INFN
Istituto Nazionale di Fisica Nucleare

中國科學院高能物理研究所
Institute of High Energy Physics
Chinese Academy of Sciences

ML Based Algorithm for Cluster Counting on Real Data And PID

Muhammad Numan Anwar

Bari (INFN) - IHEP(Beijing)
Meeting

2nd March, 2026

- **Analysis of Real Test Beam Data**
- Configuration Setup for channel 5
- Optimization of LSTM Model by Search Grid
- Peak Finding Algorithm
- CNN Clusterization result(Prelimary)

- **ML based Algorithm on for Particle Identification(Ongoing)**
- Simulation Parameters
- Number of Primary Cluster (MC)

  I) Muon

  ii) Kaon

  iii) Pion
- NN (CNN) Reconstruction Results
- Particle Identification

Main Goal of the Presentation

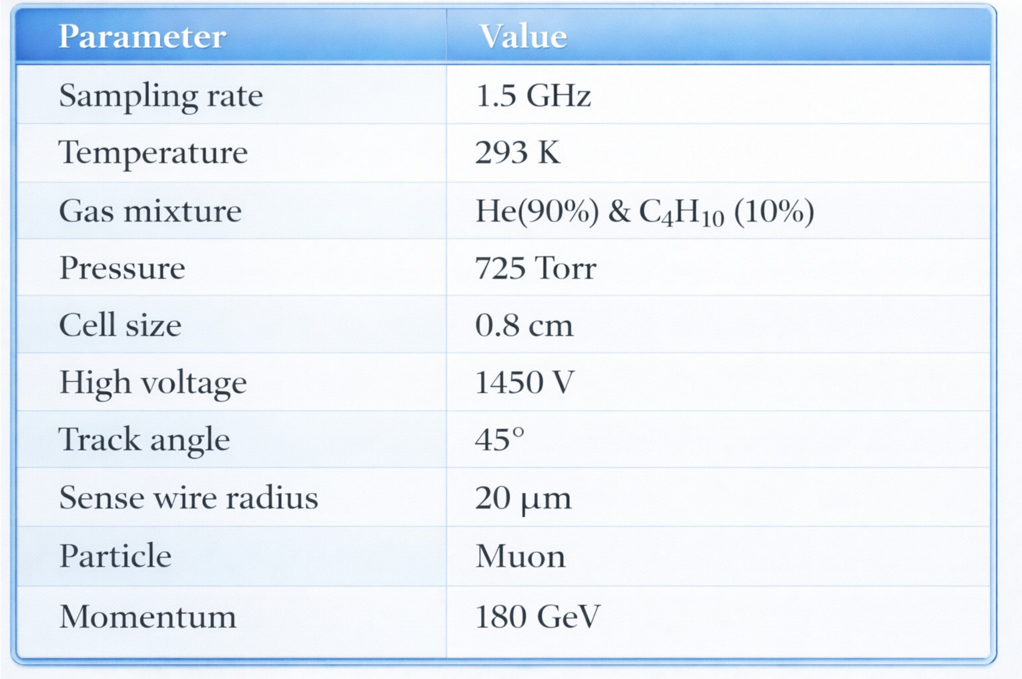
Task 1: The first task is related to apply the best trained LSTM Model (Highest Auc value among all configurations) on 180 GeV muon real data in order to detect peaks in comparison with RTA Algorithm on the behalf of different selection cuts(0.25-0.3 0.4) and then apply trained CNN model to estimate the number of primary clusters based on the detected peaks

Task 2: The second task is related to particle identification of muon and kaon. That is why first we generated some simulated samples from 2-10 GeV momenta for Kaon and pion. Then we apply best LSTM model to detect peaks in each waveform and then in the second step best CNN regression Model was applied to reconstruct the number of primary cluster on the behalf of detected peaks.

Note: I already assumed that all of you know about the two steps of cluster counting techniques. In the first step the best LSTM model find the peaks (Primary + Secondary electrons) in each waveform against the noise. While in the second step best CNN model estimate the number of primary clusters based on the detected peaks. That is why I showed the final results of Number of primary Clusters MC & CNN regression Model.

LSTM Peak Finding Algorithm on Real Data

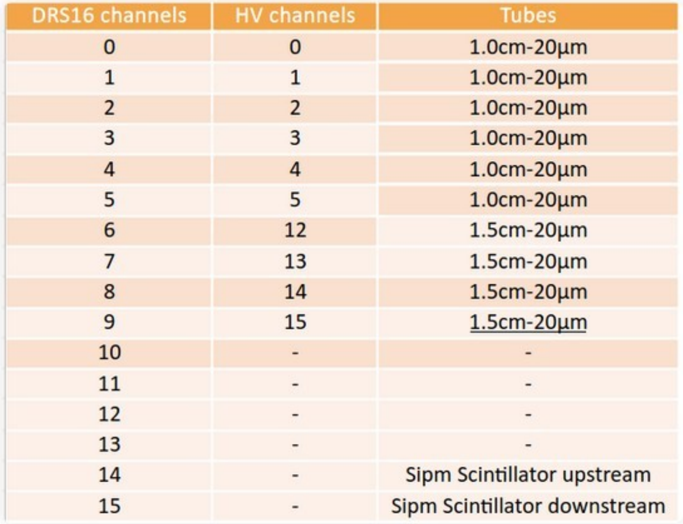
Configuration Setup for the Real Test Beam data (180 GeV & Channel 5)



Parameter	Value
Sampling rate	1.5 GHz
Temperature	293 K
Gas mixture	He(90%) & C_4H_{10} (10%)
Pressure	725 Torr
Cell size	0.8 cm
High voltage	1450 V
Track angle	45°
Sense wire radius	20 µm
Particle	Muon
Momentum	180 GeV

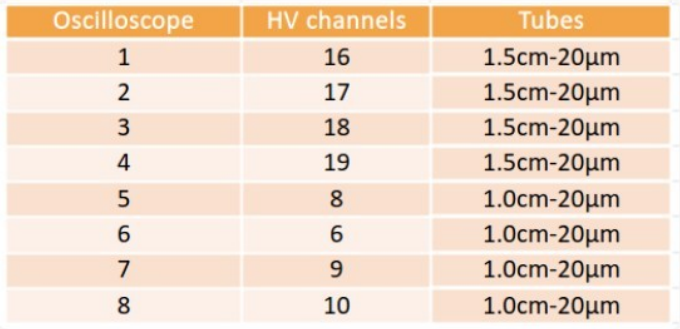
Channel 5 Setup Configuration

Confiurations and setup information



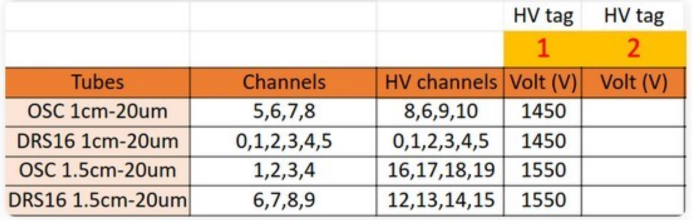
DRS16 channels	HV channels	Tubes
0	0	1.0cm-20µm
1	1	1.0cm-20µm
2	2	1.0cm-20µm
3	3	1.0cm-20µm
4	4	1.0cm-20µm
5	5	1.0cm-20µm
6	12	1.5cm-20µm
7	13	1.5cm-20µm
8	14	1.5cm-20µm
9	15	1.5cm-20µm
10	-	-
11	-	-
12	-	-
13	-	-
14	-	Sipm Scintillator upstream
15	-	Sipm Scintillator downstream

Tubes setup with DRS



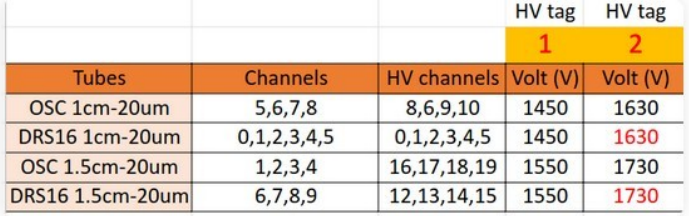
Oscilloscope	HV channels	Tubes
1	16	1.5cm-20µm
2	17	1.5cm-20µm
3	18	1.5cm-20µm
4	19	1.5cm-20µm
5	8	1.0cm-20µm
6	6	1.0cm-20µm
7	9	1.0cm-20µm
8	10	1.0cm-20µm

Tubes setup with OSC



			HV tag	HV tag
			1	2
Tubes	Channels	HV channels	Volt (V)	Volt (V)
OSC 1cm-20um	5,6,7,8	8,6,9,10	1450	
DRS16 1cm-20um	0,1,2,3,4,5	0,1,2,3,4,5	1450	
OSC 1.5cm-20um	1,2,3,4	16,17,18,19	1550	
DRS16 1.5cm-20um	6,7,8,9	12,13,14,15	1550	

HV tags with 90/10 gas mix runs

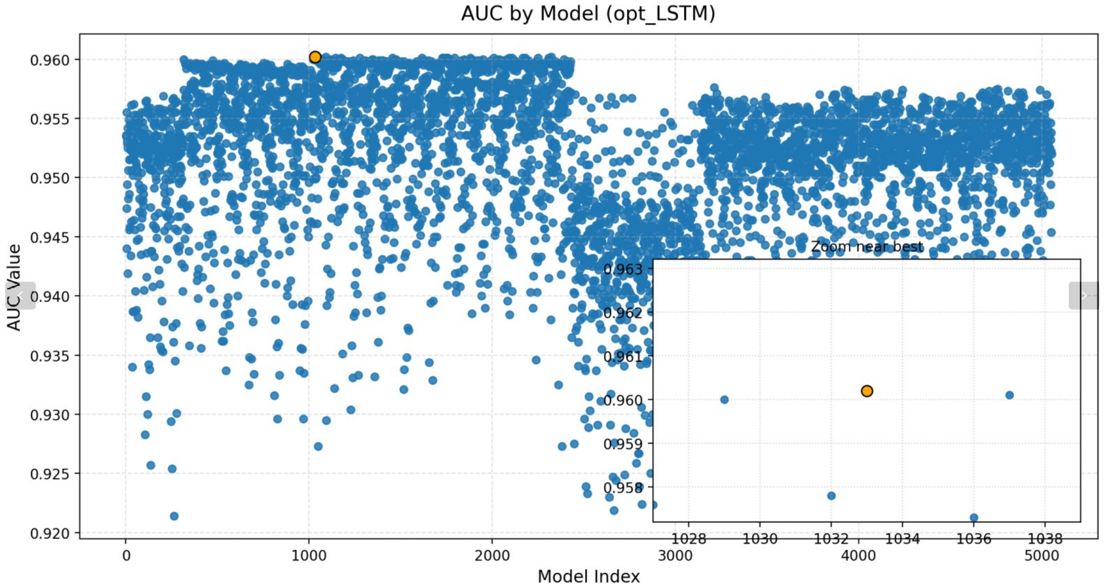


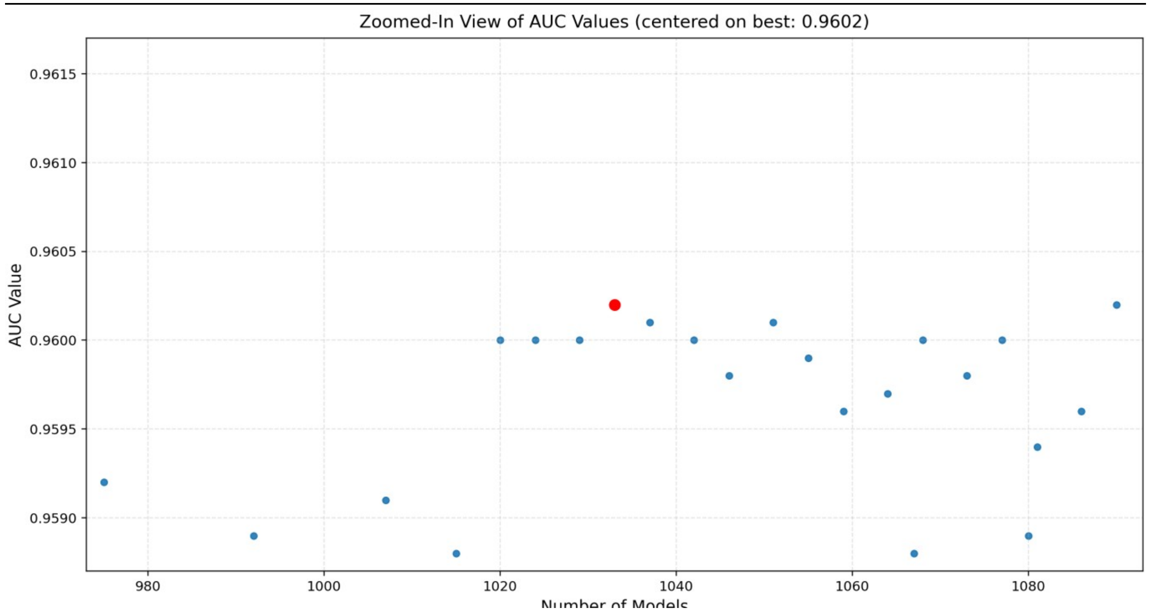
			HV tag	HV tag
			1	2
Tubes	Channels	HV channels	Volt (V)	Volt (V)
OSC 1cm-20um	5,6,7,8	8,6,9,10	1450	1630
DRS16 1cm-20um	0,1,2,3,4,5	0,1,2,3,4,5	1450	1630
OSC 1.5cm-20um	1,2,3,4	16,17,18,19	1550	1730
DRS16 1.5cm-20um	6,7,8,9	12,13,14,15	1550	1730

HV tags with 85/15 gas mix runs

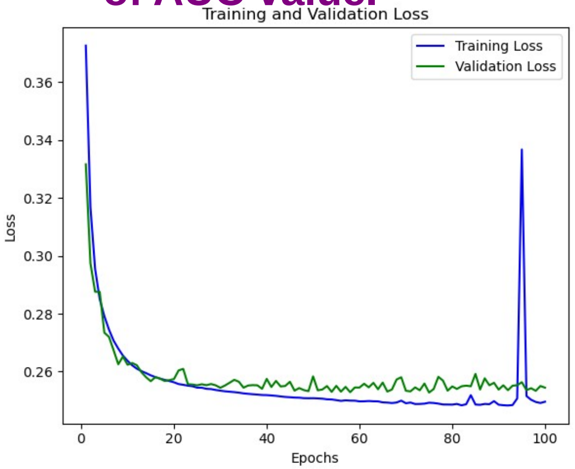
Full Configuration Setup for different channels and different gas mixtures etc

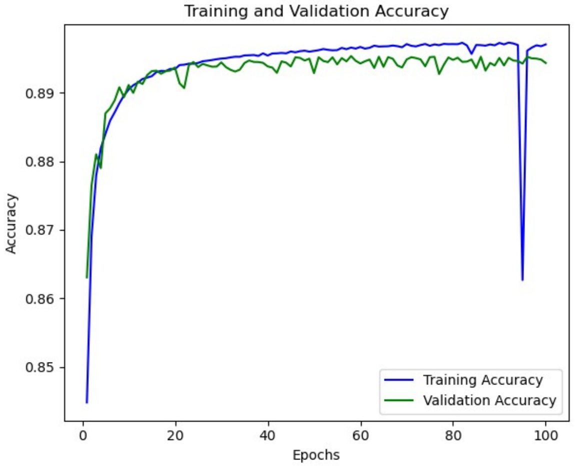
Search Grid Optimzation: BEST LSTM MODEL ON TUNE SIMULATED WAVEFORMS (50000)

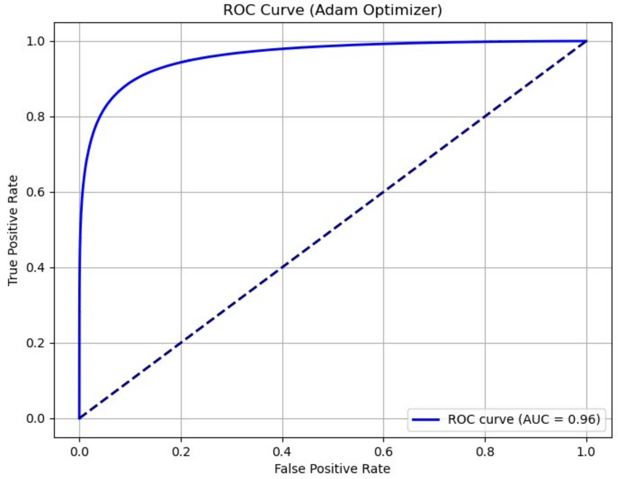


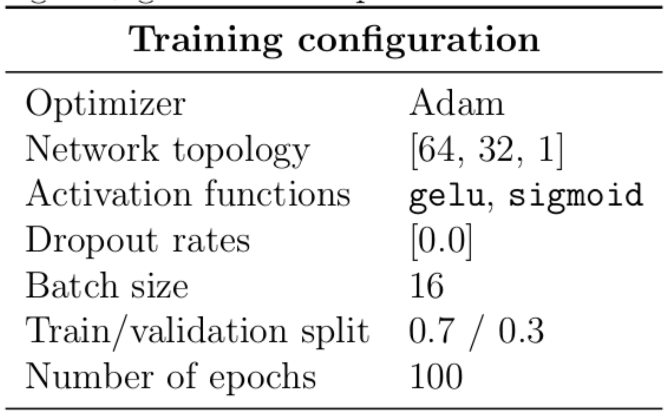


- **We used different sets of hyper-parameters like activation functions, optimizers etc to train LSTM peak finding model on 50000 tuned simulated waveforms on the simulation parameter(showed in the previous table) and then we select the best model with highest area under the curve value (0.96) among all configurations showed by red dot in left Zoom view of AUC value.**





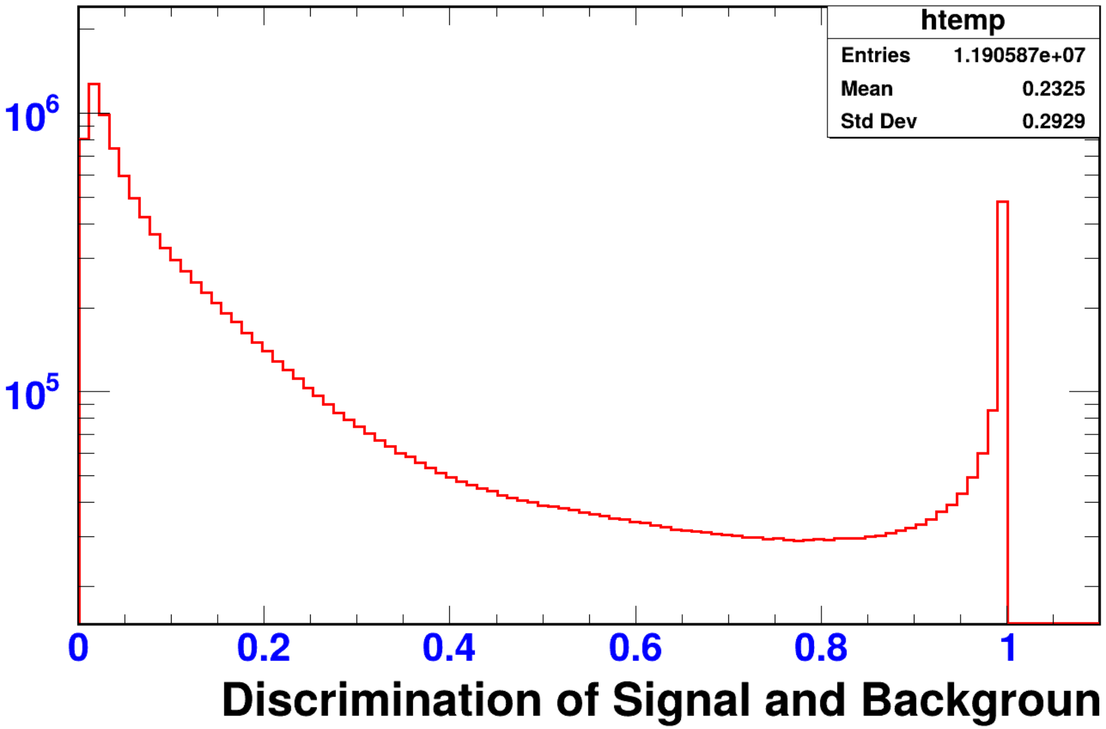


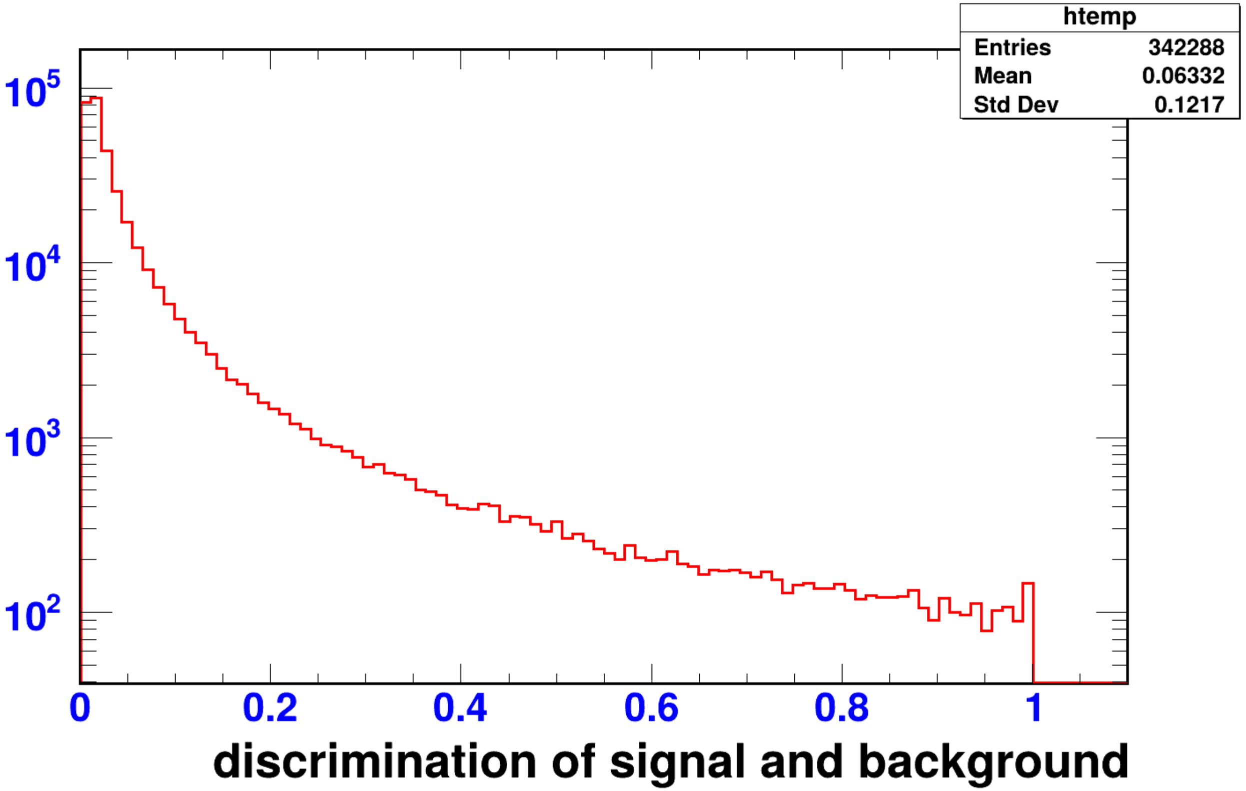


Training configuration	
Optimizer	Adam
Network topology	[64, 32, 1]
Activation functions	gelu, sigmoid
Dropout rates	[0.0]
Batch size	16
Train/validation split	0.7 / 0.3
Number of epochs	100

Hyperparameters of Best LSTM Model

TUNE SIMULATED VS Data Testing Part by using Best LSTM Model (AUC =0.96)



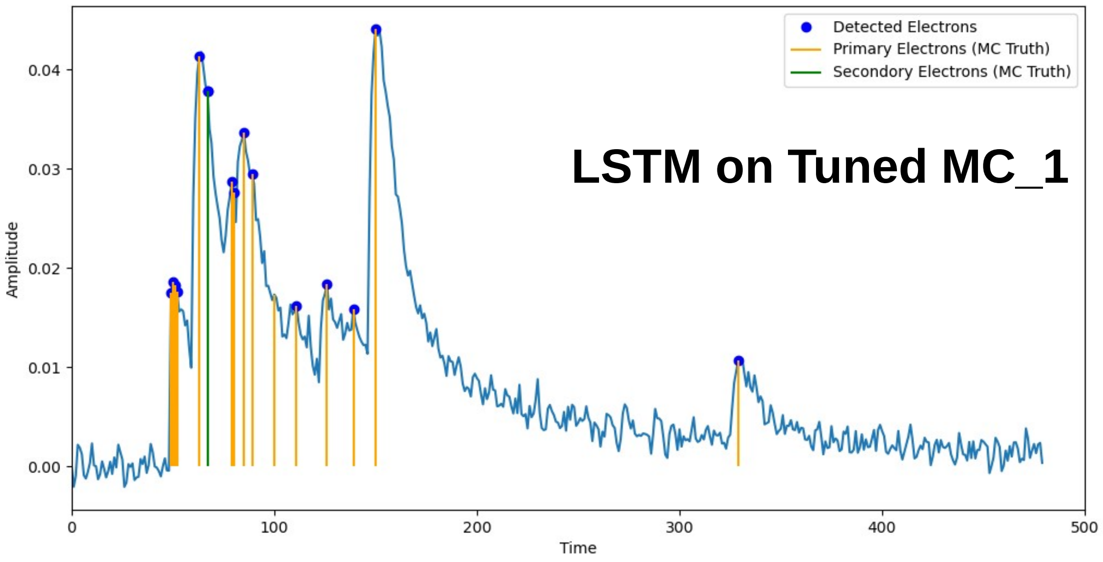


We applied best LSTM model on tune MC samples to discriminate signal and Background events

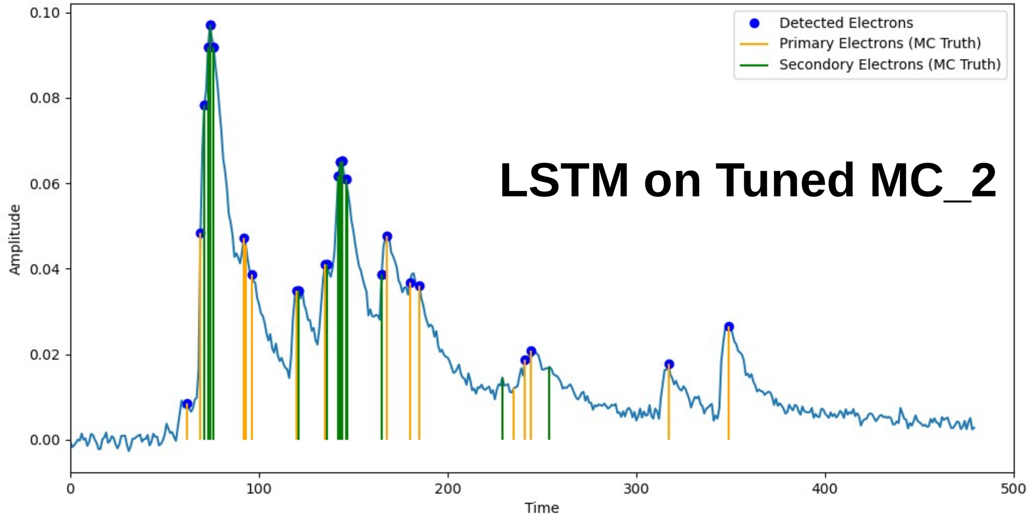
We applied best LSTM model on data to discriminate signal and Background events

LSTM Peak Finding Waveforms Results of MC & Data In Comparison with RTA/Derivative Algorithm

180 GeV Monte Carlo (MC)

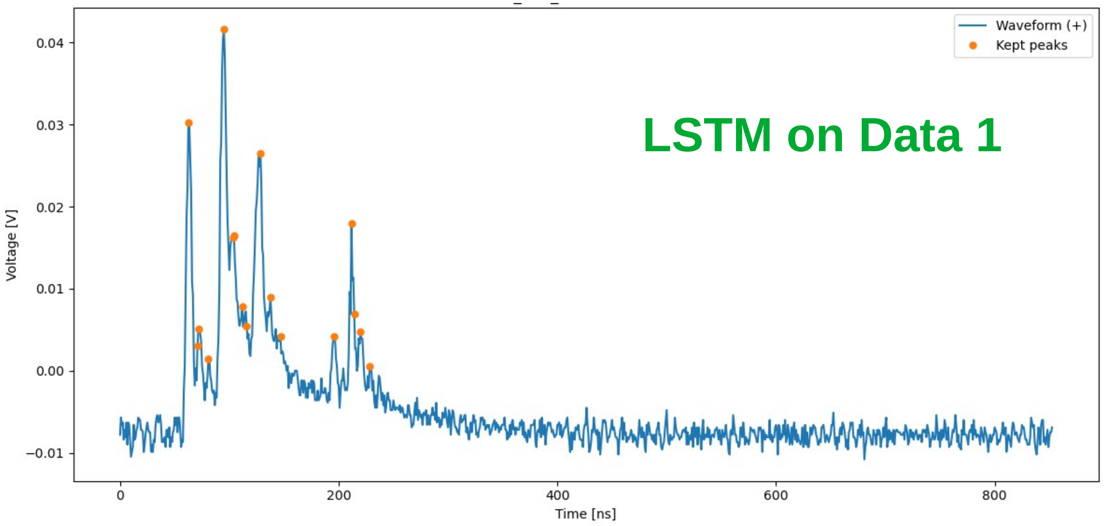


LSTM on Tuned MC_1

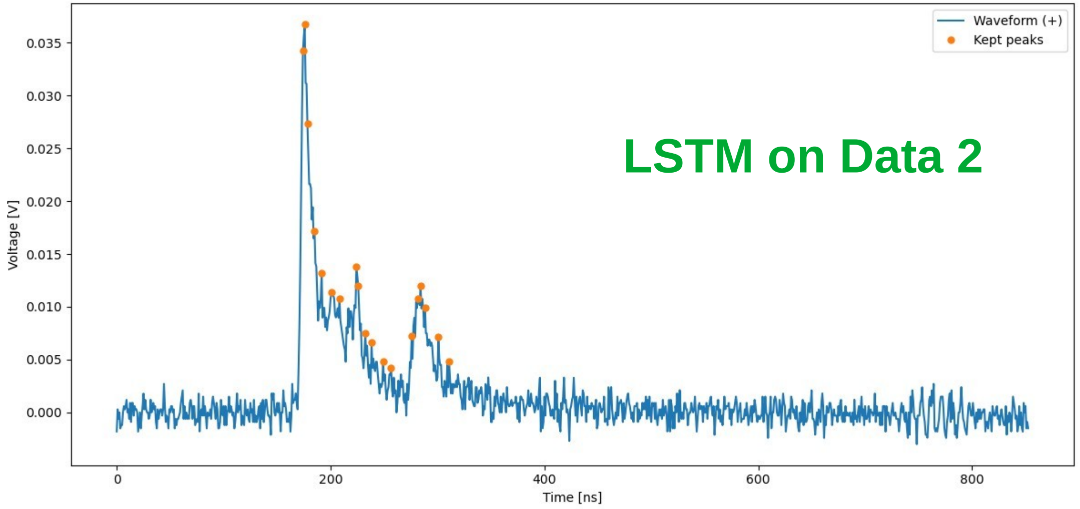


LSTM on Tuned MC_2

180 GeV Real Test Data

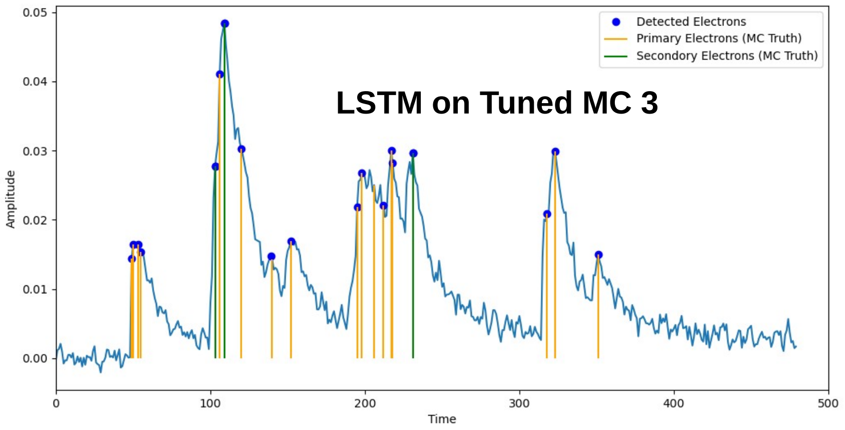


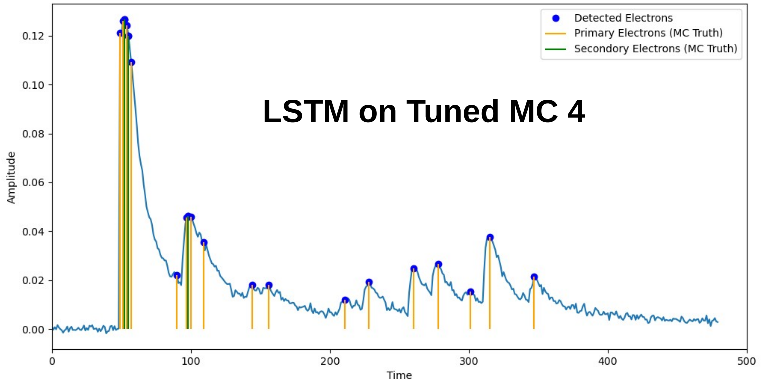
LSTM on Data 1

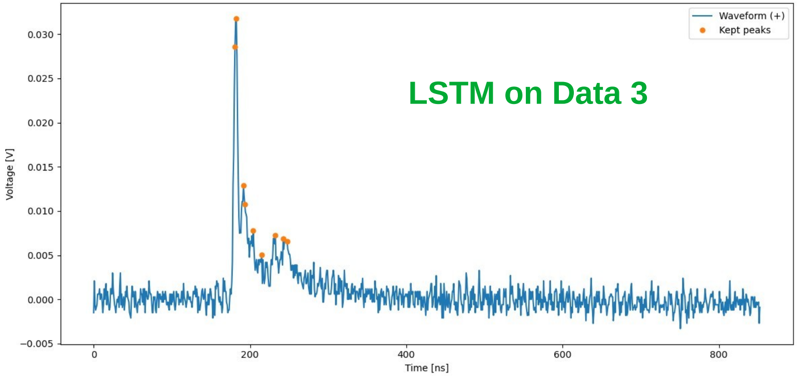


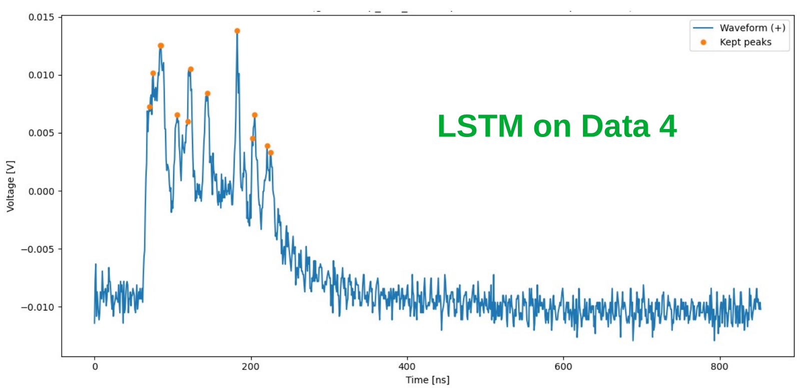
LSTM on Data 2

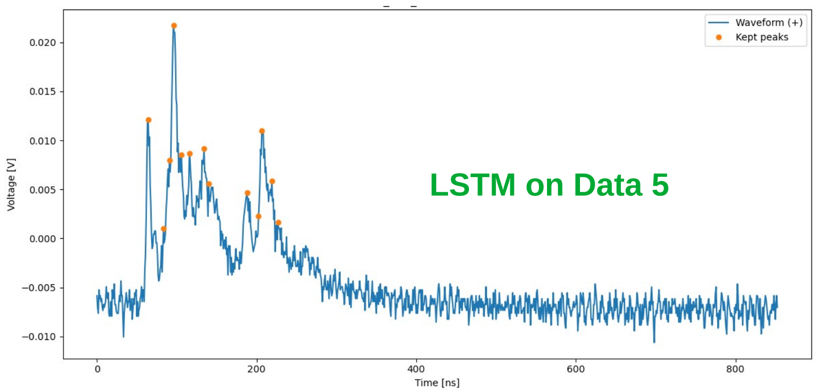
LSTM Peak Finding Results of MC & Data In Comparison with RTA/Derivative Algorithm

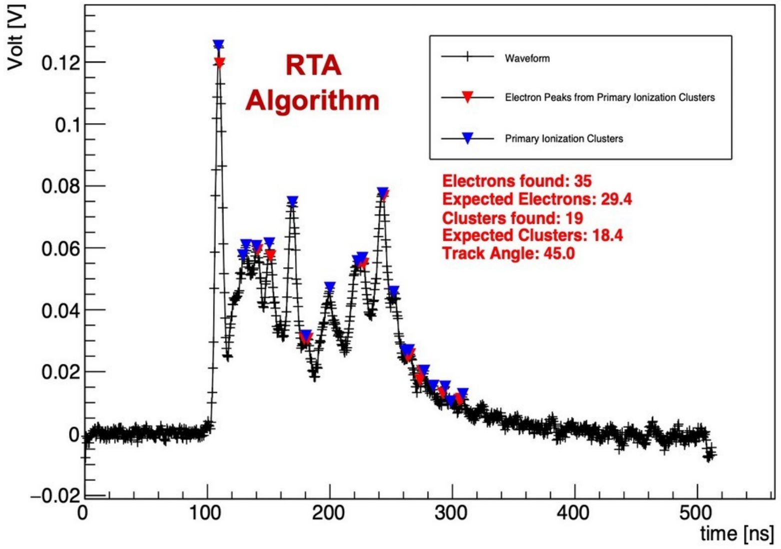

LSTM on Tuned MC 3


LSTM on Tuned MC 4

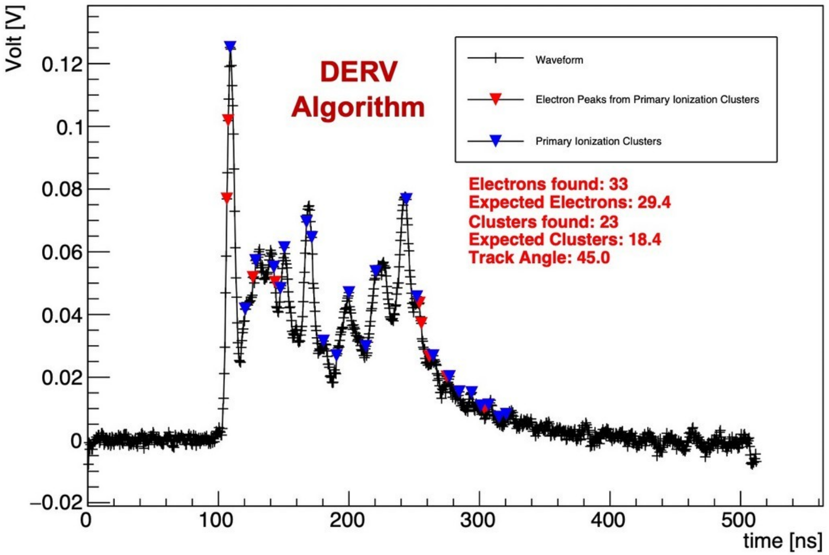

LSTM on Data 3


LSTM on Data 4


LSTM on Data 5

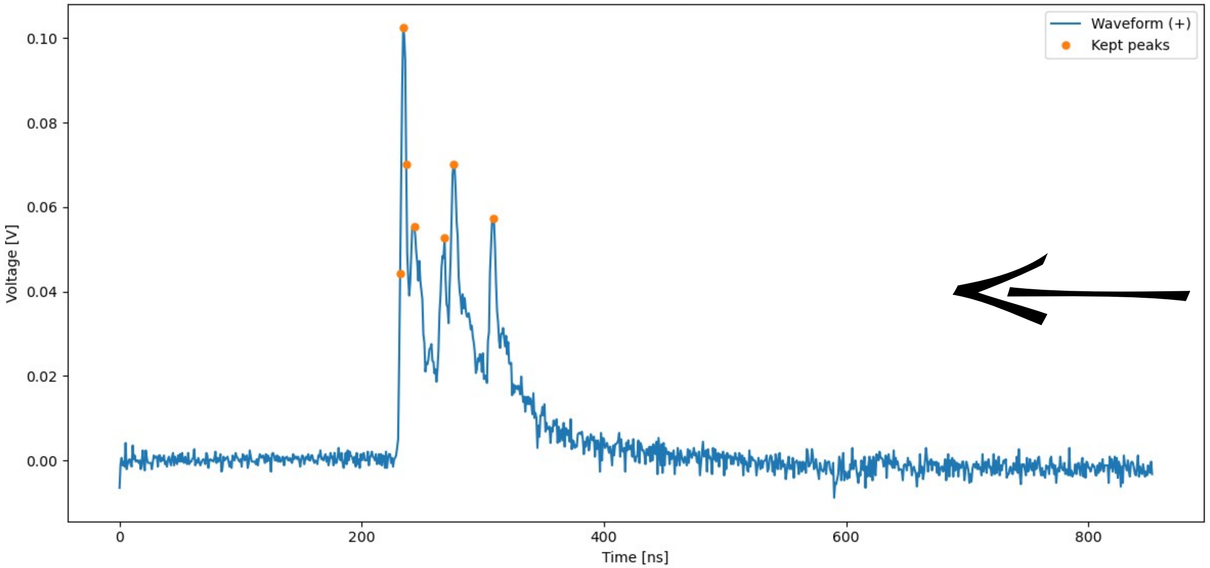

RTA Algorithm

Electrons found: 35
Expected Electrons: 29.4
Clusters found: 19
Expected Clusters: 18.4
Track Angle: 45.0

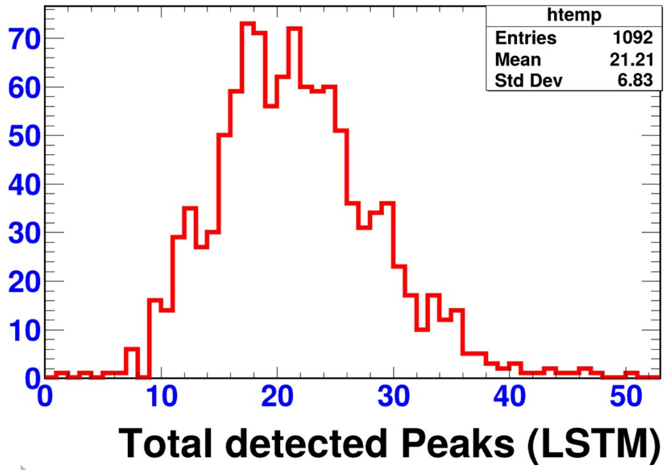

DERV Algorithm

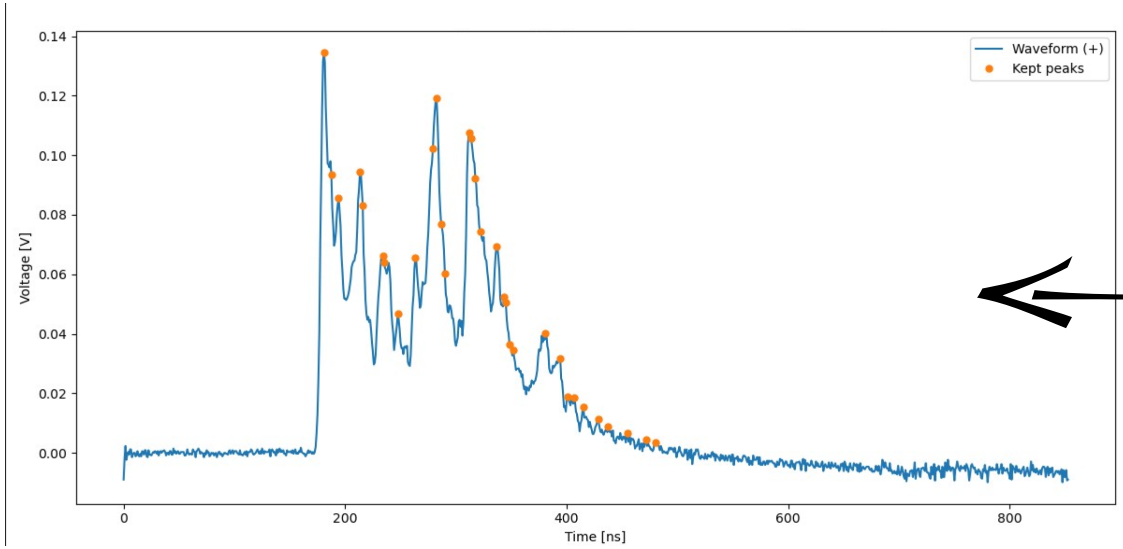
Electrons found: 33
Expected Electrons: 29.4
Clusters found: 23
Expected Clusters: 18.4
Track Angle: 45.0

Selection cuts on real data (Selection cut /Threshold)

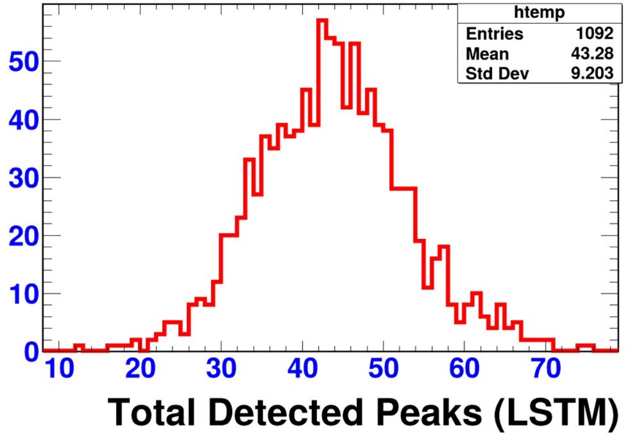


At Selection cut/Threshold = 0.4, less peaks appear as result found less mean value of the detected peaks in the distribution

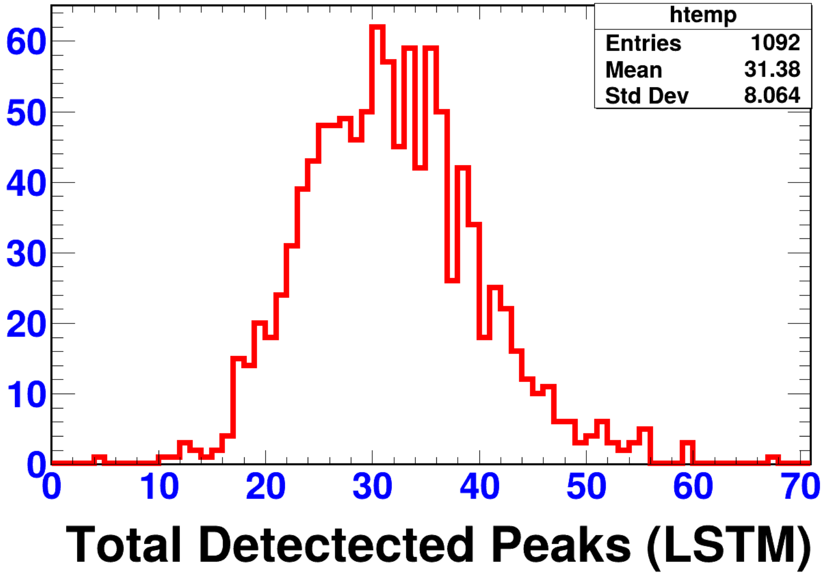


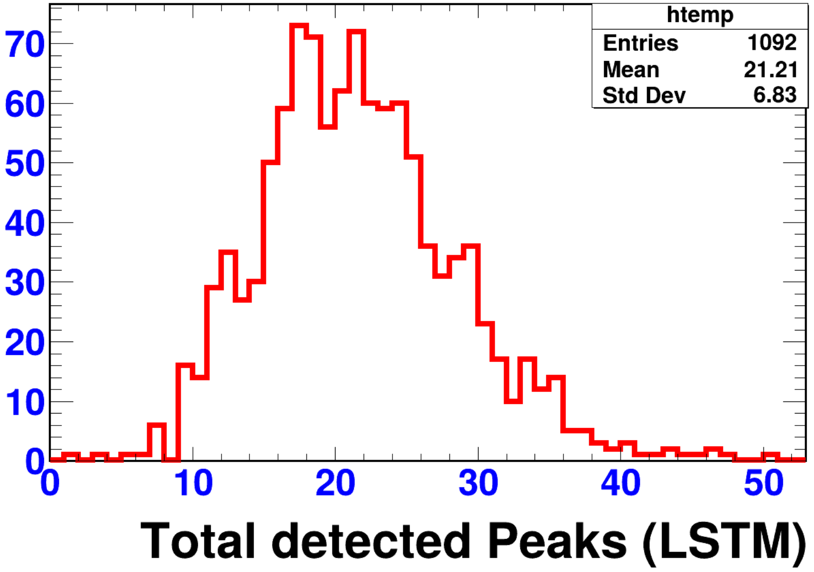


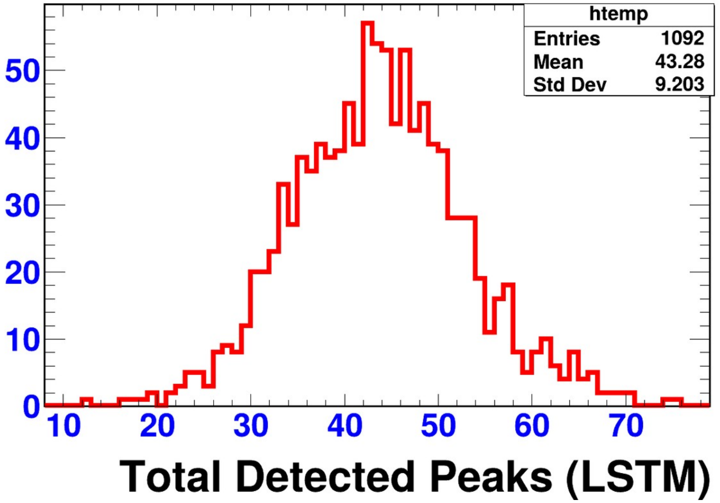
At Threshold = 0.25, some peak appear in the region of noise amplitude peaks appear as result found some fake peaks in mean value of the detected peaks in the distribution



Counting Peaks Distribution of Data with Different Selection Cuts







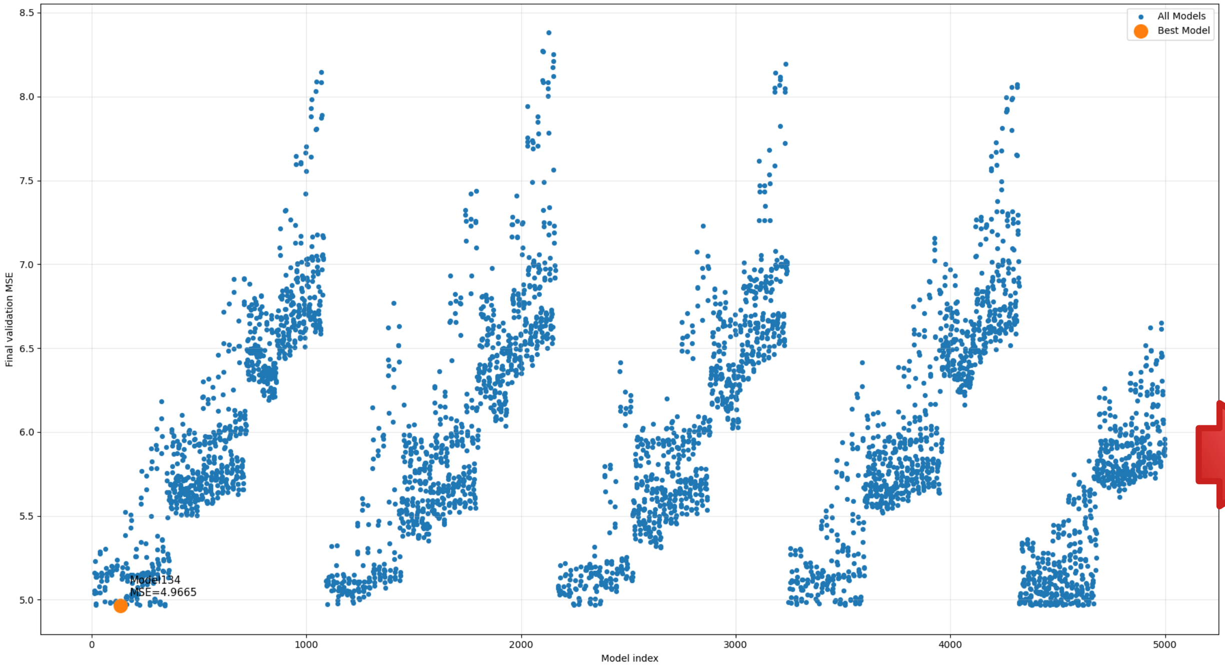
The above distrubution shows us the total detected peaks with mean value (31.38) for 1092 waveforms at selction cut 0.3

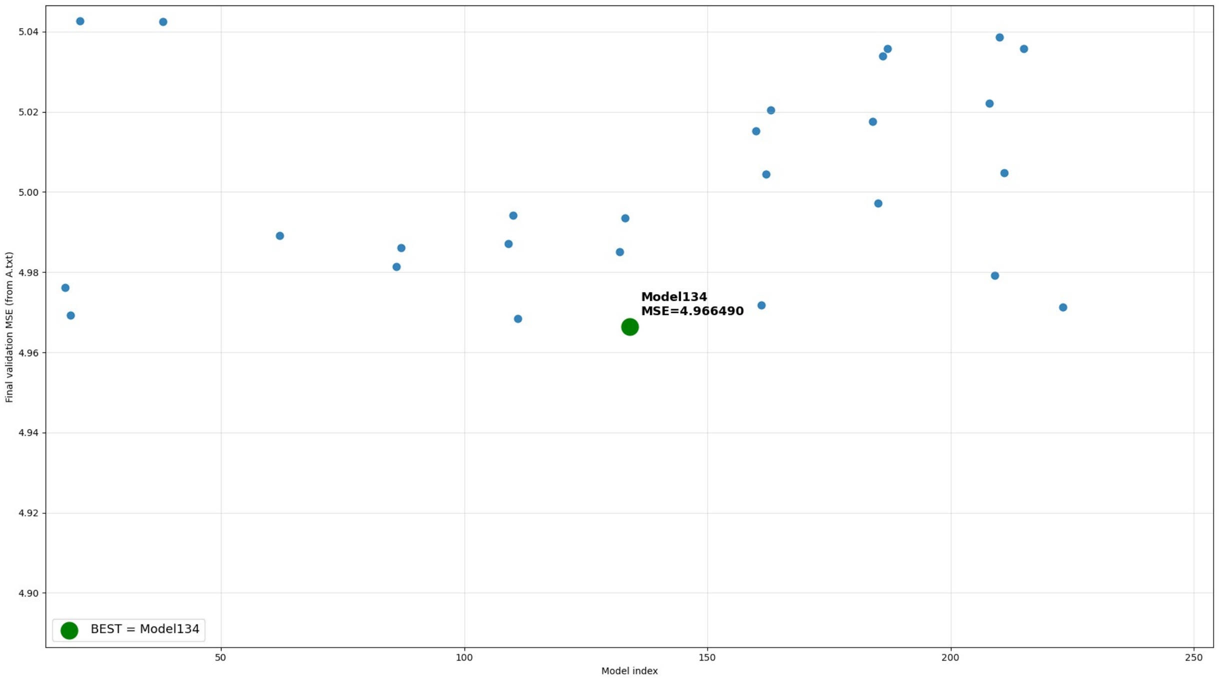
The above distrubution shows us the total detected peaks with mean value (21.21) for 1092 waveforms at selection cut 0.4

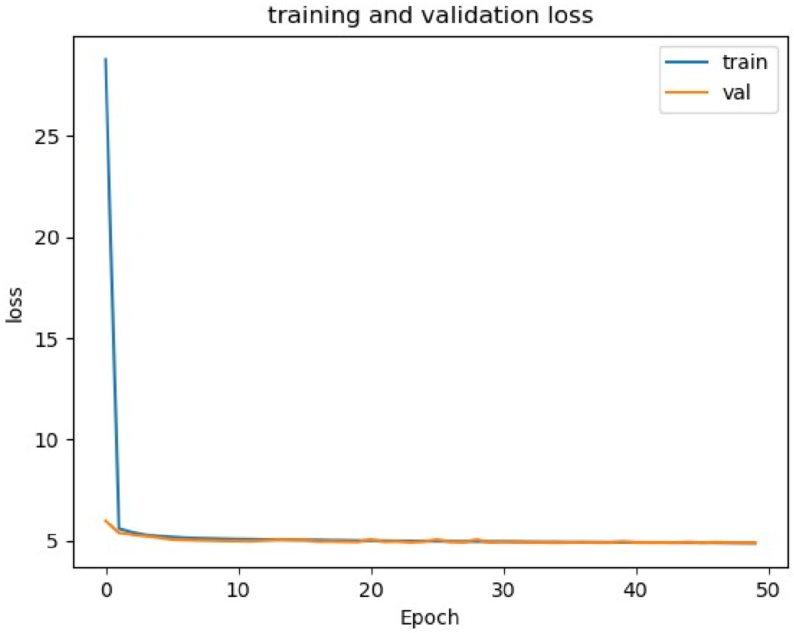
The above distrubution shows us the total detected peaks with mean value (43.28) for 1092 waveforms 0.25

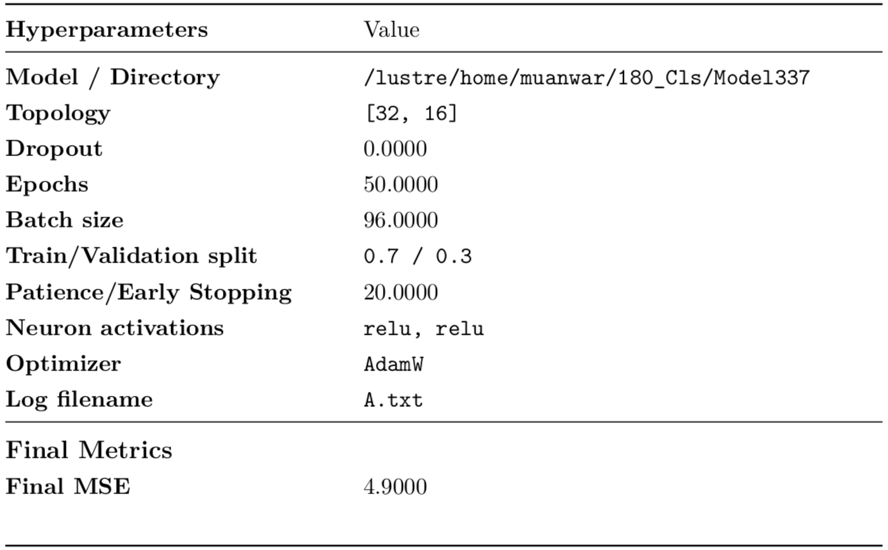
CNN Clusterization Algorithm on Real Data

Search Grid Optimzation: BEST CNN Model ON TUNE SIMULATED WAVEFORMS (50000)





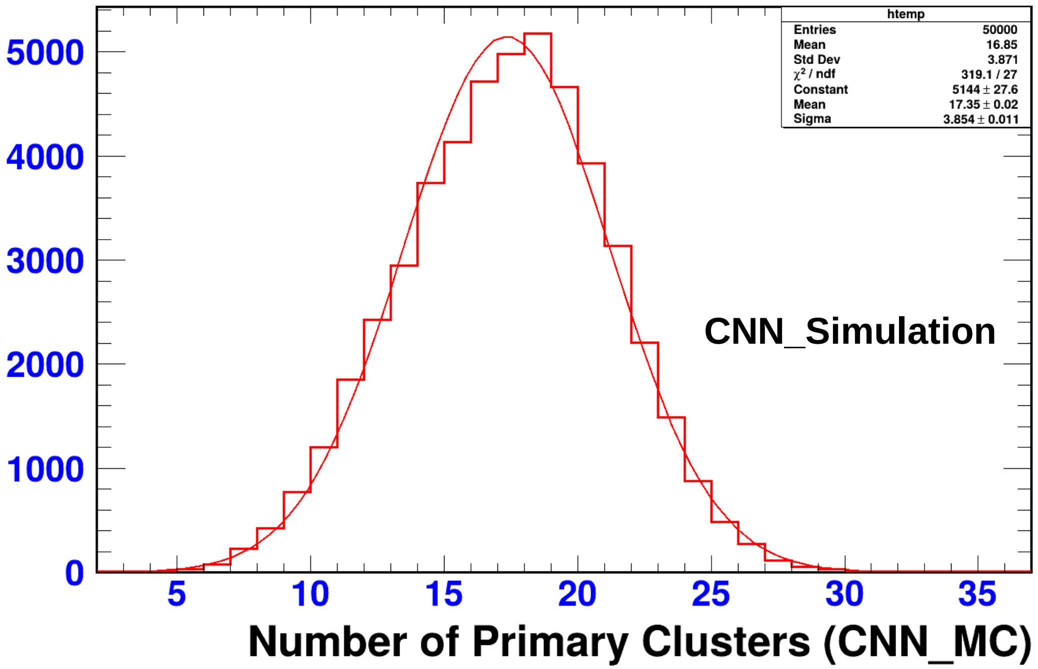




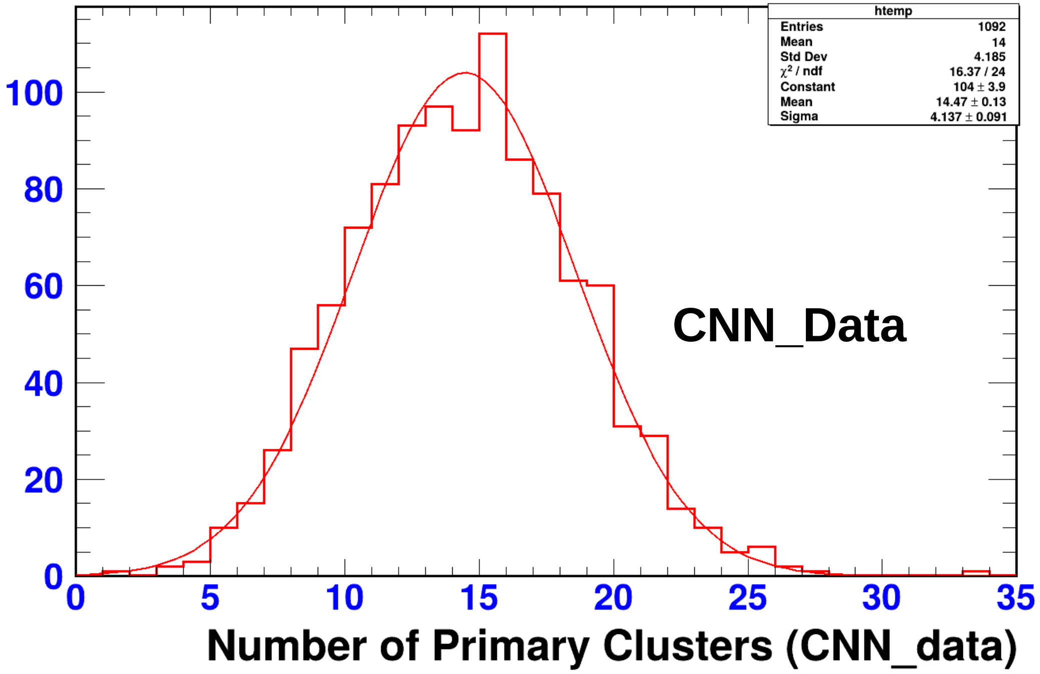
Hyperparameters	Value
Model / Directory	/lustre/home/muanwar/180_Cls/Model337
Topology	[32, 16]
Dropout	0.0000
Epochs	50.0000
Batch size	96.0000
Train/Validation split	0.7 / 0.3
Patience/Early Stopping	20.0000
Neuron activations	relu, relu
Optimizer	AdamW
Log filename	A.txt
Final Metrics	
Final MSE	4.9000

Hyperparameters of BestCNN Model (Lowest Mean Square error value)

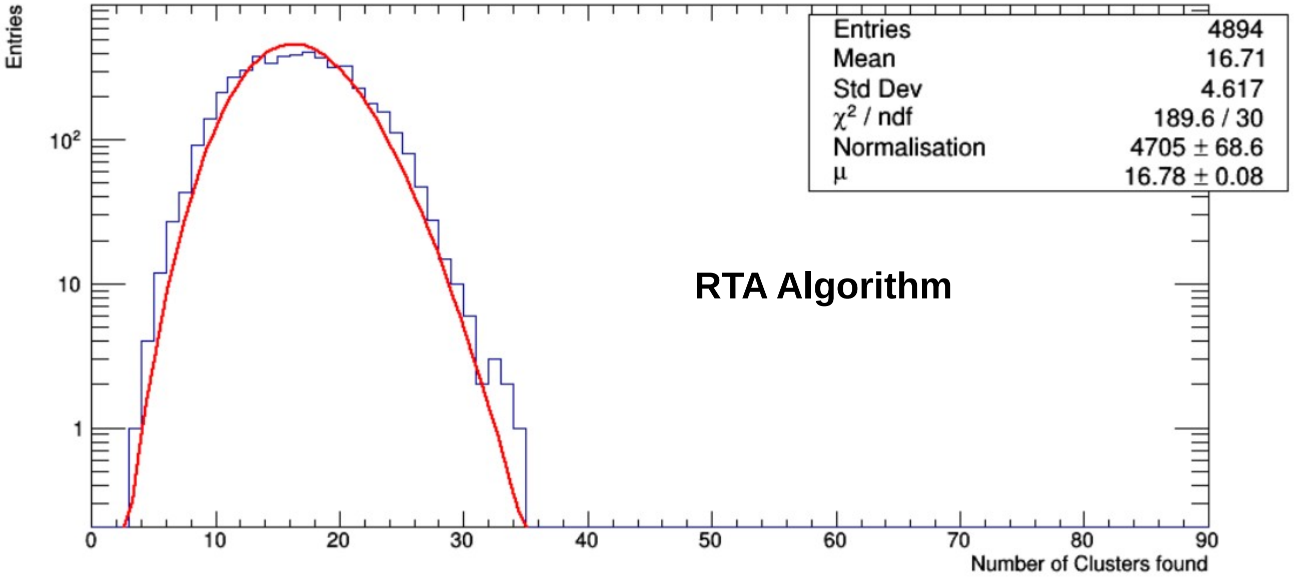
Preliminary Results of Number of Primary Clusters (CNN)



htemp	
Entries	50000
Mean	16.85
Std Dev	3.871
χ^2 / ndf	319.1 / 27
Constant	5144 ± 27.6
Mean	17.35 ± 0.02
Sigma	3.854 ± 0.011



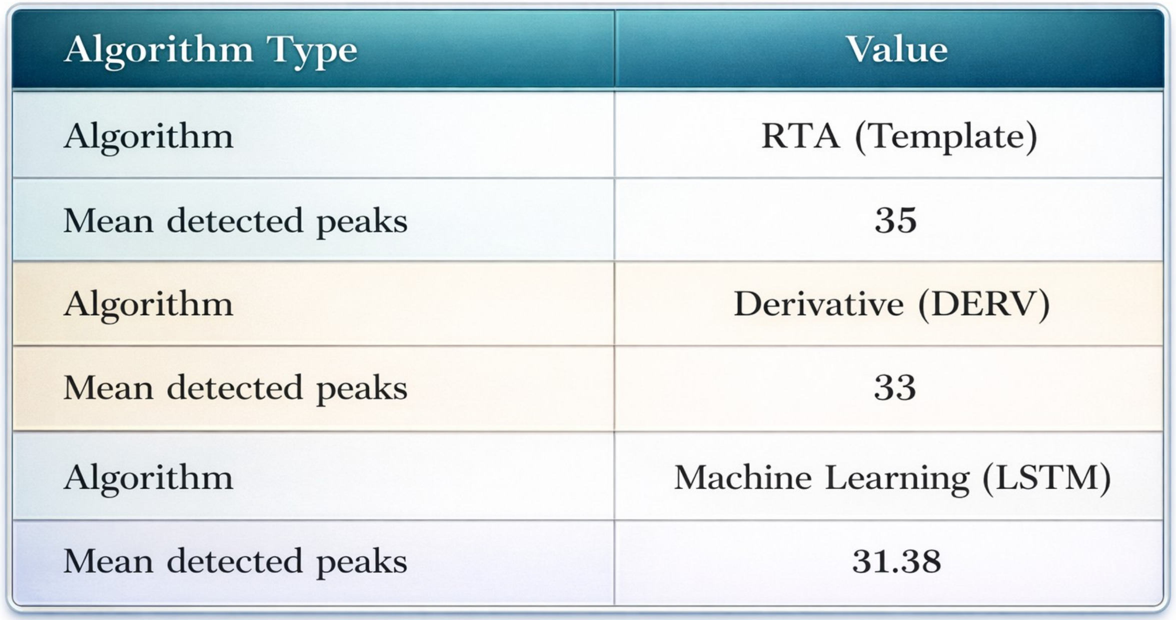
htemp	
Entries	1092
Mean	14
Std Dev	4.185
χ^2 / ndf	16.37 / 24
Constant	104 ± 3.9
Mean	14.47 ± 0.13
Sigma	4.137 ± 0.091



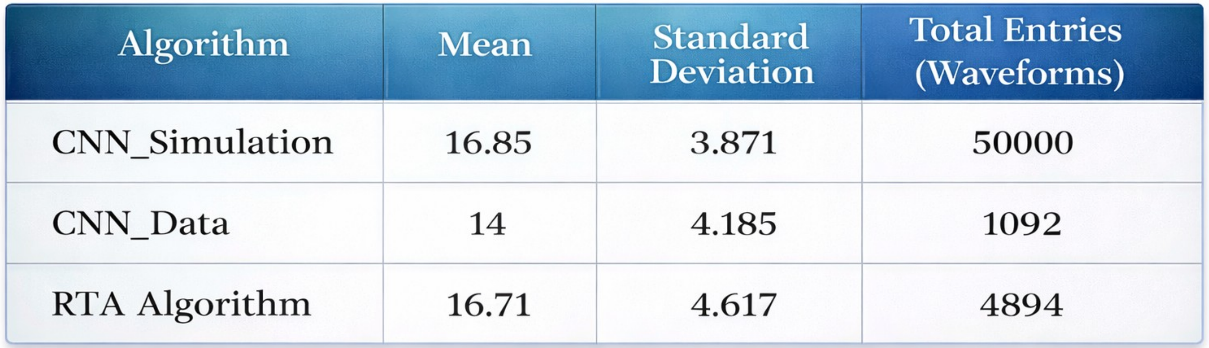
Entries	4894
Mean	16.71
Std Dev	4.617
χ^2 / ndf	189.6 / 30
Normalisation	4705 ± 68.6
μ	16.78 ± 0.08

.

Conclusion



Algorithm Type	Value
Algorithm	RTA (Template)
Mean detected peaks	35
Algorithm	Derivative (DERV)
Mean detected peaks	33
Algorithm	Machine Learning (LSTM)
Mean detected peaks	31.38



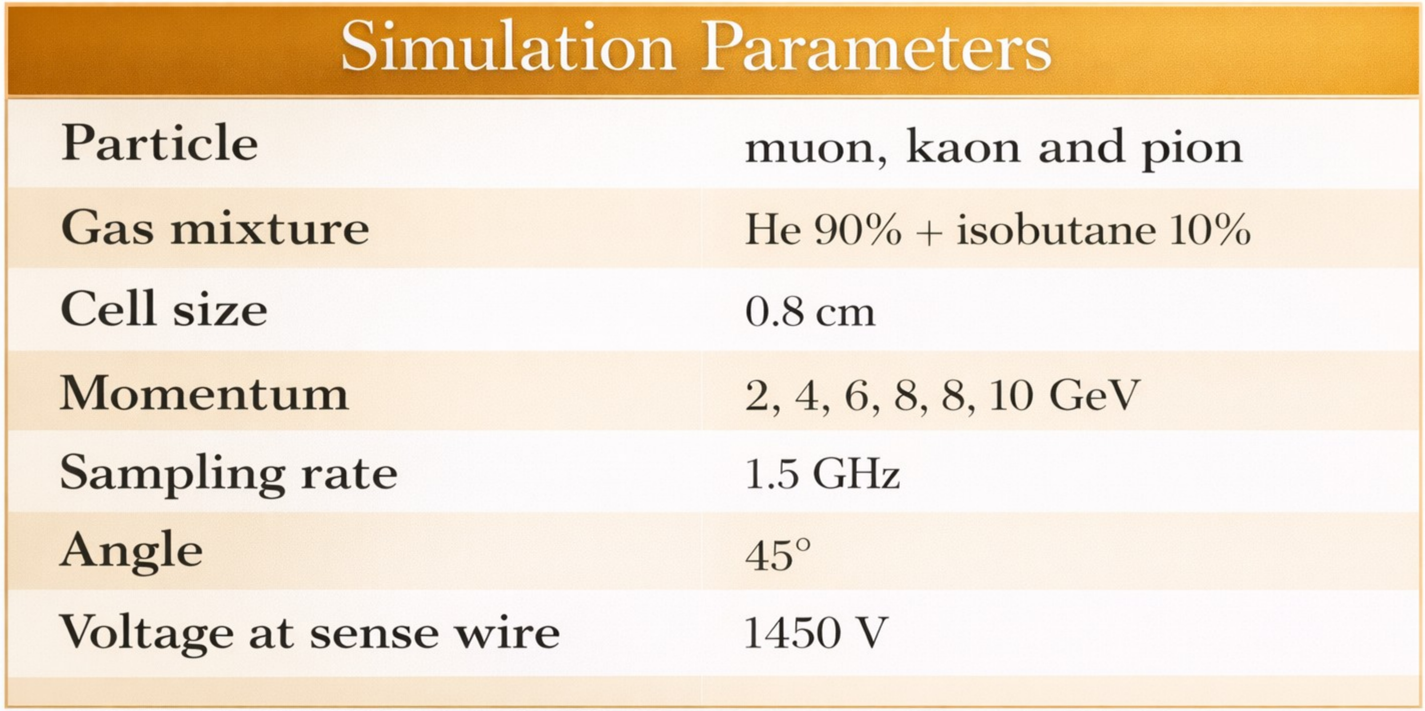
Algorithm	Mean	Standard Deviation	Total Entries (Waveforms)
CNN_Simulation	16.85	3.871	50000
CNN_Data	14	4.185	1092
RTA Algorithm	16.71	4.617	4894

Peak Finding Algorithms

Clusterization Algorithms

ML Based Algorithm for ParticleIdentification(Simulated samples + NN) (Ongoing)

ML Based Algorithm for Particle Identification: Simulation Parameters

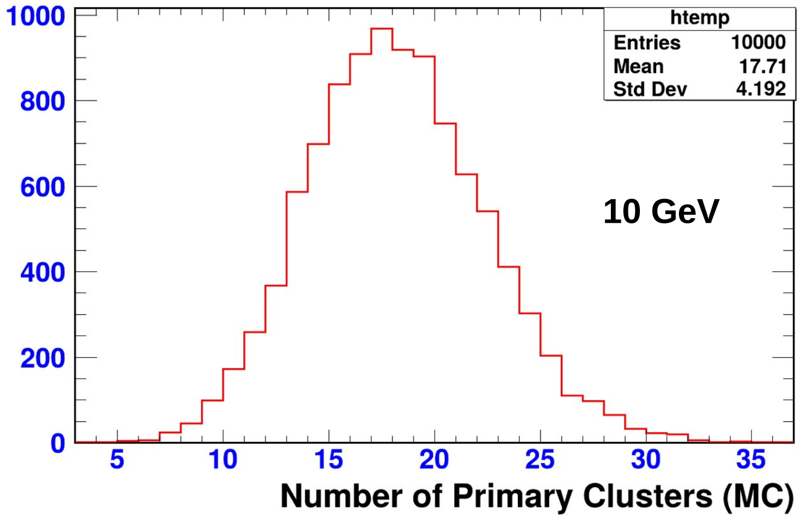


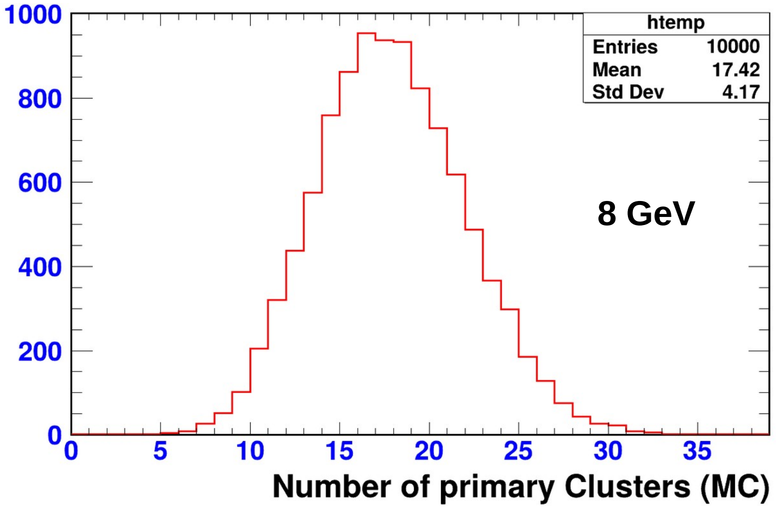
Simulation Parameters	
Particle	muon, kaon and pion
Gas mixture	He 90% + isobutane 10%
Cell size	0.8 cm
Momentum	2, 4, 6, 8, 8, 10 GeV
Sampling rate	1.5 GHz
Angle	45°
Voltage at sense wire	1450 V

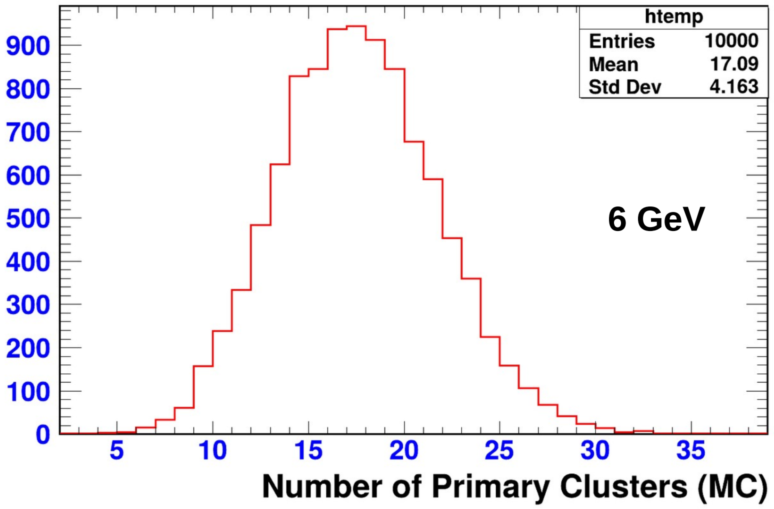
Above are Simulation parameters were used which matched the real test beam data.

Number of Primary Cluster (MC)

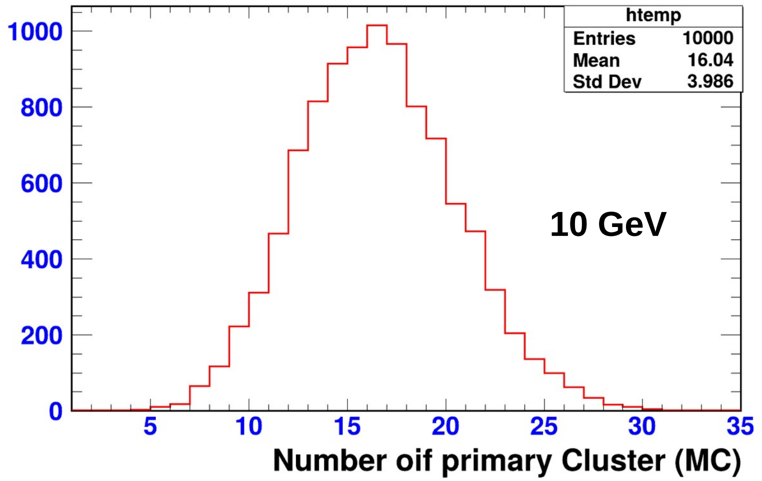
For Muon

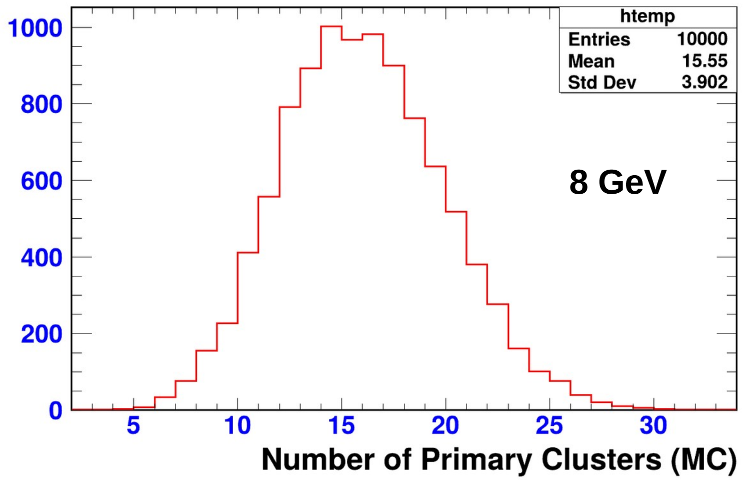


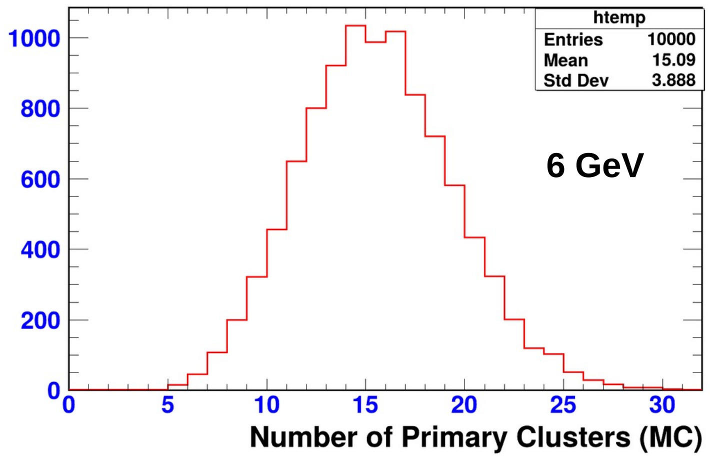




For Kaon

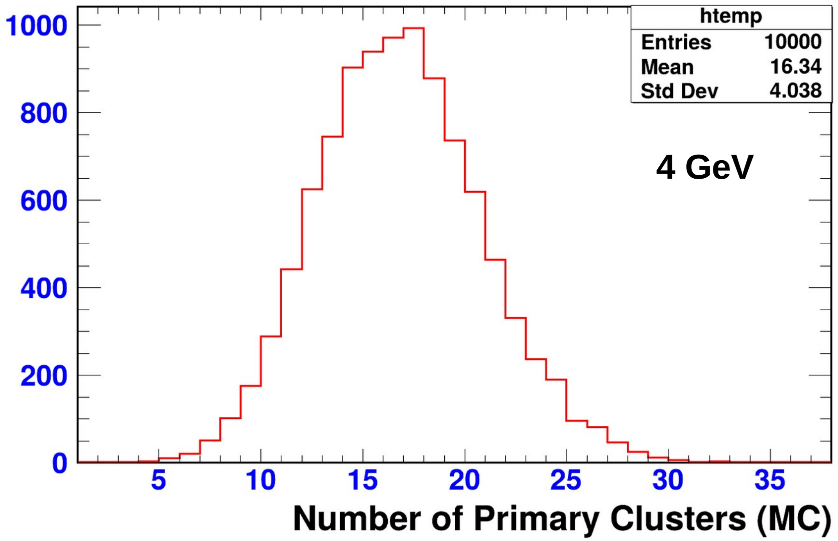


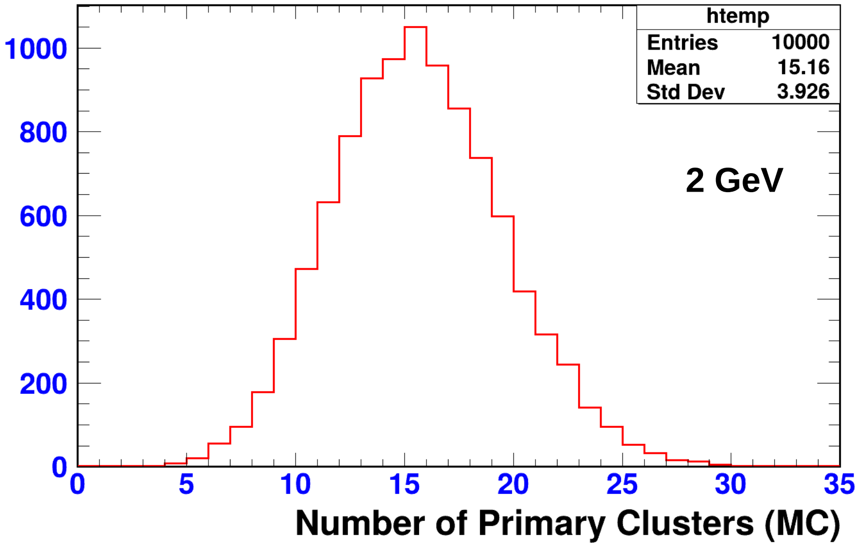




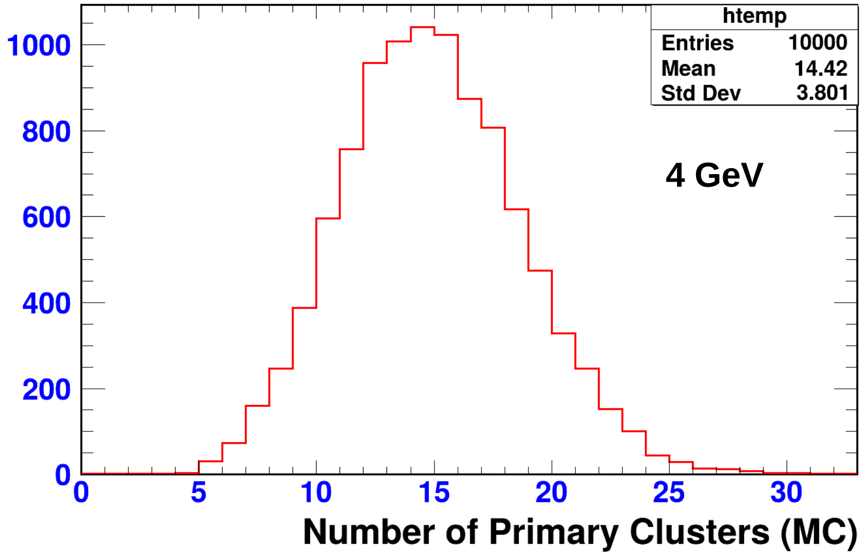
Number of Primary Cluster (MC)

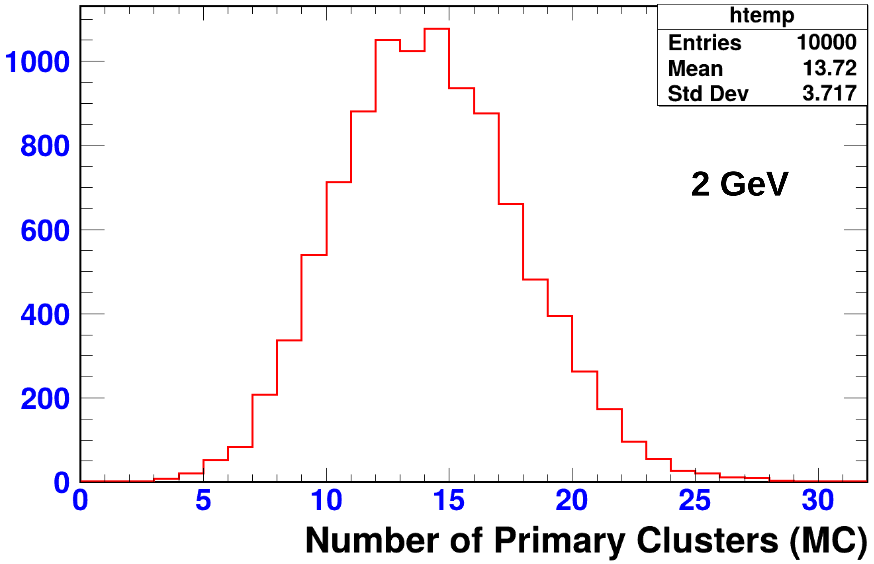
For Muon





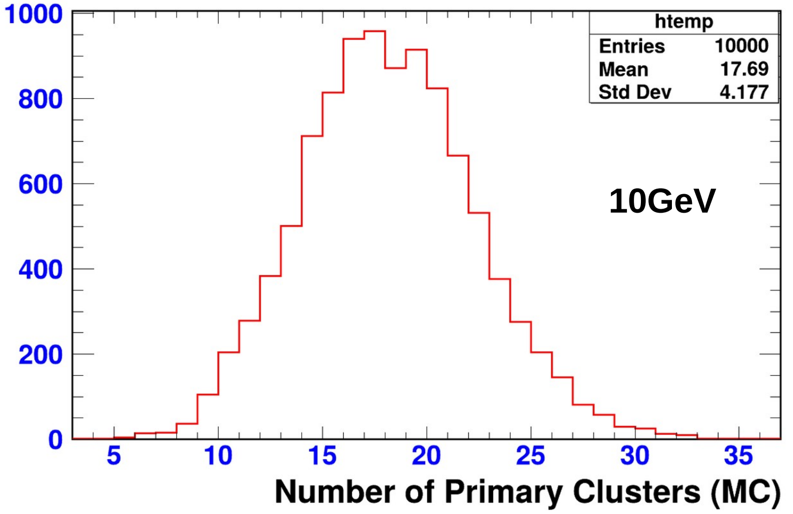
For Kaon

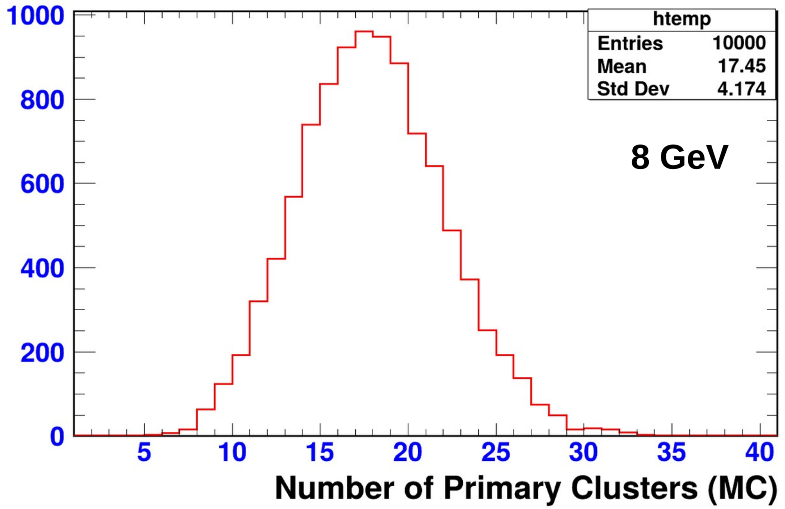


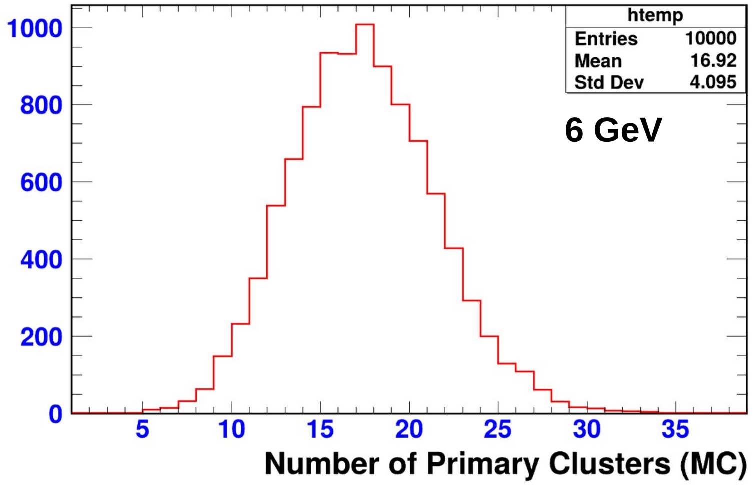


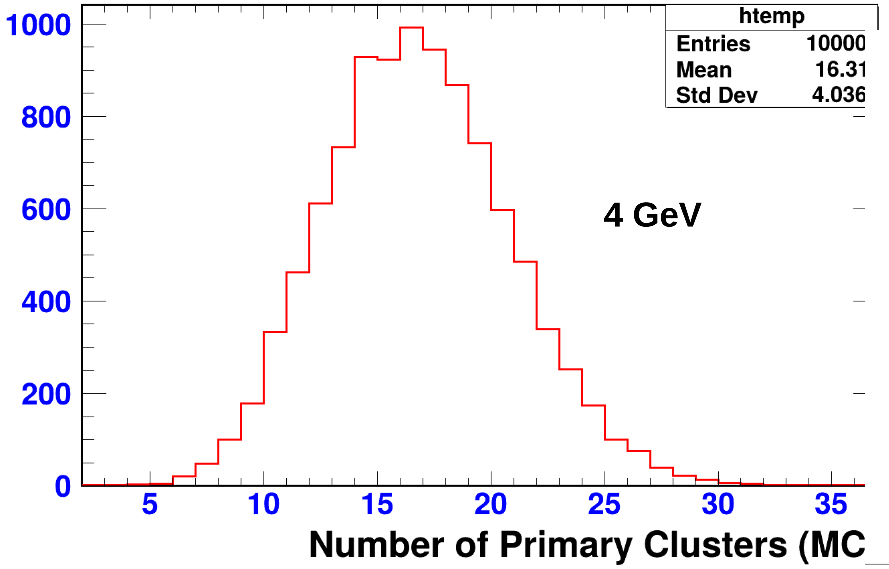
All these distributions are related to the number of primary clusters Monet carlo (MC) for the momentum (2-10 Gev) of muon and Kaon. We also need to generated it for 180 GeV momentum

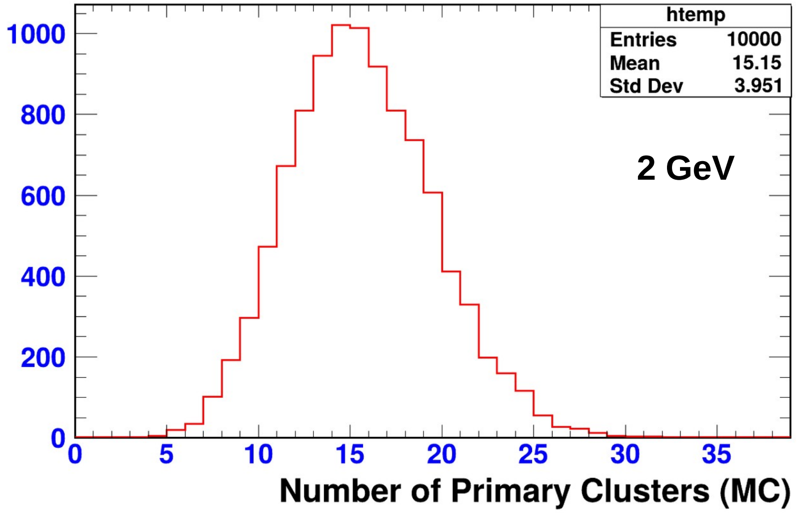
Number of Primary Cluster (MC) for Pion







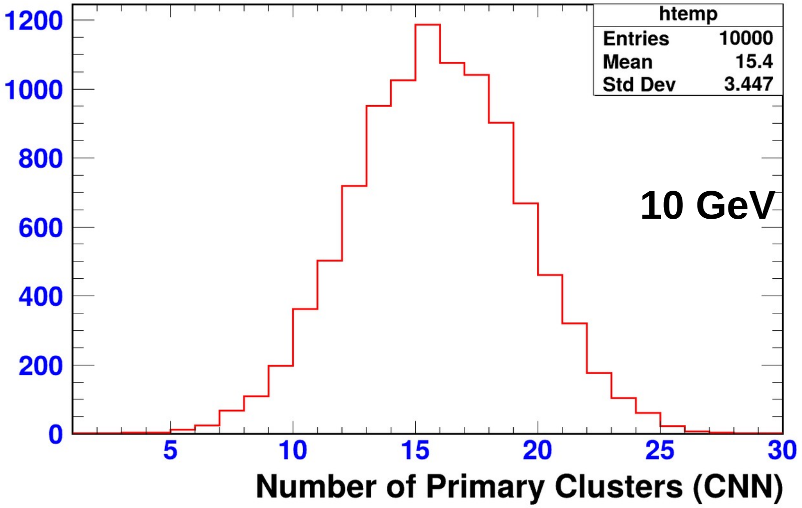


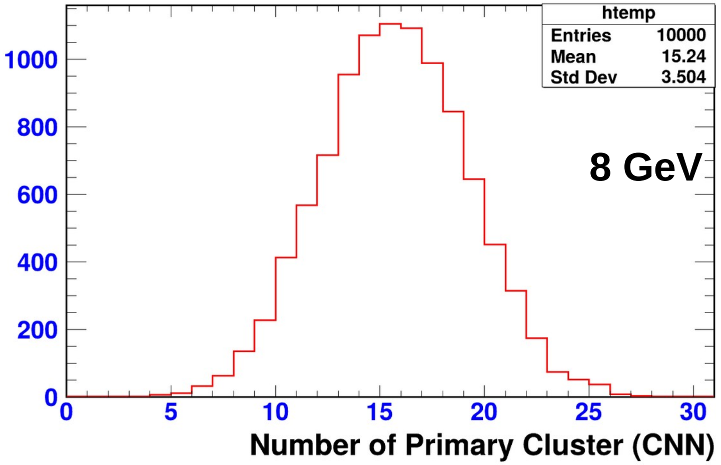


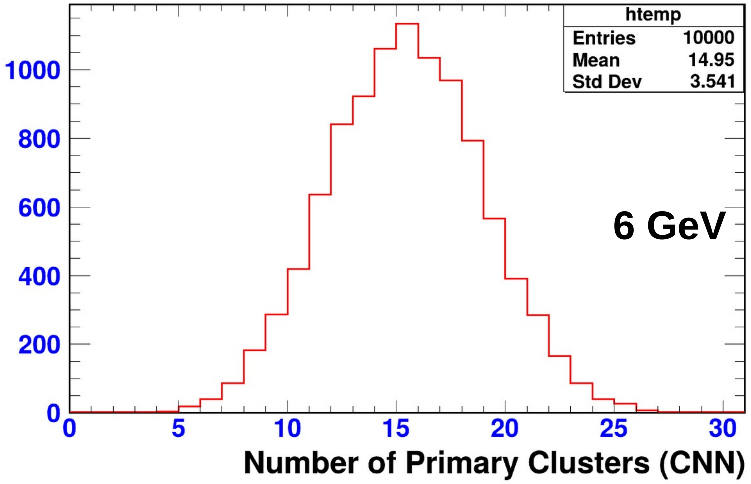
All these distributions are related to the number of primary clusters (MC) for the momentum (2-10 GeV) of Pion

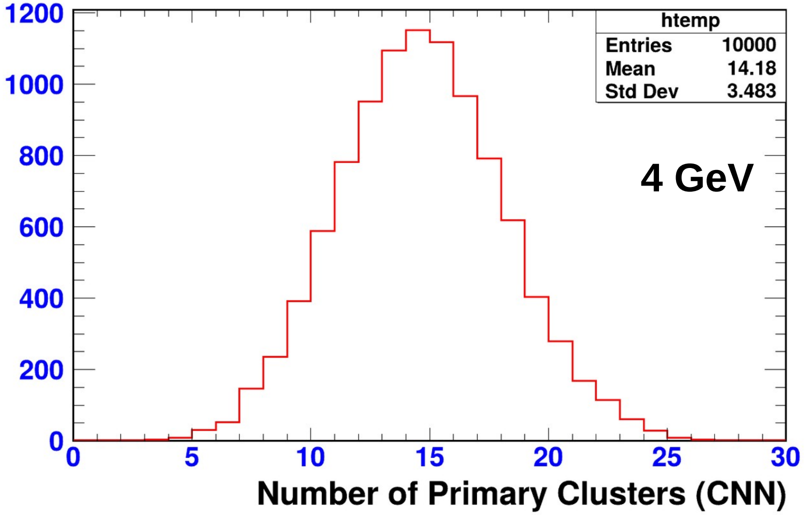
Reconstruction of Number of Primary Clusters (CNN) for Kaon

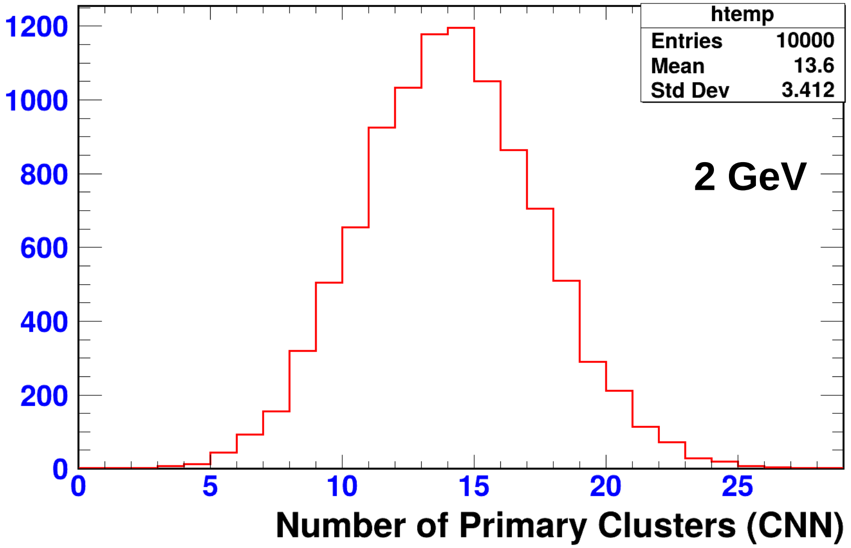
For Kaon





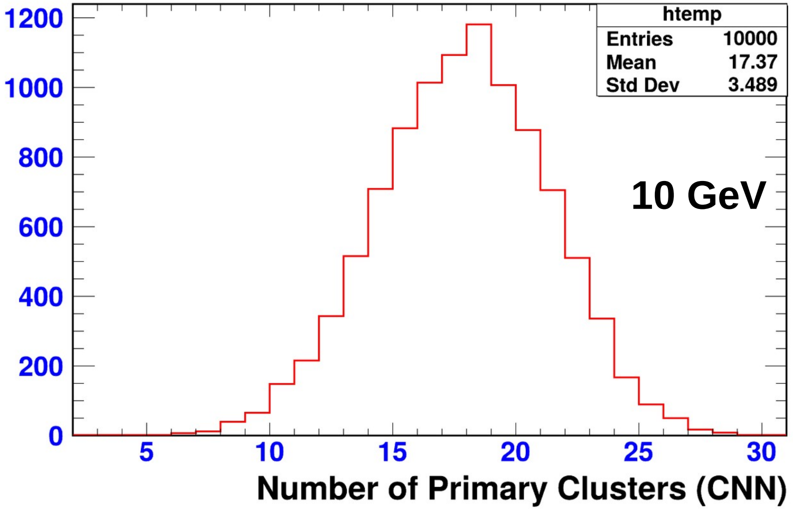


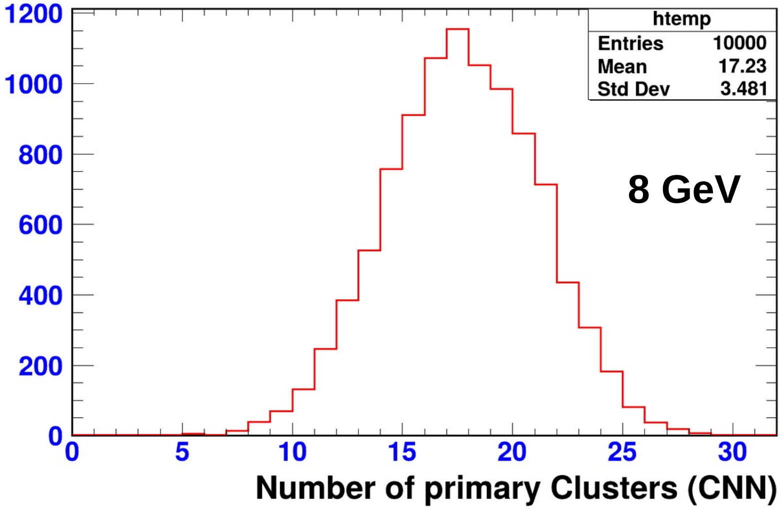


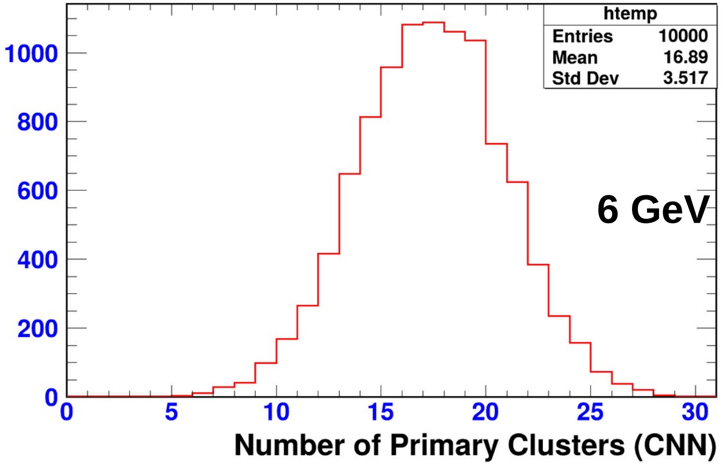


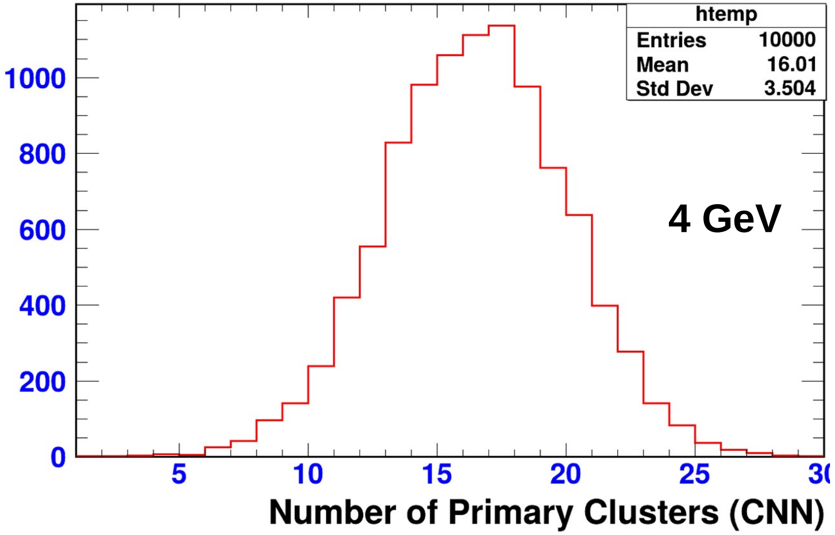
All these distributions are related to the number of primary clusters reconstructed by CNN Model for the momentum (2-10 GeV) of Kaon

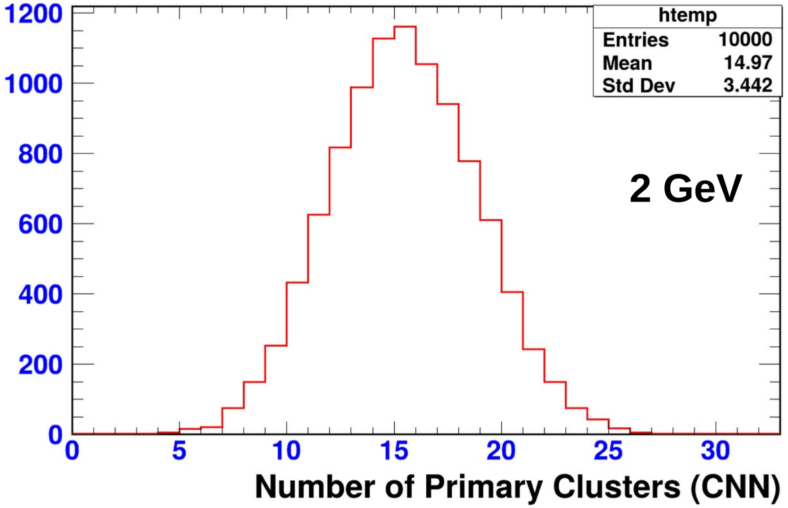
Reconstruction of Number of Primary Clusters (CNN) for Pion



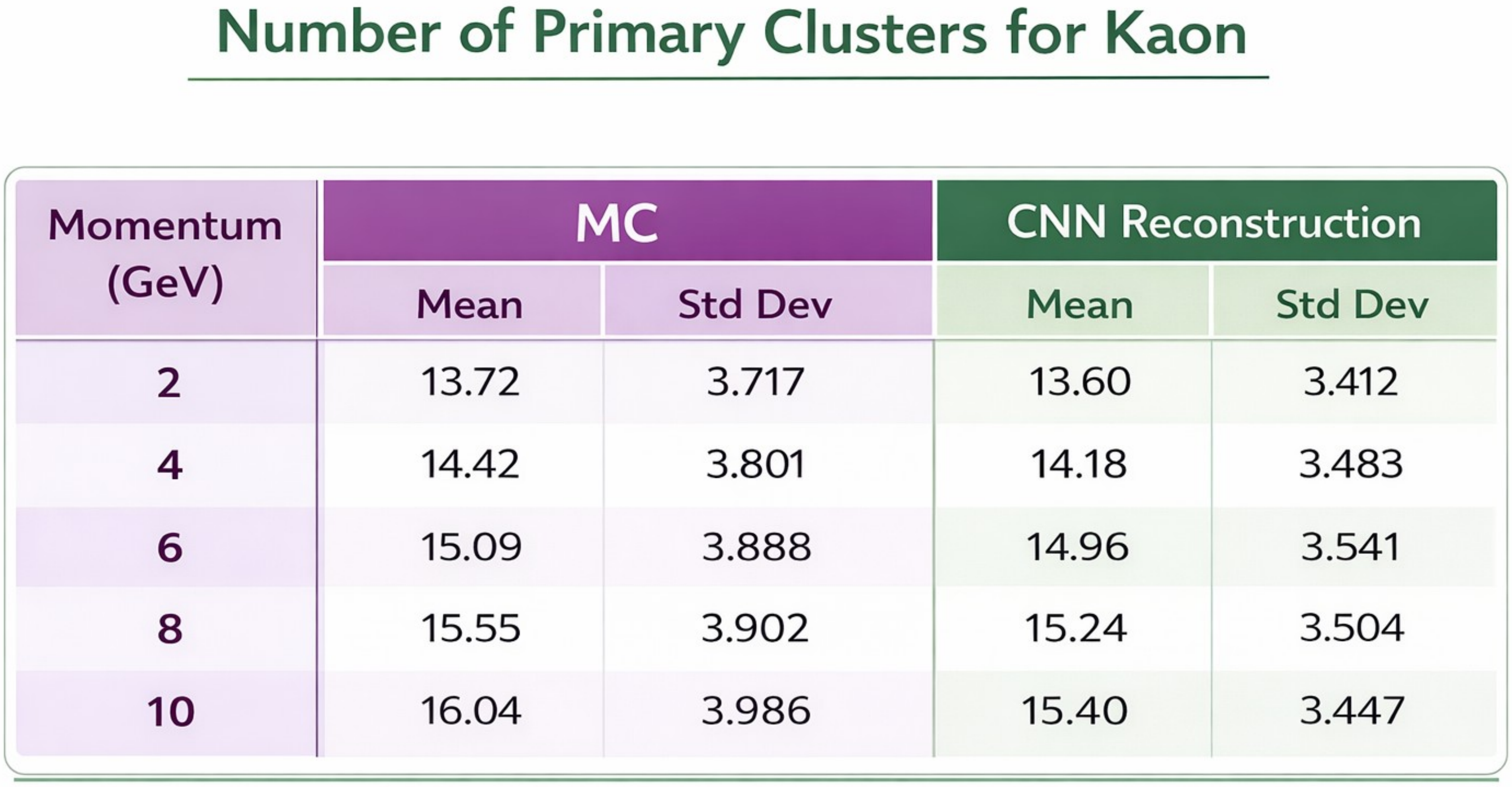








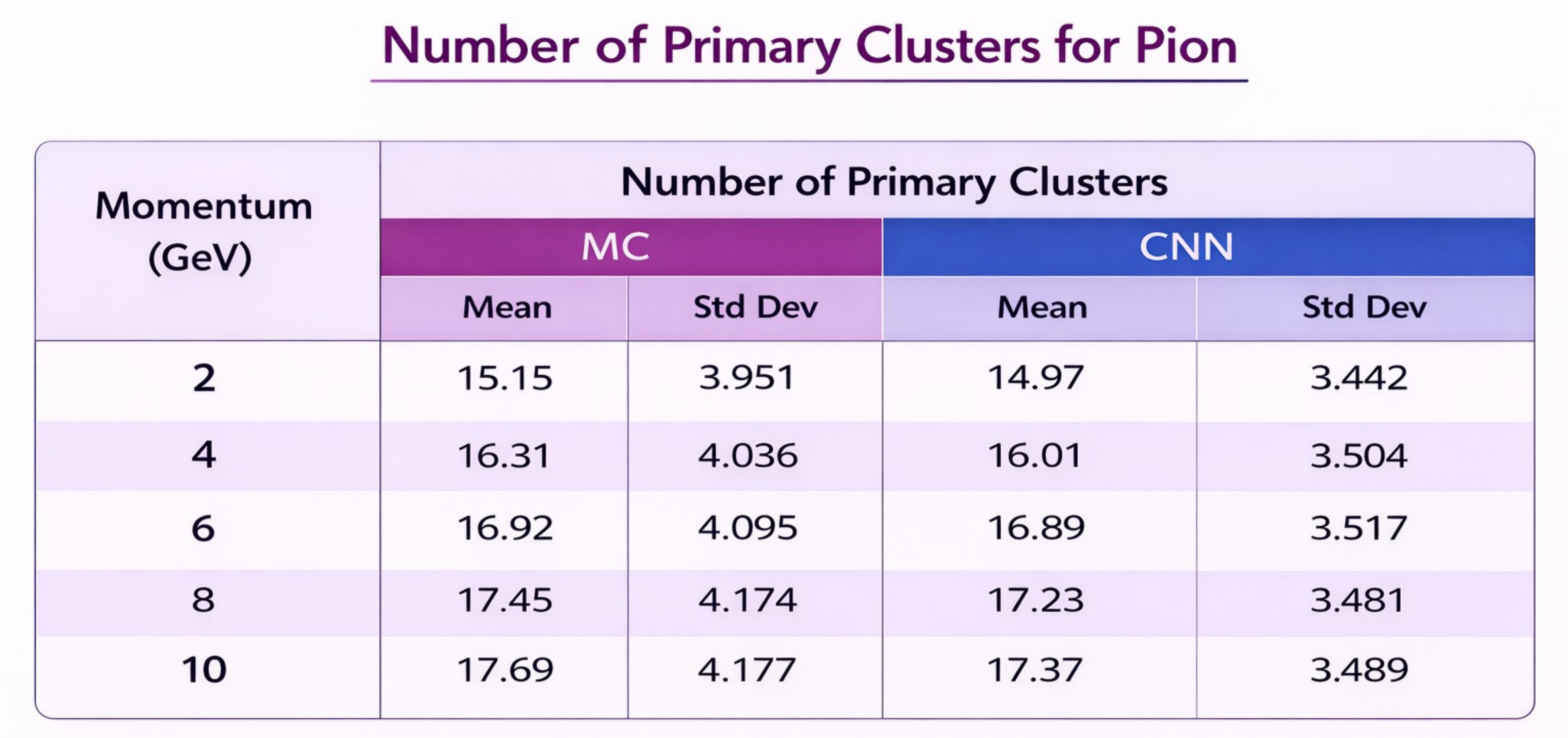
All these distributions are related to the number of primary clusters reconstructed by CNN Model for the momentum (2-10 Gev) of Pion



Number of Primary Clusters for Kaon

Momentum (GeV)	MC		CNN Reconstruction	
	Mean	Std Dev	Mean	Std Dev
2	13.72	3.717	13.60	3.412
4	14.42	3.801	14.18	3.483
6	15.09	3.888	14.96	3.541
8	15.55	3.902	15.24	3.504
10	16.04	3.986	15.40	3.447

- **The above Table shows the mean number of primary clusters (MC) and CNN reconstruction results for Kaon and will be updated soon on the behalf of an LSTM model too**

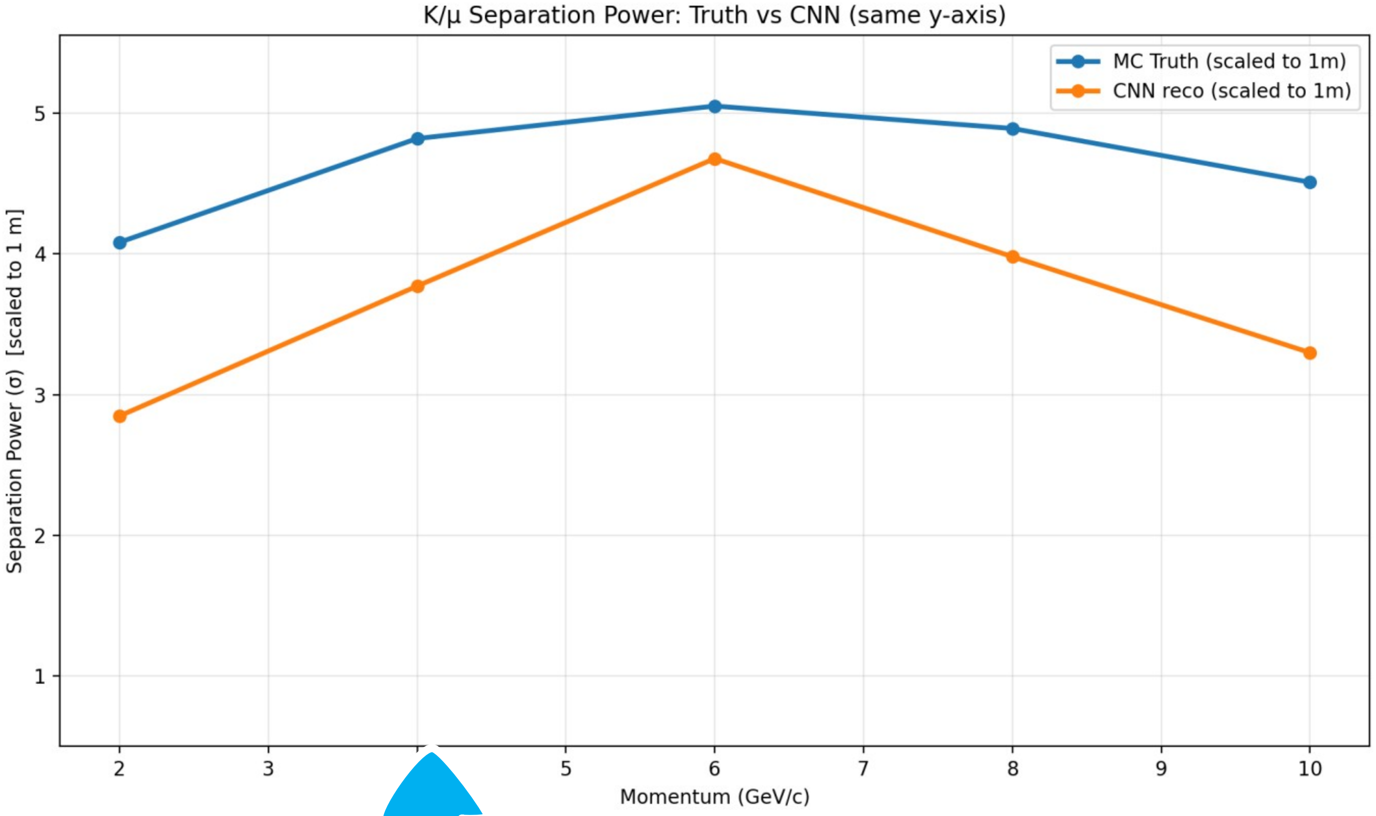


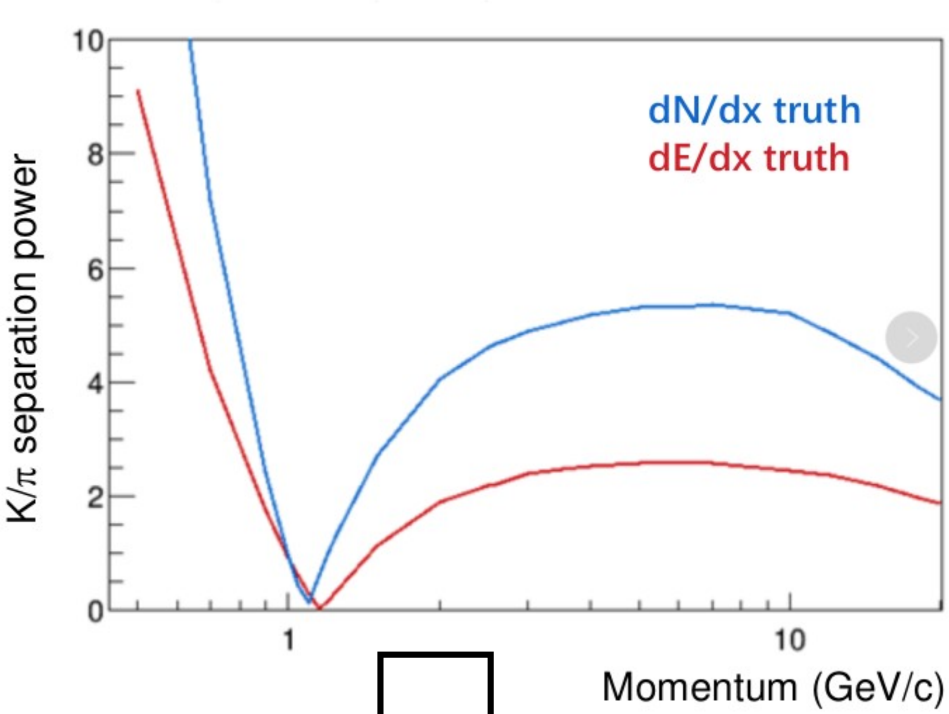
Number of Primary Clusters for Pion

Momentum (GeV)	Number of Primary Clusters			
	MC		CNN	
	Mean	Std Dev	Mean	Std Dev
2	15.15	3.951	14.97	3.442
4	16.31	4.036	16.01	3.504
6	16.92	4.095	16.89	3.517
8	17.45	4.174	17.23	3.481
10	17.69	4.177	17.37	3.489

- **The above Table shows the mean number of primary clusters (MC) and CNN reconstruction results for Pion and will be updated soon on the behalf of an LSTM model too**

PID performances for MC and NN Reconstruction



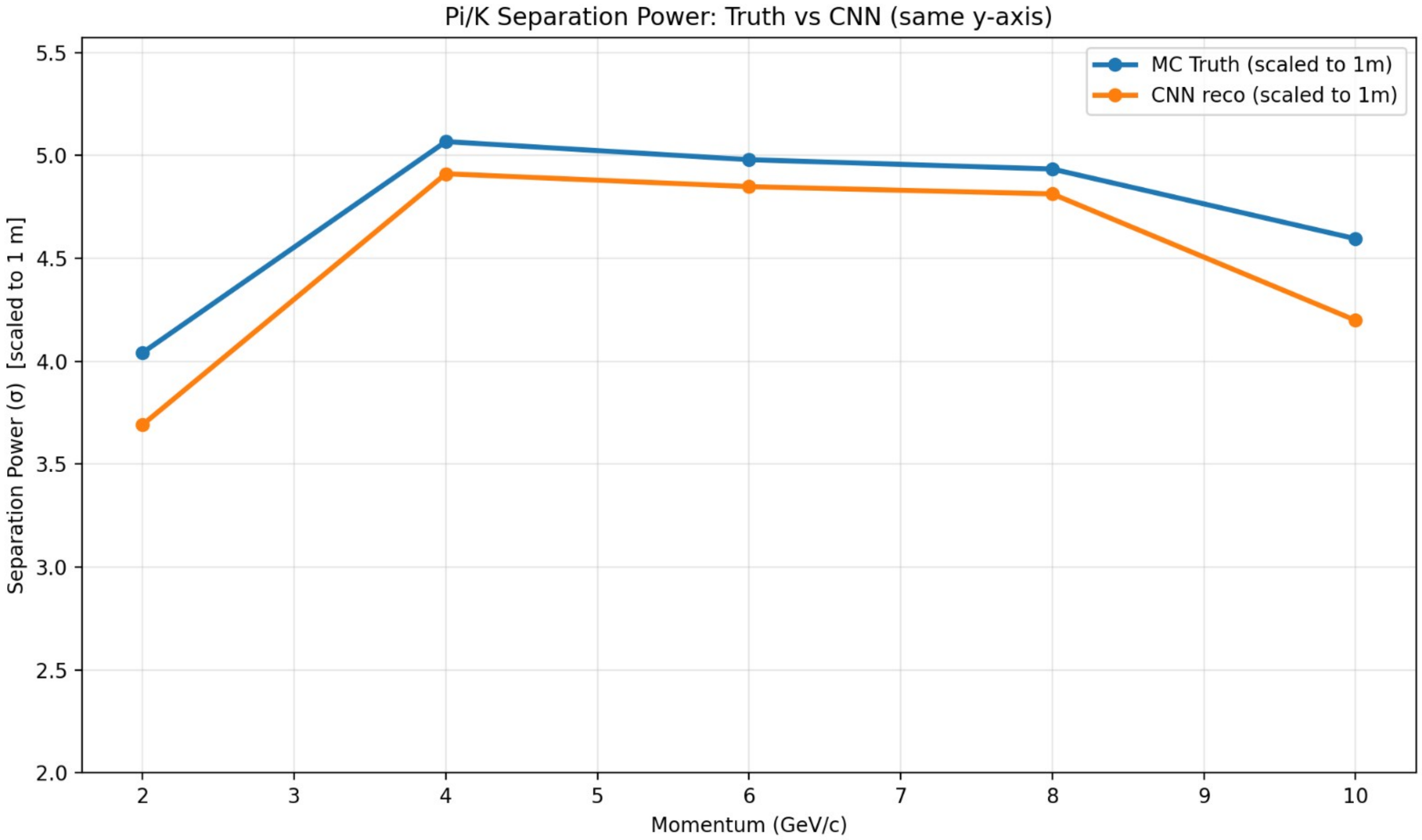


$$S = \frac{|\mu_\mu - \mu_K|}{(\sigma_\mu + \sigma_K)/2}$$

- **At best selection cut 0.55. it looks reasonable. At this selection cut the reconstruction number of primary clusters of CNN more close to MC**

- **The above plot is seperation power for kaon and pion done by Guang and his team (IHEP) from 2-10 GeV momentum too**

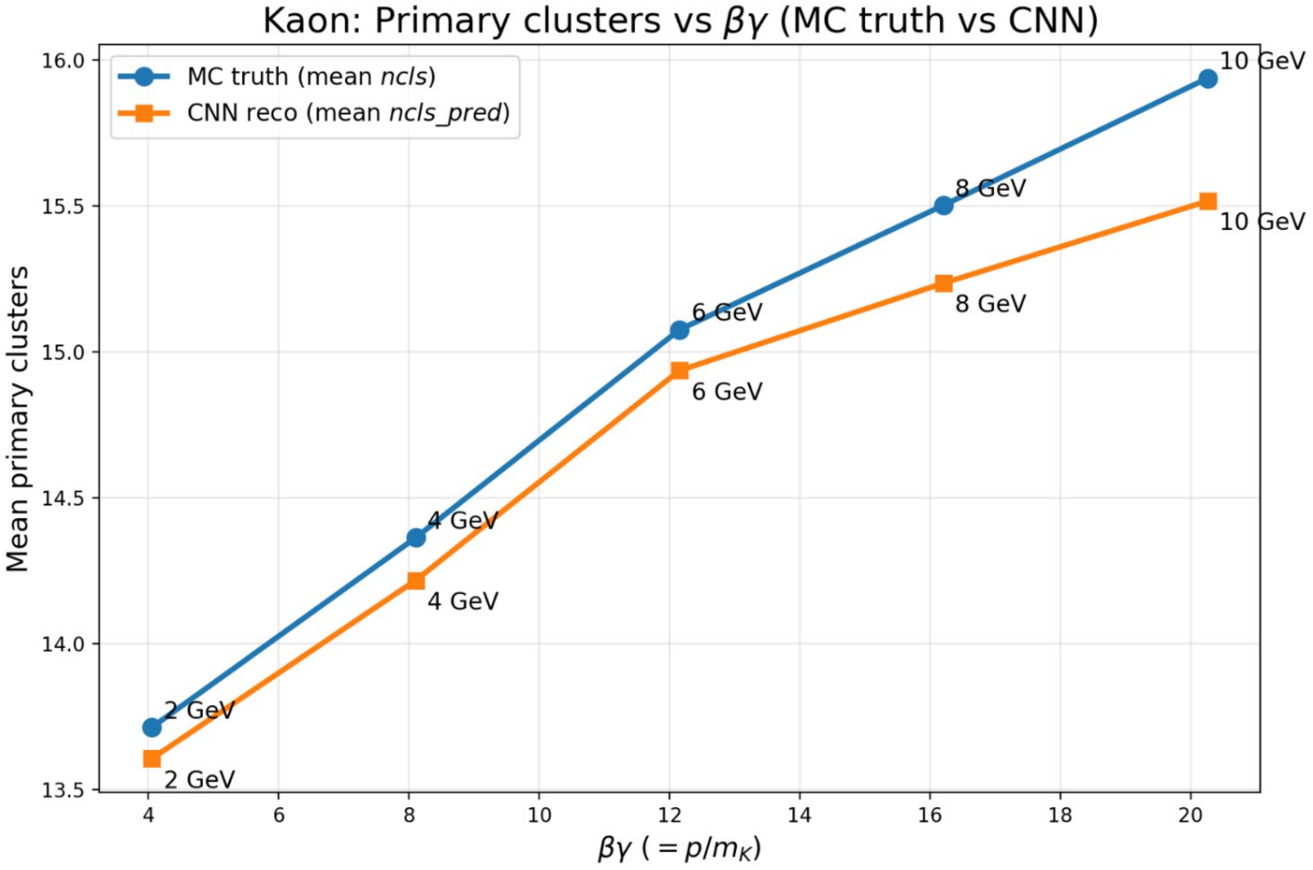
PID performances for MC and NN Reconstruction

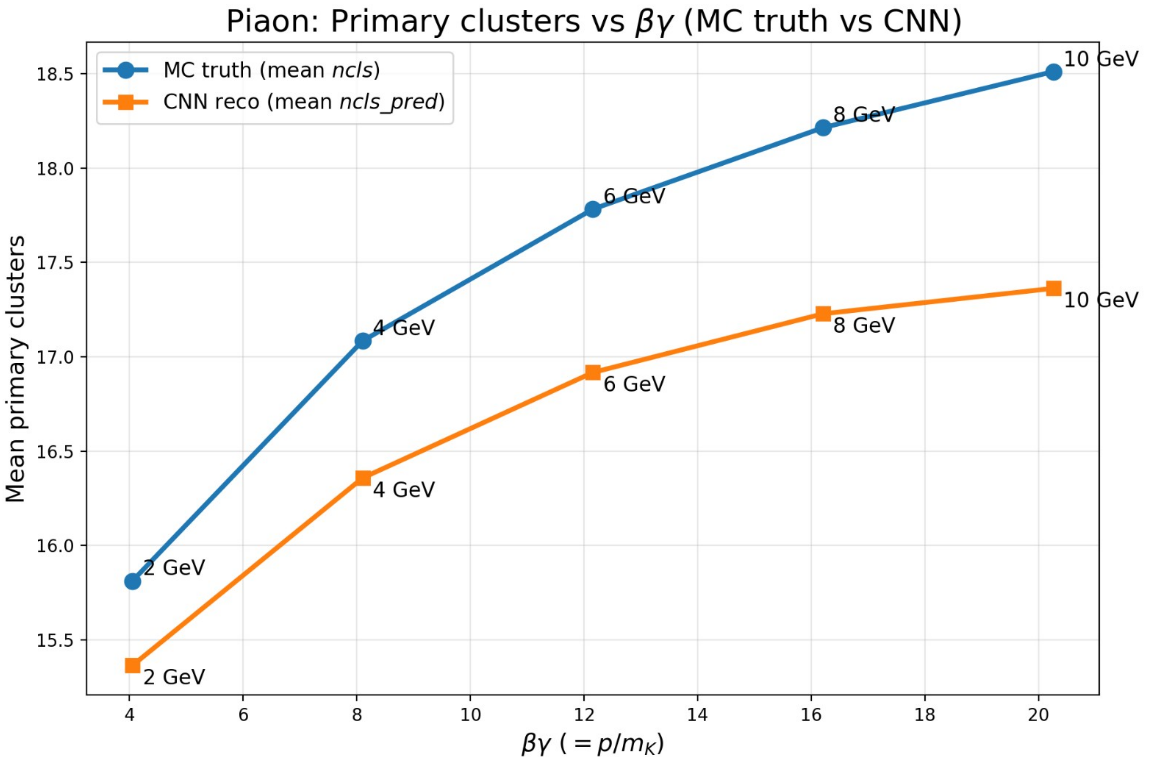


The above plot show seperation power of pion and kaon

Thank you

Some Plots(p/m)

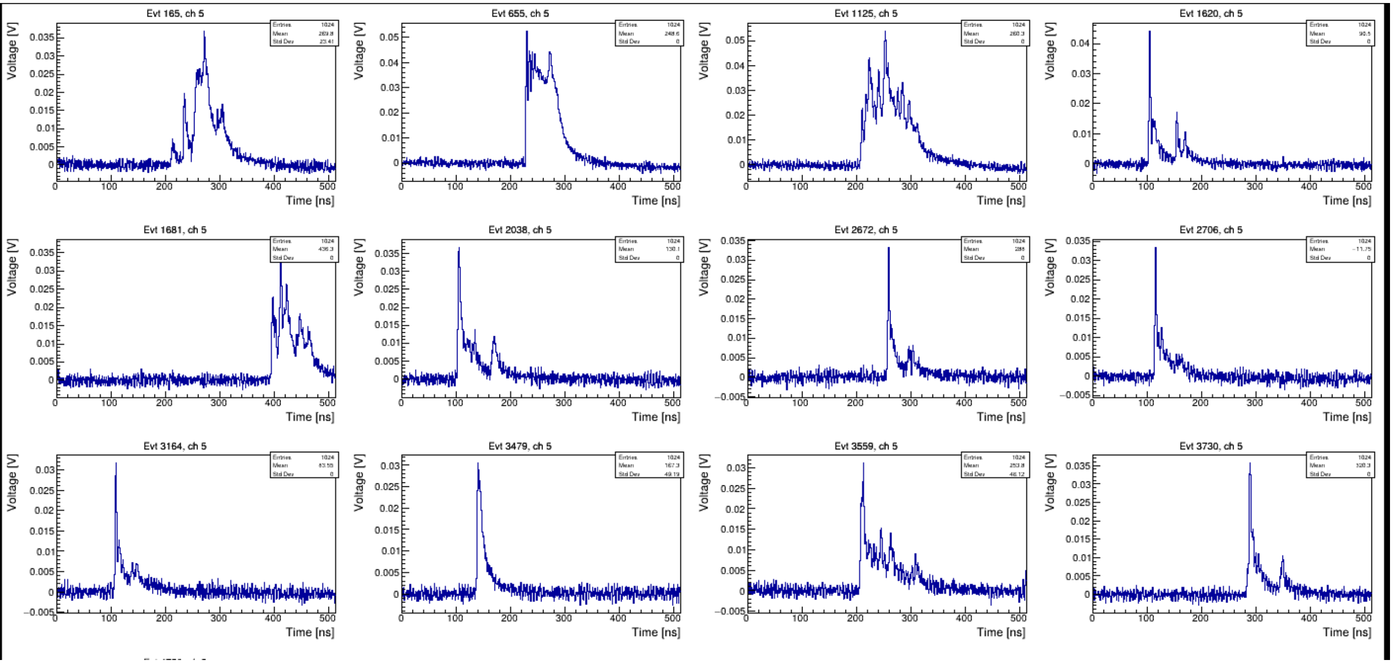




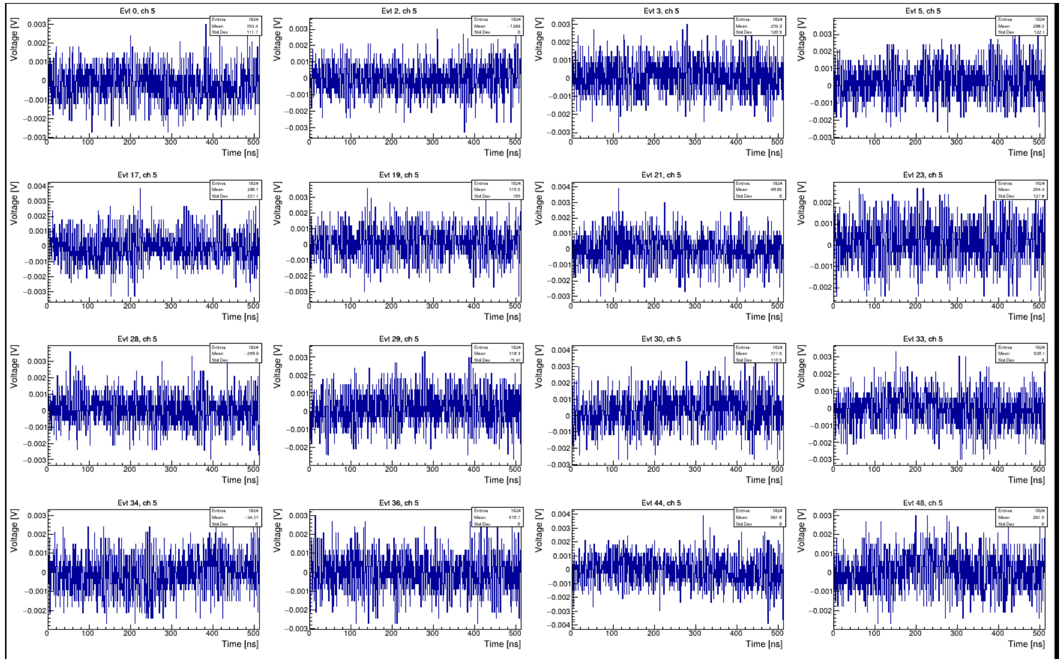
- The above plots are related to mean number of primary clusters vs βγ for MC truth and CNN for kaon and pion

Beam test data selection for Channel 5

Signal-like data

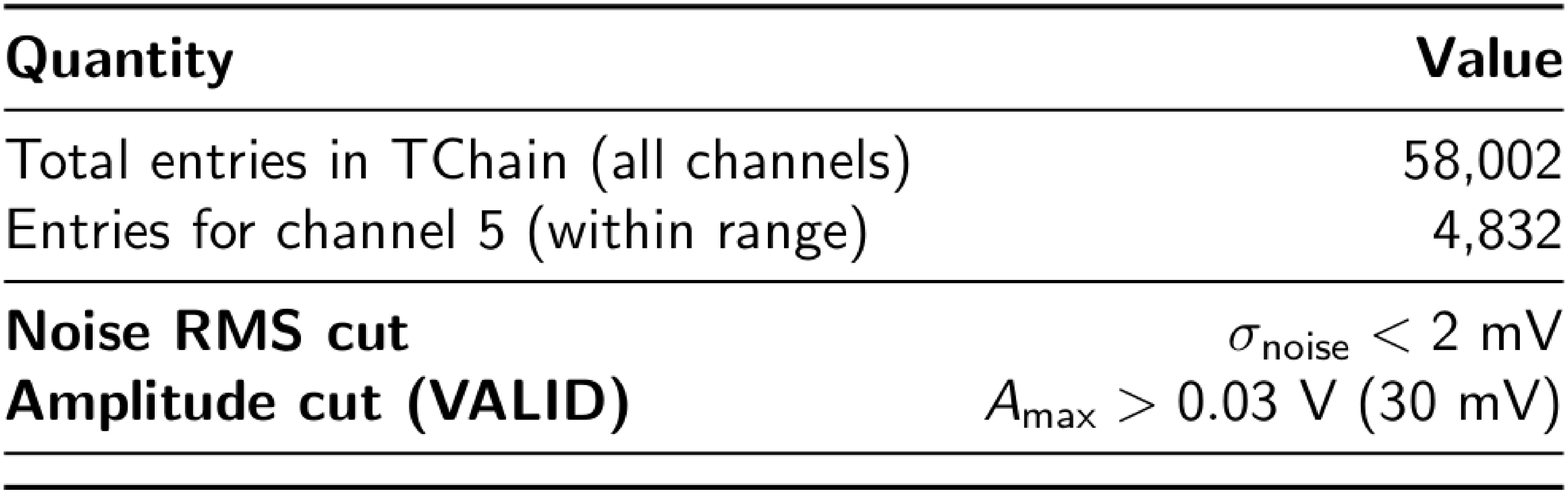


Noise-like data



- Signal-like data: peak finding
- Noise-like data: noise information

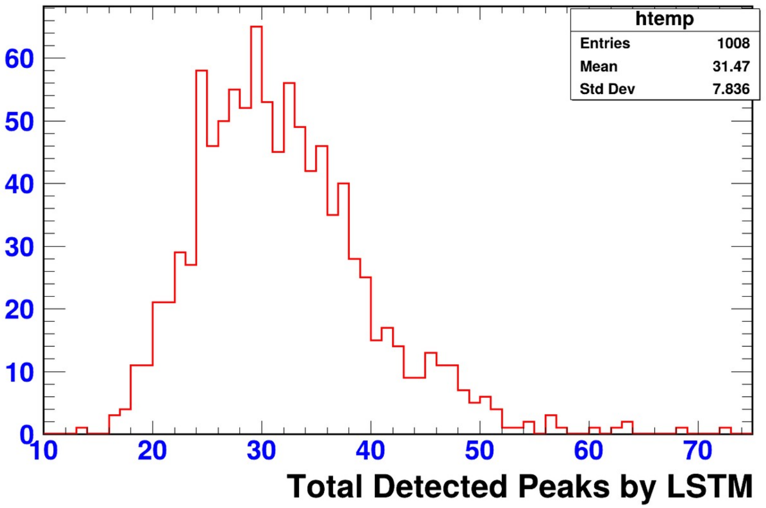
Waveform Selection Logic



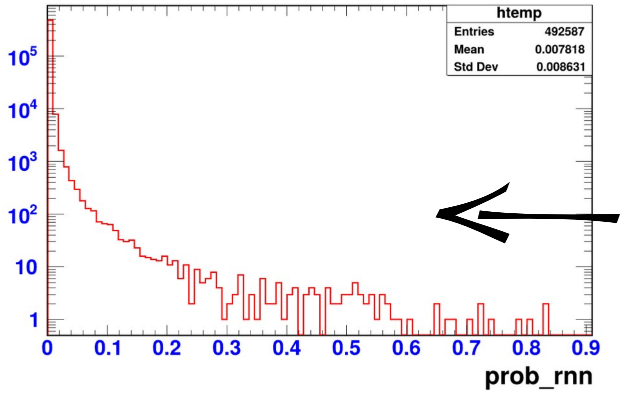
Quantity	**Value**
Total entries in TChain (all channels)	58,002
Entries for channel 5 (within range)	4,832
Noise RMS cut	$\sigma_{\text{noise}} < 2$ mV
Amplitude cut (VALID)	$A_{\text{max}} > 0.03$ V (30 mV)

Applying Untrained NN on beam test data 2022 (180 GeV) Just for Checking

Preliminary Results of Peak Finding distribution

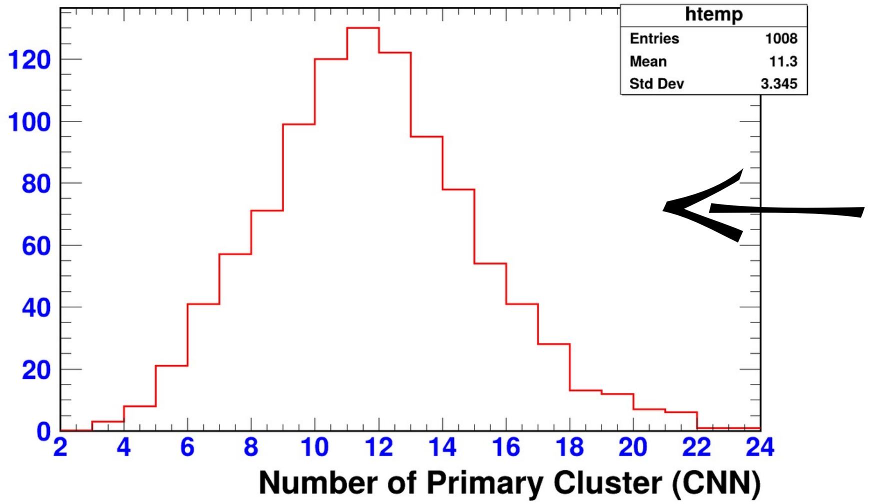


- The above distribution shows us total number of detected peaks (primary and secondary electrons) the

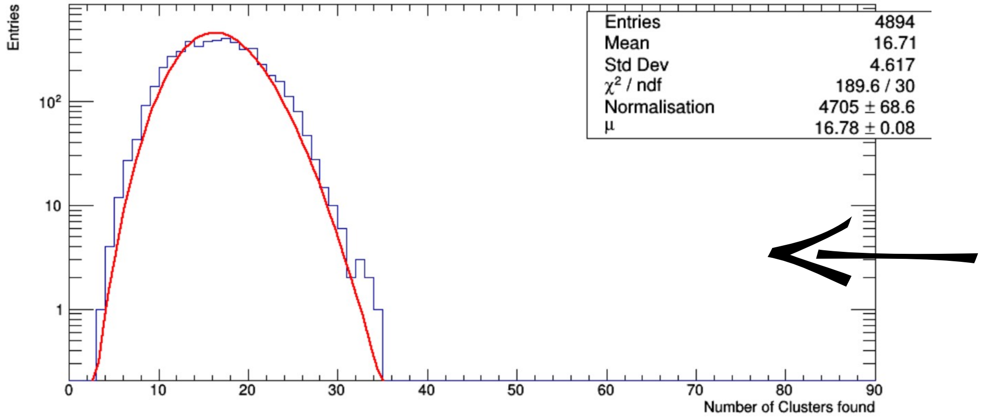


Bad Discrimination of signal and background by using trained LSTM model on untuned simulated waveforms. Because we did not see two main peaks one close to zero represent mostly background events and other close to one represent mostly signal events

Clusterization



Preliminary CNN Clusterization



RTA Algorithm

1. Peak Detection Based on Probability Cut:
The detection process involves looping over all events and applying a probability cut to decide whether a peak is considered a valid detection:

Looping Over Events: The script iterates over all entries (events) in the probability file.

Applying the Cut: For each event, it checks if the predicted probability (prob_ml) exceeds the cut threshold (cut), which is set to 0.95/0.65.

Storing Detected Peaks: If the probability exceeds the threshold, the corresponding peak time is stored in the detected_time dictionary, keyed by event number (evtno)

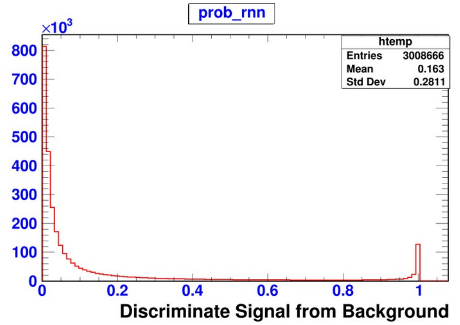
2. Matching Detected Peaks with Truth Data
After detecting peaks, the script matches these peaks with the Monte Carlo (MC) truth data to classify them as primary or secondary:

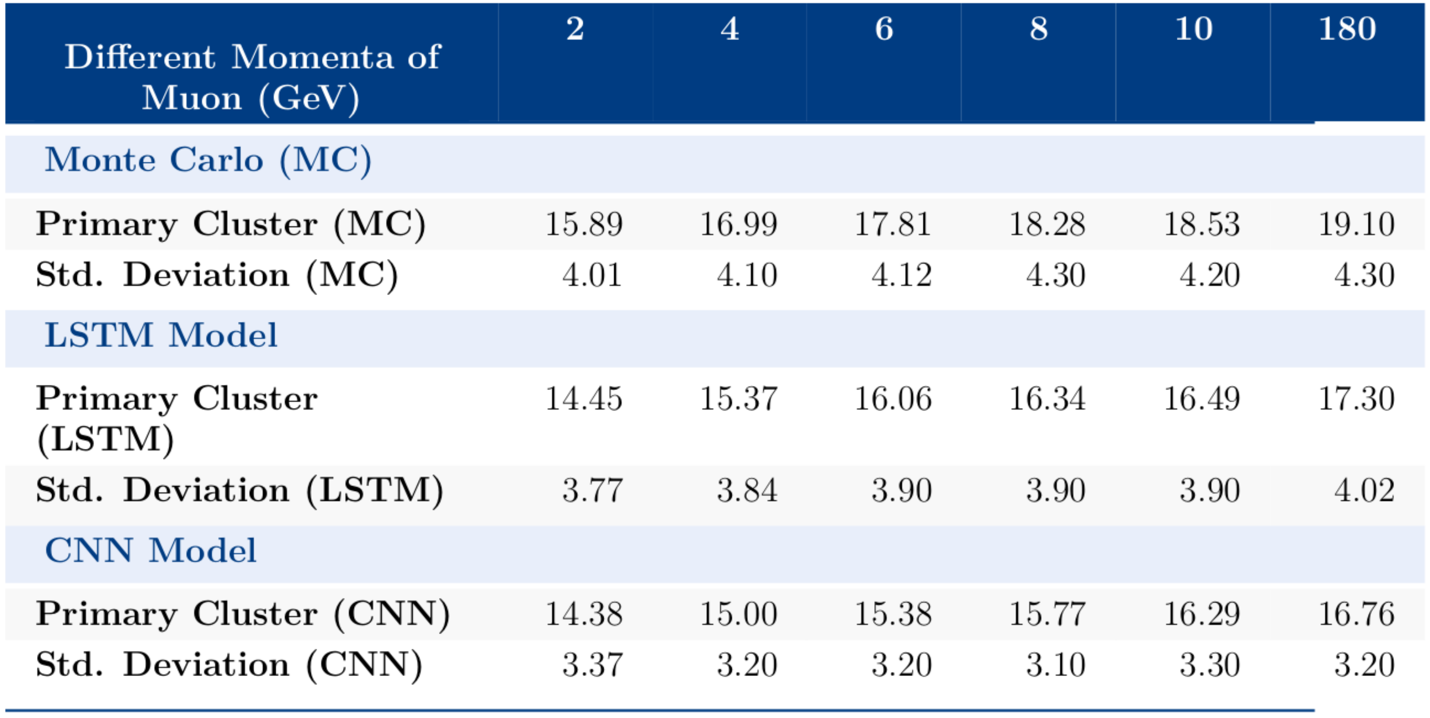
Truth Data: The truth data (truth_time, truth_tag) contains the actual times of primary and secondary peaks, labeled by truth_tag (1 for primary, 2 for secondary).

Matching Function: The match function compares detected peak times with the truth peak times. For each detected peak, it finds the closest truth peak and assigns the corresponding tag (primary or secondary) based on the truth data.

ID Assignment: The id_list array stores the classification of each detected peak as primary (1) or secondary (2)

3. Counting Primary Peaks: After classifying the detected peaks, the script counts how many of them are primary peaks:
Counting: The script iterates over the id_list and increments ncount_pri for every primary peak (tag 1)





Different Momenta of Muon (GeV)	2	4	6	8	10	180
Monte Carlo (MC)						
Primary Cluster (MC)	15.89	16.99	17.81	18.28	18.53	19.10
Std. Deviation (MC)	4.01	4.10	4.12	4.30	4.20	4.30
LSTM Model						
Primary Cluster (LSTM)	14.45	15.37	16.06	16.34	16.49	17.30
Std. Deviation (LSTM)	3.77	3.84	3.90	3.90	3.90	4.02
CNN Model						
Primary Cluster (CNN)	14.38	15.00	15.38	15.77	16.29	16.76
Std. Deviation (CNN)	3.37	3.20	3.20	3.10	3.30	3.20

Table 1: Primary cluster means and standard deviations from MC, LSTM, and CNN across different muon momenta.

1. Peak Detection Based on Probability Cut:
The detection process involves looping over all events and applying a probability cut to decide whether a peak is considered a valid detection:

Looping Over Events: The script iterates over all entries (events) in the probability file.

Applying the Cut: For each event, it checks if the predicted probability (prob_ml) exceeds the cut threshold (cut), which is set to 0.95/0.65.

Storing Detected Peaks: If the probability exceeds the threshold, the corresponding peak time is stored in the detected_time dictionary, keyed by event number (evtno)

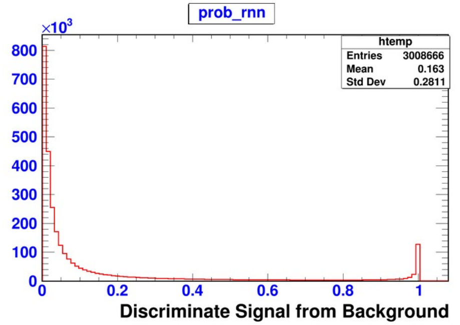
2. Matching Detected Peaks with Truth Data
After detecting peaks, the script matches these peaks with the Monte Carlo (MC) truth data to classify them as primary or secondary:

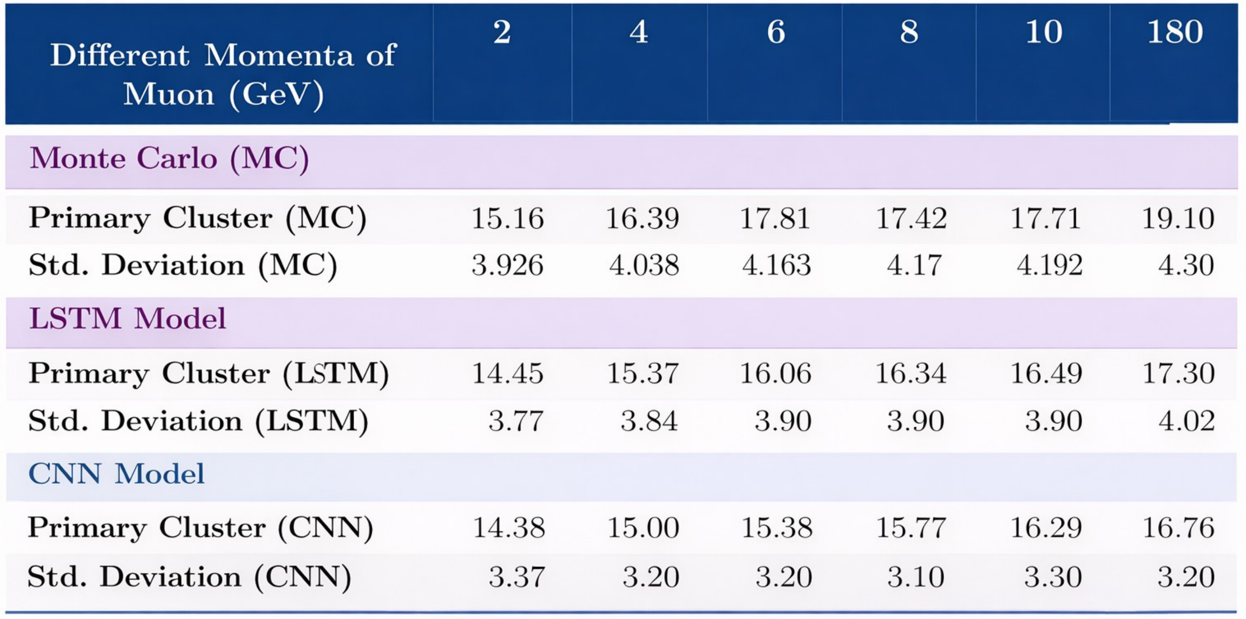
Truth Data: The truth data (truth_time, truth_tag) contains the actual times of primary and secondary peaks, labeled by truth_tag (1 for primary, 2 for secondary).

Matching Function: The match function compares detected peak times with the truth peak times. For each detected peak, it finds the closest truth peak and assigns the corresponding tag (primary or secondary) based on the truth data.

ID Assignment: The id_list array stores the classification of each detected peak as primary (1) or secondary (2)

3. Counting Primary Peaks: After classifying the detected peaks, the script counts how many of them are primary peaks:
Counting: The script iterates over the id_list and increments ncount_pri for every primary peak (tag 1)

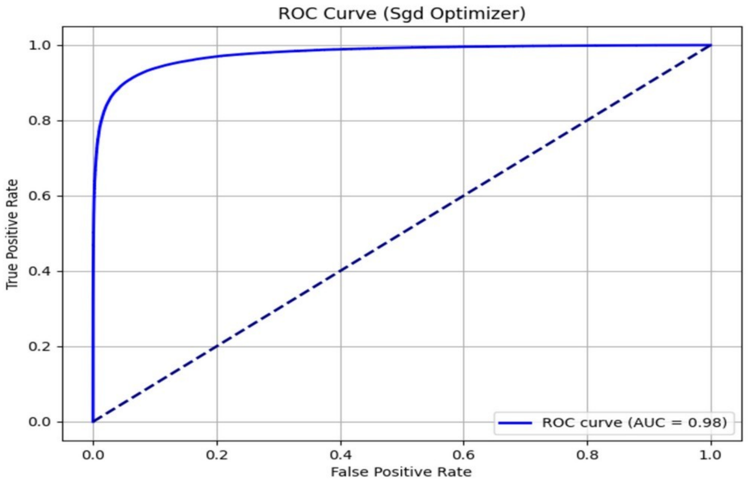


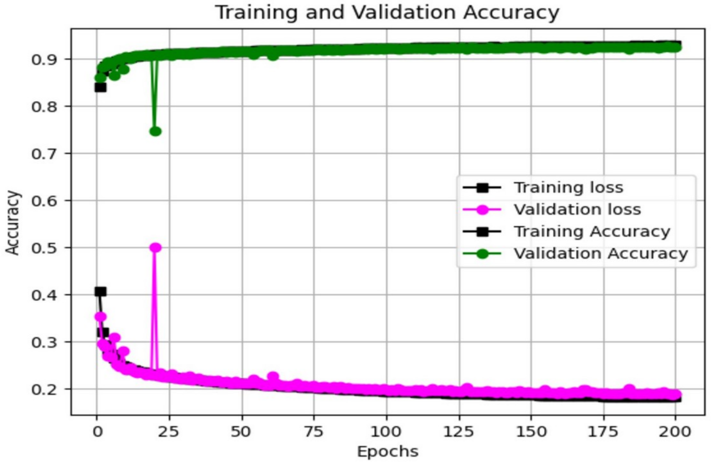


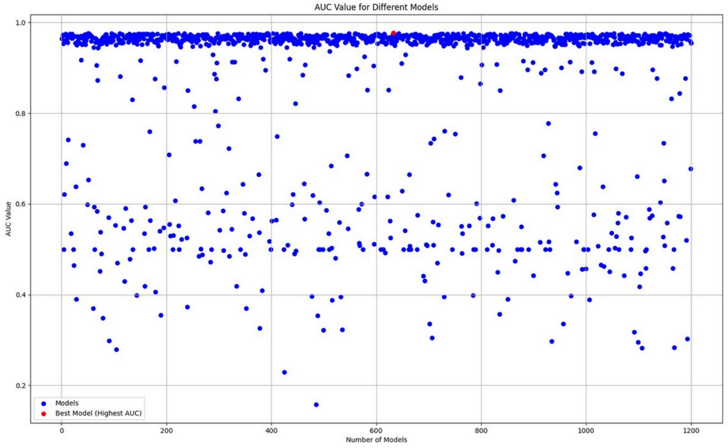
Different Momenta of Muon (GeV)	2	4	6	8	10	180
Monte Carlo (MC)						
Primary Cluster (MC)	15.16	16.39	17.81	17.42	17.71	19.10
Std. Deviation (MC)	3.926	4.038	4.163	4.17	4.192	4.30
LSTM Model						
Primary Cluster (LSTM)	14.45	15.37	16.06	16.34	16.49	17.30
Std. Deviation (LSTM)	3.77	3.84	3.90	3.90	3.90	4.02
CNN Model						
Primary Cluster (CNN)	14.38	15.00	15.38	15.77	16.29	16.76
Std. Deviation (CNN)	3.37	3.20	3.20	3.10	3.30	3.20

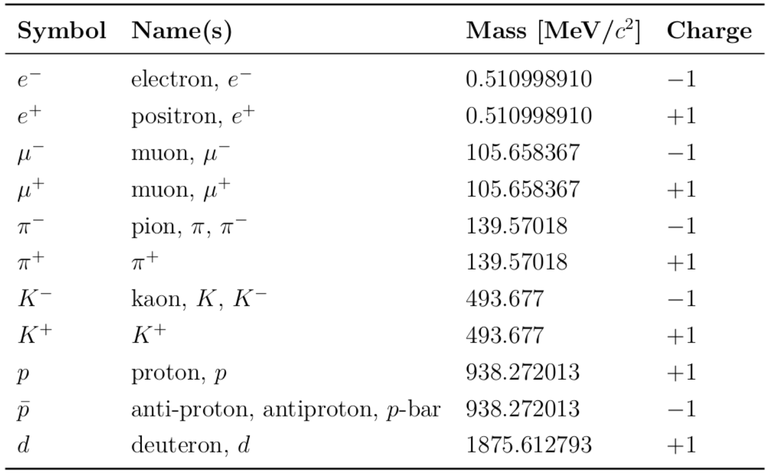
Table 1: Primary cluster means and standard deviations from MC, LSTM, and CNN across different muon momenta.

Best LSTM Peak Finding (Above 1st row) and CNN Regression Model (Below 2nd Row)

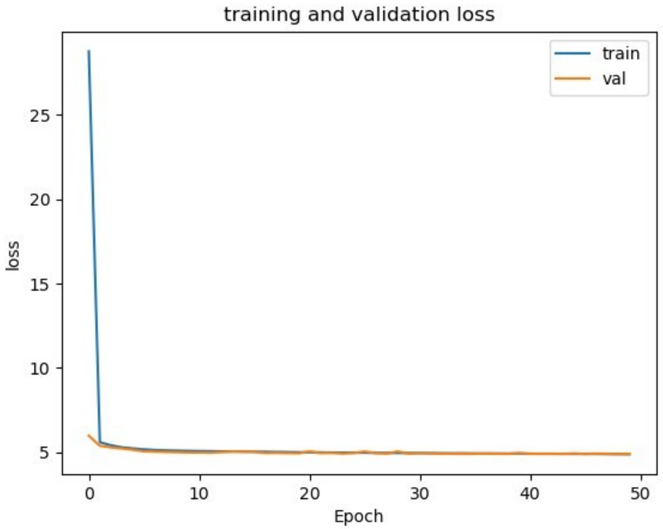


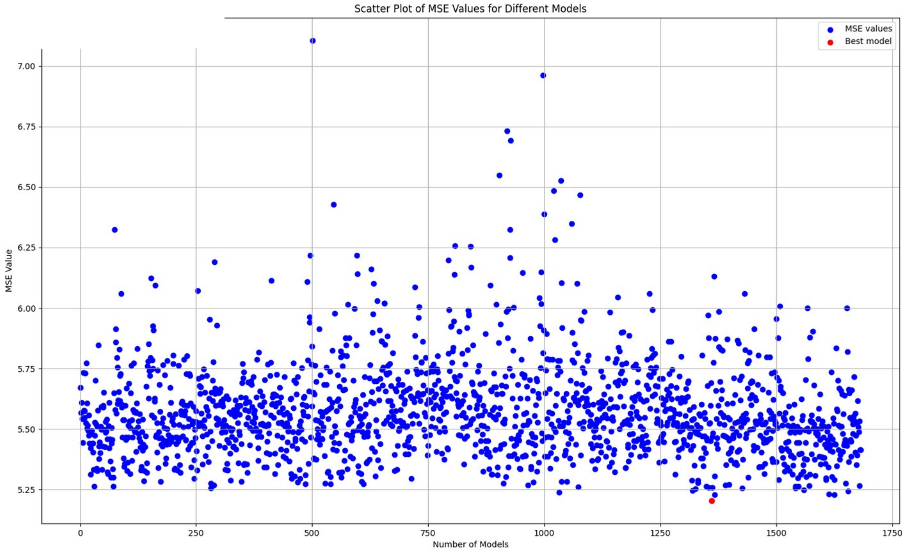




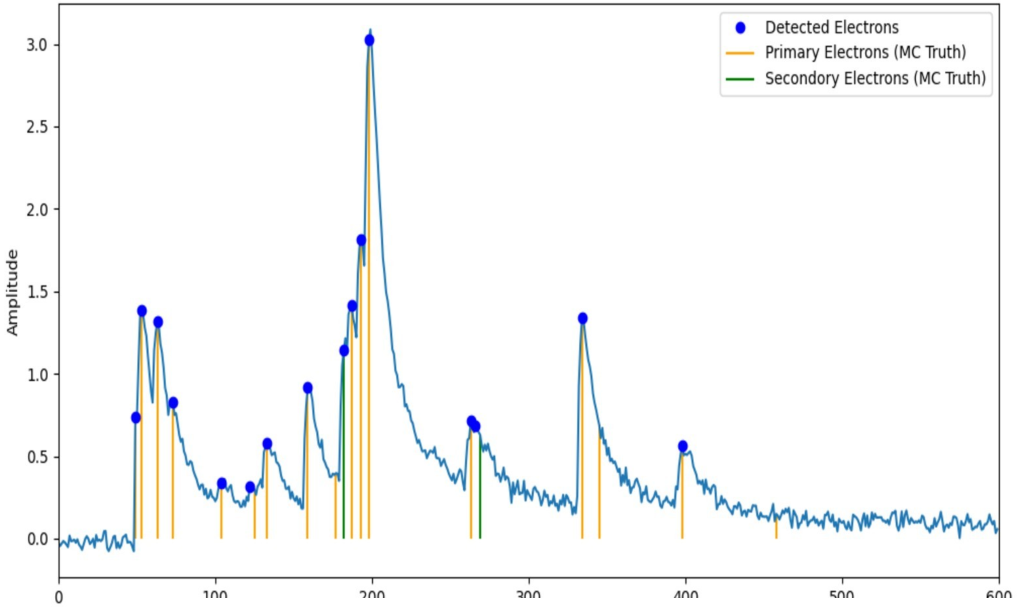


Symbol	Name(s)	Mass [MeV/c^2]	Charge
e^-	electron, e^-	0.510998910	-1
e^+	positron, e^+	0.510998910	$+1$
μ^-	muon, μ^-	105.658367	-1
μ^+	muon, μ^+	105.658367	$+1$
π^-	pion, π, π^-	139.57018	-1
π^+	π^+	139.57018	$+1$
K^-	kaon, K, K^-	493.677	-1
K^+	K^+	493.677	$+1$
p	proton, p	938.272013	$+1$
$\bar{p}$	anti-proton, antiproton, p-bar	938.272013	-1
d	deuteron, d	1875.612793	$+1$





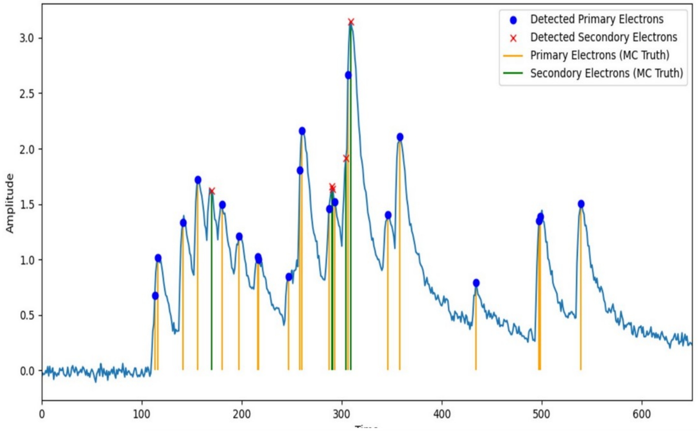
Simulated Peak Finding Waveforms and Data Distribution

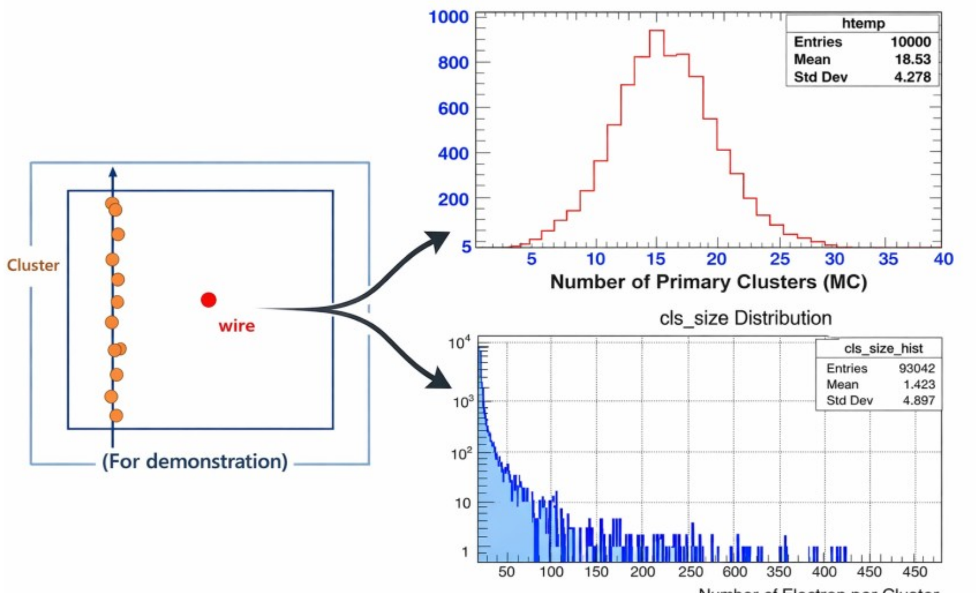


counting peaks

Expected number of electron peaks:

δ clusters/cm (m.i.p.) × 1.3 (rel. rise) × 1.6 electrons/cluster × tube size [cm] × 1/cosα





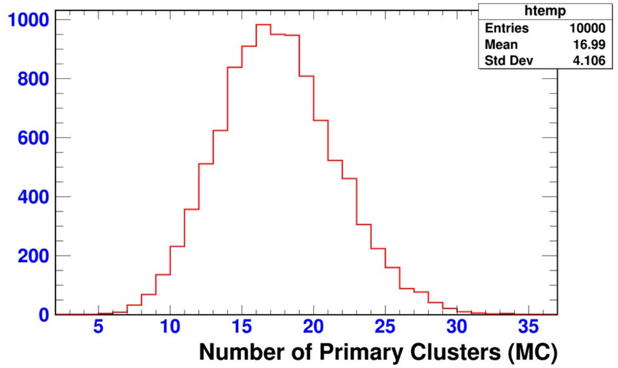
$$N_{\text{primary clusters}} \sim \text{Poisson}(\bar{N}_p)$$

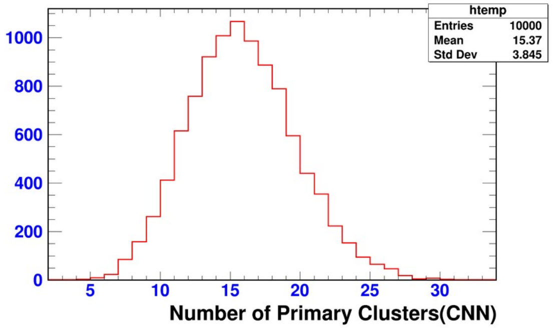
with the mean

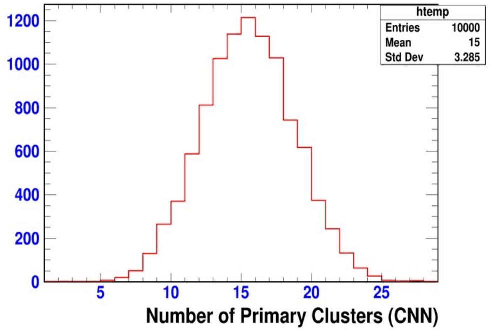
$$\bar{N}_p = \lambda L$$

- λ: average primary-cluster density (clusters per cm)
- L: track length in the gas (cm)

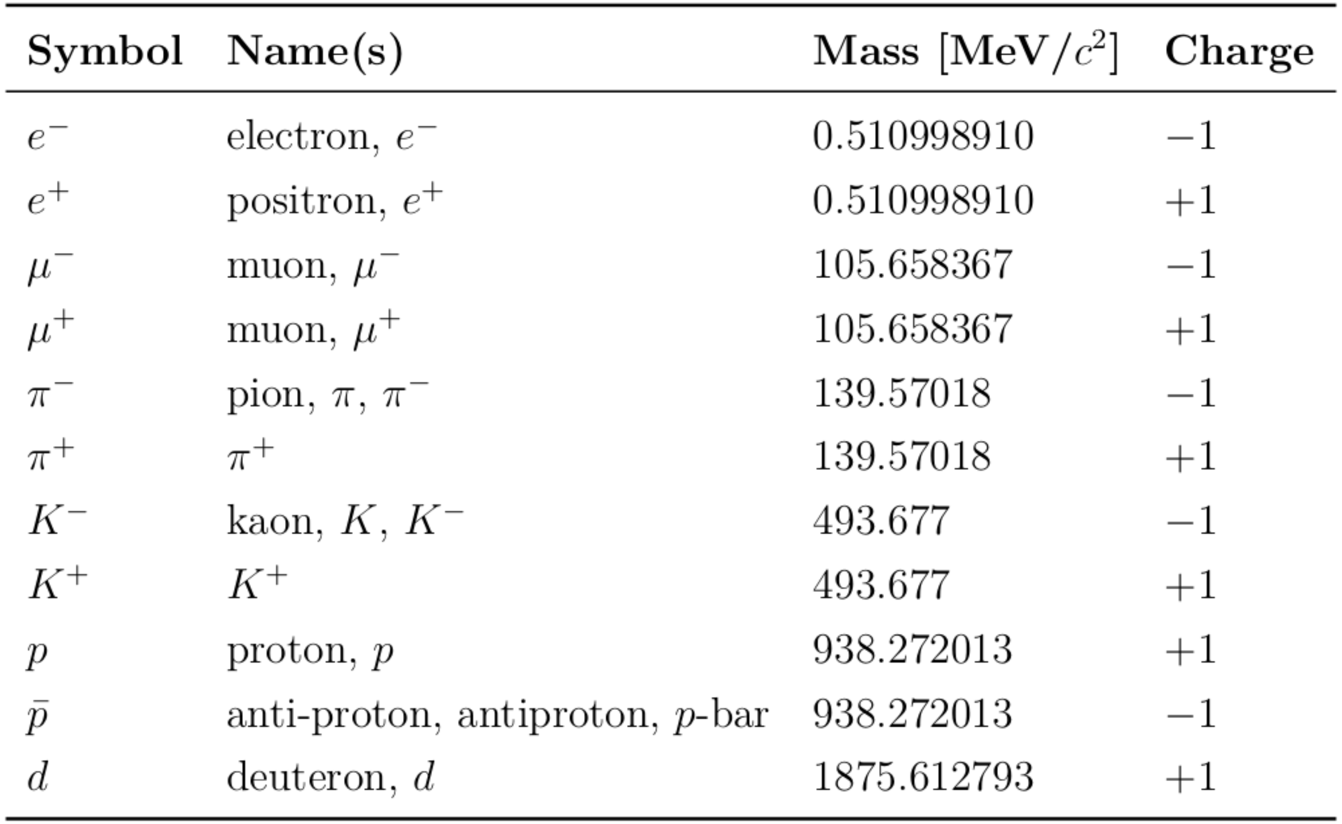
Final results of the reconstruction for 4 GeV on the base of different selection cuts



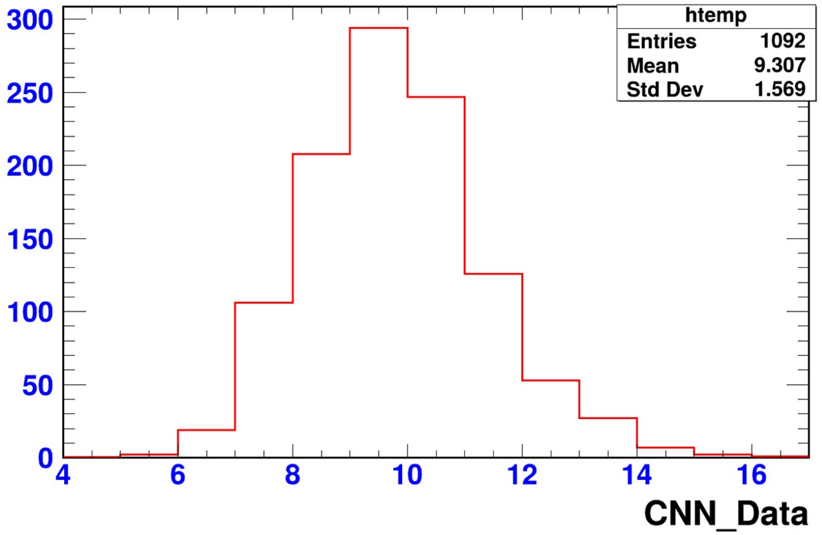


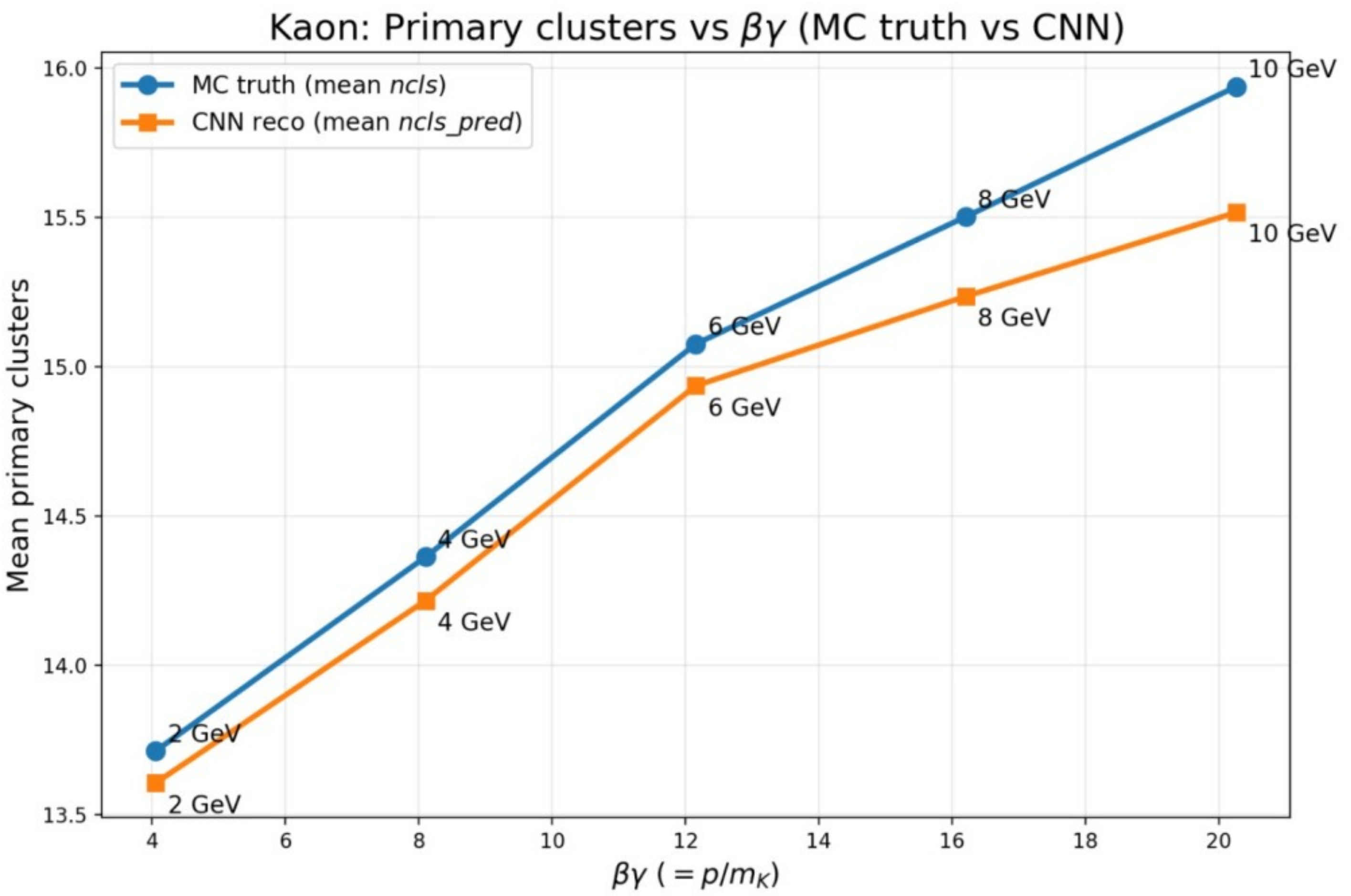


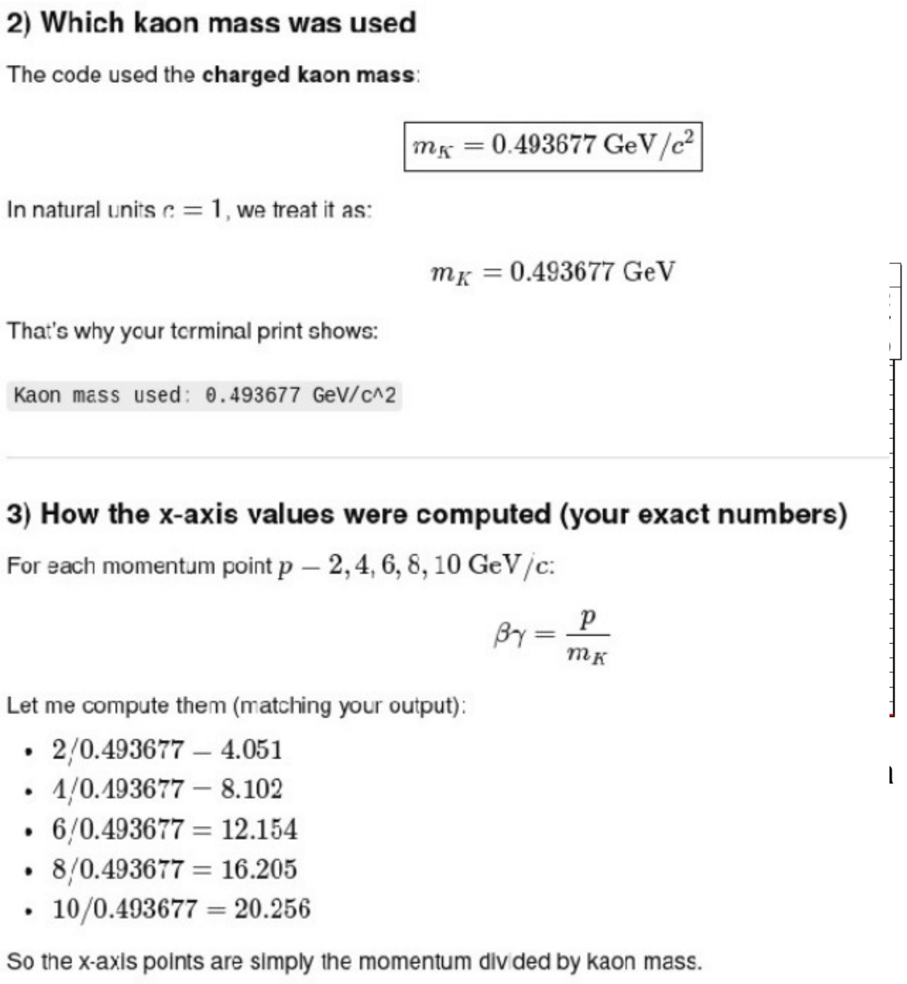
Different selection cuts for 4 GeV	MC	σ of MC	Primary Cluster(LSTM)	σ of (LSTM)	CNN	σ of (CNN)
0.55	16.99	4.1	15.37	3.84	15.	3..2
0.65	16.99	4.1	15.05	3.78	14.9	3.3
0.85	16.99	4.1	14.05	3.5	13.77	3.2
0.95	16.99	4.1	12.74	3.32	12.19	2.99



Symbol	Name(s)	Mass [MeV/c^2]	Charge
e^-	electron, e^-	0.510998910	-1
e^+	positron, e^+	0.510998910	$+1$
μ^-	muon, μ^-	105.658367	-1
μ^+	muon, μ^+	105.658367	$+1$
π^-	pion, π, π^-	139.57018	-1
π^+	π^+	139.57018	$+1$
K^-	kaon, K, K^-	493.677	-1
K^+	K^+	493.677	$+1$
p	proton, p	938.272013	$+1$
$\bar{p}$	anti-proton, antiproton, p-bar	938.272013	-1
d	deuteron, d	1875.612793	$+1$







2) Which kaon mass was used

The code used the **charged kaon mass**:

$$\boxed{m_K = 0.493677\ \mathrm{GeV}/c^2}$$

In natural units $c = 1$, we treat it as:

$$m_K = 0.493677\ \mathrm{GeV}$$

That's why your terminal print shows:

```
Kaon mass used: 0.493677 GeV/c^2
```

3) How the x-axis values were computed (your exact numbers)

For each momentum point $p = 2, 4, 6, 8, 10\ \mathrm{GeV}/c$:

$$\beta\gamma = \frac{p}{m_K}$$

Let me compute them (matching your output):

- $2/0.493677 = 4.051$
- $4/0.493677 = 8.102$
- $6/0.493677 = 12.154$
- $8/0.493677 = 16.205$
- $10/0.493677 = 20.256$

So the x-axis points are simply the momentum divided by kaon mass.