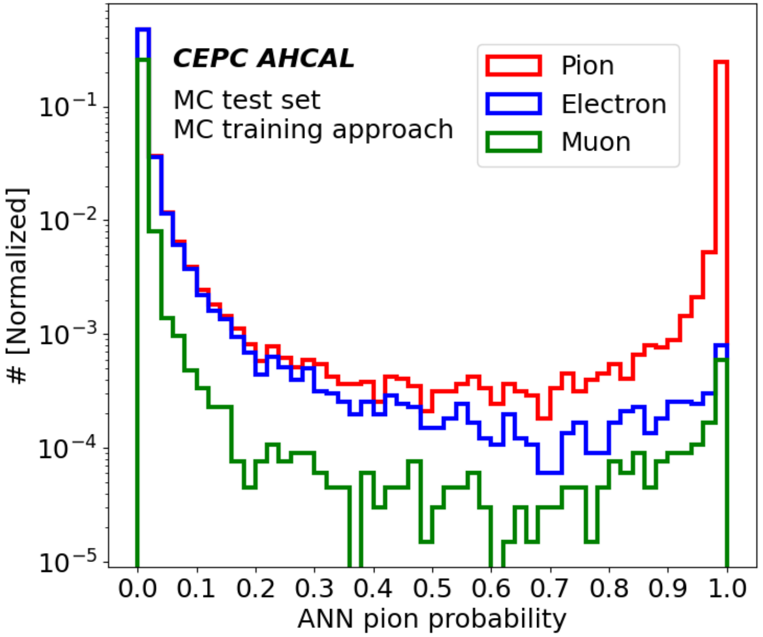


ANN Output

Separation power achieved. e.g. Pions get higher probability classified as pions.







Backgrounds :Signal =2:1

PID based on BDT

- **Apply Extreme Gradient Boosting**
- **Reconstruct 12 input variables**
  - **Shower density**
  - **Shower start**
  - **Shower length**
  - **Hits number**
  - **Shower radius**
  - **Fractal dimension**
    - FD_1, FD_6
  - **Fired layers**
  - **Shower layers**
  - **Shower layer ratio**
  - **Z width**

MC samples

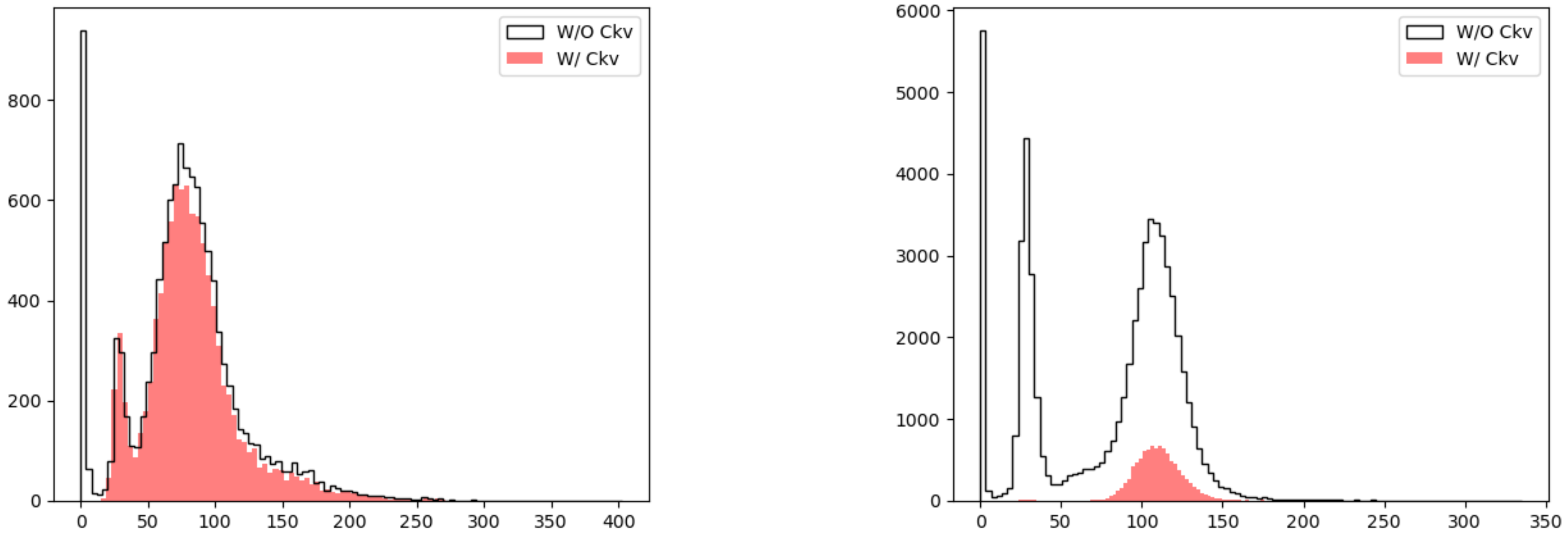
Rank: Variable	Variable weight
1: Shower radius	0.377
2: Shower layers	0.232
3: Hits number	0.088
4: Fired layers	0.083
5: Shower start	0.080
6: Shower density	0.049
7: Z width	0.034
8: FD_6	0.017
9: FD_1	0.015
10: Shower layer ratio	0.014
11: Shower end	0.006
12: Shower length	0.006

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

- The sensitivity of the Cherenkov detector in the low energy range (<30 GeV) could allow the collection of electron and pion samples.

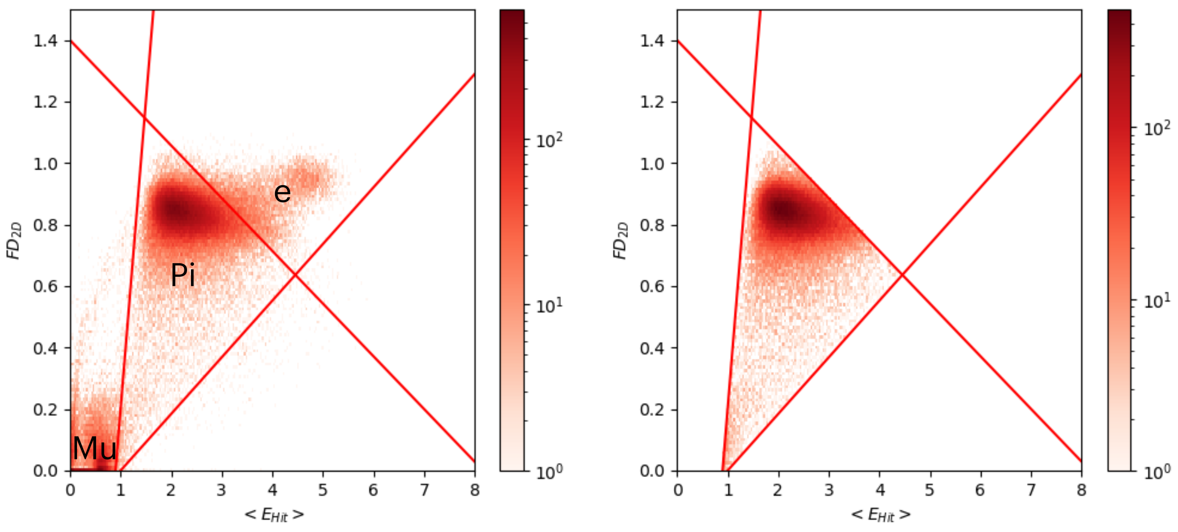


- Cherenkov cut. 5GeV pion and 5 GeV electron TB data.

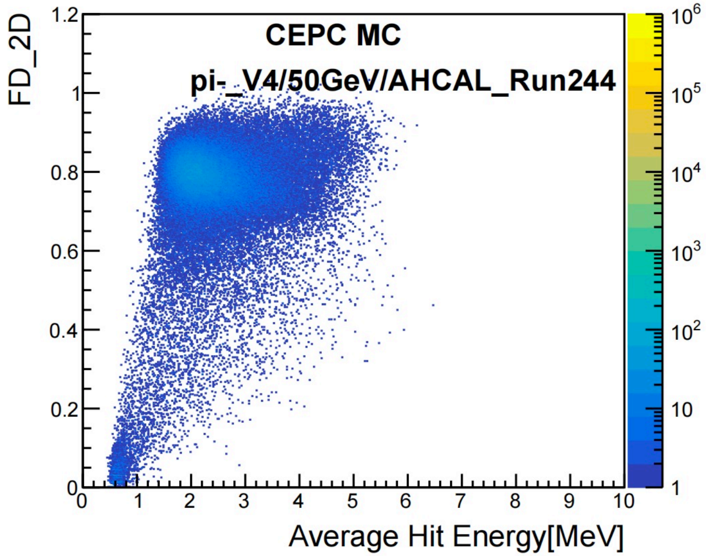
For data collected in 2023, SPS and PS

Data pre-selection

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



FD cut

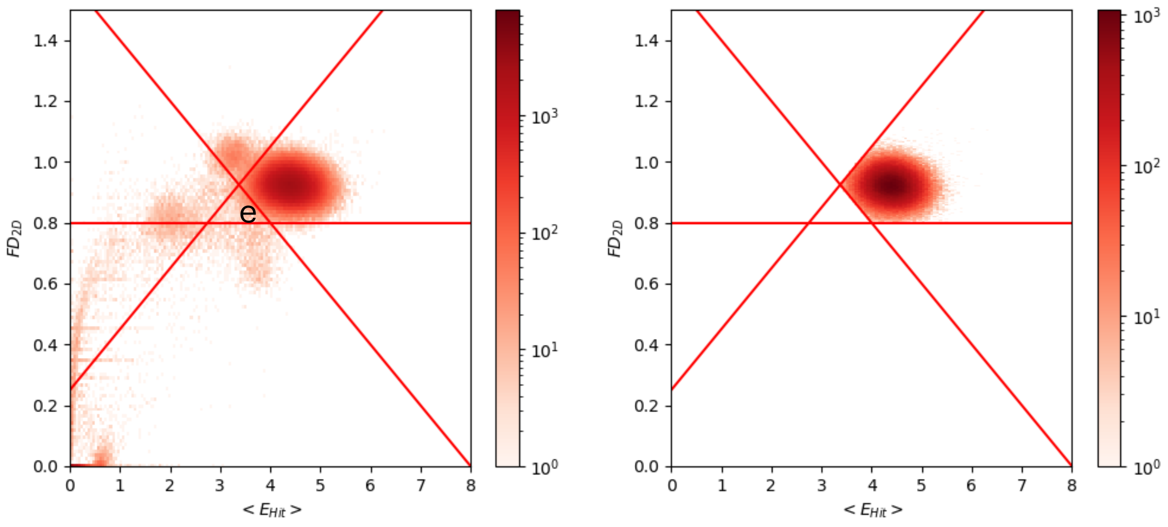


MC

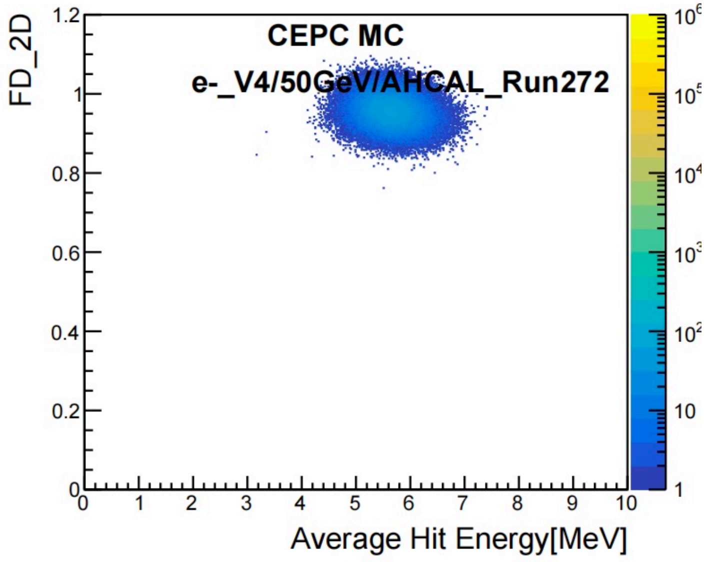
For data collected in 2023, SPS and PS

Data pre-selection

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



FD cut

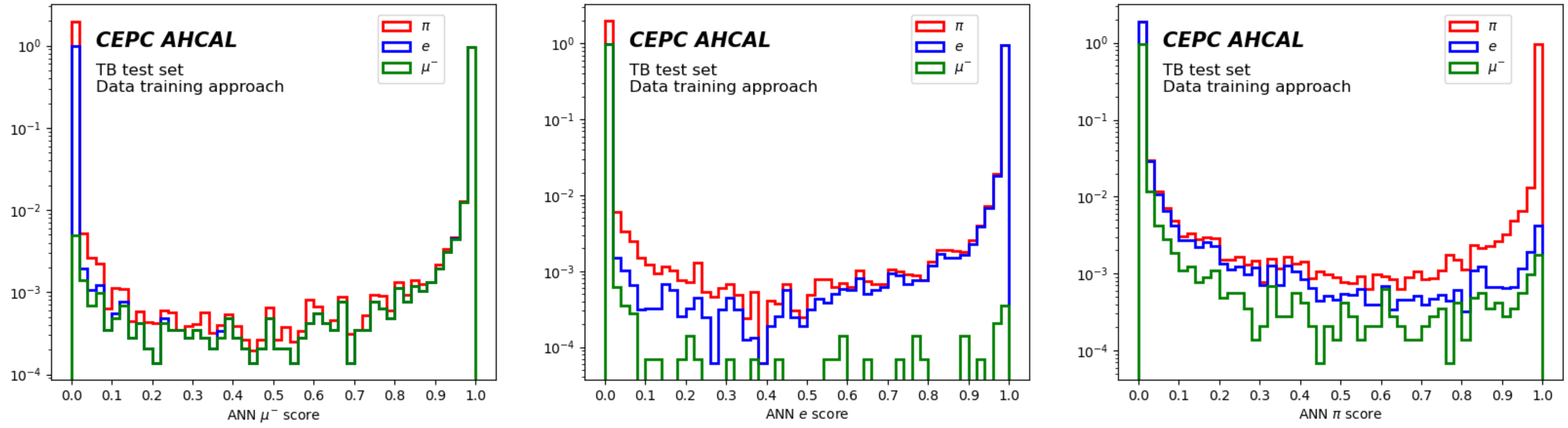


MC

For data collected in 2023, SPS and PS

Evaluated on TB test set.

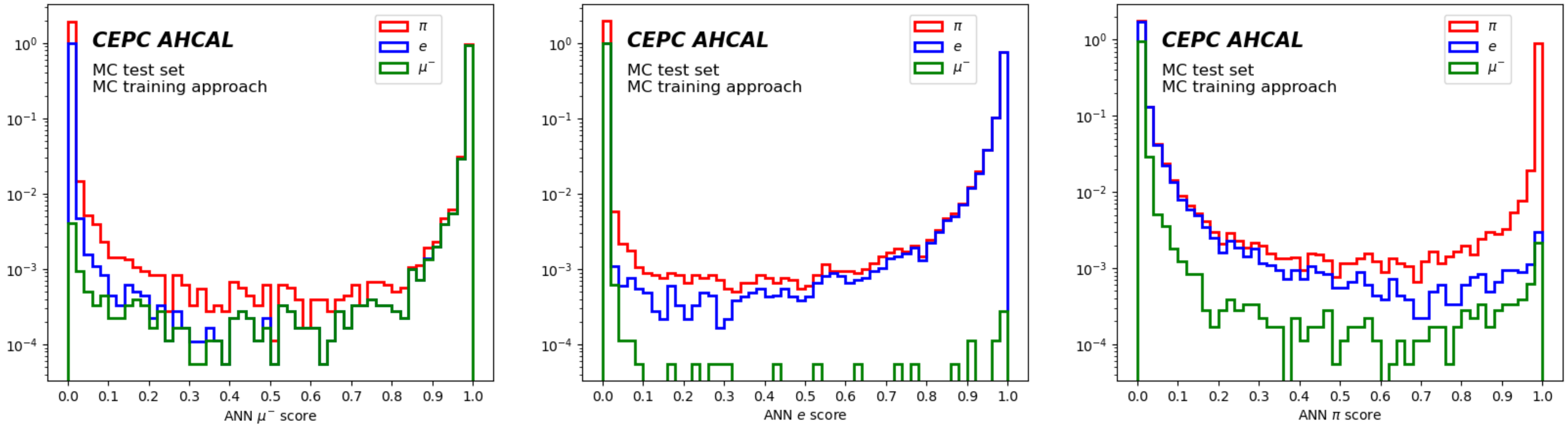
- Signals would get higher scores (closer to 1).
    - E.g. a pion would get higher ANN Pion score.
- An additional threshold ANN threshold cut would help to reject backgrounds.



Output of ANN (data training approach).

Evaluated on MC test set.

- Signals would get higher scores (closer to 1).
  - E.g. a pion would get higher ANN Pion score.
- An additional threshold ANN threshold cut would help to reject backgrounds.



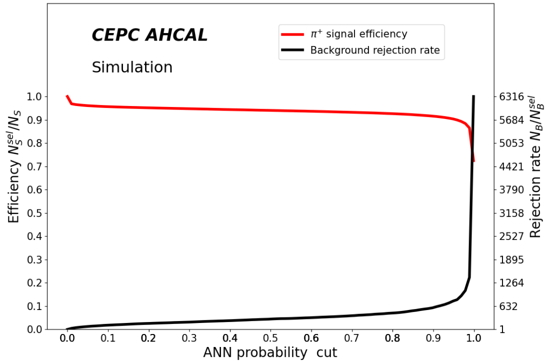
Output of ANN (data training approach).

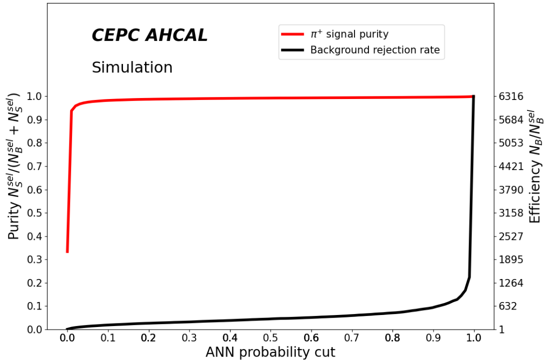
Evaluated on MC test set

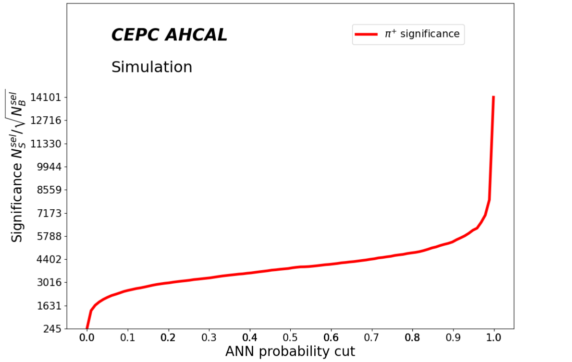
Pion signal VS Backgrounds

Tested on Mixed MC Test Set.

- **120k Pi + 120k mu + 120k positron**







ANN Score Cut	Pi Purity $(N_S^{sel}/(N_B^{sel}+N_S^{sel}))$	Pi Efficiency (N_S^{sel}/N_S)	Background Rejection Rate (N_B/N_B^{sel})
0.1	0.98	0.96	117
0.3	0.99	0.95	200
0.7	0.994	0.93	376
0.9	0.996	0.91	618

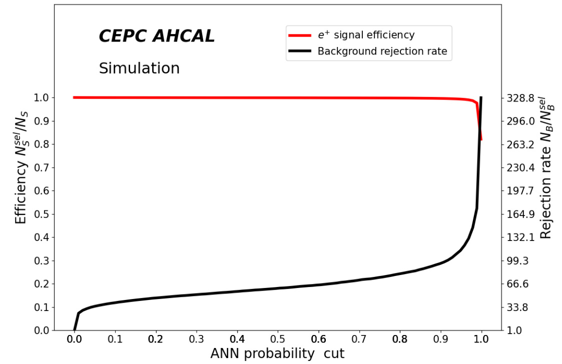
Evaluated on MC test set

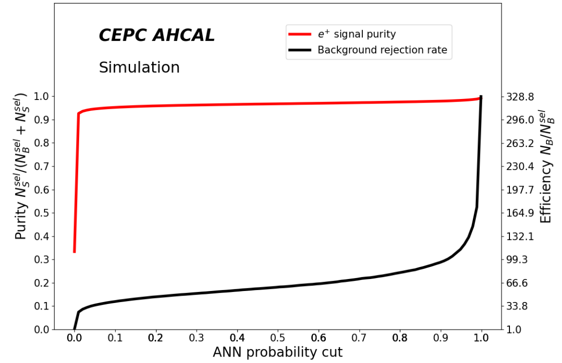
Electron signal VS Backgrounds

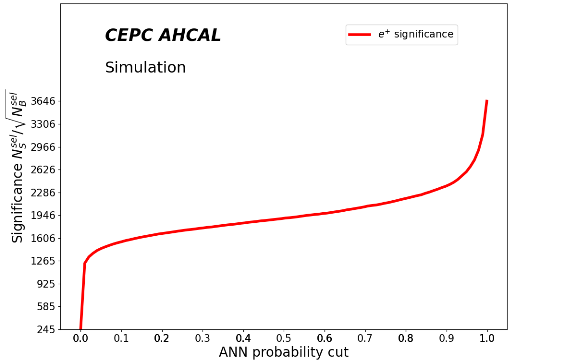
Tested on Mixed MC Test Set.

- **120k Pi + 120k mu + 120k positron**

ANN Score Cut	E Purity $(N_S^{sel}/(N_B^{sel}+N_S^{sel}))$	E Efficiency (N_S^{sel}/N_S)	Background Rejection Rate (N_B/N_B^{sel})
0.1	0.95	0.999	40
0.3	0.96	0.999	52
0.7	0.97	0.998	72
0.9	0.98	0.996	97







Evaluated on MC test set

Muon signal VS Backgrounds

Tested on Mixed MC Test Set.

- **120k Pi + 120k mu + 120k positron**

ANN Score Cut	Mu Purity $(N_S^{sel}/(N_B^{sel}+N_S^{sel}))$	Mu Efficiency (N_S^{sel}/N_S)	Background Rejection Rate (N_B/N_B^{sel})
0.1	0.97	0.99	59
0.3	0.97	0.99	69
0.7	0.98	0.99	80
0.9	0.98	0.98	86