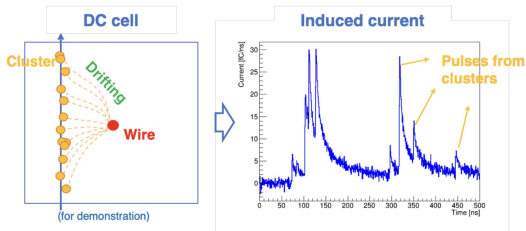# Introduction



- PID with a drift chamber is a key feature for the 4th conceptual detector of CEPC.

- PID technique:
  - $dE/dx$ : Energy loss per unit length, Landau distribution, large fluctuation.
  - $dN/dx$ : Number of primary ionization clusters per unit length, Poisson distribution, small fluctuation.

- An improved PID method:

  $N_p$(# of primary ionization clusters) measurement$\rightarrow$ $dN/dx$ technique
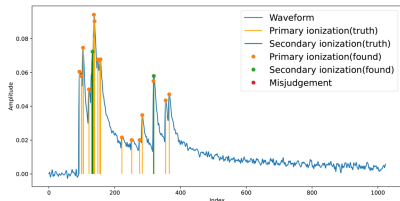
# Cluster Counting Algorithm

- **Cluster counting**: Cluster counting algorithm is to find the <span style="color:red"># of clusters</span> from the waveform.
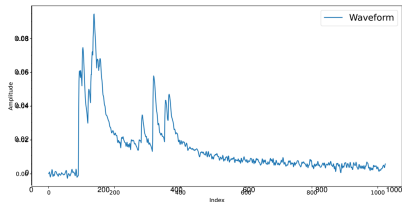


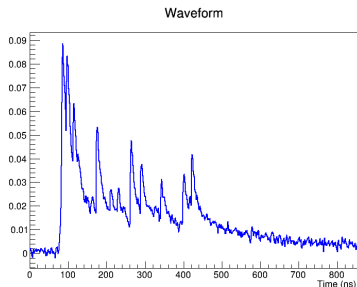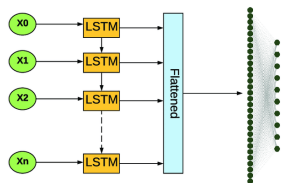Peak Finding → Find primary and secondary ionization peaks from the waveform



- Two steps: Peak finding and $N_p$ determination.

- Machine learning can use full information of the waveform.

- ML tools: TensorFlow, Keras, etc.

$N_p$ determination → Determine the number of clusters by the peaks found in previous step

The number of primary ionization clusters ($N_p$)

# Peak Finding Algorithm

**Peak Finding: find all ionization peaks from current waveform.**



Waveform



RNN (Recurrent Neural Network)
LSTM (Long Short-Term Memory)

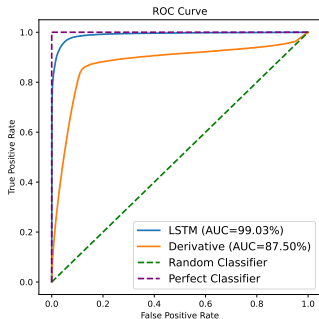- A classification problem to classify signal and background in the waveform.
- The waveform is time sequence data.

- With feedback loops, RNN has "memory". LSTM is a special kind of RNN.
- Well-suited to classifying based on time sequence data.

# Peak Finding Results



ROC Curve



D2 algorithm

LSTM algorithm

Note: ROC curve is frequently used for evaluating the performance of binary classification algorithms. ROC curve with larger area under-curve (AUC) is better.
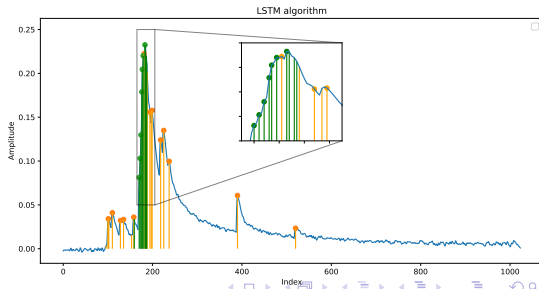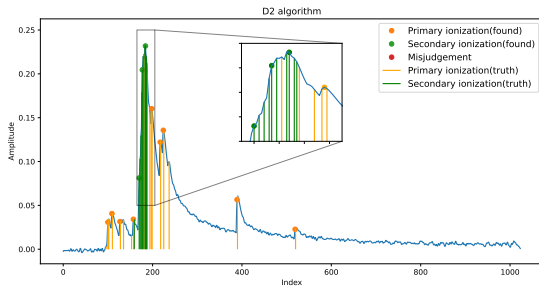
# $N_{\mathrm{p}}$ Determination Algorithm

$N_{\mathbf{cluster}}$ **determination**: Determine the # of clusters according to the ionization peaks found by peak finding algorithm.
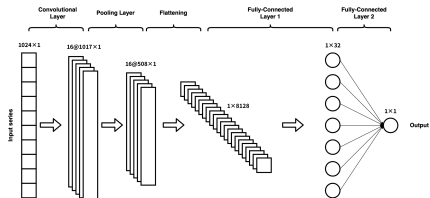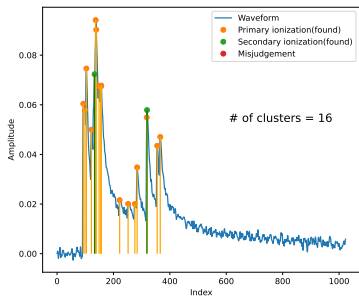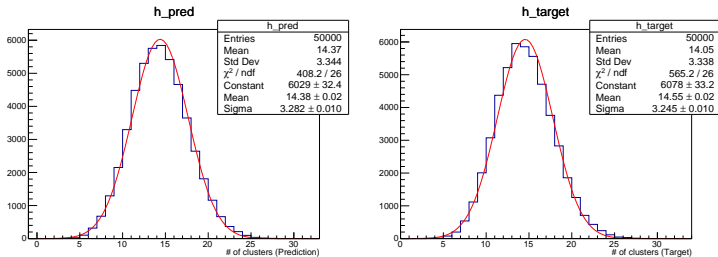




CNN (Convolutional Neural Network)

- A regression problem to predict # of clusters.
- The features are detected time, and the labels are # of clusters.

- Extracting features form local input patches.
- 1D CNN can handle sequence data.

# $N_p$ Determination Results and Summary



Model predicted distribution and true distribution of # clusters

- Single cell resolution($\sigma/\mu$) $\sim 22.8\%$ ($22.3\%$ in truth).
- Very good Gaussian distribution.
- The relative error is quite similar to the truth value, which implies stable efficiency.

**Summary**

- A two-step cluster counting algorithm with Machine Learning is developed.
- The algorithm is able to achieve a resolution close to the truth level, which is better than the algorithm based on derivatives.
- NEXT TO DO: Make the full evaluation of the algorithm and apply the algorithm to the experimental data.