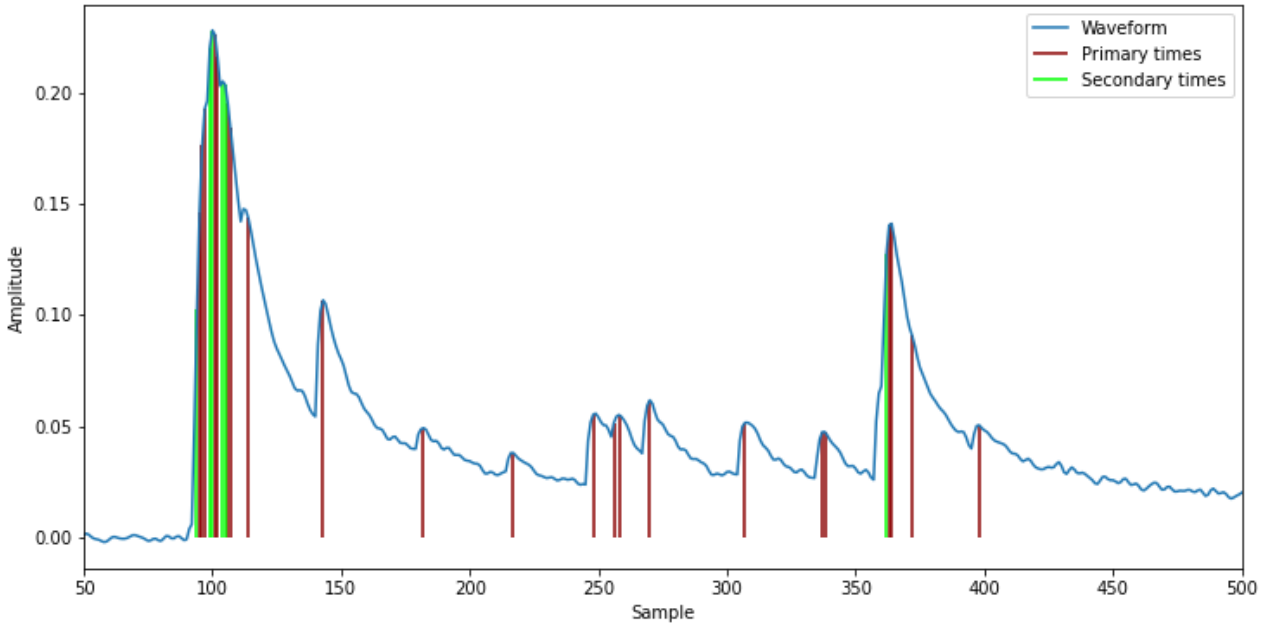
Cluster finding algorithm based on machine learning

Guang Zhao

zhaog@ihep.ac.cn

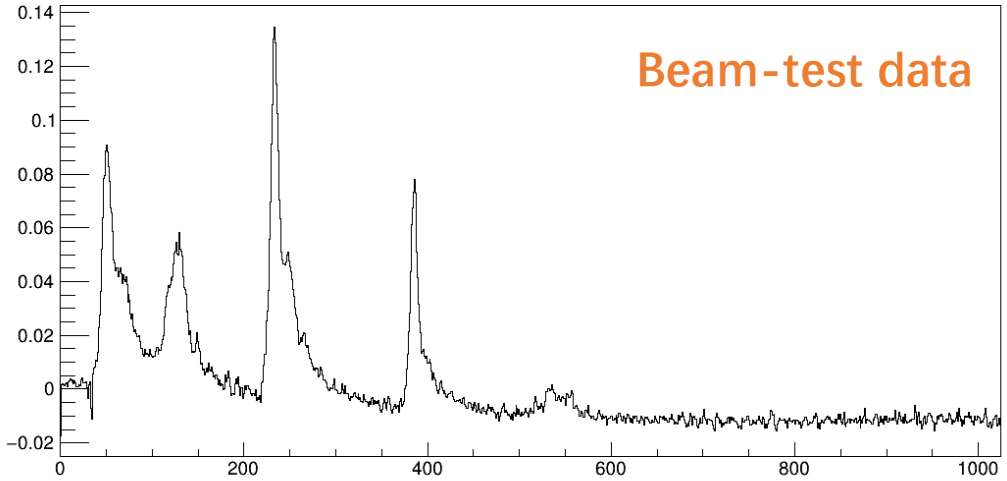
Problem description

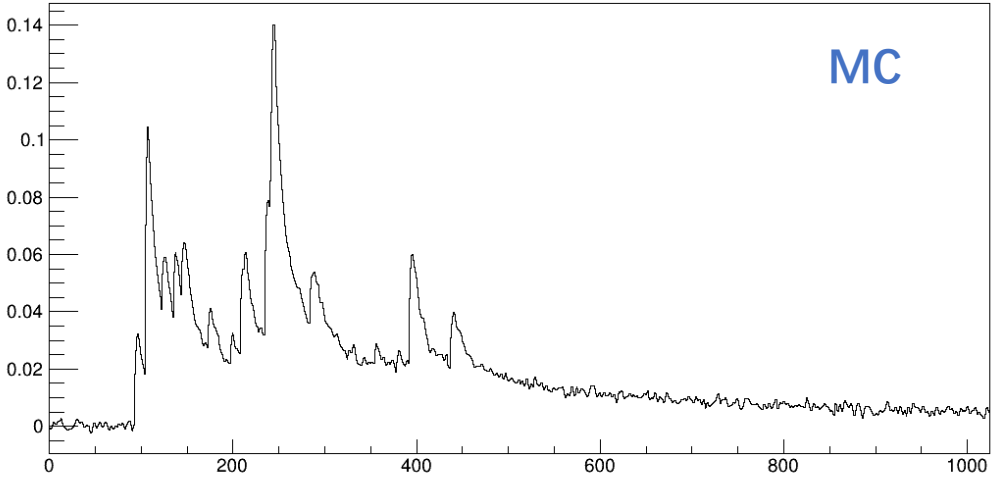


- **High-efficient cluster counting algorithm is essential for the dN/dx measurement**
- **Goal of cluster counting:**
  - Both **primary electrons** and **secondary electrons** contribute peaks on the waveform
  - Find the number of peaks from primary electrons
- **Cluster counting = Peak finding + Clustering**
  - Peak finding: Find all peaks
  - Clustering: Discriminate the # of primary ones

Dataset

- **Labelled MC samples**
  - ~20 primary ionizations per waveform
  - Tuned based-on beam test waveforms: noise model, amplitude, peak rising-time

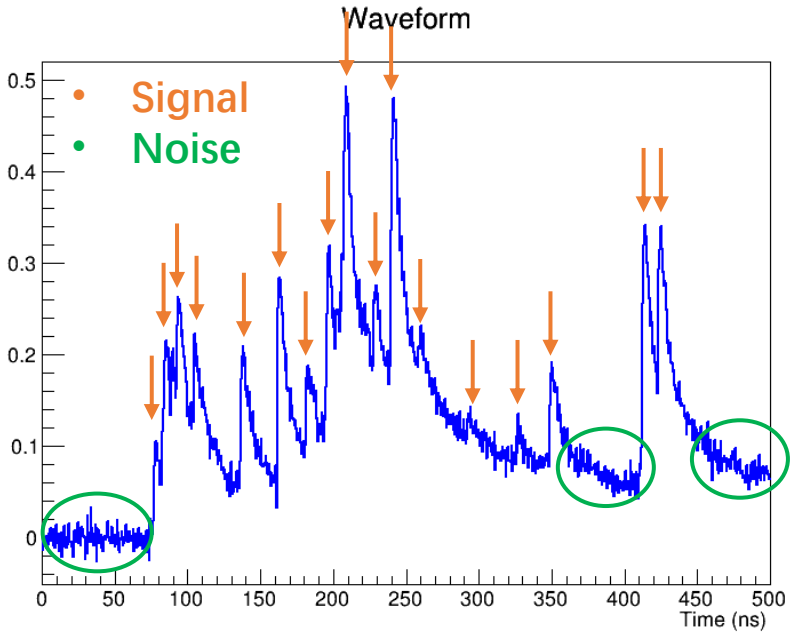




Peak finding algorithm

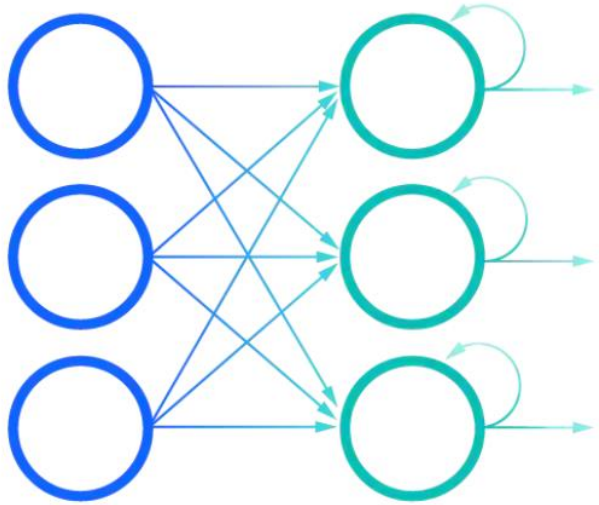
Problem:

- Peak detection of waveforms
- Supervised-classification: “signal” and “noise”
- Data in time-sequence form



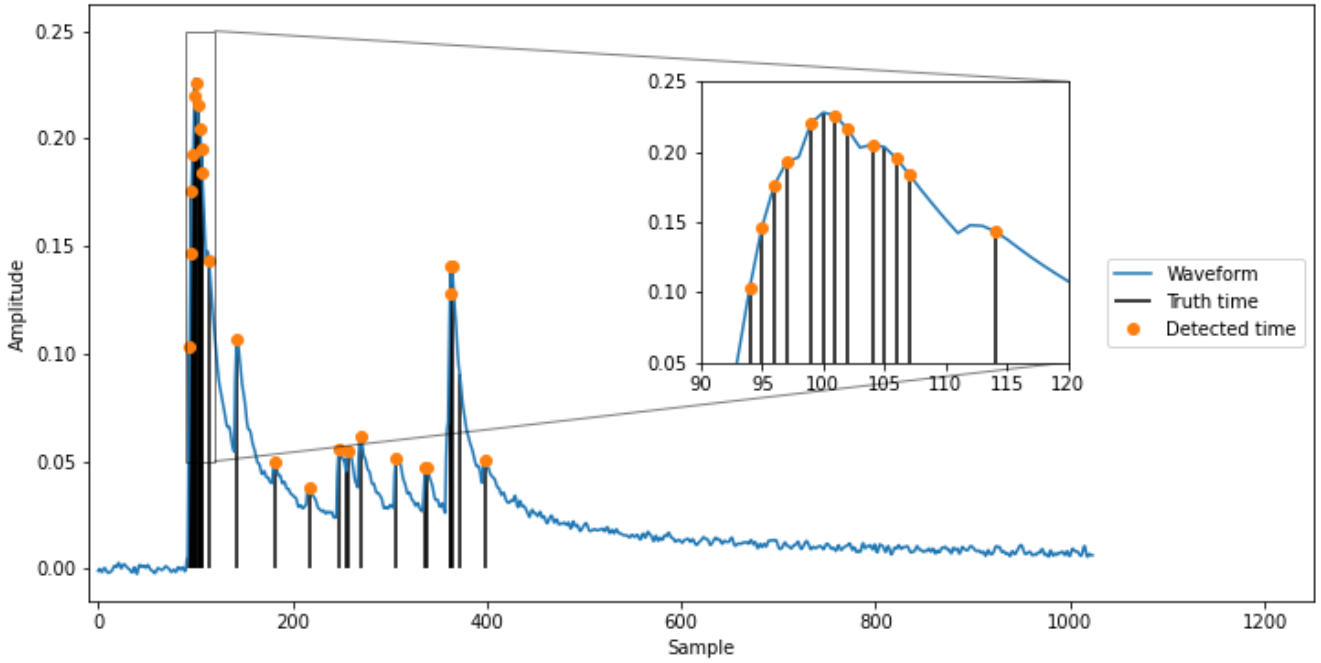
Recurrent Neural Network (RNN):

- “Memory” structure: internal loops over sequence elements
- Powerful to handle time-sequences

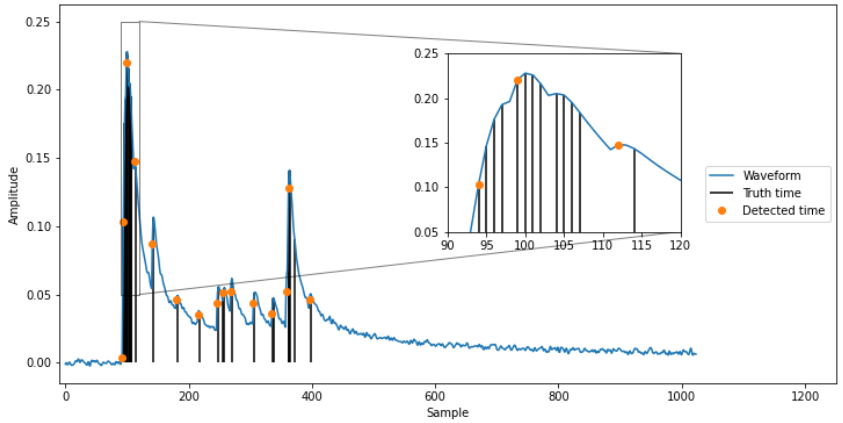


An example for demonstration

Neural Network

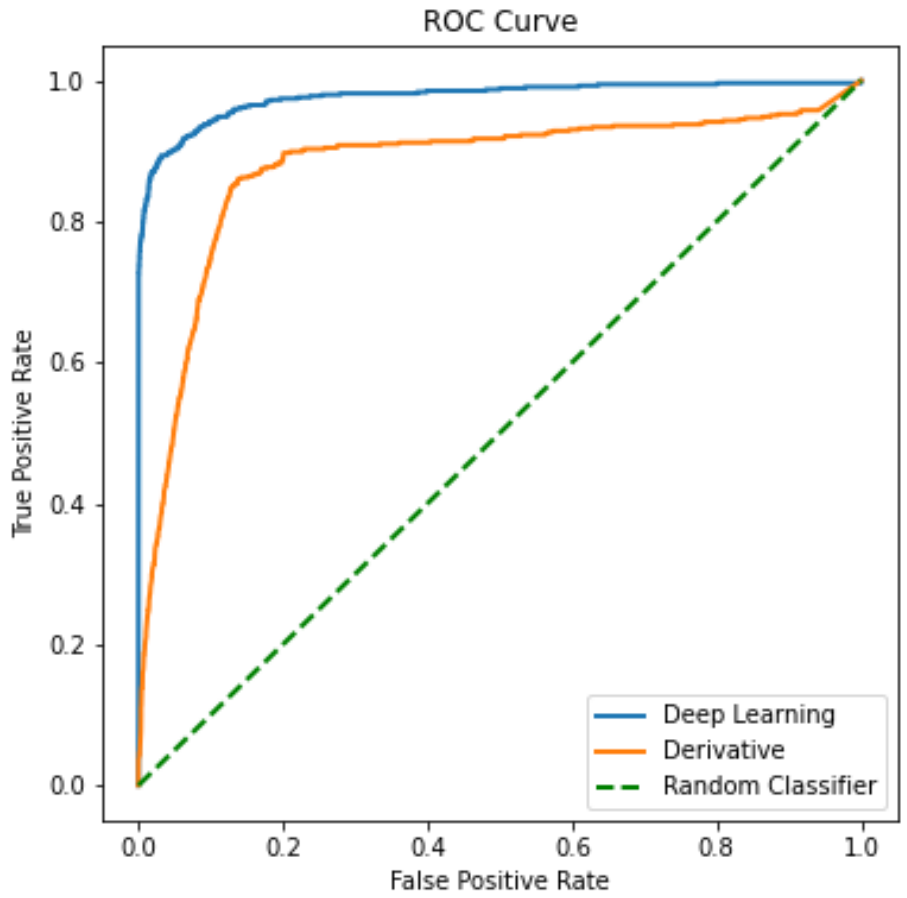


Derivative



The NN can find the peaks more effective!

Receiver Operating Characteristic (ROC)



NN is a better binary classifier than the derivative method

Note: ROC curve is a standard tool for evaluation binary classifiers. ROC curve with larger area-under-curve (AUC) is better

Clustering algorithm

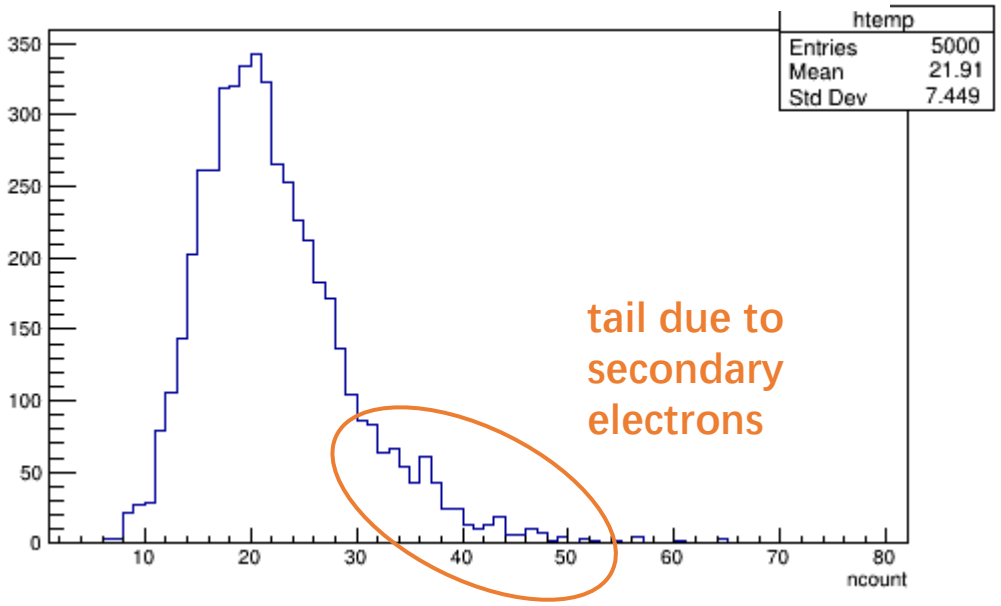
Problem:

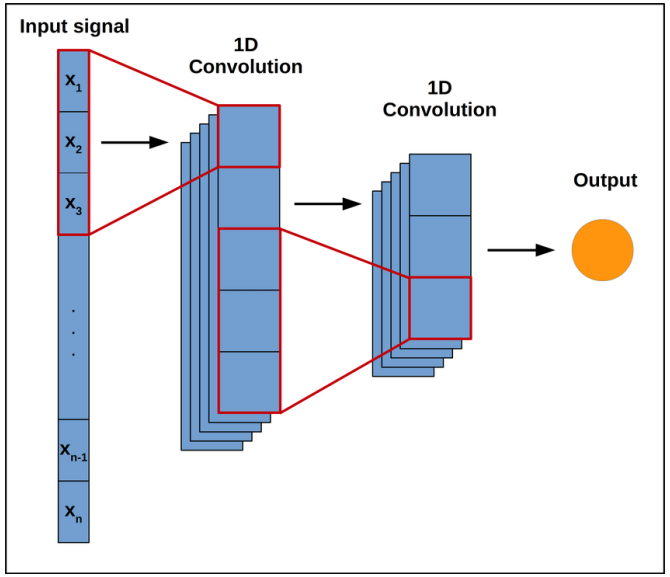
- Determine the # of peaks from primary electrons
- Supervised regression

1D Convolutional Neural Network:

- Extracting features form local input patches
- 1D version of CNN is highly relevant to sequence processing

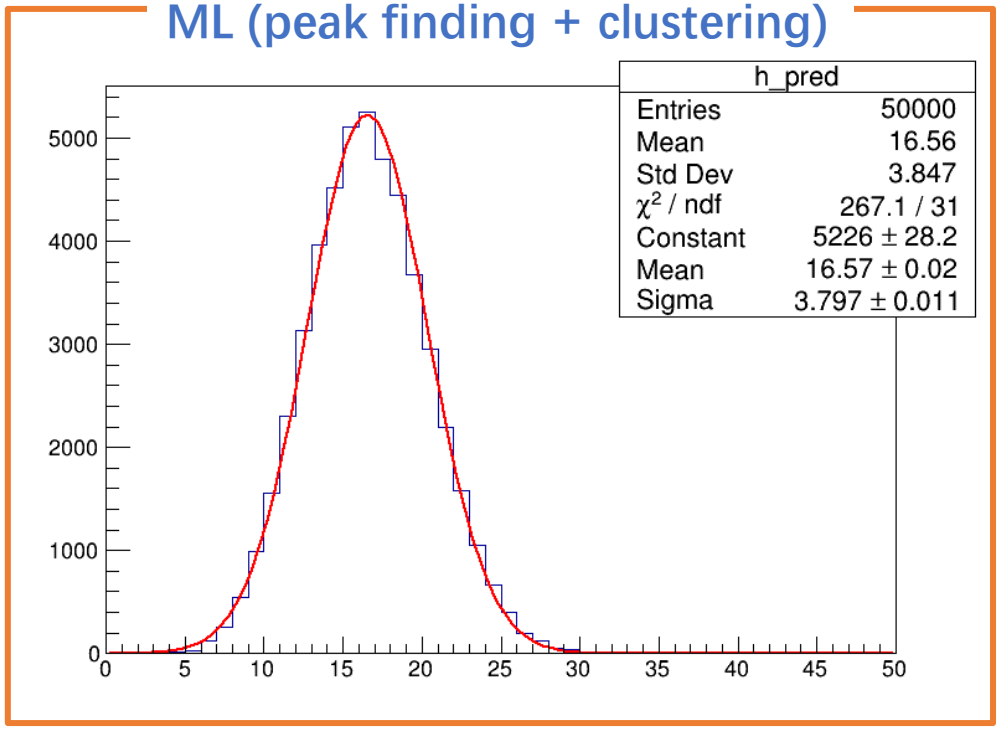
Detected peaks from the peak finding





N_{cls} distributions

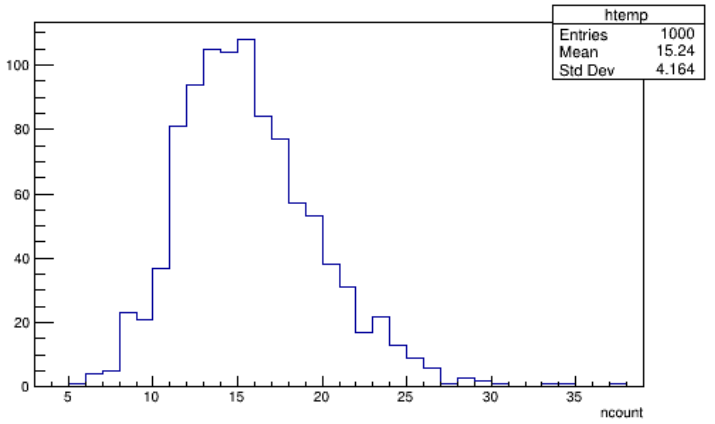
ML (peak finding + clustering)



Resolution ~23%

- ✓ Very good Gaussian-like distribution
- ✓ The resolution is very close to the truth value (~21%), which implies possible improvement on PID

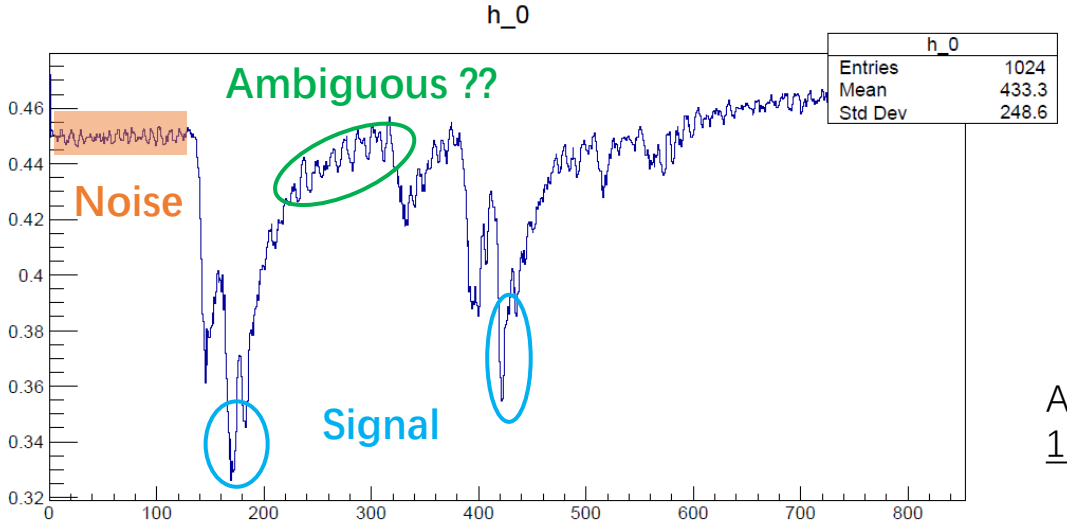
Derivative (no clustering)



- ✓ RMS/MEAN ~ 27%

Summary

- ✓ **The deep learning algorithm shows promising results from a MC sample**
- ➢ **For full simulation:**
  - ➢ To train the network with datasets of larger momentum range
- ➢ **For beam test data (in progressing):**
  - ➢ Try to label the data
  - ➢ In progress to better understand the waveform and improve the preprocessing of the training dataset



A 3cm tube waveform from
11Nov_0angle_HVnominal_1p2GSPS_5k.root