PID study with cluster counting on the drift chamber of CEPC 4th detector

Shuiting Xin on behalf of DC PID working group

2023.6.14

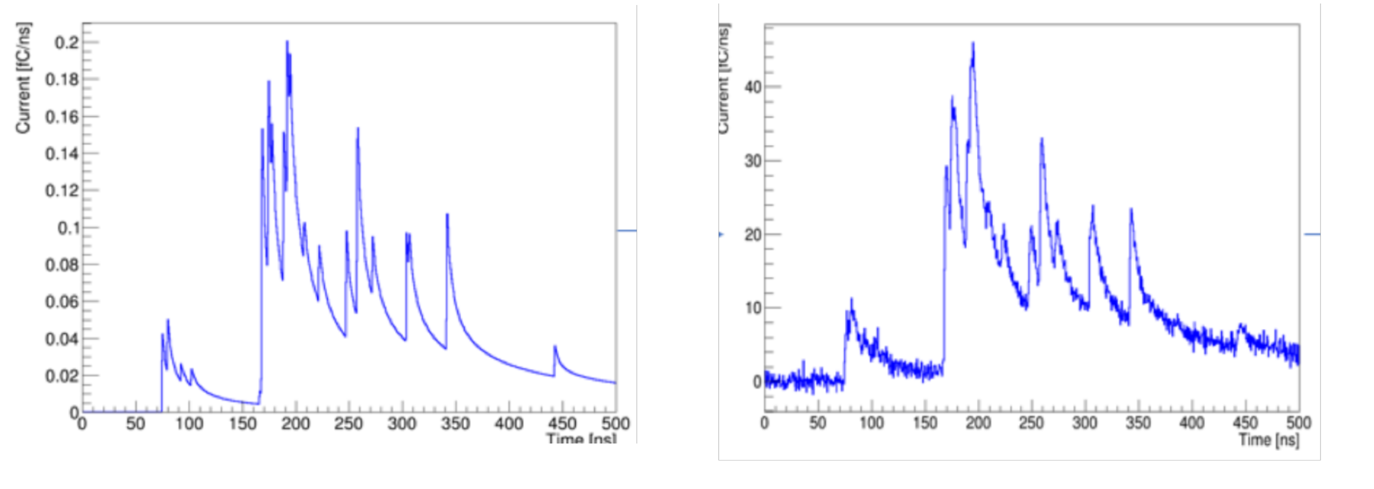
Introduction

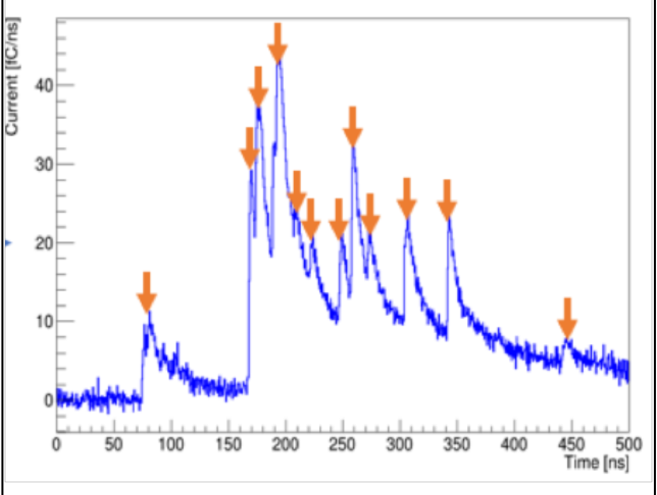
- **Full simulation is the foundation of dN/dx PID study.**
- **Major challenges**
  - Garfield++ simulation of waveform is super time comsuing.
  - Need a more realistic model from the test beam data
- **A full simulation package is developed considering the challenges**
- **Performed PID analysis using full simulation with high statistics.**
- **Updated a new machine learning algorithm for cluster reconstruction.**

Full simulation

From simulation to waveform analysis

The full simulation package





Signal generation

speed up Garfield++ simulation

Digitization

extracted from data

- Noise: FFT analysis
- Pre-amplifier response
- Amplitude scale

Waveform analysis

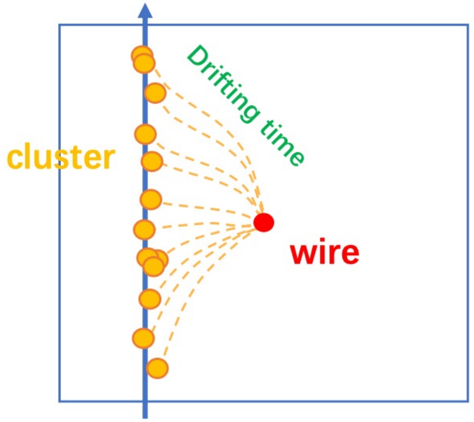
- Drivative method
- Machine learning method

Signal generation: An effective model

Details

Speed up of signal generation: Replacements of the amplification and signal creation of Garfield++ simulation.

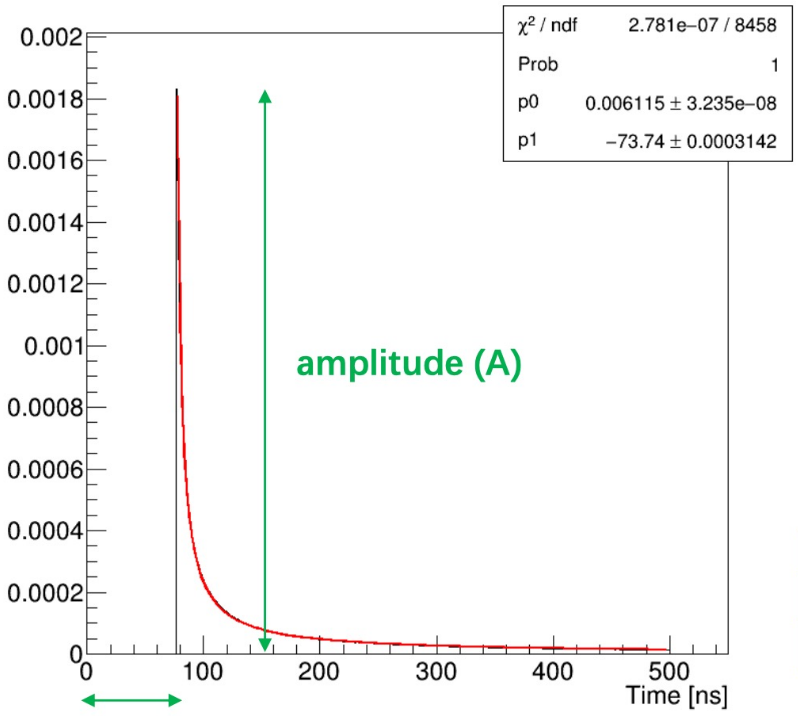
Electrons from ionization:
drift/diffusion ➔ avalanche ➔ induce current



(for demonstration)

Very time consuming in Garfield++
➔ Need parameterization

Single pulse: pulse(A, t)



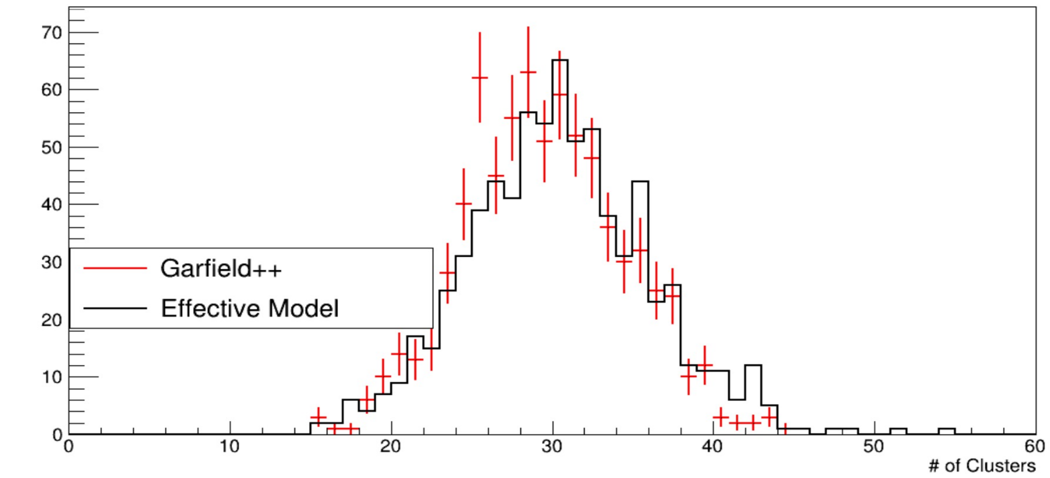
starting time (t)

Parameterization:
- Amplitude
- Starting time
- Pulse shape

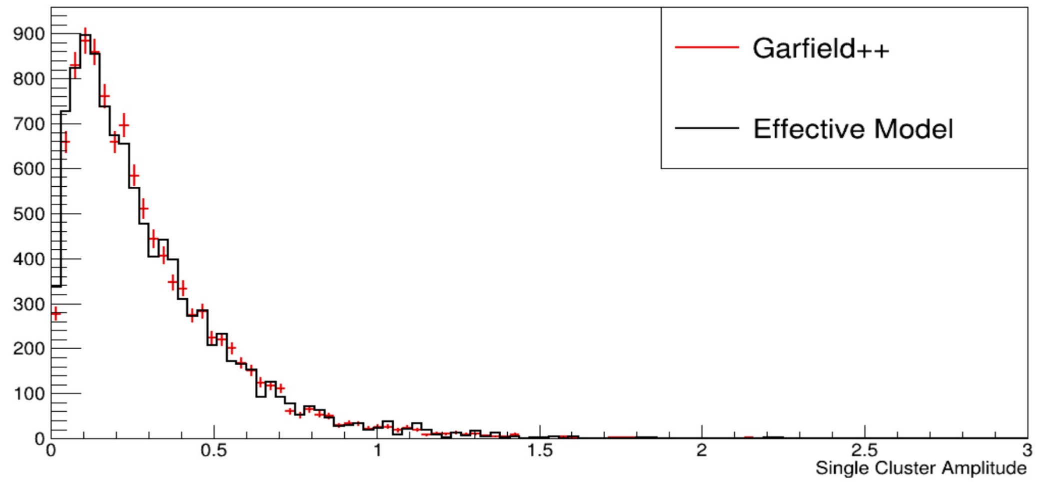
Need to extract information from Garfield++

Signal generation: validation

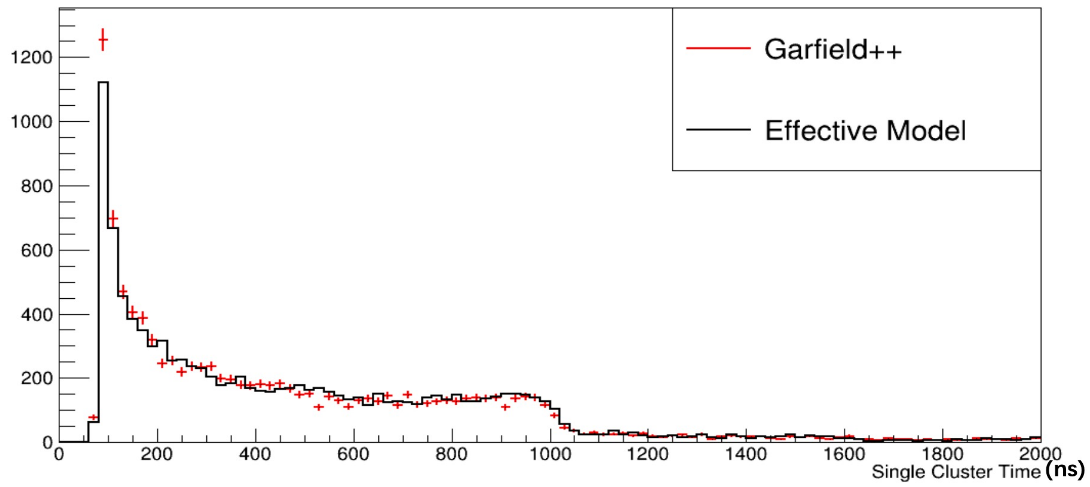
of primary ionizations

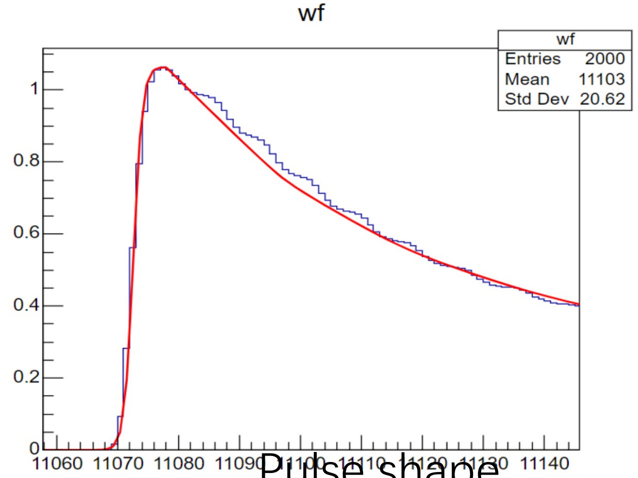


Single-pulse amplitude



Single-pulse time





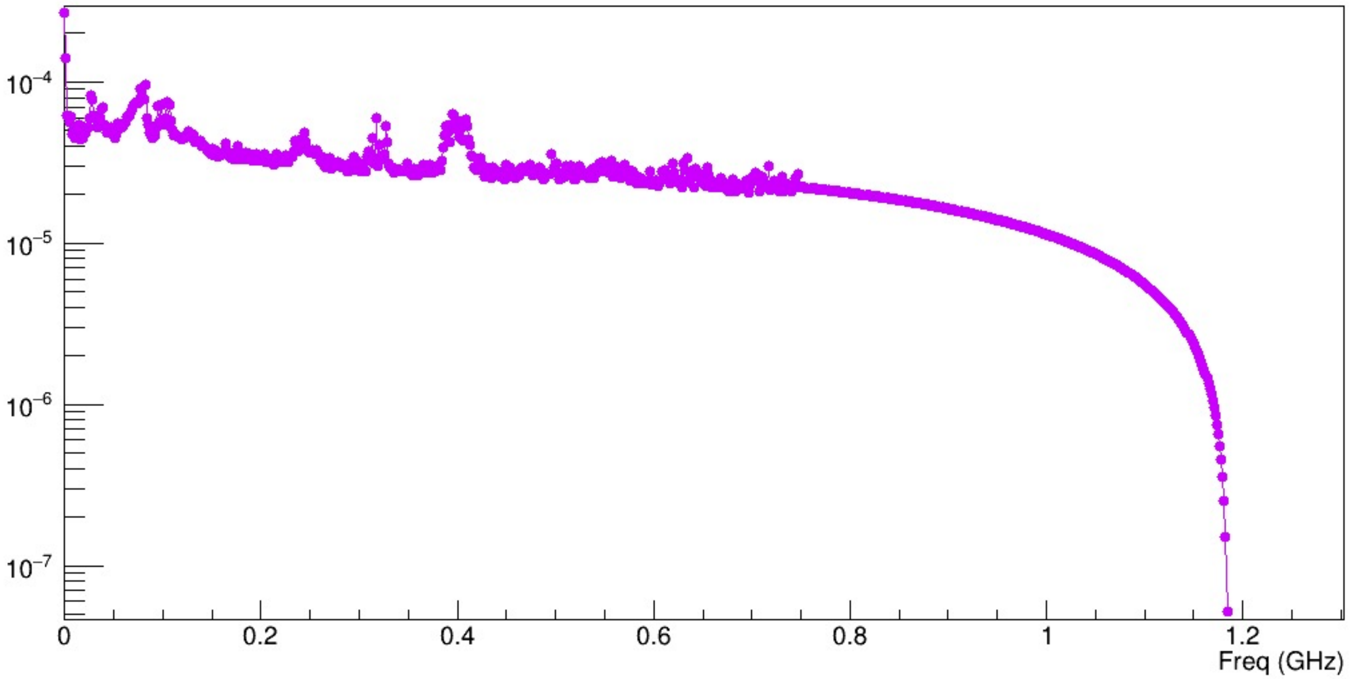
Pulse shape

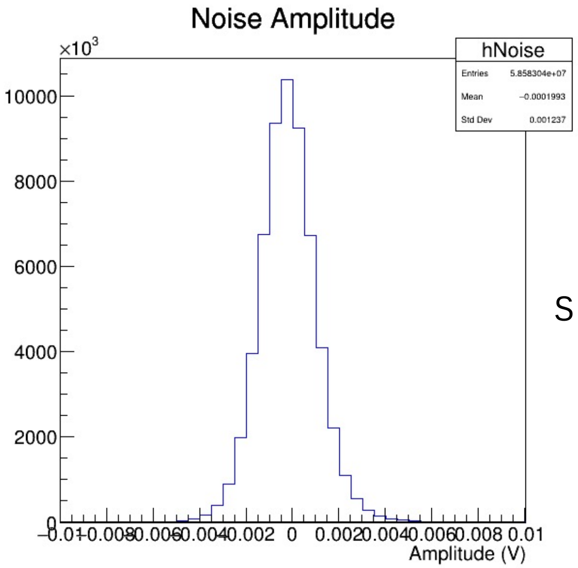
The model is well consistent to the Garfield++ simulation

Digitization: Noise generation

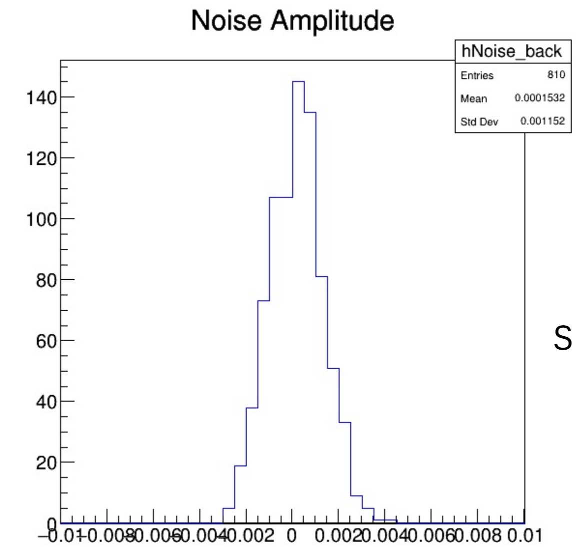
- **Data sample used in MC tuning**
  - Run 15/16/17, DRS channel 5, Gas mixture: 90/10 (He/iC_4H_{10}), Cell size: 1 cm, Sampling rate: 1.5 GHz
- **Generating noise with FFT.**

Frequency response





Data
sigma: 0.00124



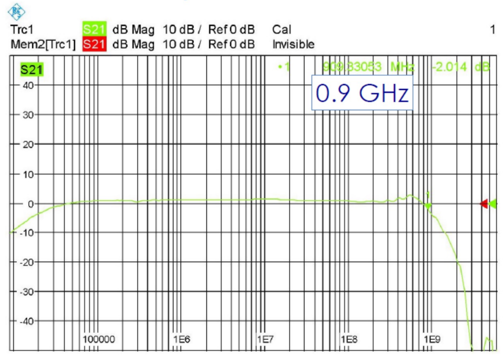
Sim
sigma: 0.00115

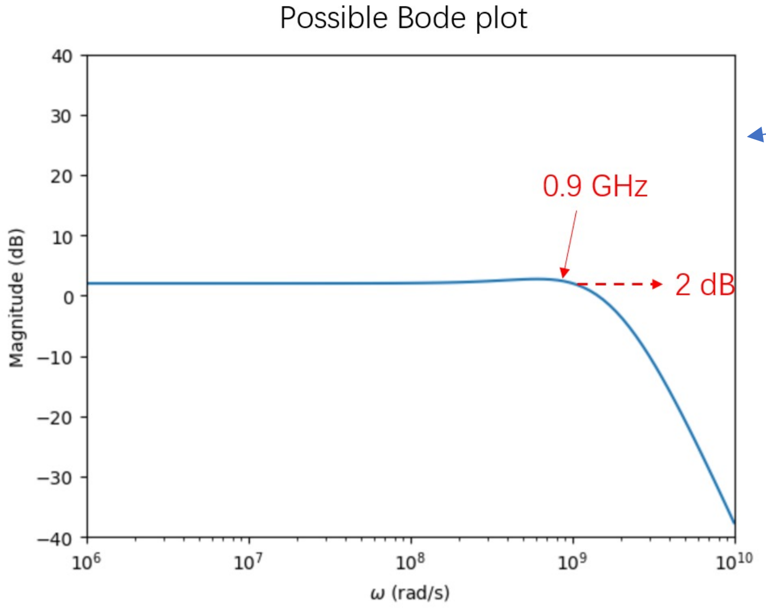
Level of noise ratio in experiment data: 5%.

Digitization: Pre-amplifier response

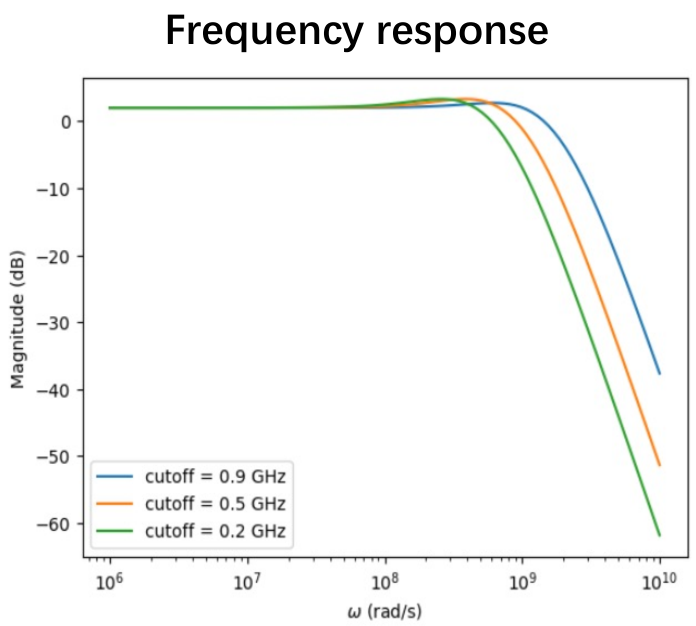
- Parameterized using beam test data inputs (provided by Gianluigi)
- Use cutoff ~0.5 GHz / (risetime ~5-6 ns) in digitization

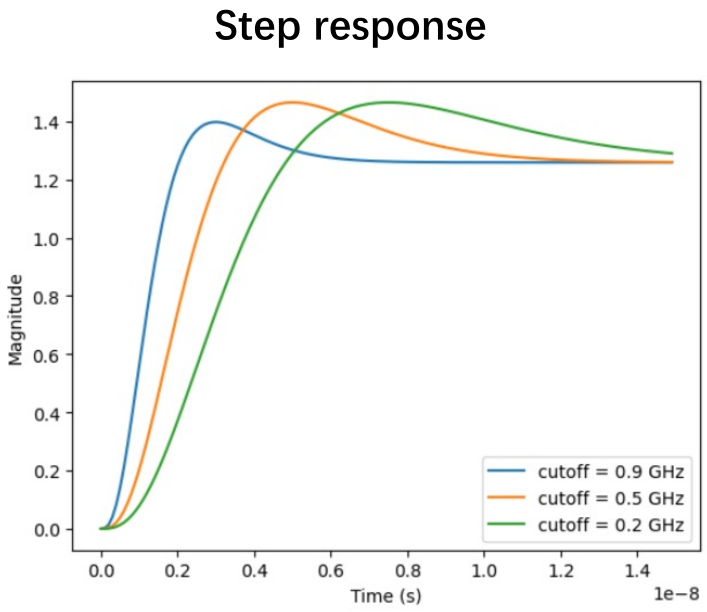
Preamp channel – Gain 1





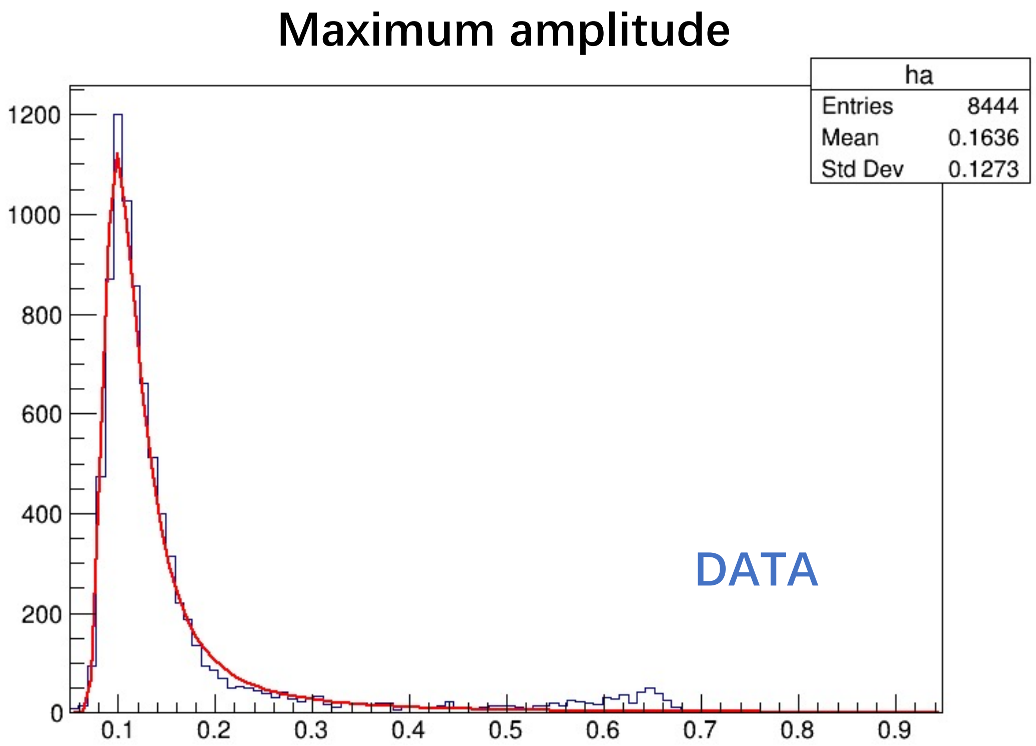
Solution: $$H(s) = \frac{1.4 \times 10^{28} \times (s + 6.0 \times 10^8)}{(s + 1.6 \times 10^9)^4}$$



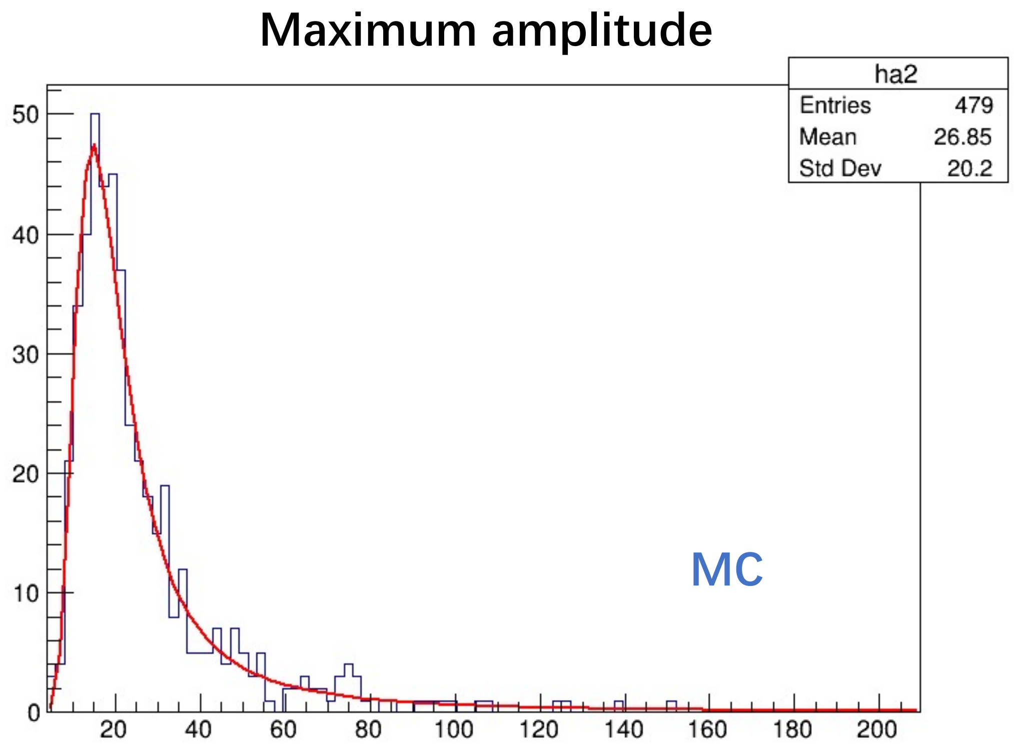


Digitization: Scaling of the amplitude

Scale the MC to data by 0.0065 (max amp)

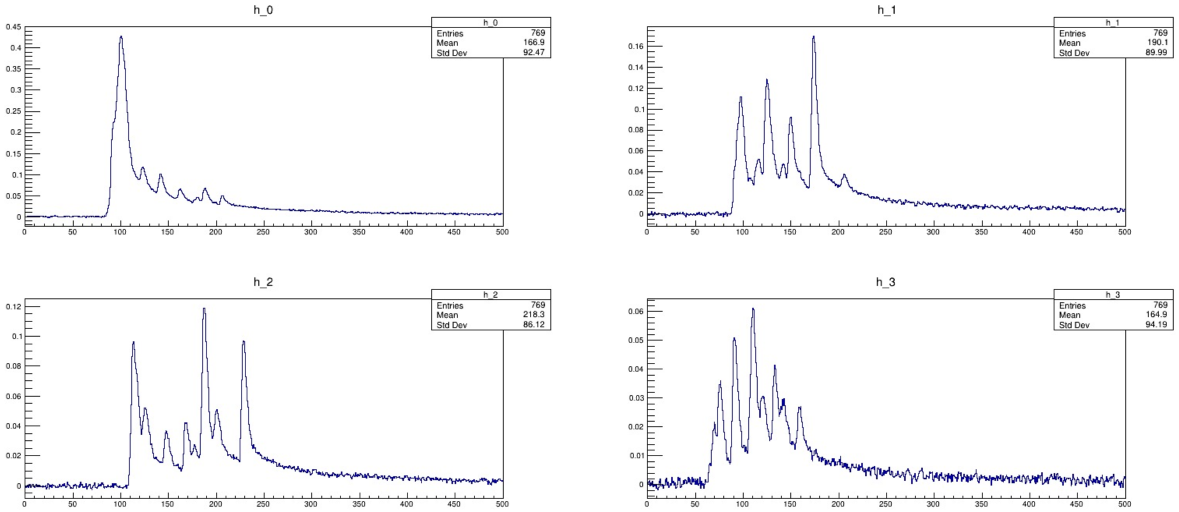


EXT NO.	PARAMETER NAME	VALUE	ERROR	STEP SIZE	FIRST DERIVATIVE
1	Constant	6.20453e+03	1.05194e+02	-1.66577e+00	1.14770e-05
2	MPV	1.01780e-01	2.80470e-04	2.11655e-06	-2.14508e-01
3	Sigma	1.17640e-02	1.53617e-04	1.53108e-05	-2.52383e-01

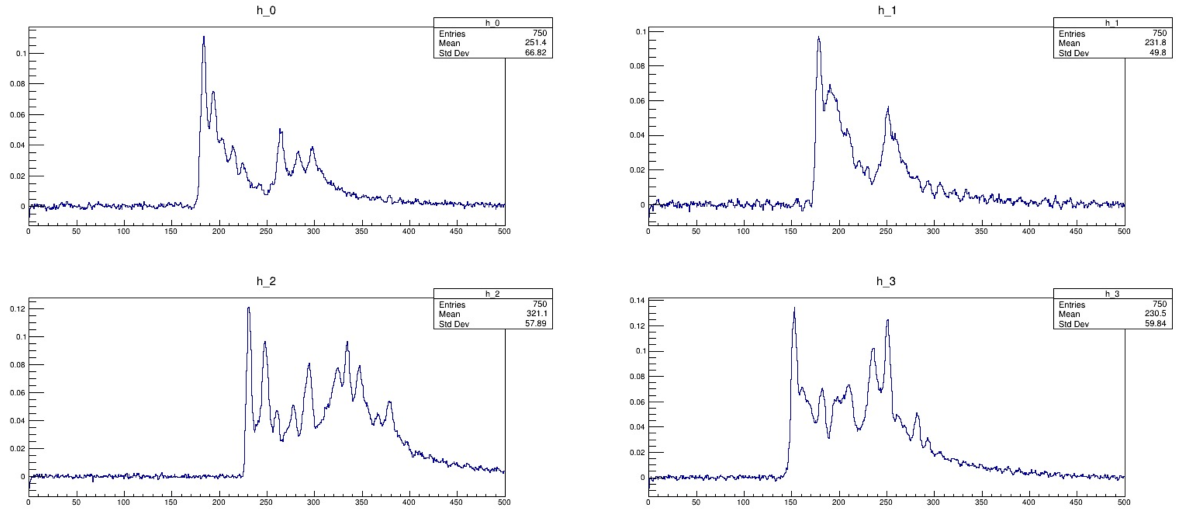


EXT NO.	PARAMETER NAME	VALUE	ERROR	STEP SIZE	FIRST DERIVATIVE
1	Constant	2.63111e+02	1.85036e+01	9.21563e-03	7.21129e-07
2	MPV	1.56630e+01	3.76337e-01	1.71205e-04	1.11828e-04
3	Sigma	3.74788e+00	2.10173e-01	7.09253e-06	-3.55048e-03

Waveform shapes



Simulation



Beam test data

PID analysis

PID analysis

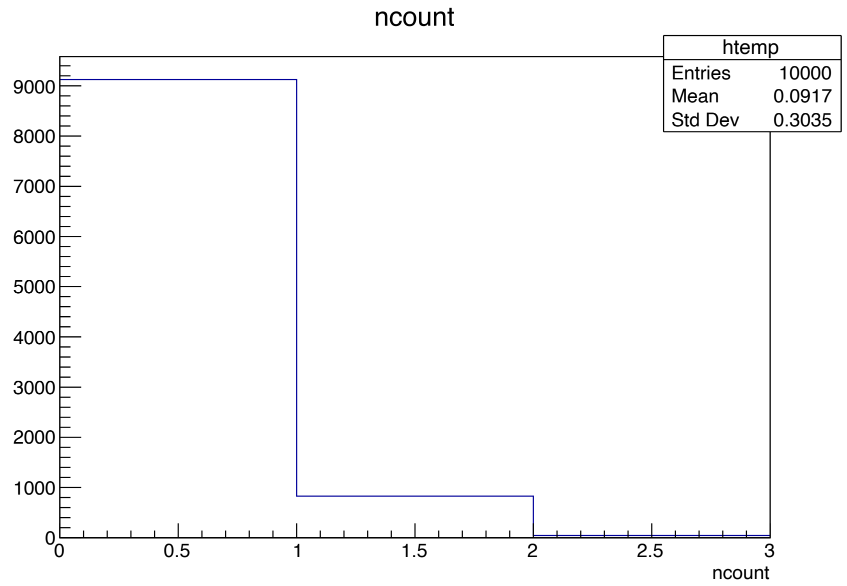
- Evaluating K/pi separation from cluster reconstruction, two steps:
  - Peak finding: find all ionization peaks from waveform.
  - Clusterization: determine the number of clusters (N_{cls}) according to the ionization peaks.

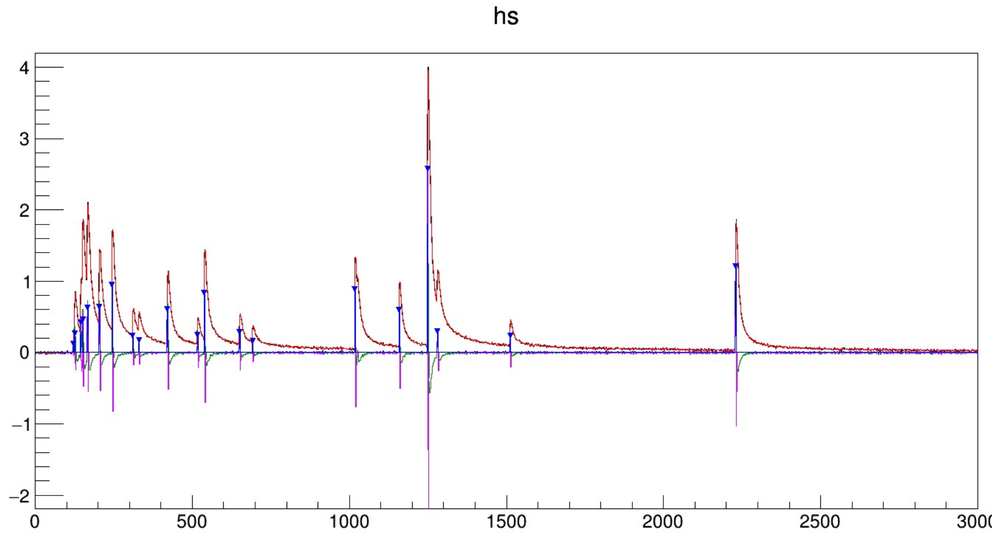
- Simulation data:
  - 20k events of k/pi tracks (previous Garfield++ events: 1k).
  - 2%/5% noise ratio (noise tuned from data).
  - New Pre-amplifier, old current-sensitive pre-amplifier.
  - Cell size: 1.8 cm.
  - Gas mixture: 90% He 10% iC_4H_{10}.

Peak finding

- **Derivative algorithm is used.**
- **Adjusted parameters to minimize the fake rates**
  - Performed peak finding on pure noise samples
- **Parameters**
  - Moving average = 1
  - Threshold = 0.04

Pure noise sample

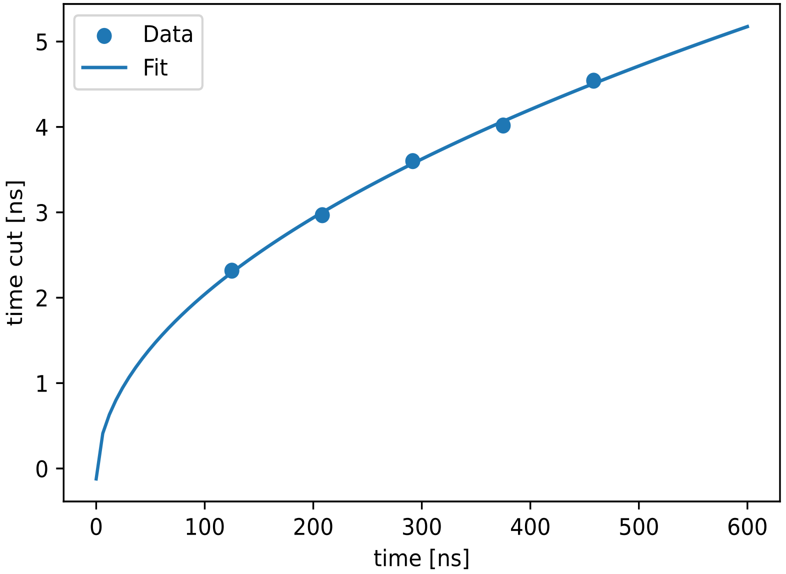


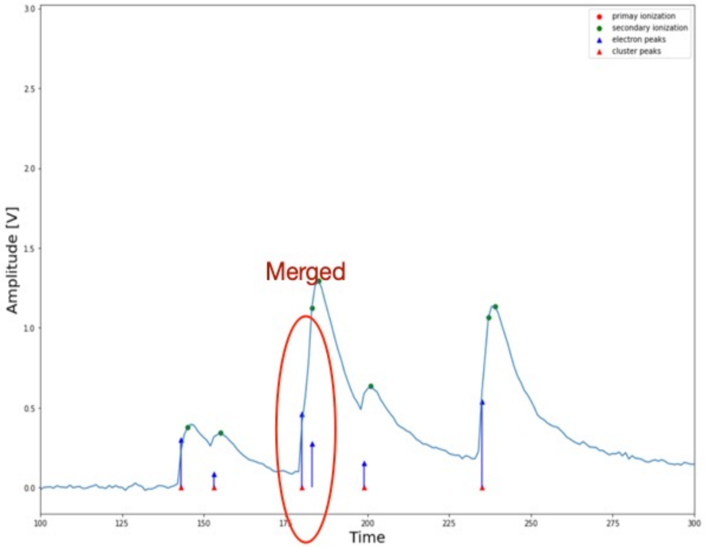


Waveform with 2% noise ratio sample

Clusterization

- **Merging algorithm using timing information of continuous peaks**
  - The time cut is measured from MC. Fitted function: $t_{cut} = a * \sqrt{t_{drift}} + b$
- **Clusterization steps:**
  - $t_{cluster}$ definition: time of the middle position of a cluster
  - If $\Delta t = abs(t_{cluster,i} - t_{peaks,j}) < t_{cut}$, one merges peak-j to cluster-i, updates $t_{cluster,i}$



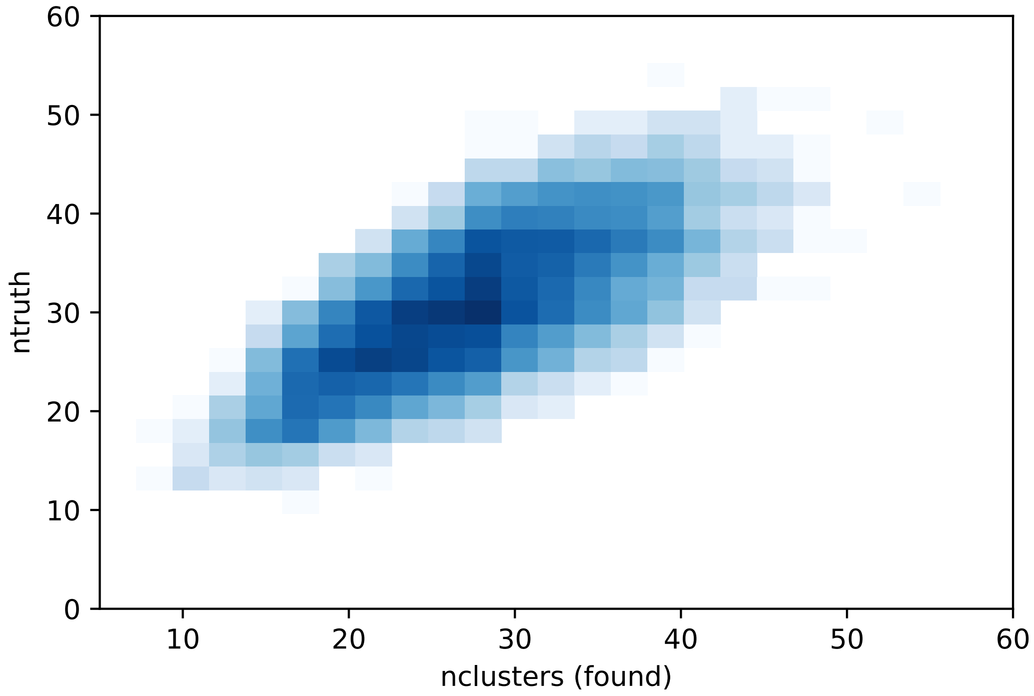


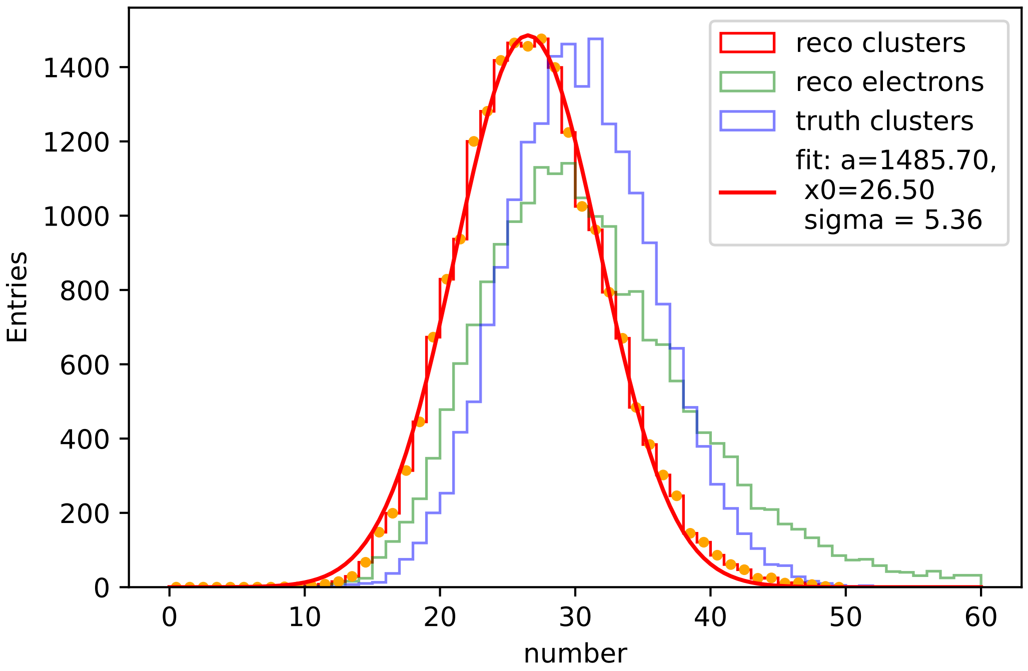
2% noise ratio sample

Cluster reconstruction performance

Preliminary performance:

- Cluster reconstruction efficiency: eff = #reco cls/ #truth cls = 92.5%,
- 1m resolution ~ 2.7%

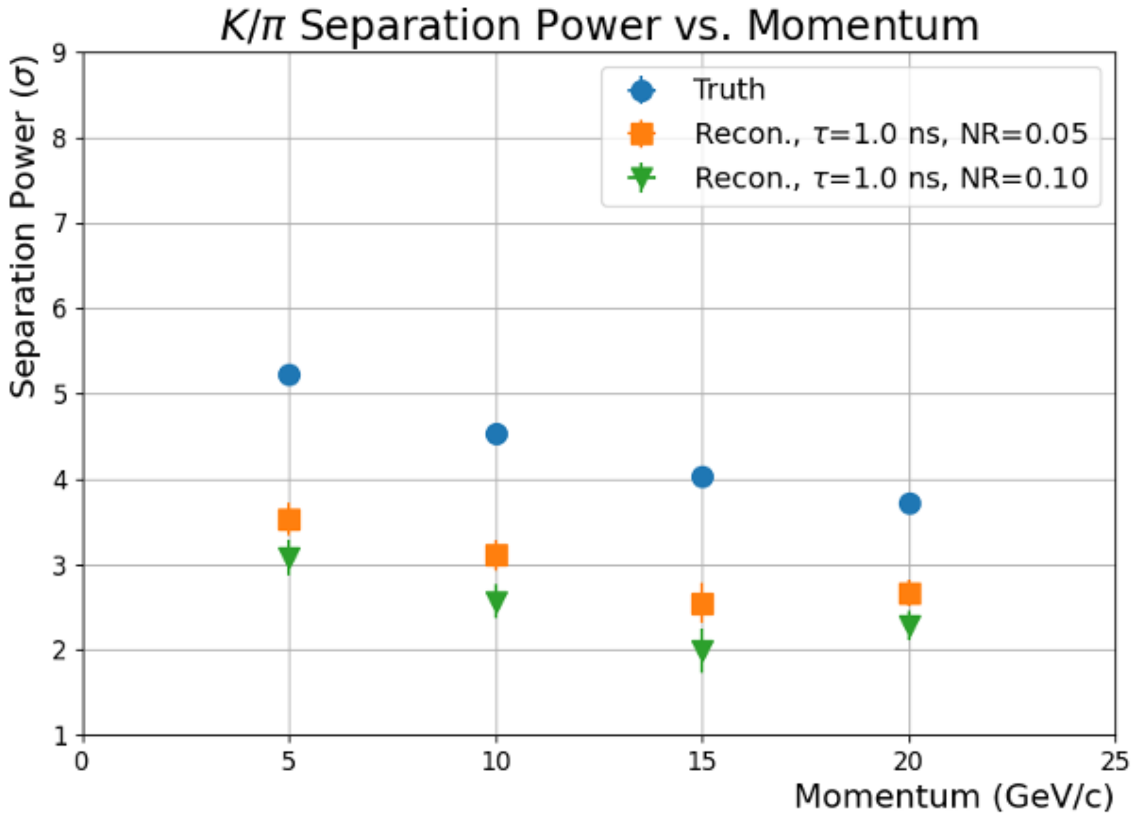




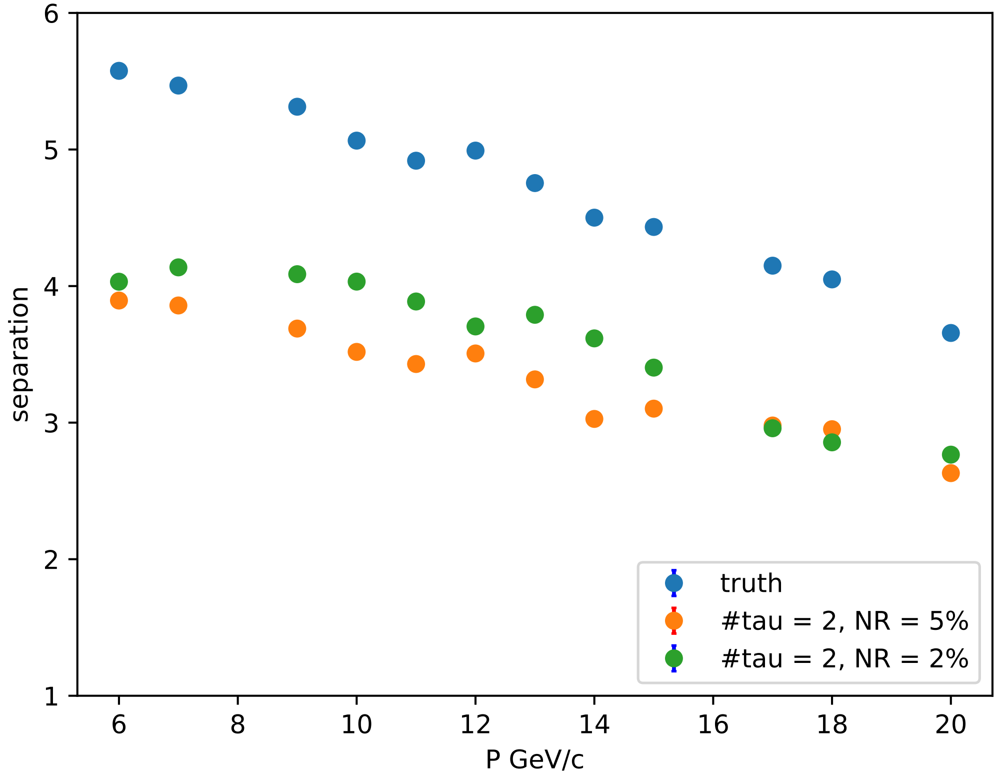
2% noise ratio sample

Preliminary PID performance

- K/pi separation using old current-sensitive pre-amplifier (tau = 2), noise ratio = 2%/5% condition.
- New results: 2.75 (2.65) σ with 2% (5%)noise at momentum = 20GeV.
- Reconstruction algorithm parameters are suboptimal.



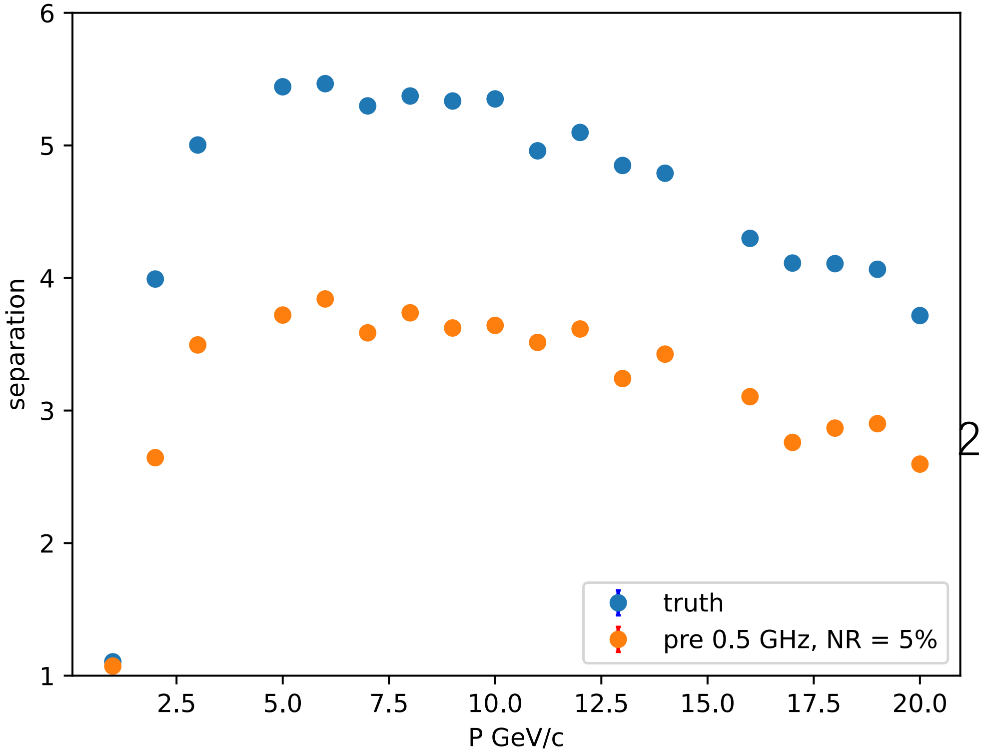
Previous results (Garfield++ simulation, white noise, 1k events)



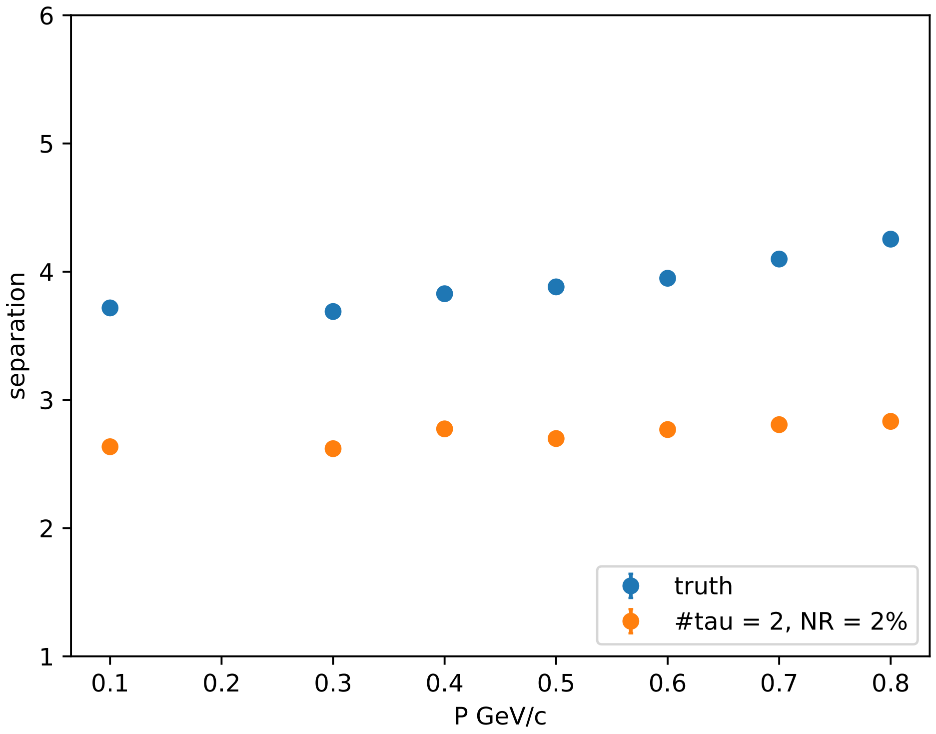
New results (experimental noise, 20k events)

Preliminary PID performance

K/pi separation using new pre-amplifier (cut-off 0.5 GHz), noise ratio = 5% condition



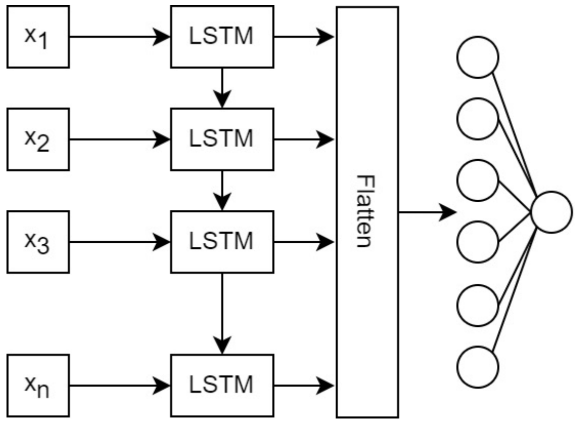
2.6 σ



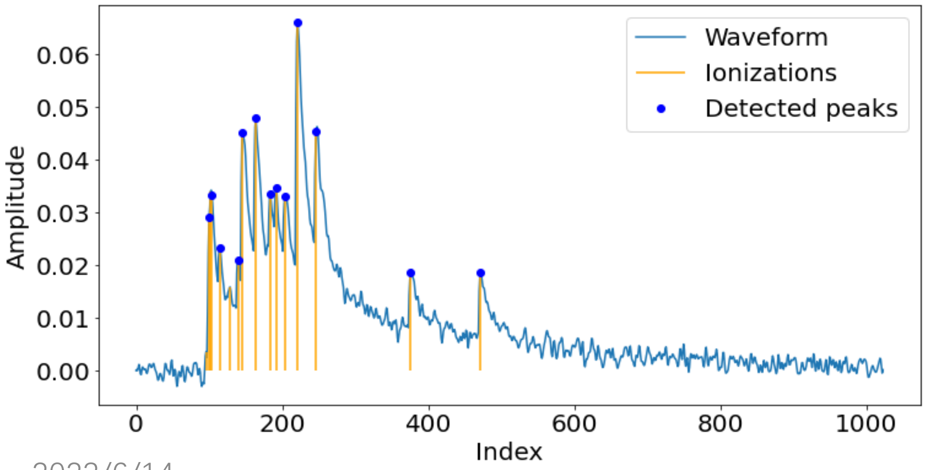
Reconstruction efficiency: 78% to 74%
(with cos(theta) increasing)

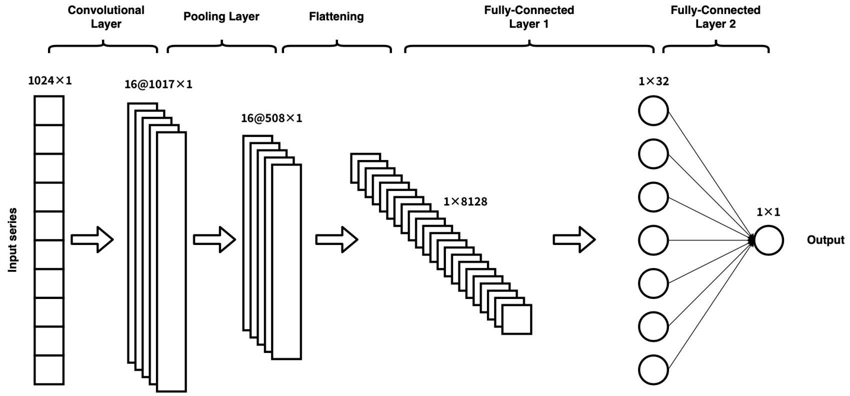
Updates of machine learning algorithm

Review of previous LSTM+CNN method

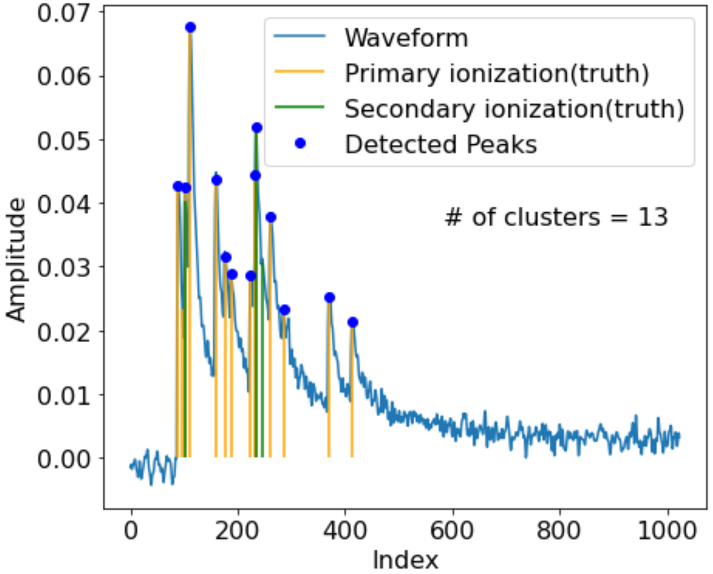


Step1: A classification problem to classify ionization signals and backgrounds in the waveform using LSTM.





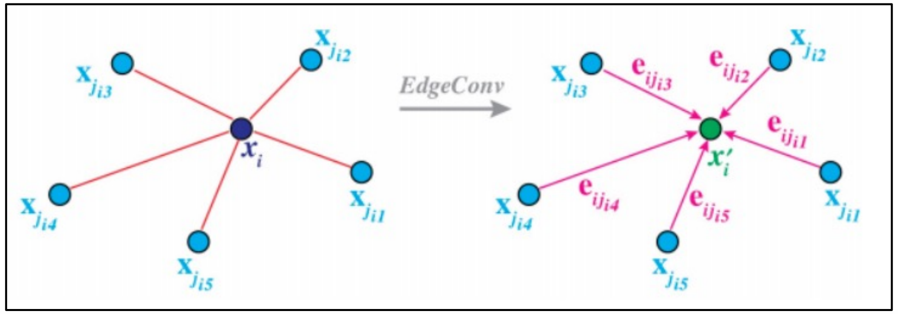
Step2: A regression problem to predict N_{cls} using CNN.

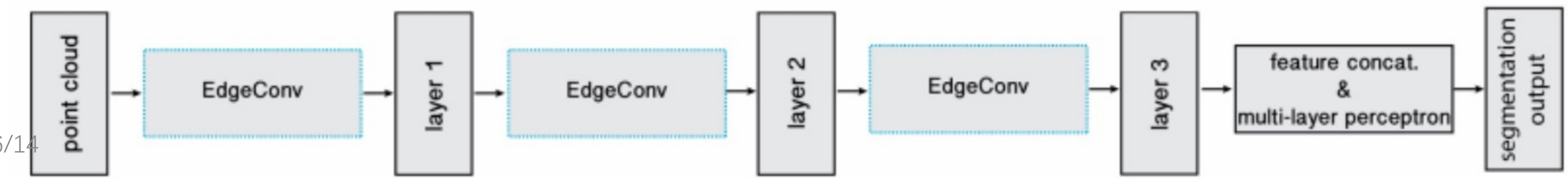


Attempt to train clusterization algorithm with DGCNN

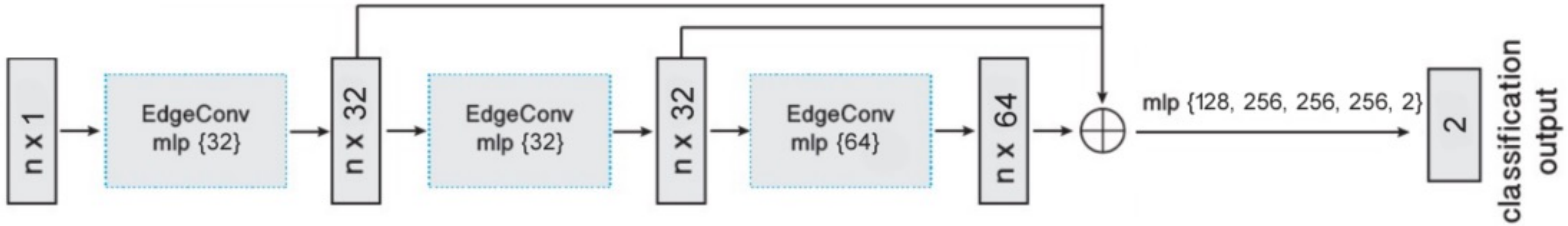
- Graph Neural Network (GNN):
  - based on graph-structured data, capture the dependencies and relationships between nodes in the graph.
  - A group of ionization peaks as a graph, where the peaks are nodes and the relationships between them are edges.
- Dynamic Graph CNN (DGCNN):
  - Dynamically construct the graph at each layer: connect k-NN nodes.
  - Better capture local geometric features.
  - Already been used in high energy physics → ParticleNet.

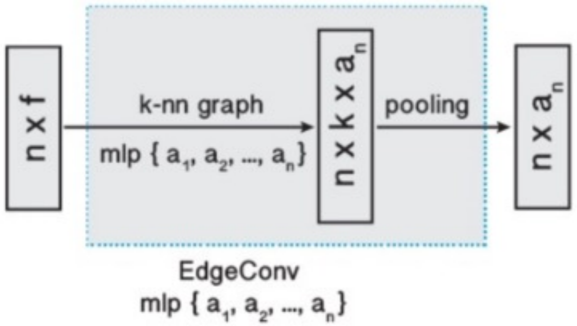
EdgeConv:





DGCNN structure

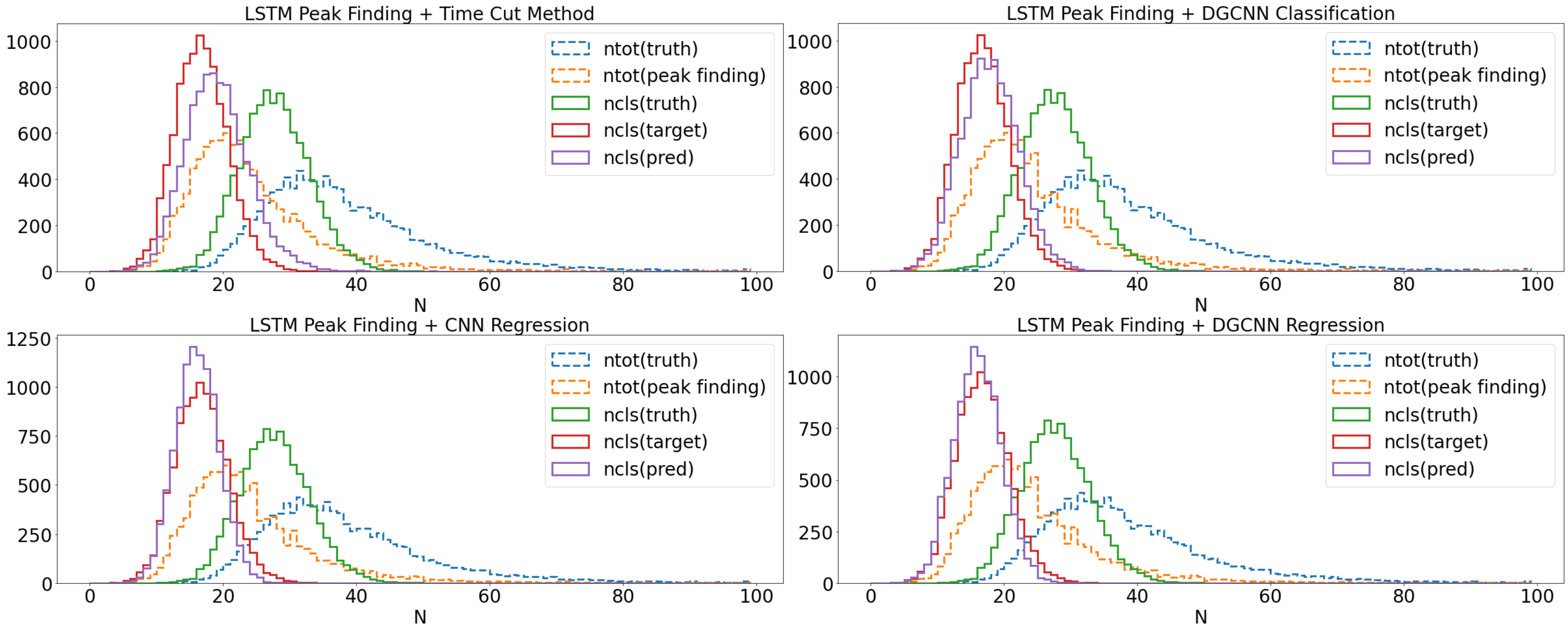




- Graph: Each waveform corresponds to a graph
- Node: Ionization peaks found by LSTM model
- Node feature: Positions (time) of the ionization peaks on the waveform
- Edge: Dynamically computed.
- Labels: Types of ionization peaks (1 for primary ionization peaks, 0 for non-primary ionization peaks)
- Loss: Cross entropy loss (Log softmax + NLL Loss)

⇒ Node classification

Clusterization results



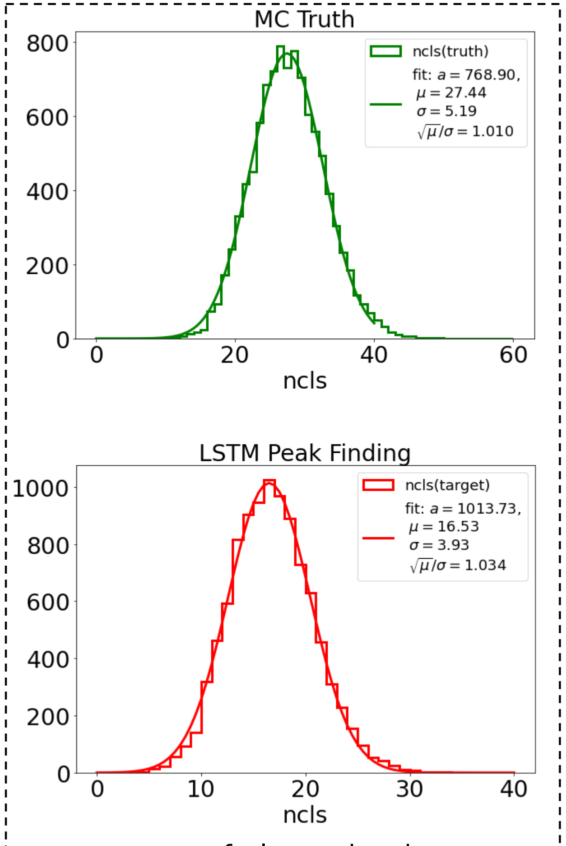
ntot(truth)：Total ionization peaks in MC truth

ntot(peak finding): Total ionization pmeaks after Peak Finding step1

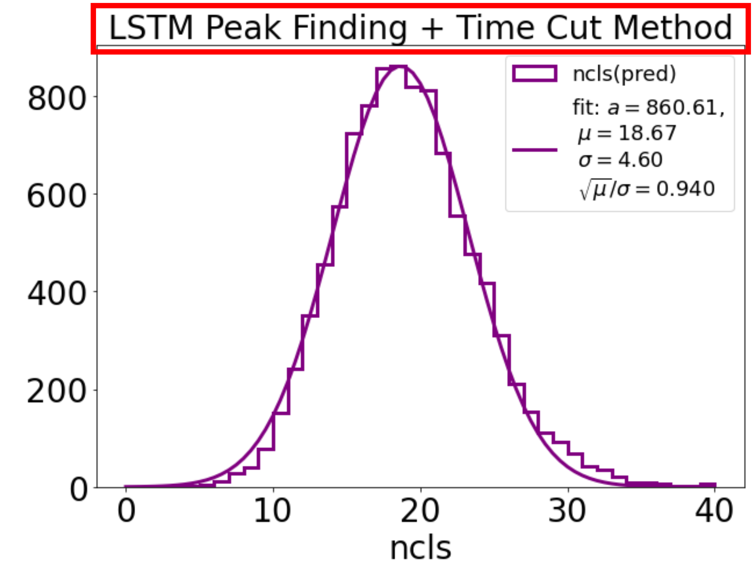
ncls(truth)：Number of clusters in MC truth

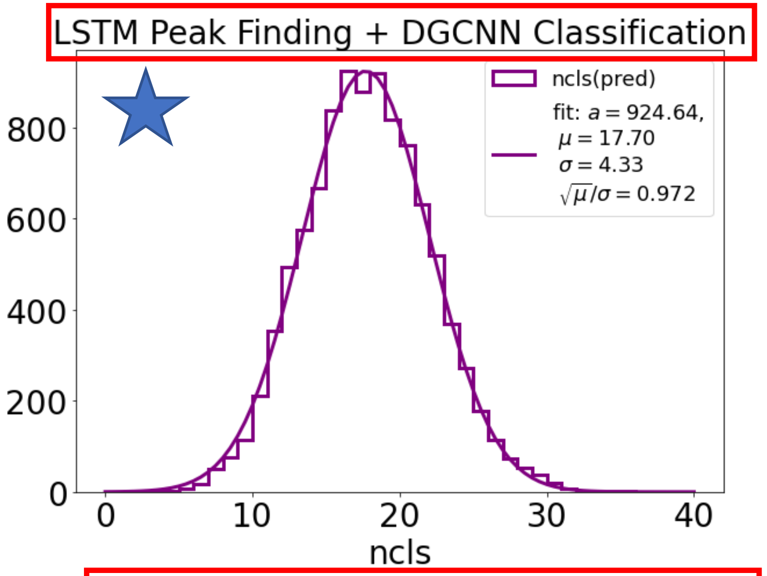
ncls(target): Number of clusters after Peak finding (from MC truth), target of clusterization algorithm
ncls(pred): Number of clusters predicted by clusterization algorithm step2

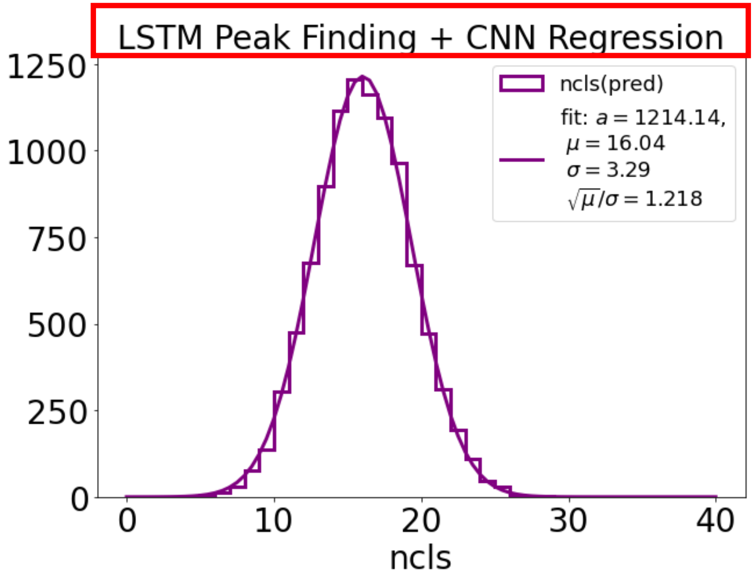
Clusterization results

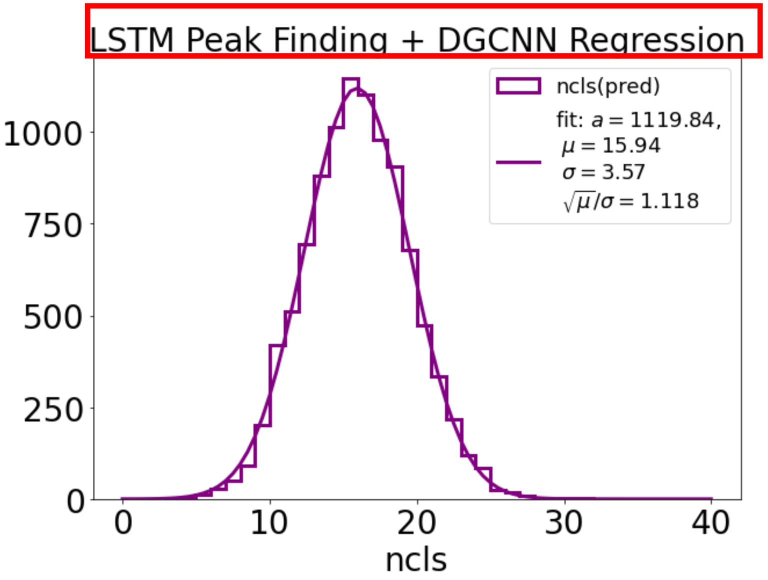


target of clusterization









Step1: Peak finding results as inputs

DGCNN classification has better Ncls distribution among four methods.

Clusterization performance of different methods

Method	N_{mean}	σ	$\sqrt{N_{\text{mean}}}/\sigma$	σ/N_{mean}
MC Truth	27.44	5.19	1.010	18.9%
Target	**16.53**	**3.93**	**1.034**	**23.8%**
Time Cut	18.67	4.60	0.940	24.6%
CNN Regression	16.04	3.29	1.218	20.5%
DGCNN Regression	15.94	3.57	1.118	22.4%
DGCNN Classification	17.70	4.33	0.972	24.4%

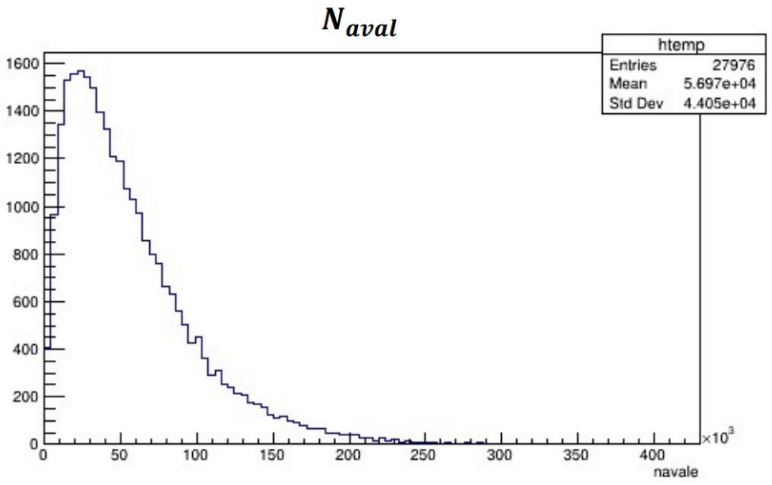
- DGCNN classification has better Ncls distribution than traditional time cut method and CNN.
- Considering combination of the loss functions of DGCNN Regression and DGCNN Classification.

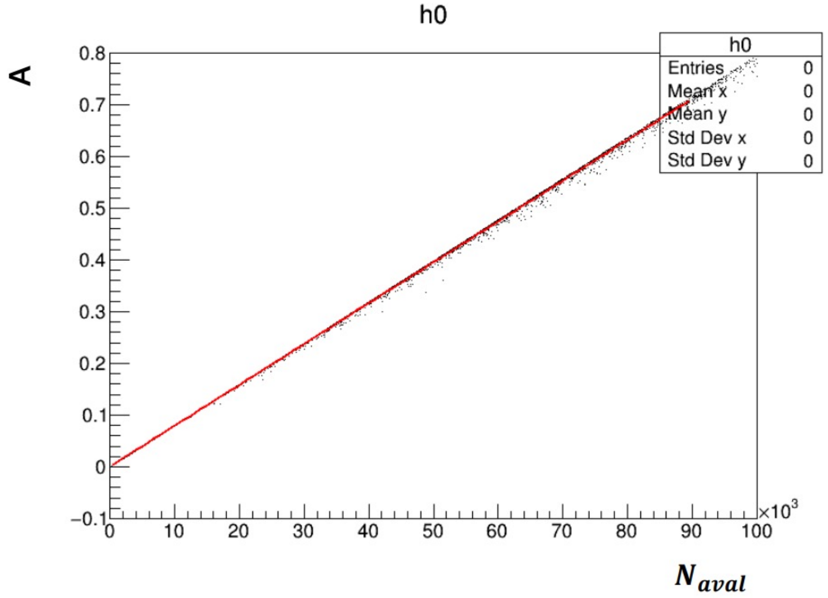
Summary

- The full simulation package is updated and works well.
  - Effective models are implemented to speed up the simulation
  - Simulation of electronics and noises are tuned with data.
- A PID analysis is performed using events from the simulation package.
  - Preliminary result with experimental noise and new pre-amplifier is given.
  - Better than 2.6 σ K/pi separation at 20 GeV.
- Cluster counting algorithms using ML are developed.
  - DGCNN classification method gives a better Ncls distribution than others.

Backup slides

pulse amplitude model



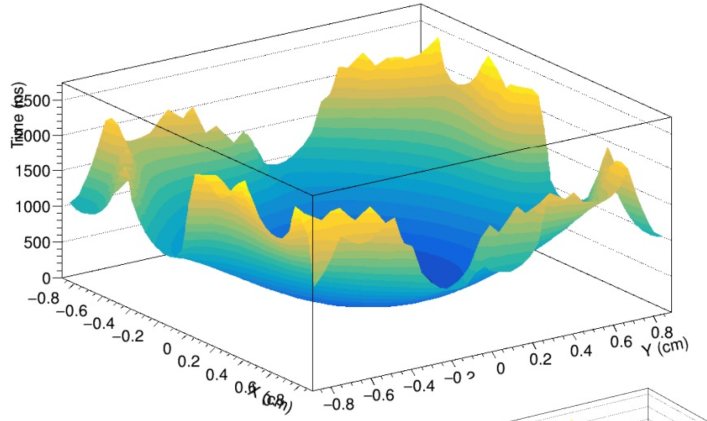


- Strong inhomogeneous field around a thin wire yields **Polya** distributions
- Obtain N_{aval} distribution from Garfield simulation

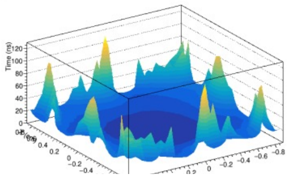
- Induced current $\propto -\frac{N_{aval}}{t+t_0}$
- Pulse height $A \propto N_{aval}$
- Linear fit:
  - $A(N_{aval}) = p_0 + p_1 \times N_{aval}$

Pulse time model

μ

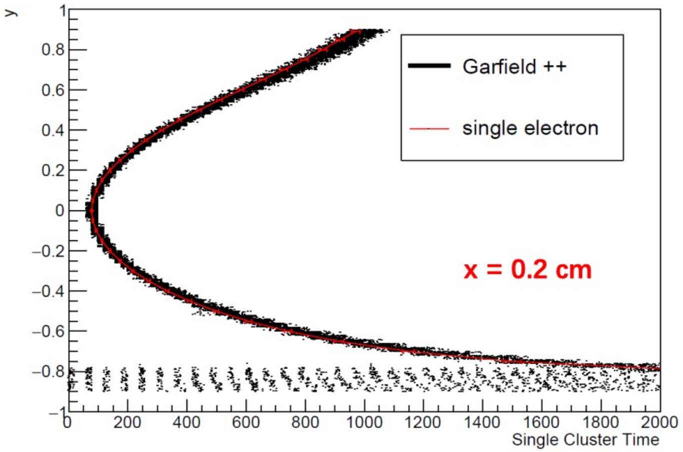


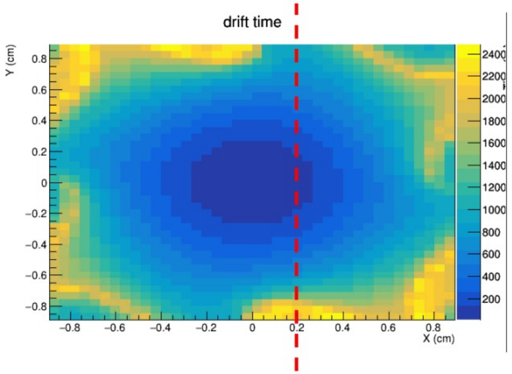
σ



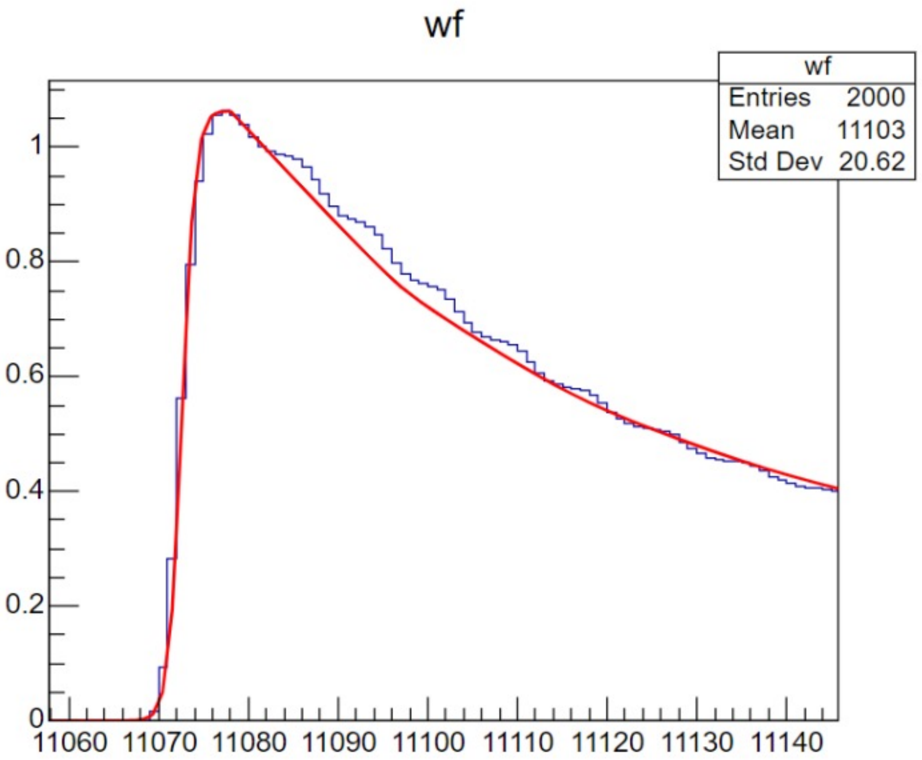
- For a fixed electric/magnetic field:
  - t is mainly determined by initial position of the electron
- Measure the relationship from Garfield++ simulation
  - $t(x, y) = Gauss(\mu(x, y), \sigma(x, y))$

Comparison to Garfield++



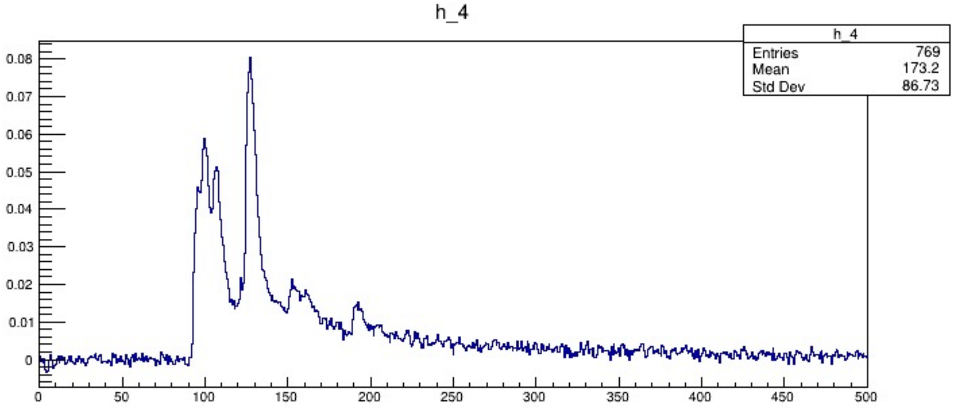


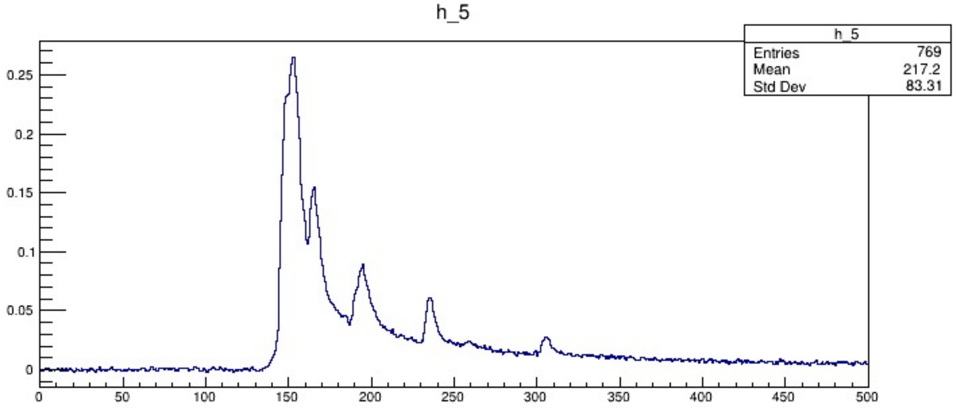
Pulse shape model

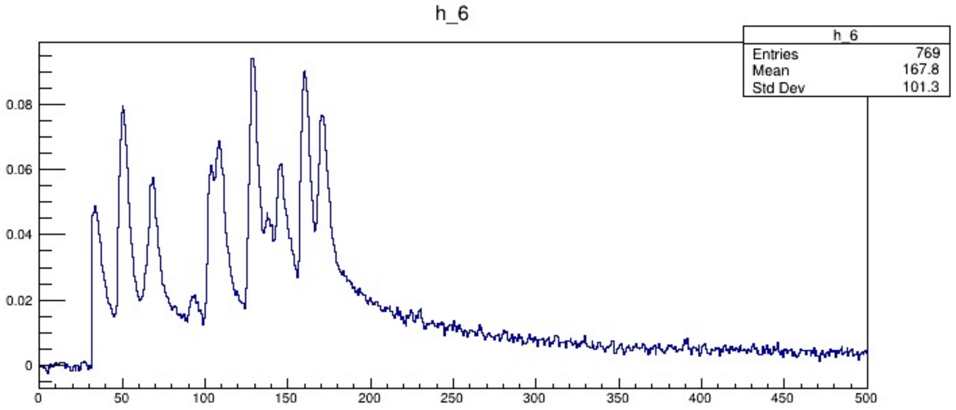


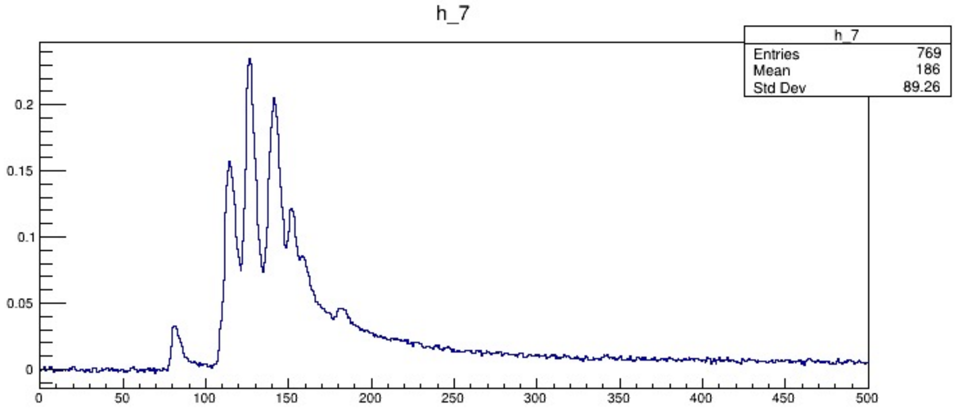
- **Fit the Garfield pulse by:**
- $$f(x|A,t) = \begin{cases} p_0 \times \frac{e^{-p_1(x-p_2)}}{1+e^{-\frac{t-p_3}{p_4}}}, x < t \\ A \times \frac{p_5^{p_6}}{(x-t)^{p_6}+p_5^{p_6}}, x \geq t \end{cases}$$

Simulation waveforms (2)

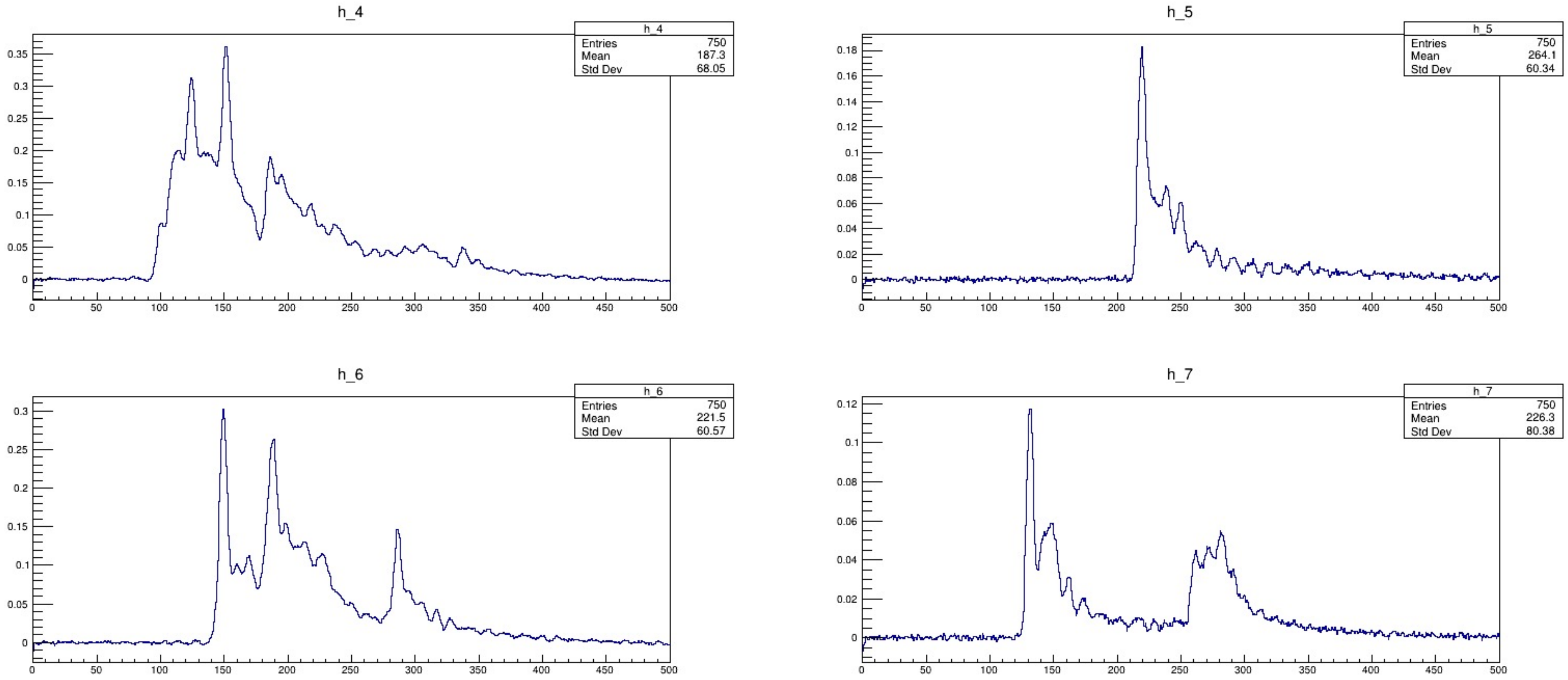




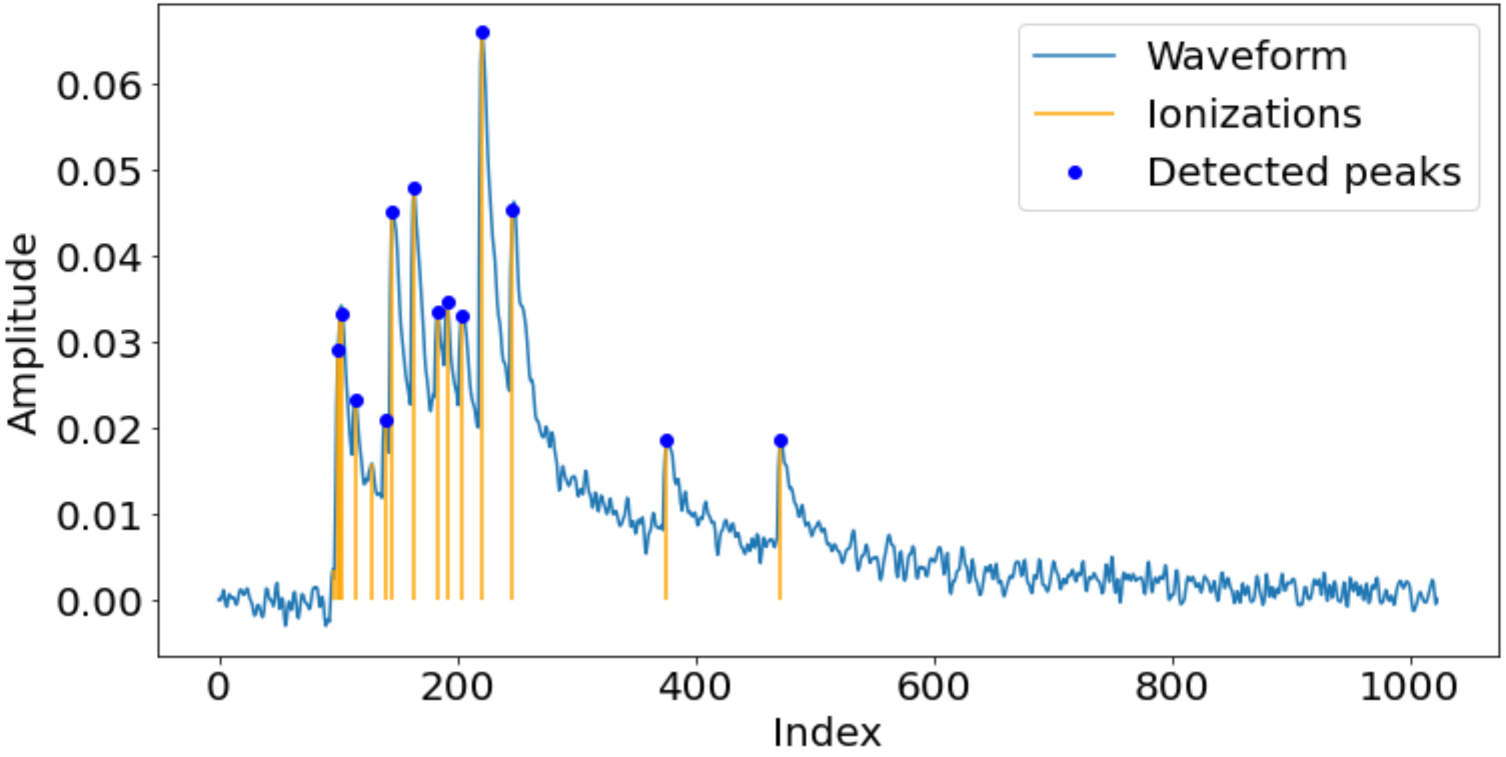


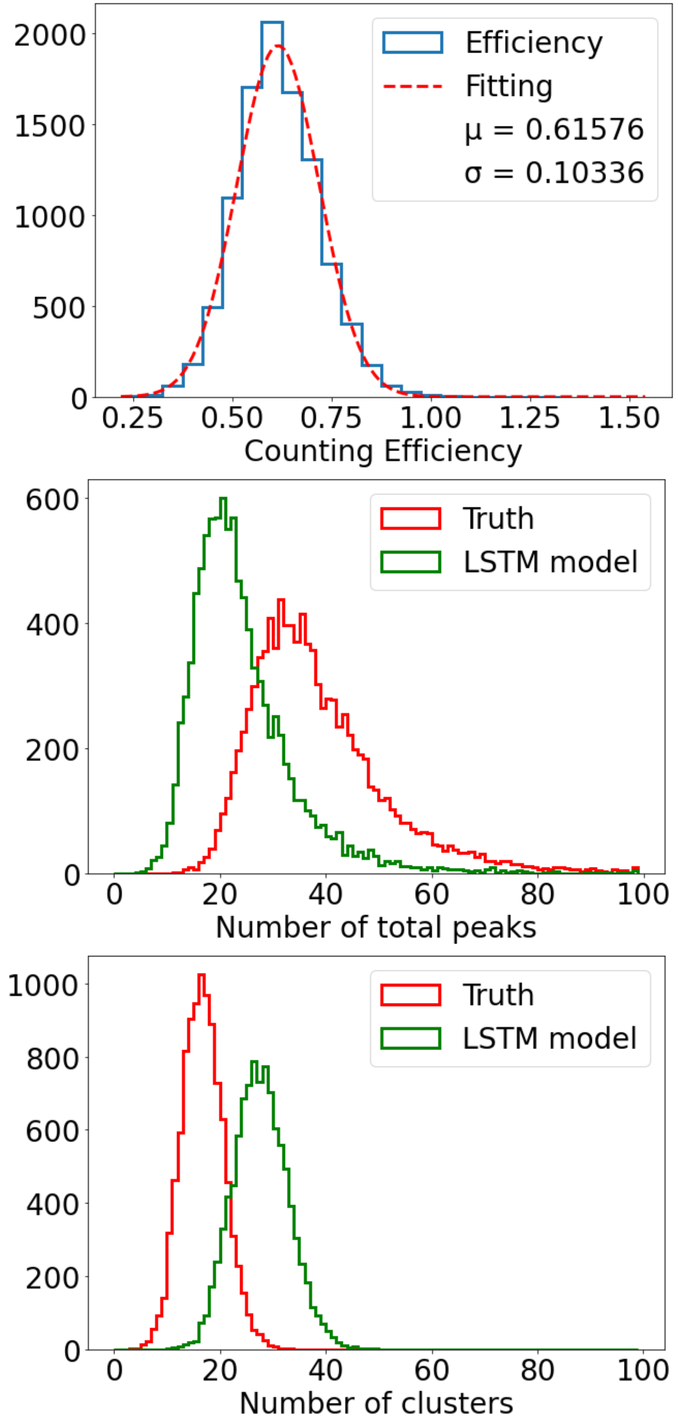


Data waveforms (2)



- **Peak finding results:**





The efficiency of peak searching is about 60%, but the primary ionization numbers obtained after peak finding still have a good Gaussian shape.