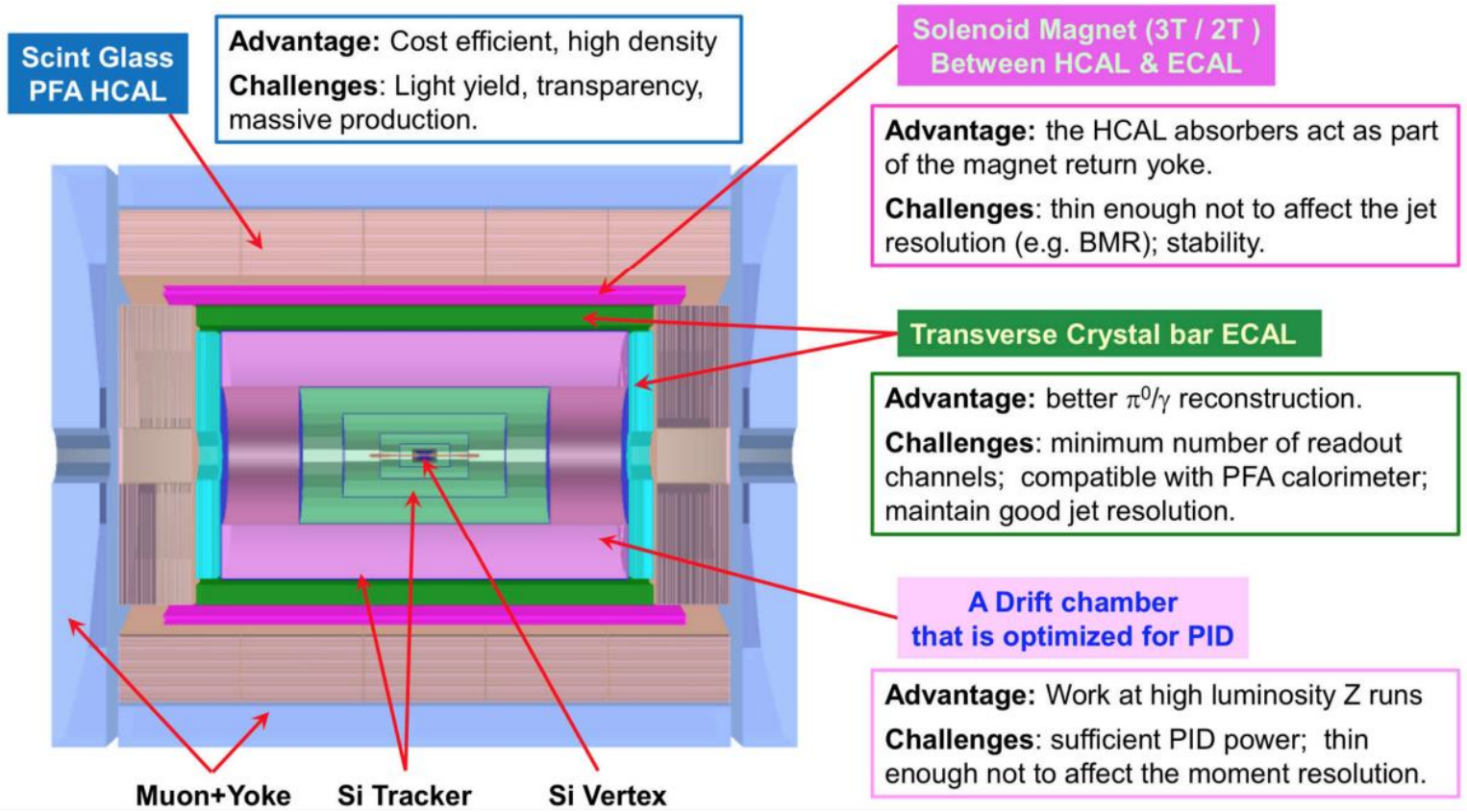
基于深度学习的漂移室电离计数重建算法的研究

田喆飞 (tianzhefei@whu.edu.cn) [1]，赵光[2]，董明义[2]，
刘帅毅 [2]，孙胜森[2]，伍灵慧[2]，辛水艇[2]，张振宇[1]，周详[1]

1.武汉大学
2. 中国科学院高能物理研究所

第二十届全国科学计算与信息化会议
2023年7月11日

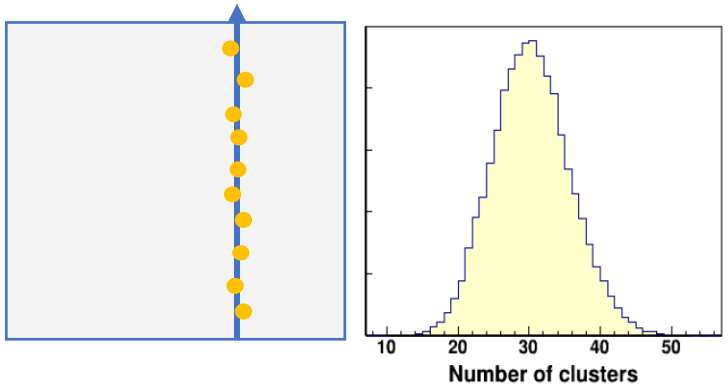
CEPC第四个探测器概念



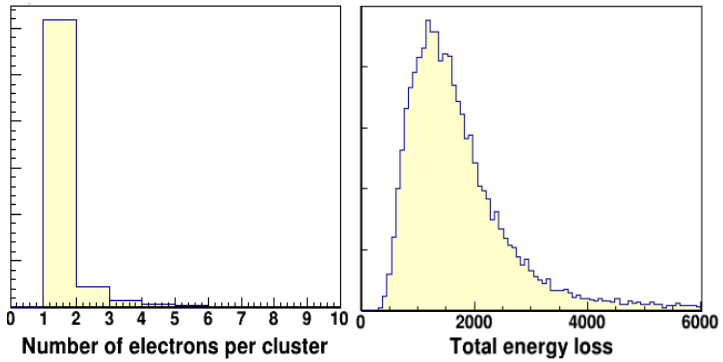
- 硅探测器和漂移室相结合的径迹探测器设计。
- 探测器应用电离簇团计数 (cluster counting) 进行粒子鉴别 (PID)。
- CEPC实验的物理目标要求在20 GeV/c处K/π鉴别能力达到2σ。

dE/dx vs dN/dx

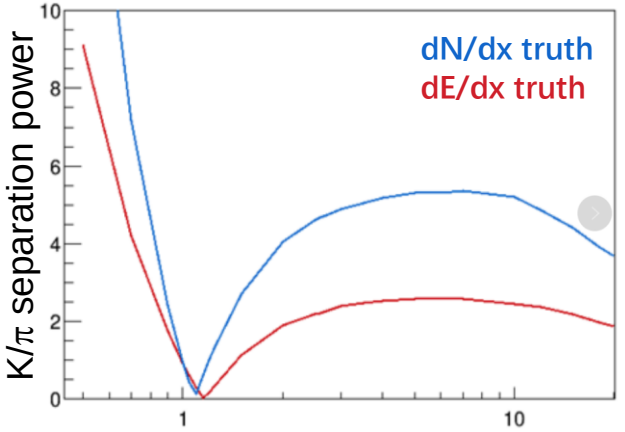
Primary ionization



Secondary ionization

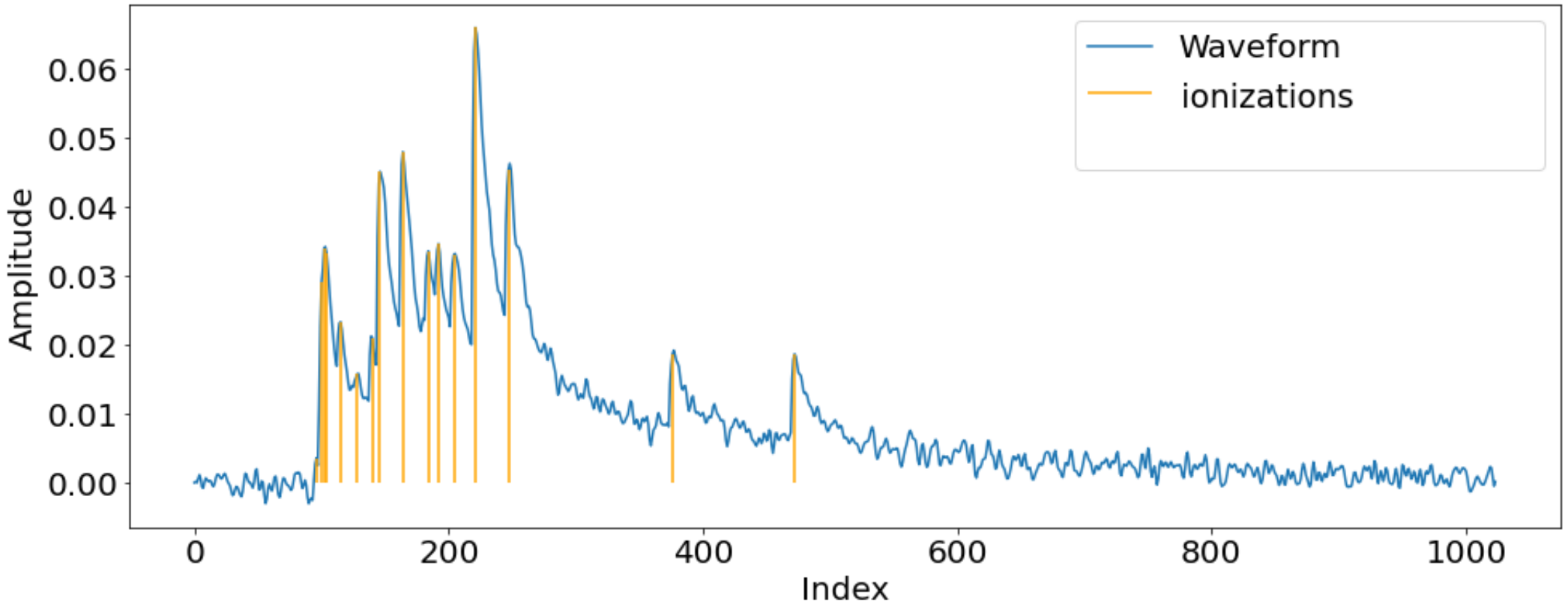


K/π separation power
dN/dx vs dE/dx



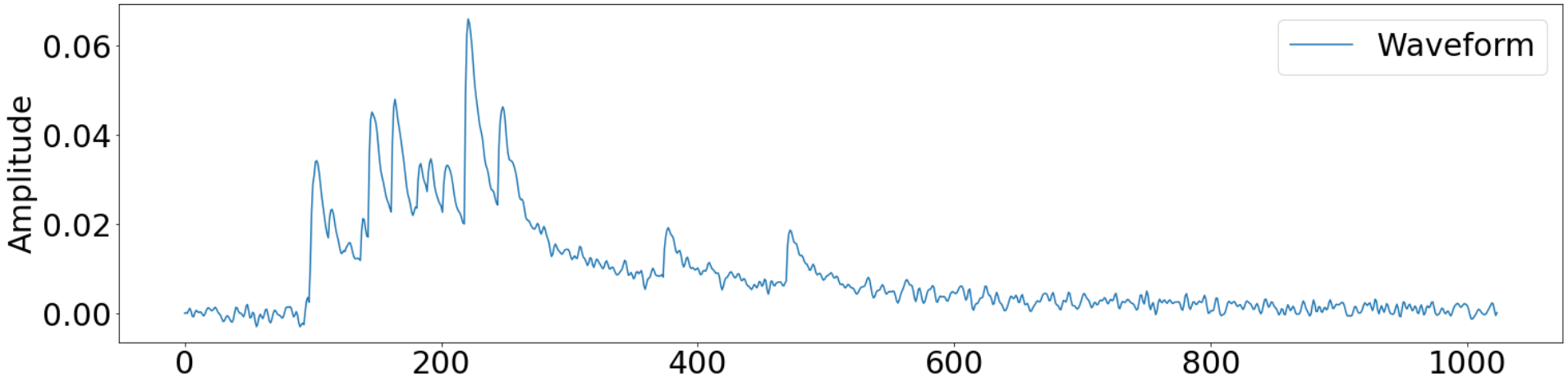
- 带电粒子穿过漂移室，电离工作气体产生原初电离电子，电子在电场作用下向信号丝漂移，并在漂移过程中继续电离产生次级电离电子。两种电子均漂移至信号丝产生信号。
- **传统PID方法：**测量单位距离能损dE/dx。
- 与dE/dx相比，dN/dx拥有更好的PID性能
  - $\boldsymbol{dE/dx}$**: 单位距离能损，朗道分布，涨落大**
  - $\boldsymbol{dN/dx}$**: 单位距离原初电离数，泊松分布，涨落小**
- **电离簇团 (cluster)：**由**原初电离 (primary ionization)** 和相应的**次级电离(second ionization)**组成。
- 测量原初电离数=测量电离簇团数(N_{cls}) ⇒ **电离计数算法 (Cluster Counting)**

电离计数算法

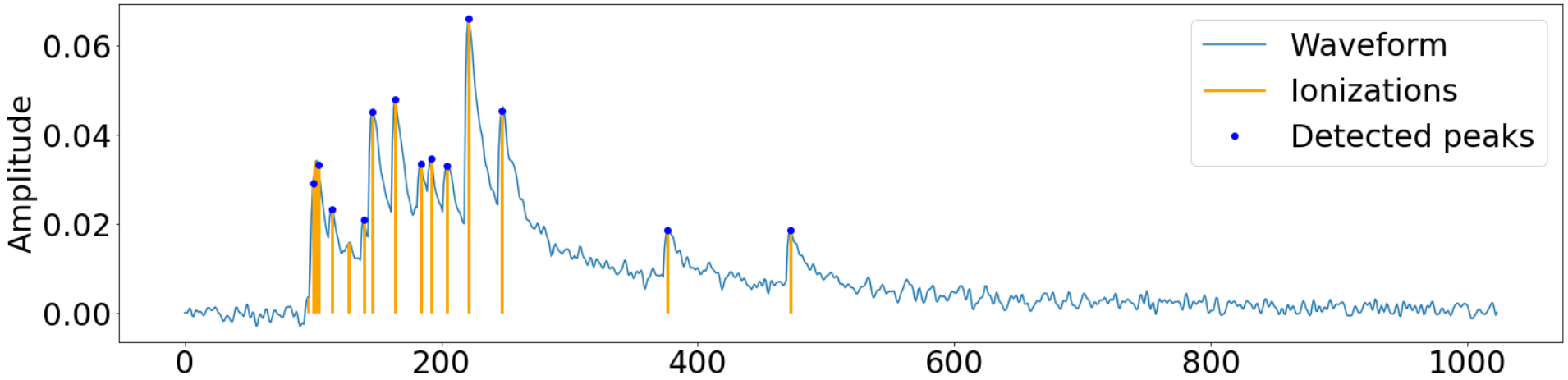


- **电离计数算法:** 从漂移室单元信号丝上产生的感应电流波形中获得电离簇团数N_{cls}
- 原初电离和次级电离均对波形有贡献，目标是得到原初电离对应的N_{cls}
- 两步走: **寻峰 (Peak finding)** & **合并 (clusterization)**
- 可以使用成熟的机器学习工具：TensorFlow, Keras, PyTorch, …

Workflow



⬇ **寻峰:** 从波形中找出电离信号峰



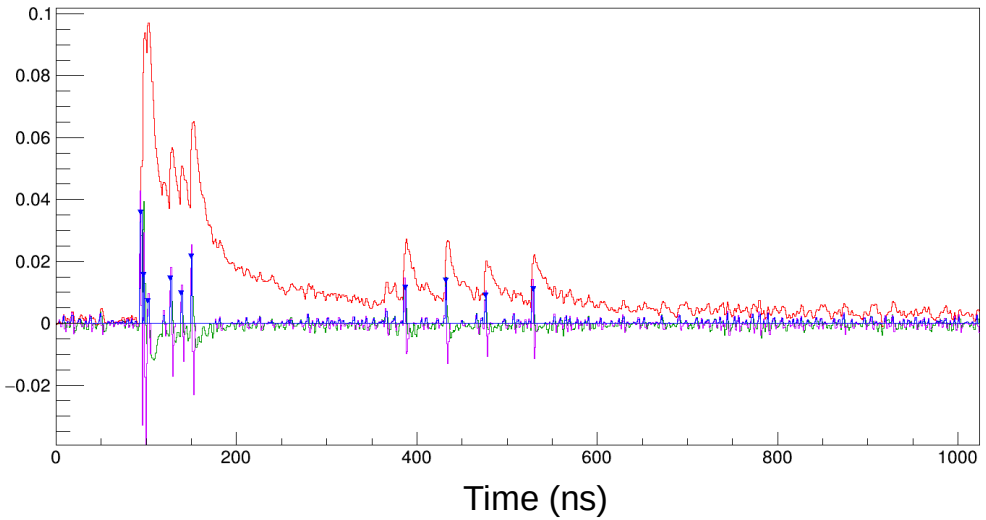
⬇ **合并:** 获得电离簇团数 N_{cls}

N_{cls}: 电离簇团数/原初电离数

传统方法：导数寻峰+时间差合并

寻峰（Peak finding）

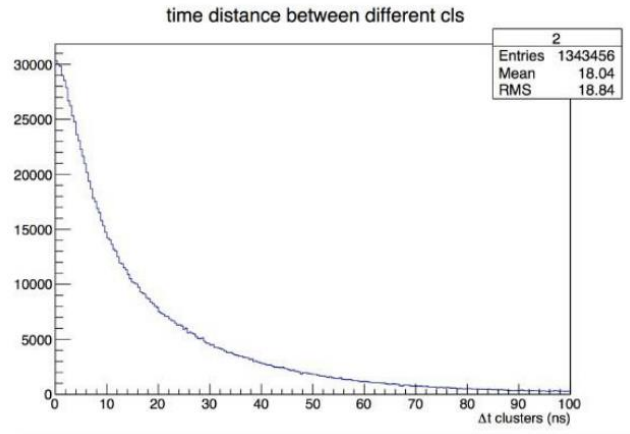
- 利用一阶导数和二阶导数
- 通过上升沿的斜率变化来寻峰

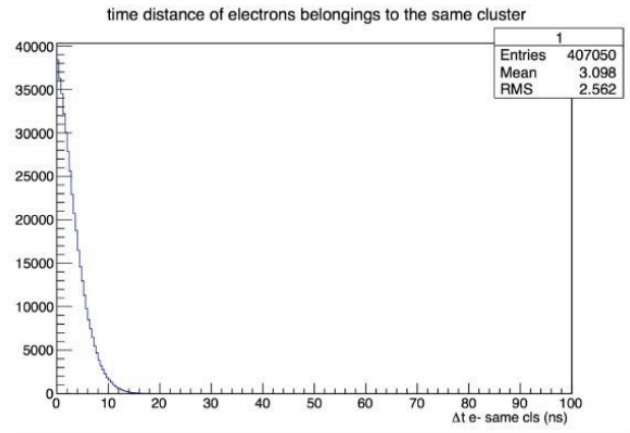


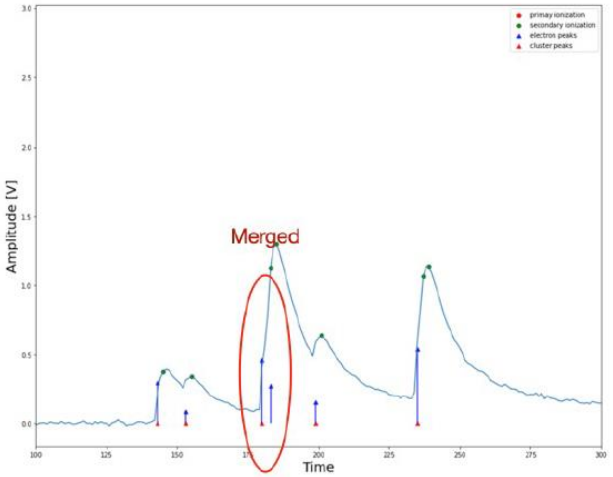
- **优点:** 快速、高效
- **缺点:** 难以应对噪声和重叠的信号峰

合并（Clusterization）

- 通过Δt合并电离峰
- 通过MC样本估计合并的cut条件





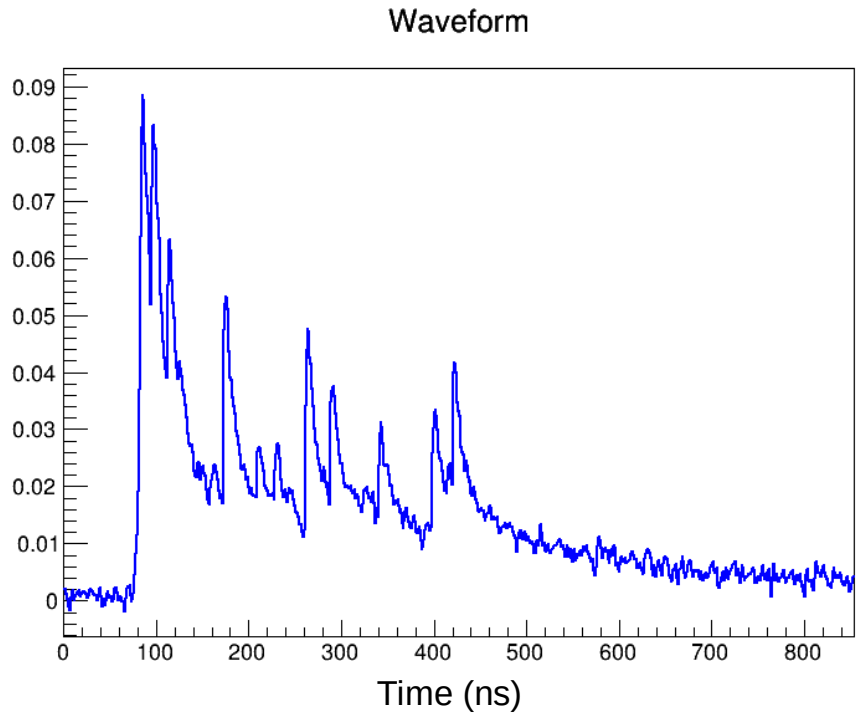


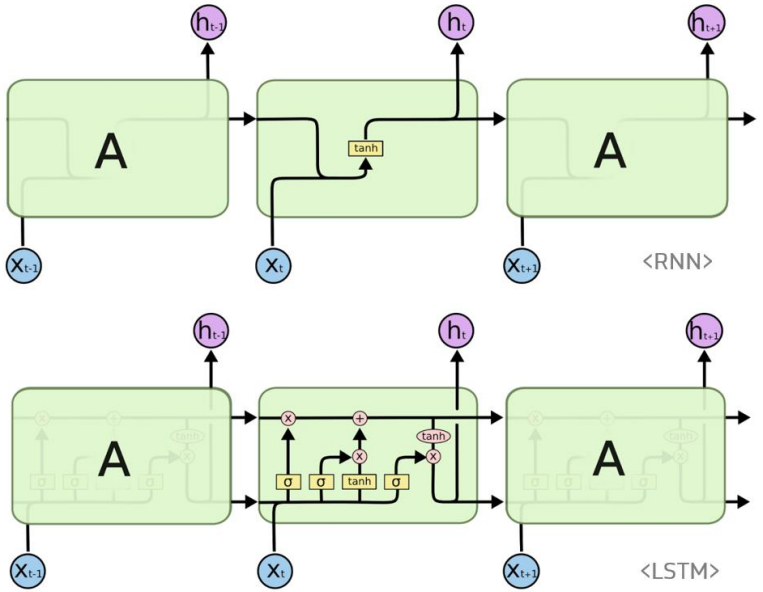
使用机器学习的理由

- **传统算法:** 人类给出规则，让机器根据规则来达成目的
- **机器学习:** 机器从大量数据中学习规则
- 对于电离计数:
  - 机器学习可以使用波形的全信息，而非像导数算法那样只使用导数的上升沿。
  - 机器学习或许可以学习得到波形、信号峰、时间分布和N_{cls}之间的隐藏关系
  - 问题可以被建模为分类或回归问题⇒适用于成熟的机器学习工具，如TensorFlow、Keras、PyTorch

第一步：寻峰

- **寻峰 (Peak finding): 从波形中提取电离信号峰**





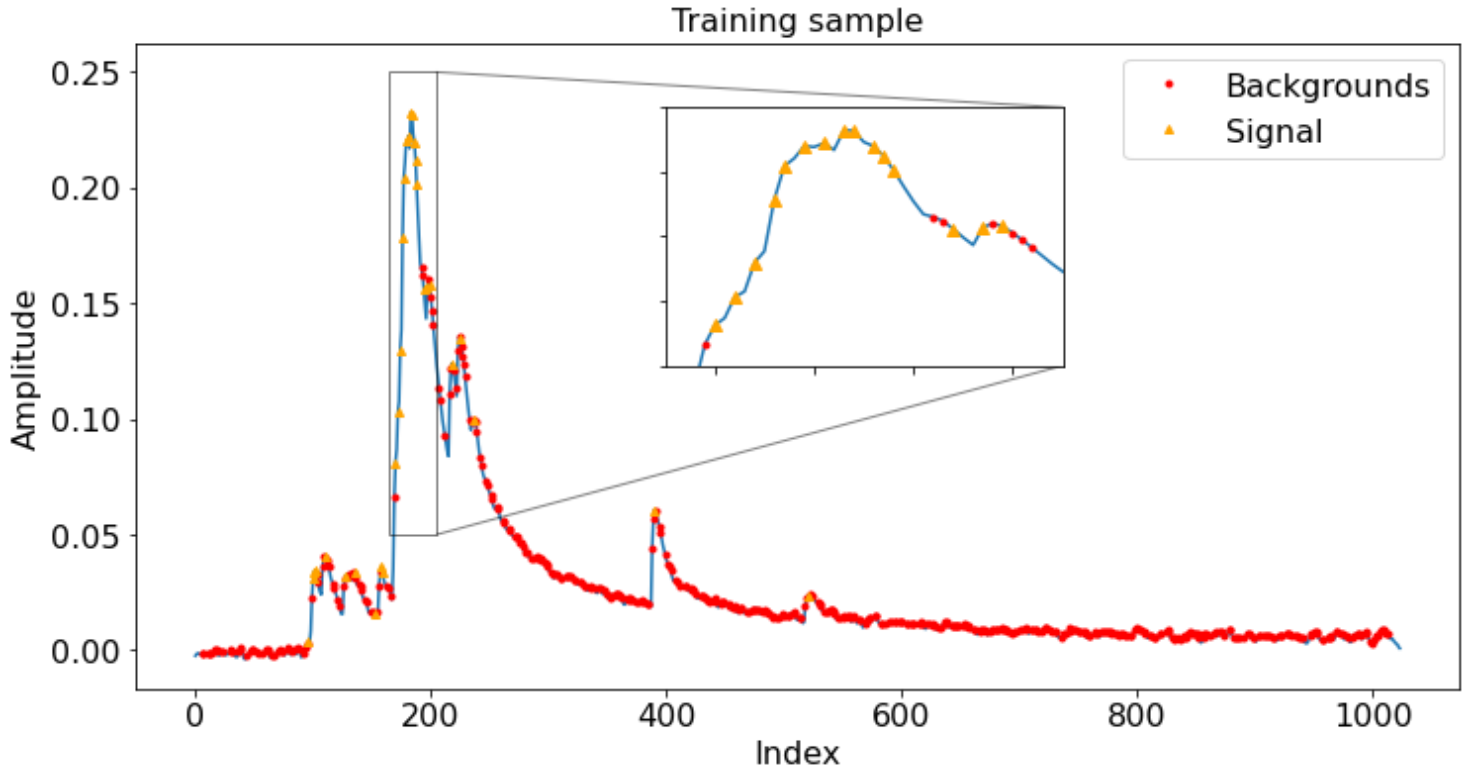
- **分类问题：**分类信号峰和噪声
- 波形数据是时序数据，适用于循环神经网络 (RNN)，尤其是长短期记忆网络 (LSTM)。

循环神经网络 (Recurrent Neural Network)

- 拥有反馈循环，有“记忆”功能。
- 最常用的一类RNN为拥有长期记忆功能的LSTM。

训练样本

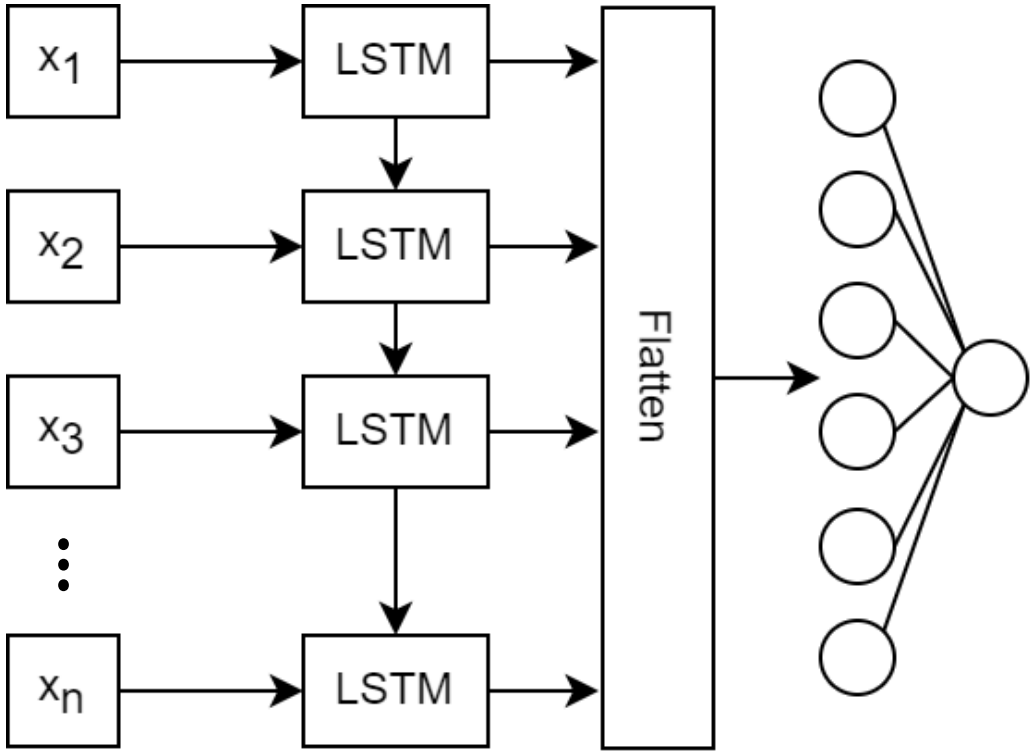
- **训练集：**Monte Carlo模拟产生的漂移室单元信号波形样本（π粒子，动量范围为0.2GeV~20GeV，5%噪声）



- 从波形中提取可能为信号峰的小波段作为输入样本
  - 小波段：全部斜率下降点附近的**(−5, +9)小波段**
  - 根据MC Truth对小波段打**信号/噪声**的标签
  - 数据集中噪声远多于信号，通过**欠采样**进行平衡

网络结构

- **网络结构: 长短期记忆 (LSTM) 模型**

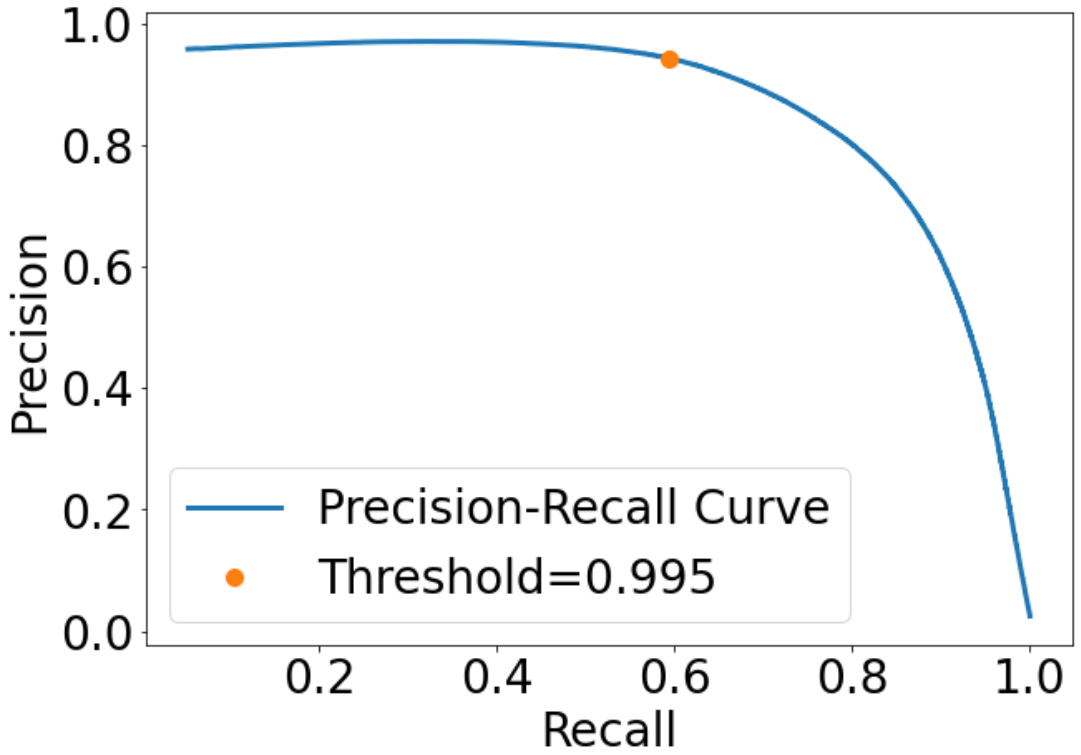


- 数据集: 波形中的候选信号峰（斜率降低点）
- 标签: 信号 / 本底
- 特征: 信号峰候选附近的15个点组成的小波段的时间 (x) 和振幅 (y)
- 损失函数: BCE loss

⇒ 二元分类问题

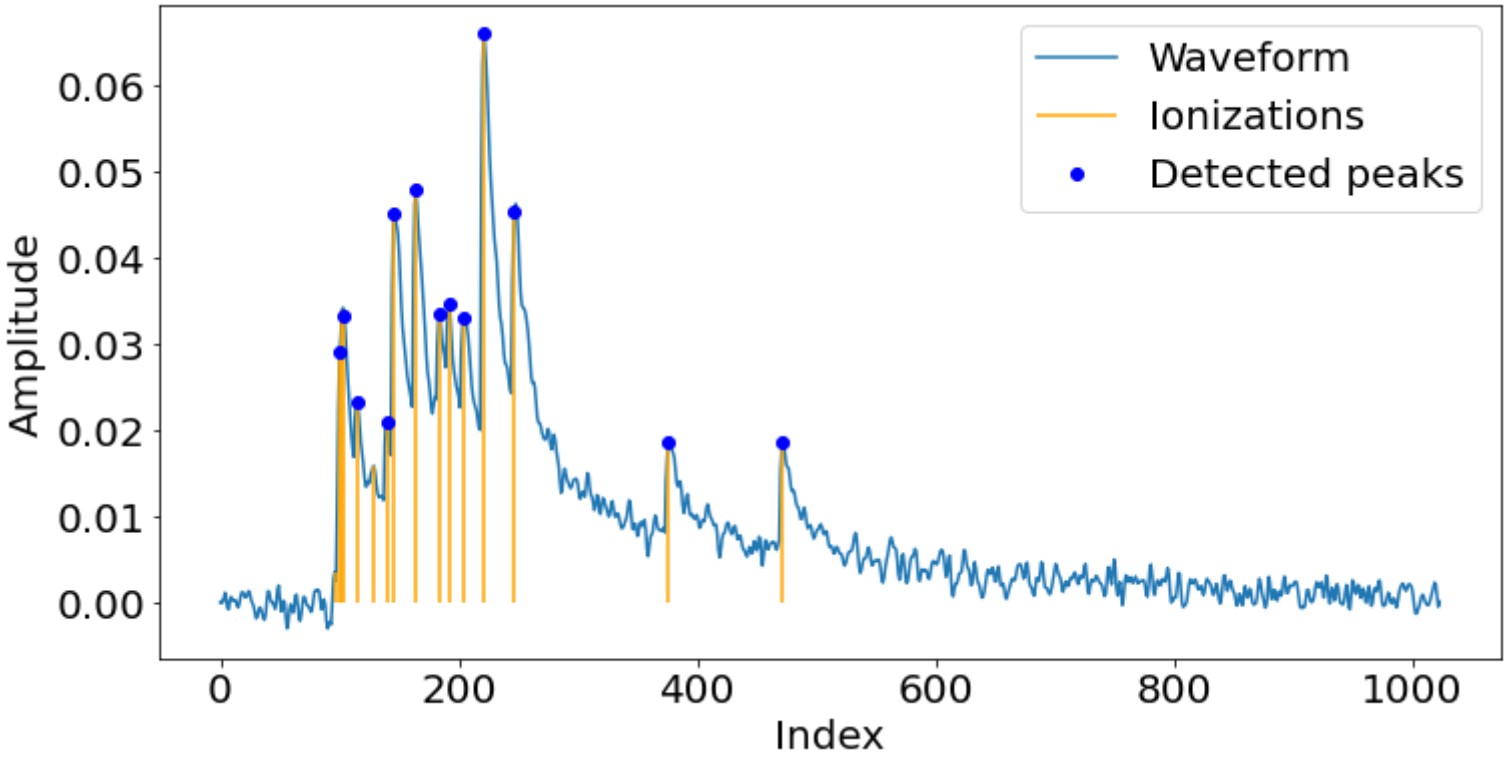
算法评估

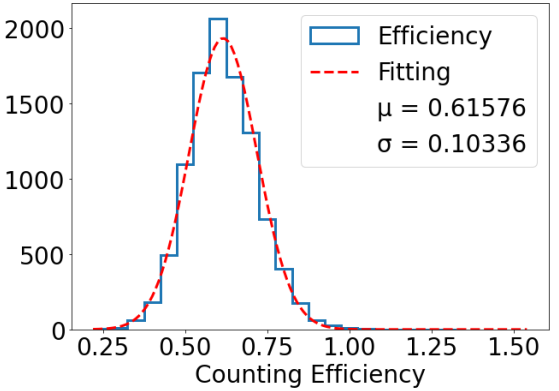
- 测试样本：信号波形样本（π粒子，动量范围为0.2GeV~20GeV，5%噪声）
- 效率 (Recall) = TP/(TP+FN) = 59.5%
- 纯度 (Precision) = TP/(TP+FP) = 94.3%
- 目标是得到信号峰，要求纯度尽量高 ⇒ 提高分类器的概率阈值以确保纯度。

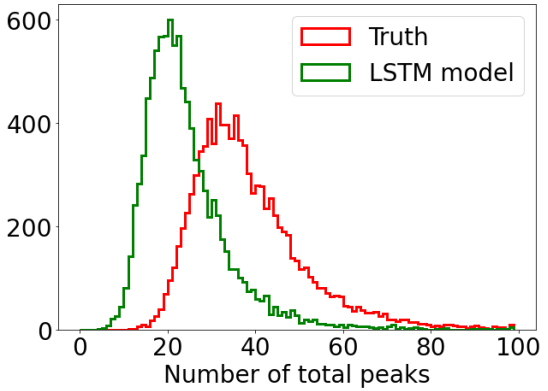


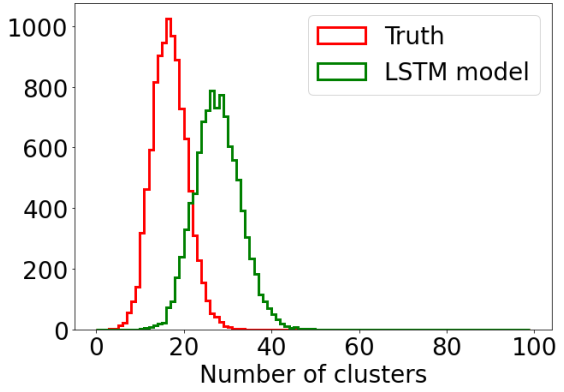
		真实	
		信号	噪声
预测	信号	TP	FP
	噪声	FN	TN

寻峰结果



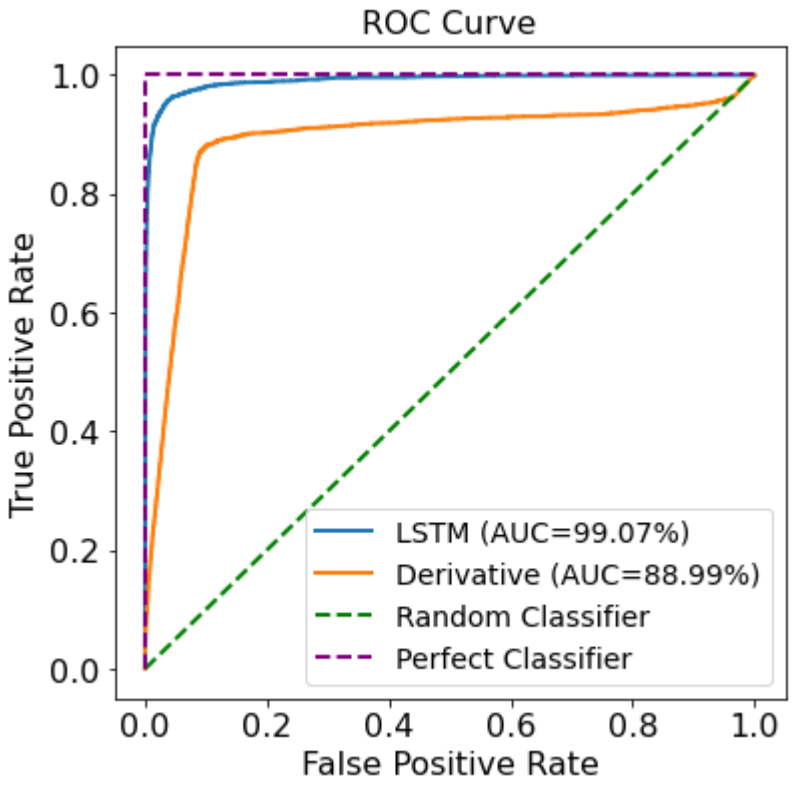




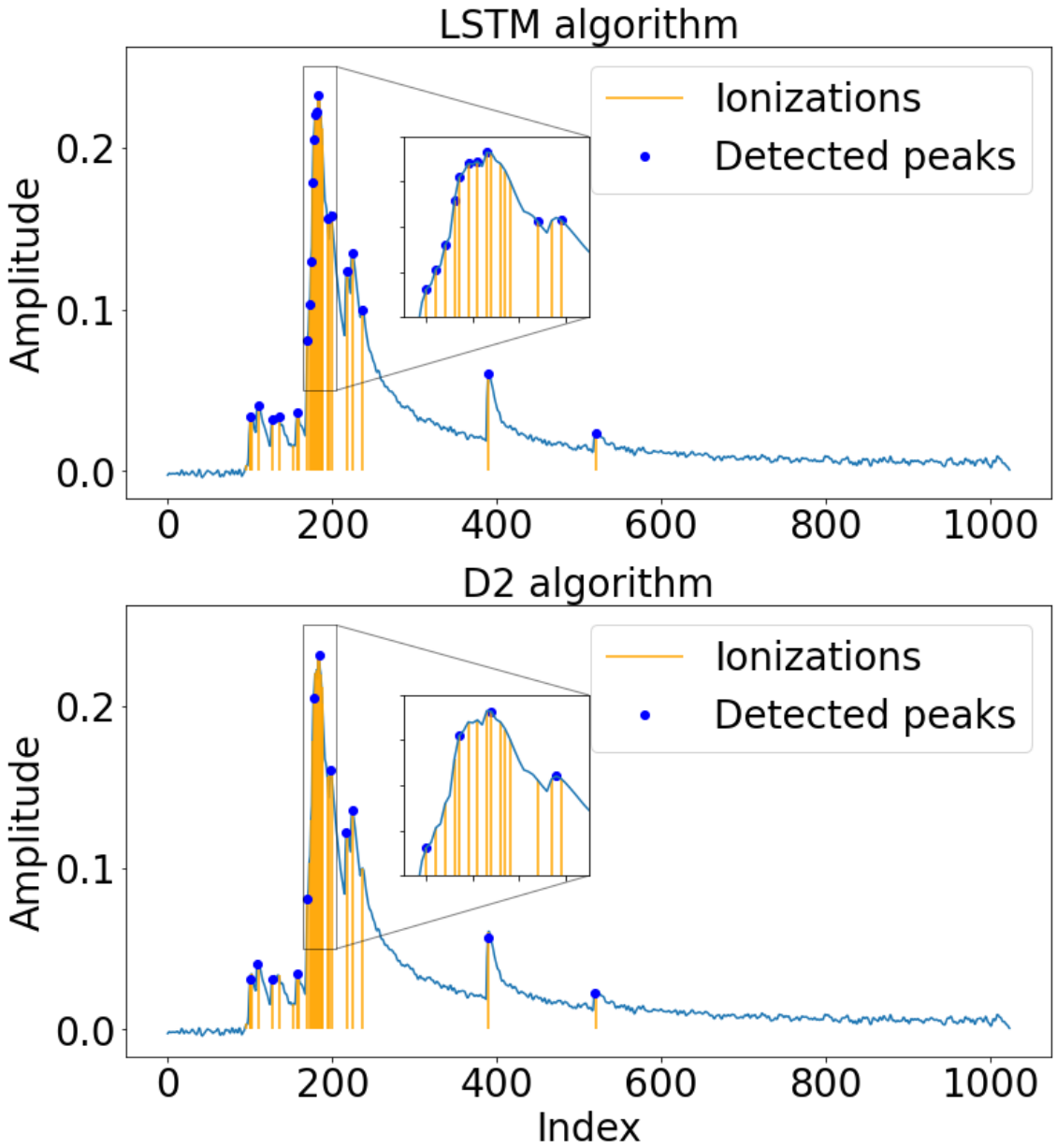


寻峰效率约为60%，对于不同动量的样本效率稳定。
寻峰后得到的原初电离分布仍有良好的泊松形状。

寻峰算法对比

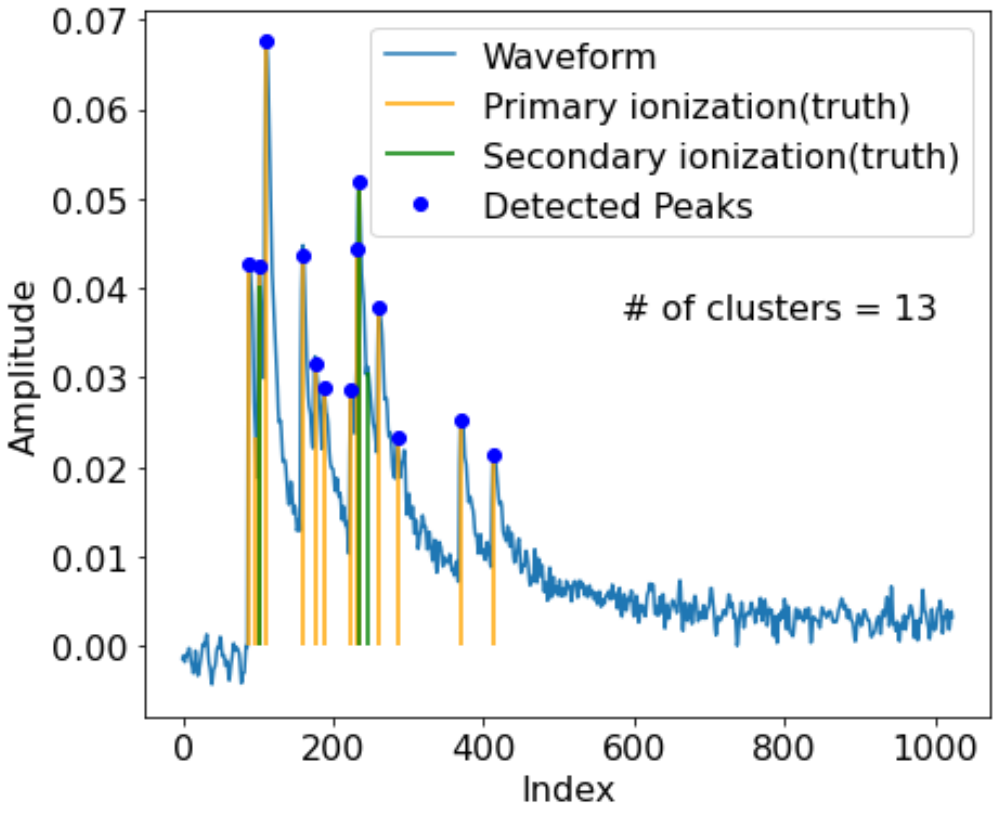


ROC曲线的曲线下面积（AUC）越高的分类器越好。LSTM寻峰算法对堆叠信号更灵敏，优于基于导数的寻峰算法。

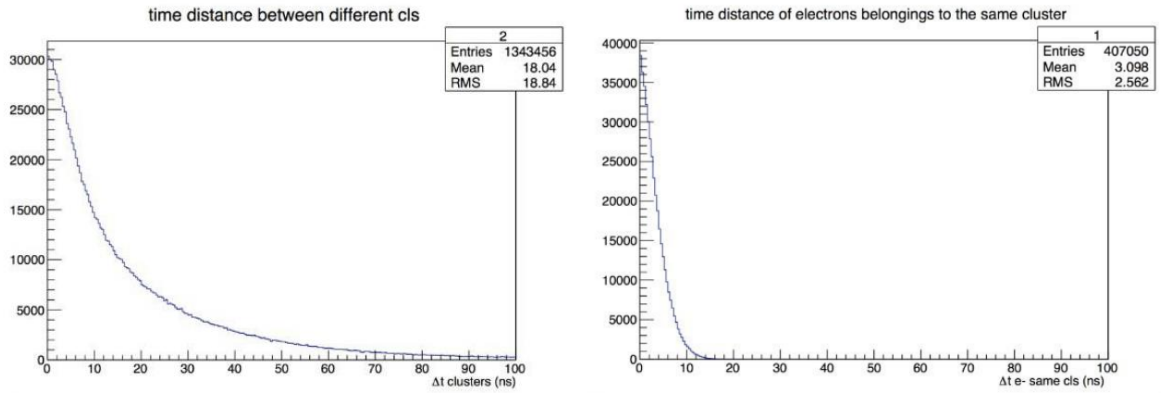


第二步：合并

- **合并 (Clusterization): 从寻峰算法得到的电离峰中得到N_{cls}**

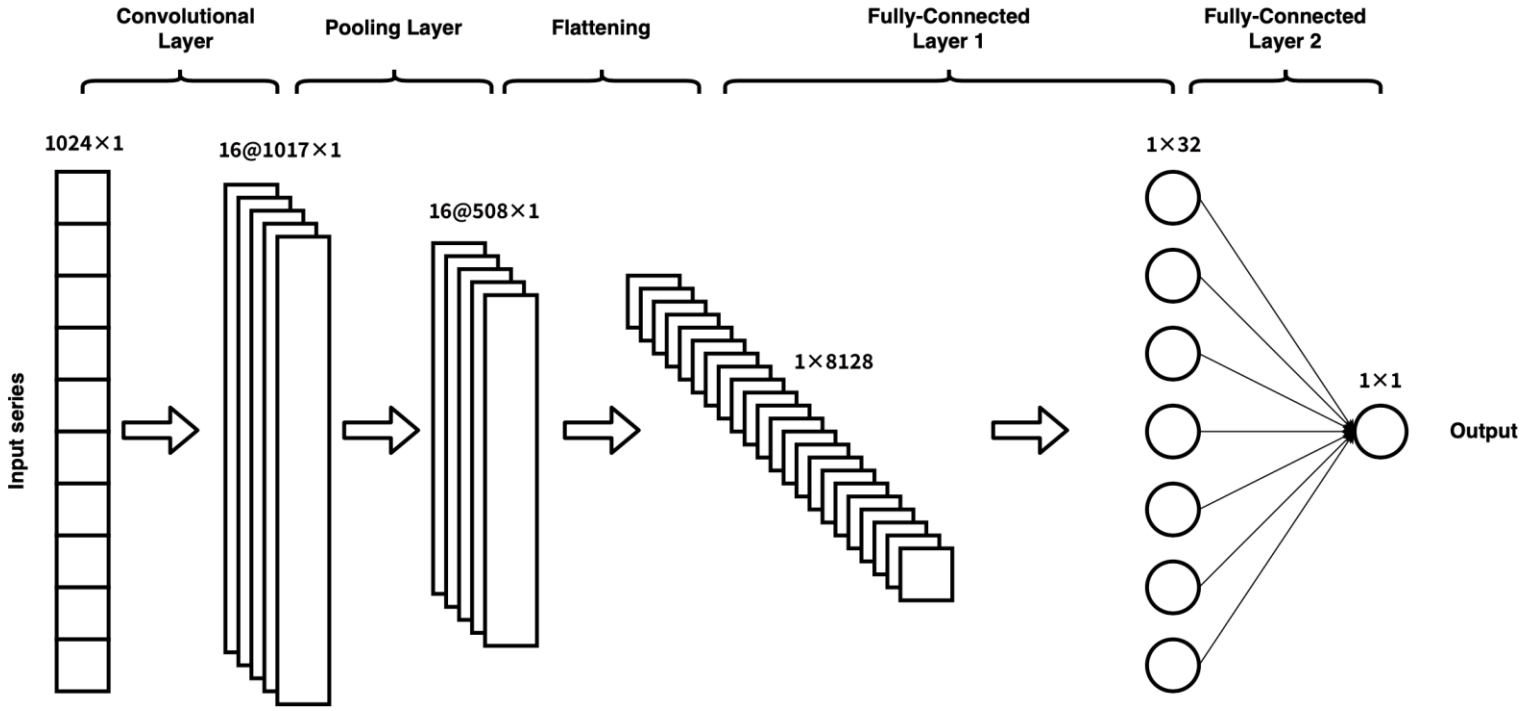


- 以寻峰算法得到的电离信号峰为训练样本。



- 目标：从寻峰得到的信号峰中提取N_{cls}
- 物理上的依据：电离峰之间的时间差
- 思路① CNN回归：
  - 通过信号峰分布预测N_{cls}
  - 输入数据结构：1D序列
- 思路② GNN分类：
  - 判断每个信号峰是否为原初电离，进一步求出原初电离数N_{cls}
  - 输入数据结构：图结构数据（点云）

思路①：回归 (CNN)



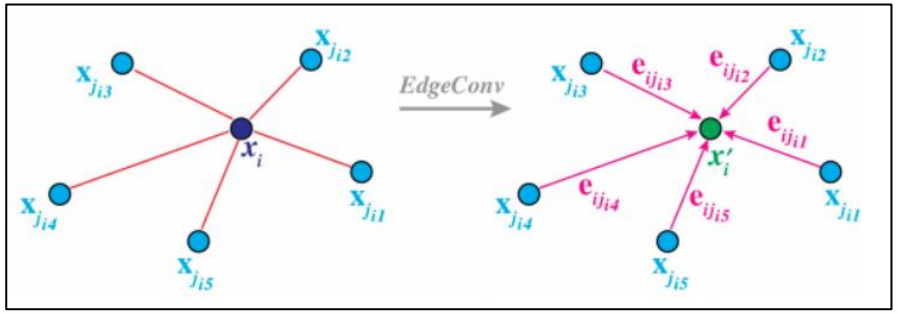
- 数据集 (Dataset)：寻峰算法处理后的波形
- 标签 (Label)：电离簇团数（原初电离数）N_{cls}
- 特征 (Feature)：形状为(n_channels, 1)的数组，数组中电离峰对应的位置的值为1，其余位置值为0
- 损失 (Loss): Mean Squared Error

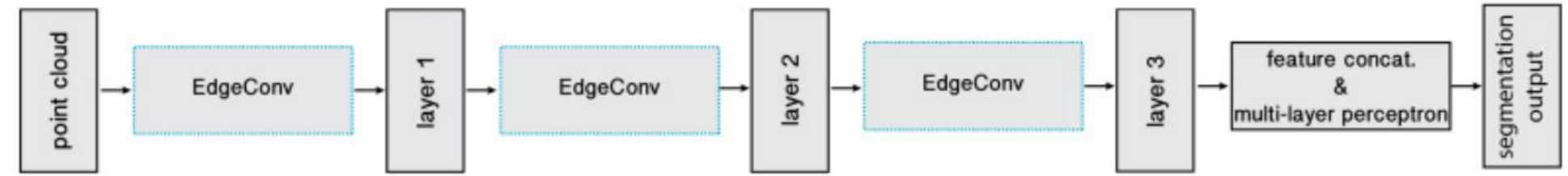
⇒ 回归问题

思路②：分类 (DGCNN)

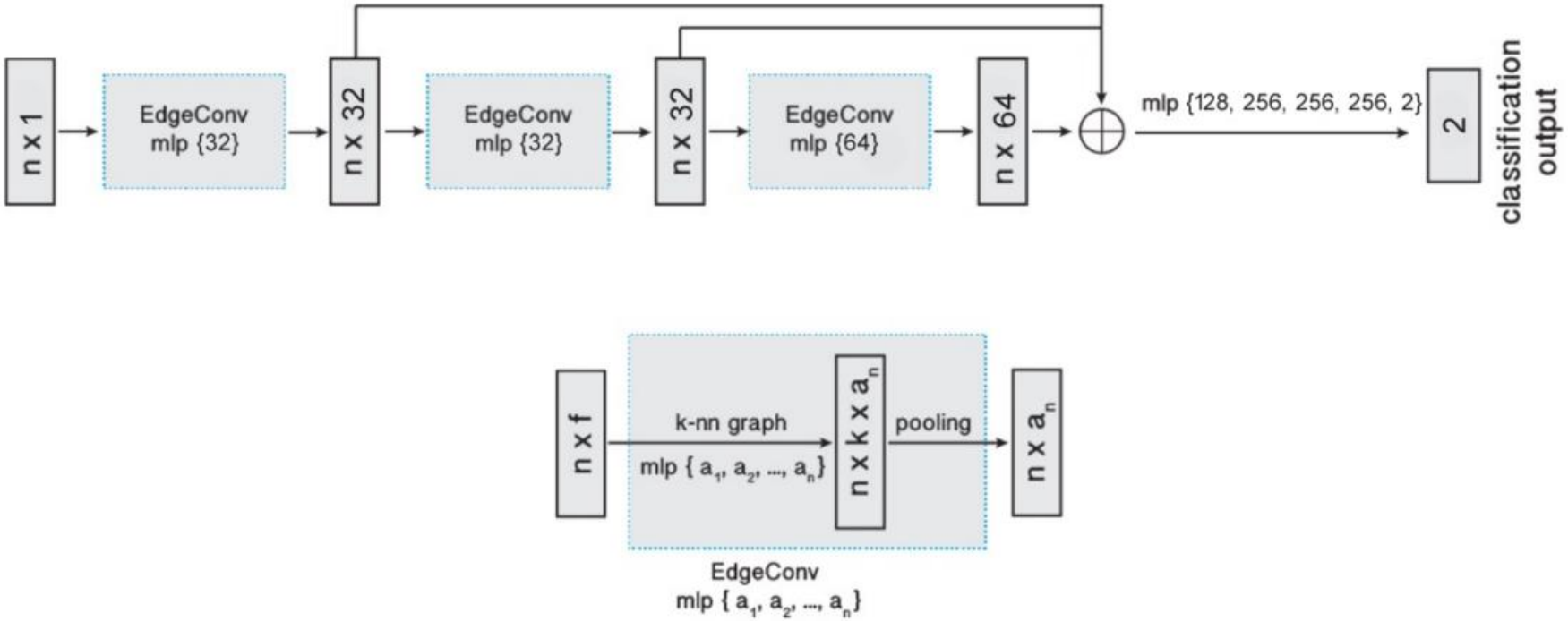
- 图神经网络 (Graph Neural Network, GNN)：
  - 捕捉图结构数据中节点之间的关系。
  - 事例中的一组电离峰⇒图 (Graph)，每个电离峰⇒节点 (Node)，电离峰的时间⇒节点特征 (Node feature)，电离峰之间的关系⇒边 (Edge)
  - 优点：无序、直观、更泛用
- 动态图卷积神经网络 (Dynamic Graph CNN, DGCNN)：
  - 通过k近邻在每一层特征空间动态建立图
  - 可以更好地捕捉局部特征⇒用以捕捉电离簇团特征
  - 已于高能物理领域获得应用⇒ParticleNet.

EdgeConv:





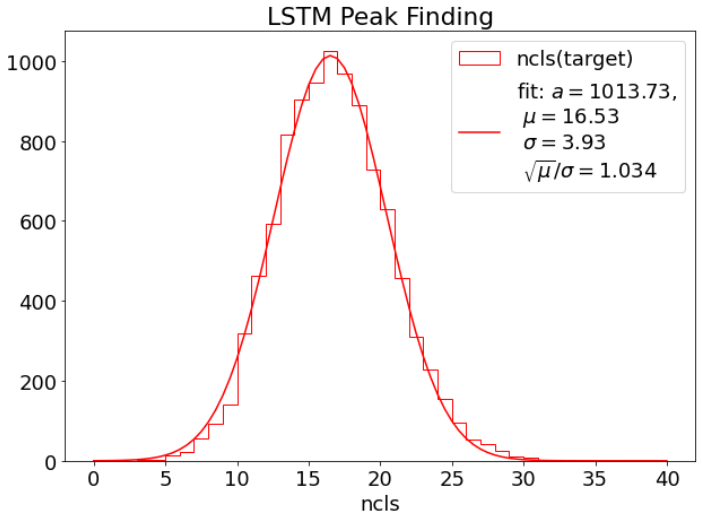
思路②：分类 (DGCNN)

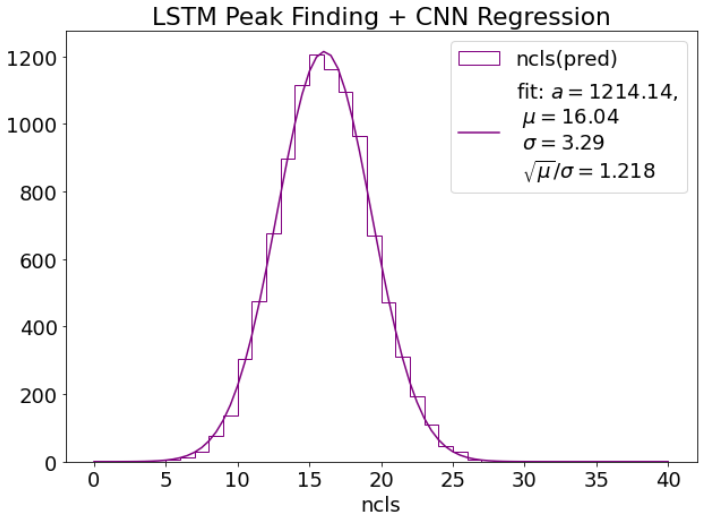


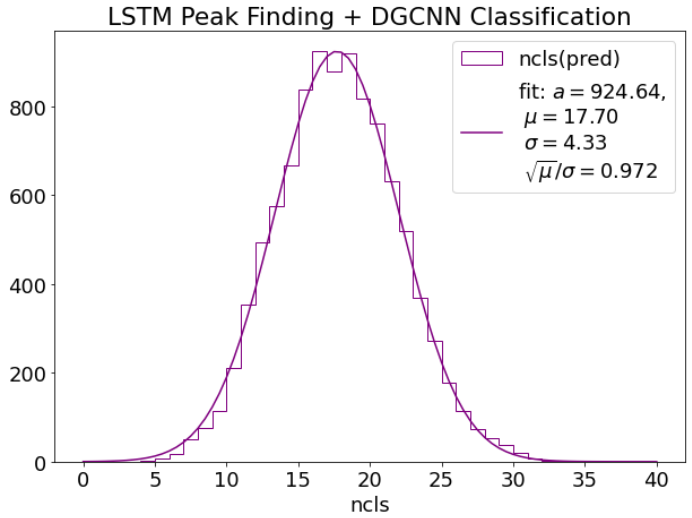
- 图 (Graph): 每个波形对应一张图
- 节点 (Node): 寻峰算法得到的电离峰
- 节点特征 (Node feature): 电离峰的位置（时间）
- 边 (Edge): EdgeConv层动态计算
- 标签 (Label): 电离峰的种类（原初电离为1，非原初电离为0）
- 损失 (Loss)：BCE loss

⇒节点分类（Node classification）问题

算法评估





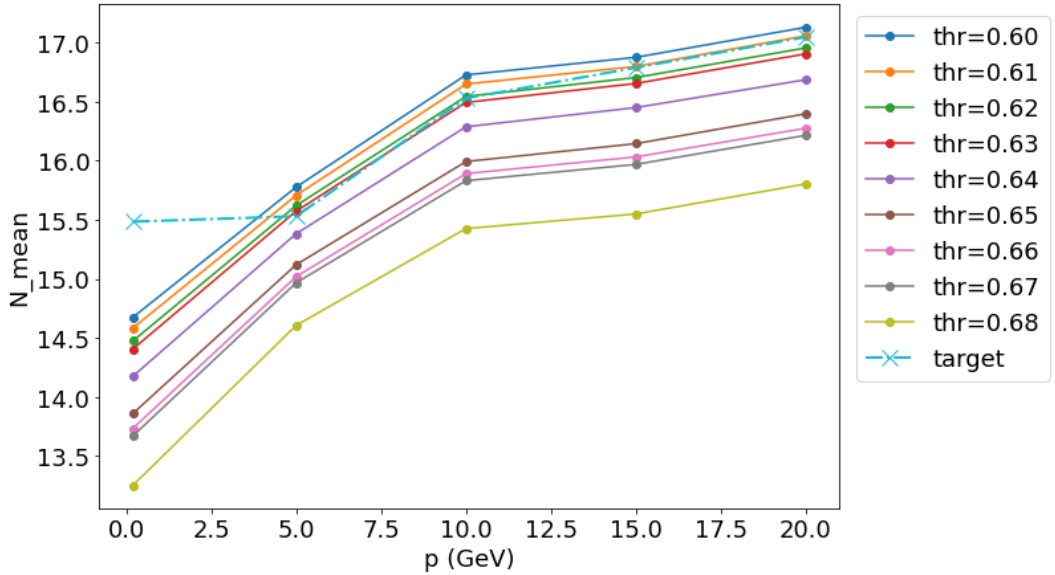


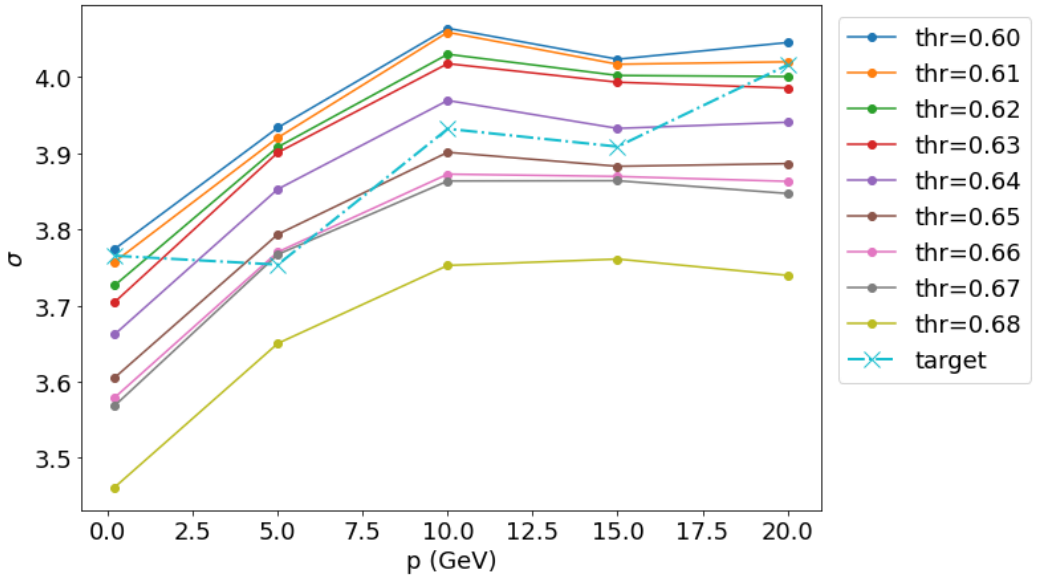
测试样本：信号波形样本（π粒子，动量为固定值，5%噪声）

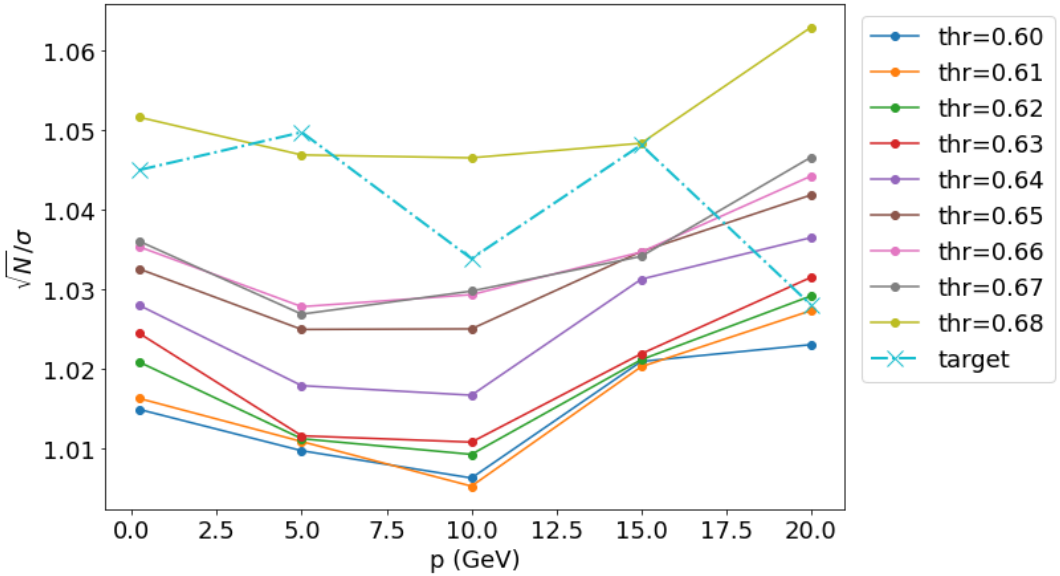
合并算法得到的N_{cls}即电离计数算法的最终结果。算法的评估基于最终得到的N_{cls}分布。N_{cls}应该遵守泊松分布，满足$\sqrt{N_{mean}}/\sigma \approx 1$。

回归法：分布变窄，中心值变低
分类法：分布变宽，中心值变高

DGCNN分类阈值优化

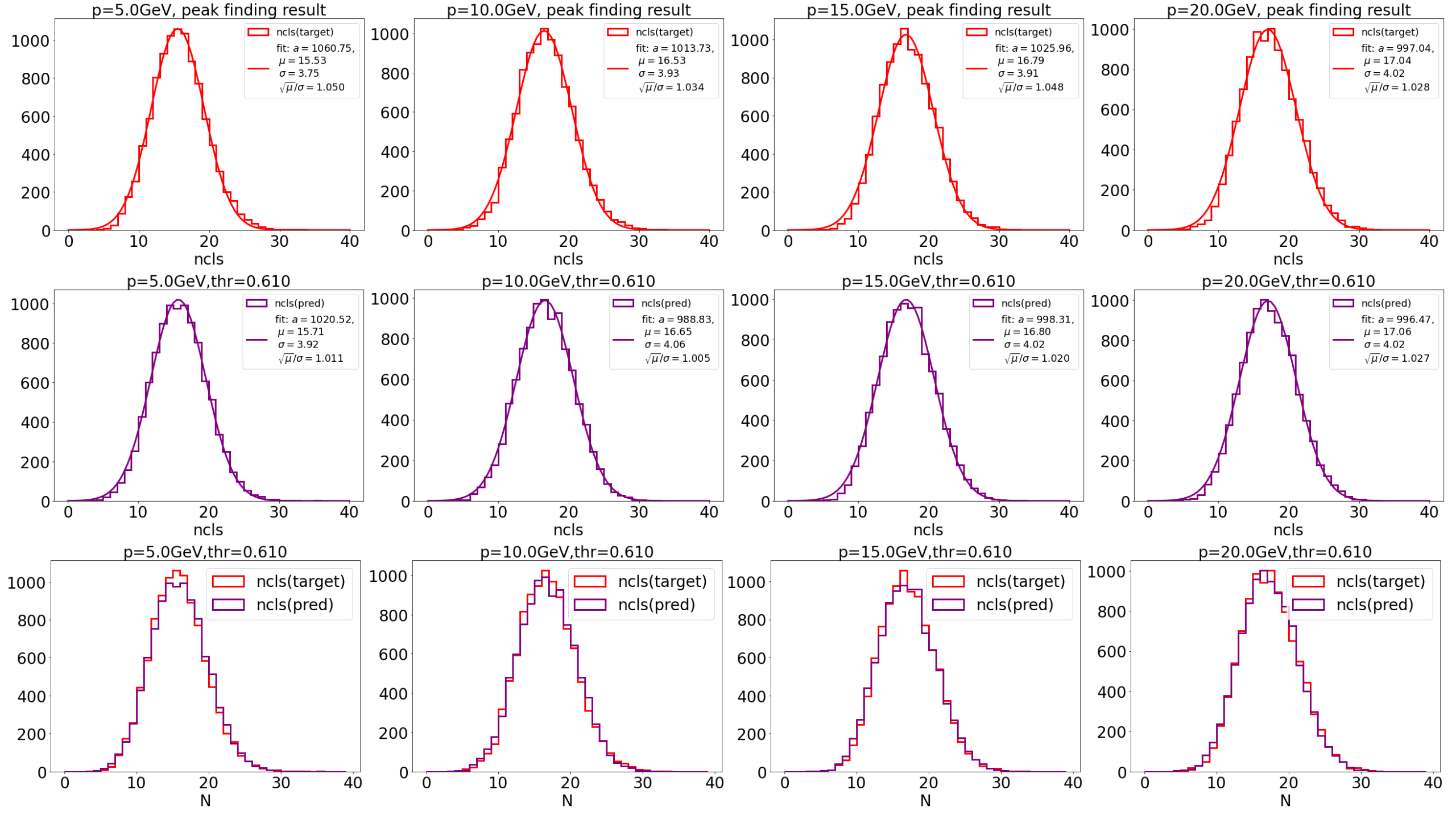




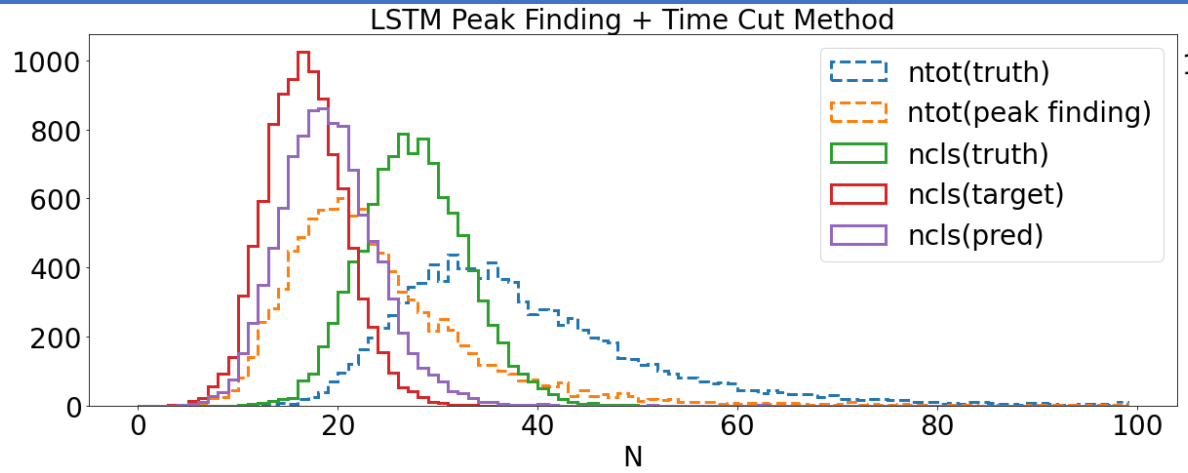


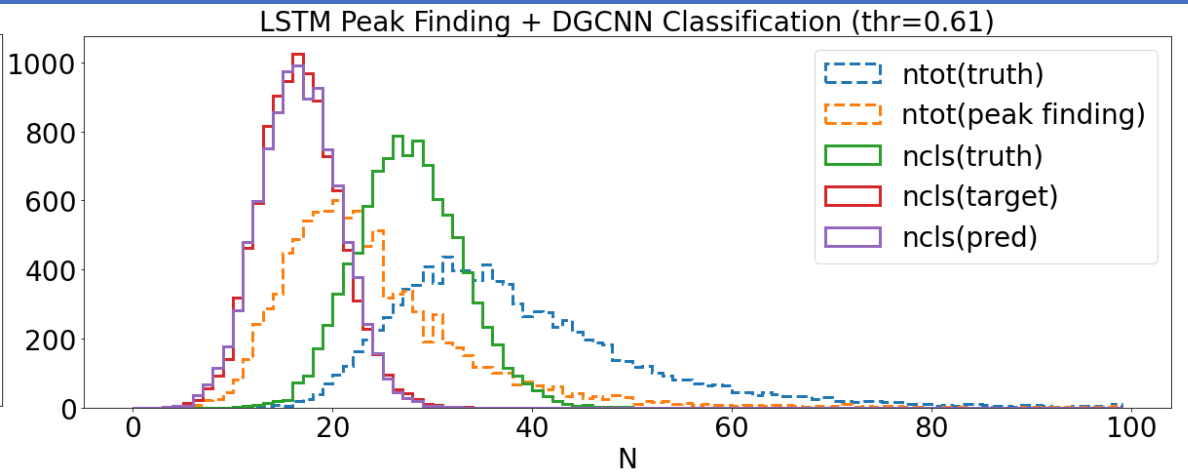
阈值优化：

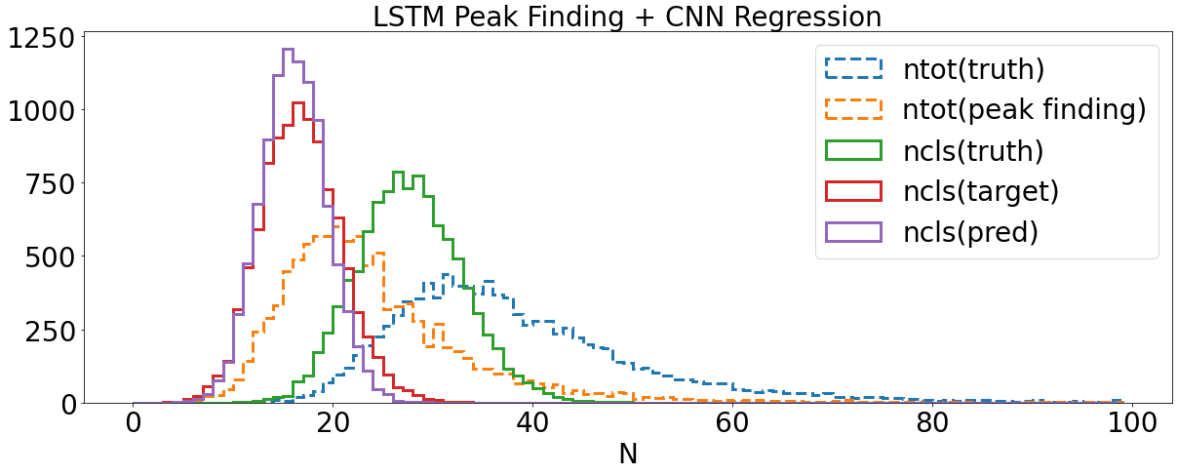
- 扫描阈值，并比较不同阈值下预测值和目标的原初电离总数N_{cls}分布差异。
- 评估参数：N_{mean}，σ，$\sqrt{N_{\mathrm{mean}}}/\sigma$

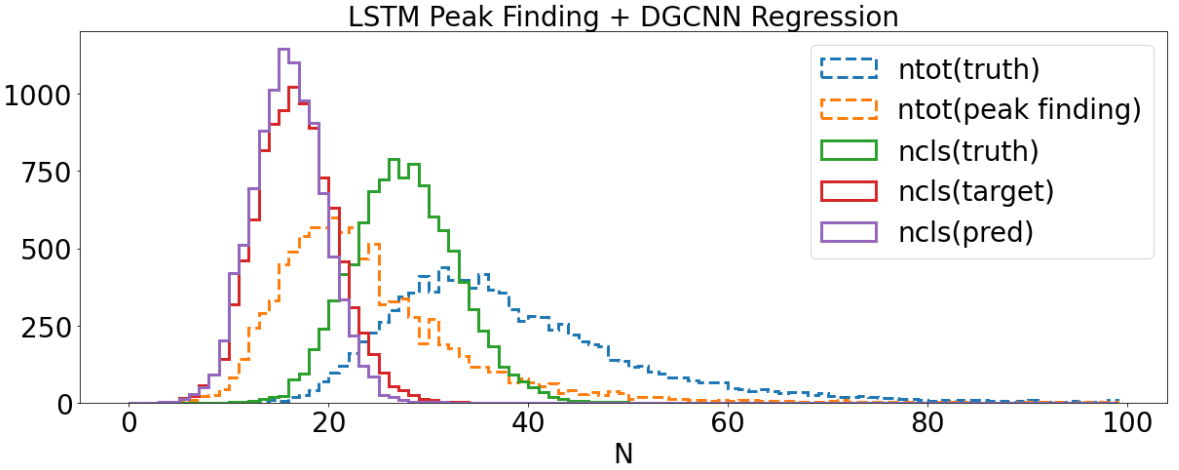

p=5.0GeV, peak finding result
ncls(target)
fit: $a = 1060.75$, $\mu = 15.53$ $\sigma = 3.75$ $\sqrt{\mu}/\sigma = 1.050$
p=10.0GeV, peak finding result
ncls(target)
fit: $a = 1013.73$, $\mu = 16.53$ $\sigma = 3.93$ $\sqrt{\mu}/\sigma = 1.034$
p=15.0GeV, peak finding result
ncls(target)
fit: $a = 1025.96$, $\mu = 16.79$ $\sigma = 3.91$ $\sqrt{\mu}/\sigma = 1.048$
p=20.0GeV, peak finding result
ncls(target)
fit: $a = 997.04$, $\mu = 17.04$ $\sigma = 4.02$ $\sqrt{\mu}/\sigma = 1.028$
p=5.0GeV,thr=0.610
ncls(pred)
fit: $a = 1020.52$, $\mu = 15.71$ $\sigma = 3.92$ $\sqrt{\mu}/\sigma = 1.011$
p=10.0GeV,thr=0.610
ncls(pred)
fit: $a = 988.83$, $\mu = 16.65$ $\sigma = 4.06$ $\sqrt{\mu}/\sigma = 1.005$
p=15.0GeV,thr=0.610
ncls(pred)
fit: $a = 998.31$, $\mu = 16.80$ $\sigma = 4.02$ $\sqrt{\mu}/\sigma = 1.020$
p=20.0GeV,thr=0.610
ncls(pred)
fit: $a = 996.47$, $\mu = 17.06$ $\sigma = 4.02$ $\sqrt{\mu}/\sigma = 1.027$
p=5.0GeV,thr=0.610
p=10.0GeV,thr=0.610
p=15.0GeV,thr=0.610
p=20.0GeV,thr=0.610
ncls(target)
ncls(pred)
ncls
N

结果对比



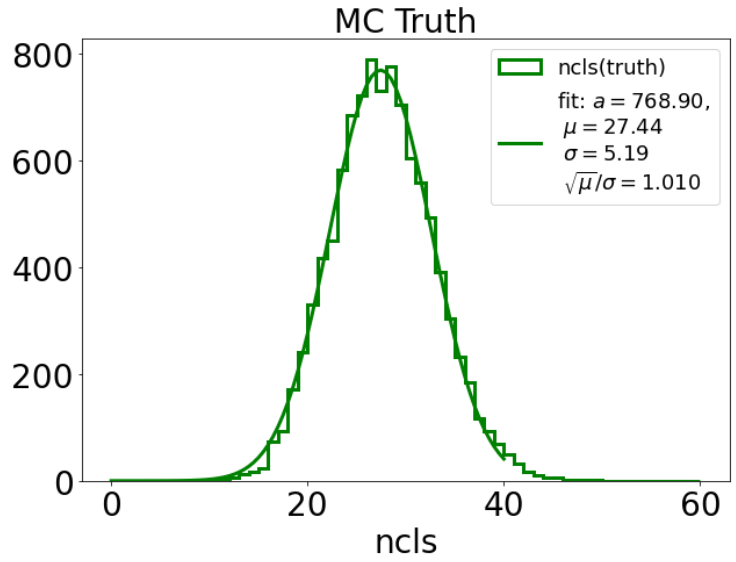


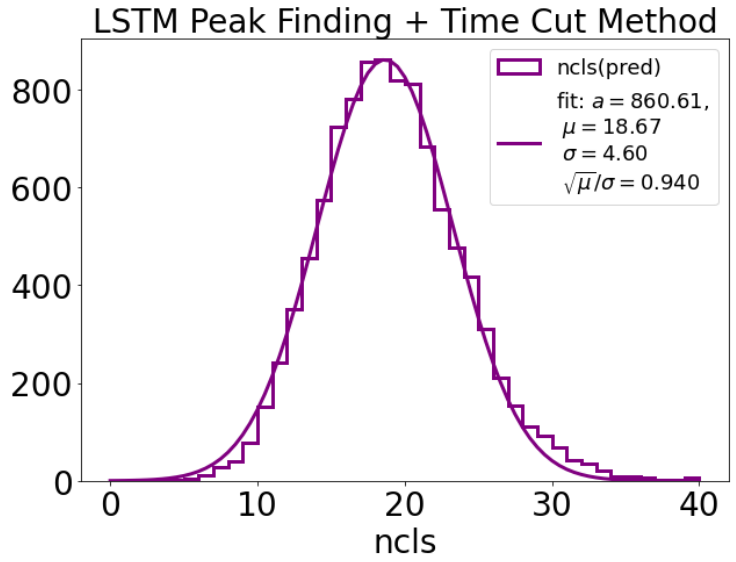


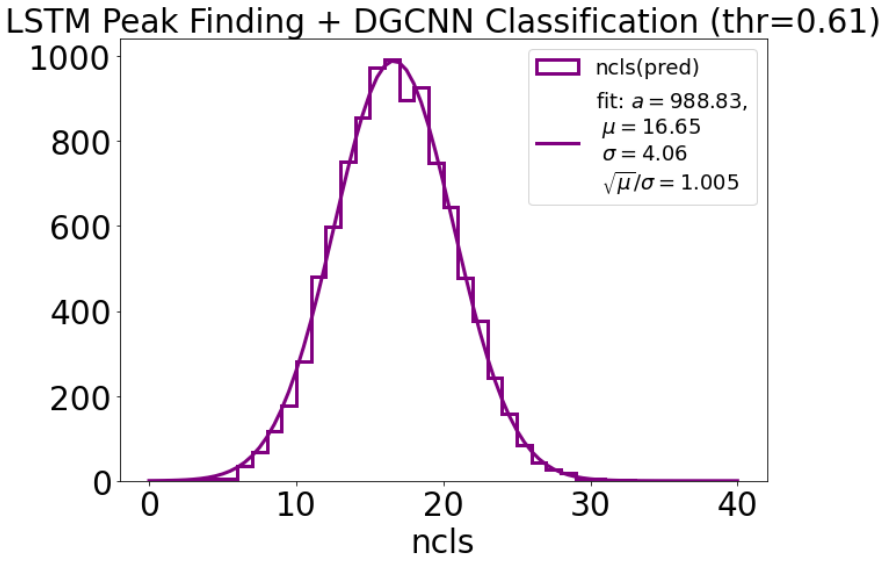


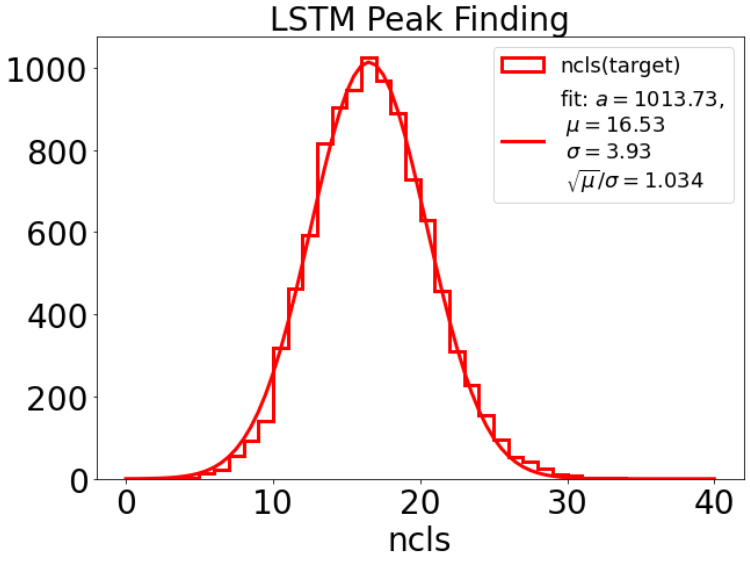
ntot(truth)：MC truth中总电离数；
ntot(peak finding)：寻峰算法得到的总电离数；
ncls(truth)：寻峰算法得到的总电离数；
ncls(target)：寻峰算法得到的总电离数中的原初电离数，合并算法的目标值；
ncls(pred)：合并算法预测得到的原初电离数。

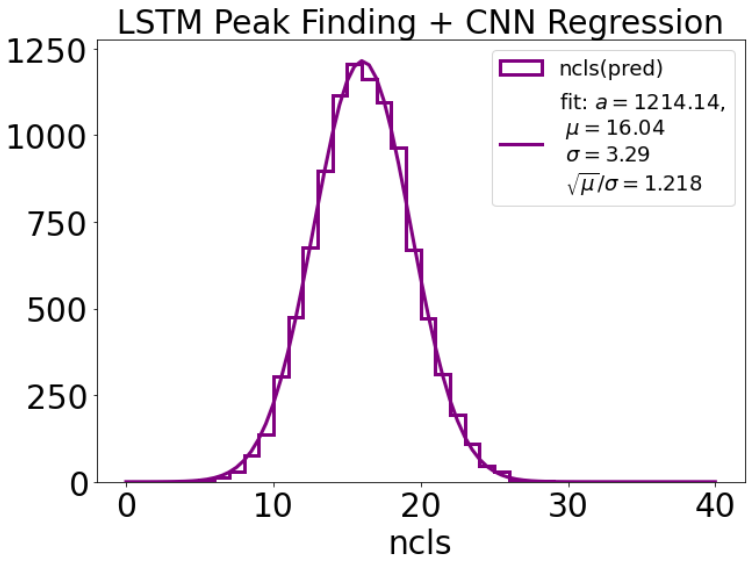
结果对比

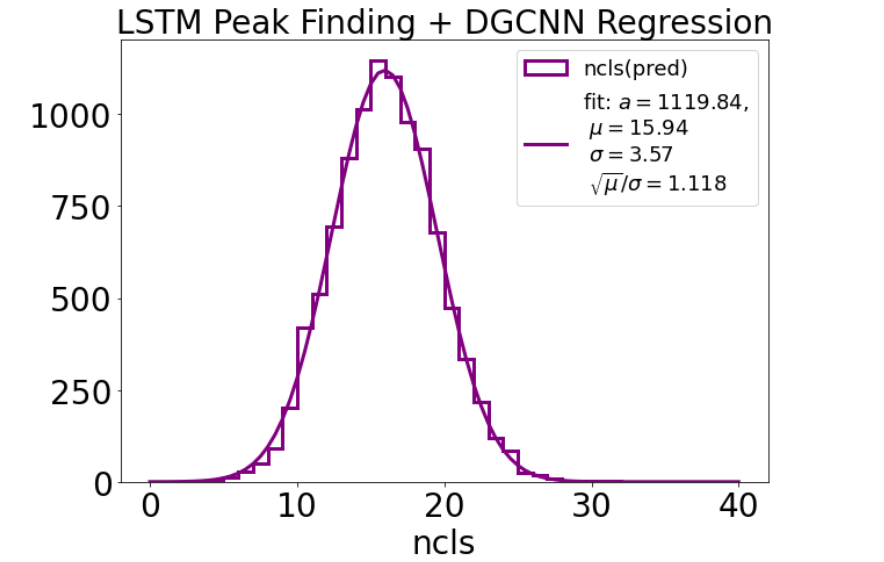












相比回归方法和传统方法，DGCNN分类算法获得了与目标值最接近的N_{cls}分布。

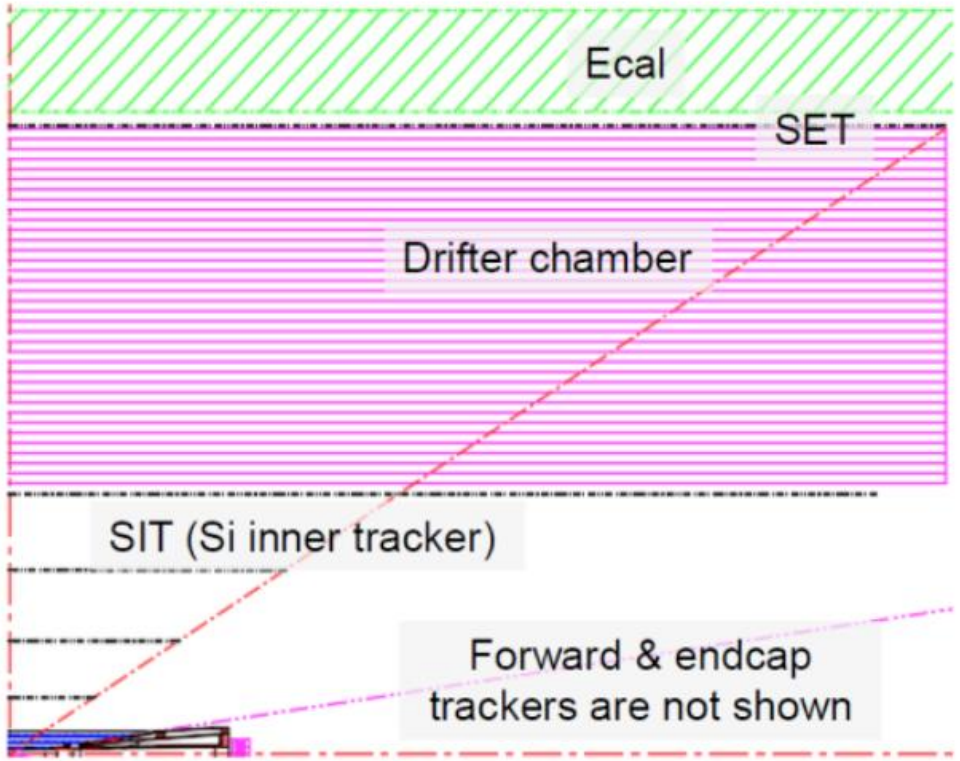
总结与展望

Method	N_{mean}	σ	σ/N_{mean}
Target (MC Truth)	**16.53**	**3.93**	**23.8%**
Classical Method	18.67	4.60	24.6%
CNN Regression	16.04	3.29	20.5%
DGCNN Regression	15.94	3.57	22.4%
DGCNN Classification (thr=0.61)	**16.65**	**4.06**	**24.4%**

- 基于深度学习的电离计数算法较传统算法有较大提升。
  - LSTM寻峰算法的寻峰纯度和灵敏度高于传统导数算法
  - DGCNN合并算法可以获得比传统时间合并算法更好的N_{cls}分布。
- 未来将基于beam test真实数据进行进一步的研究。

BACKUP

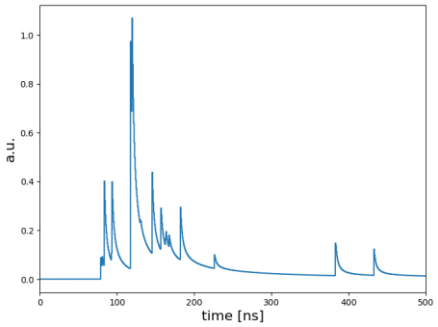
漂移室设计



Preliminary DC parameters	
Inner radius	800mm
Outer radius	1800mm
Cell size	18 mm ✕ 18 mm
Gas mixture	He/iC4H10=90:10
Length of outermost wires (cosθ=0.82)	5143mm

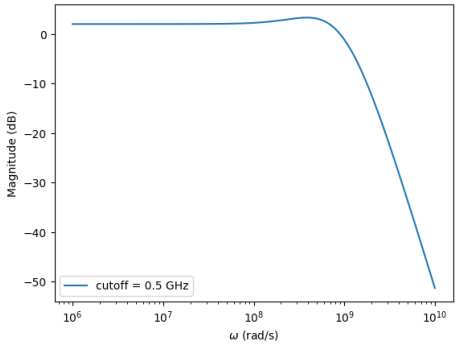
波形模拟

Induced signal



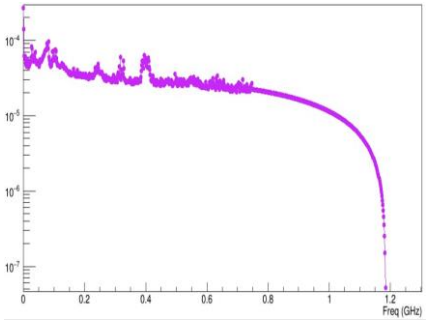
Fast simulation based on Garfield++ study

Electronics response



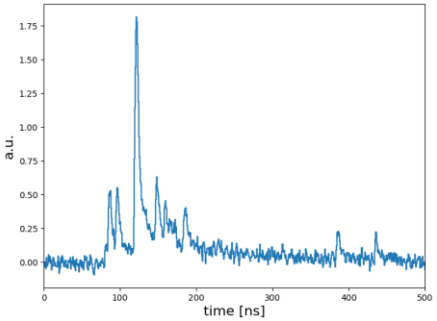
From electronics using in the beam test

Noise



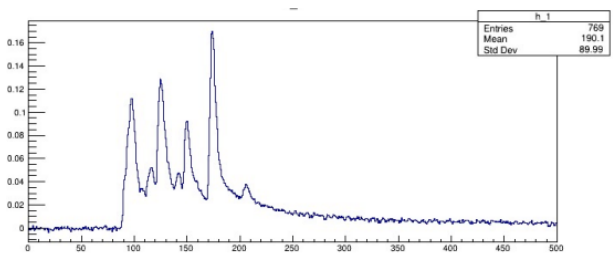
From the beam test

Waveform

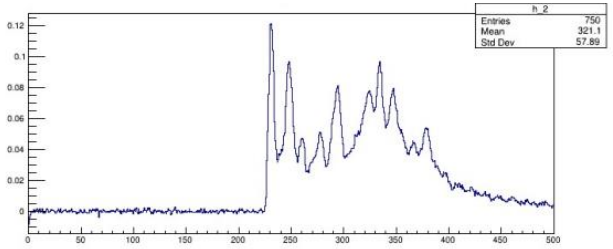


SamplingRate(GHz): 1.5,
TimeWindow(ns): 2000,
CellSize(cm): 1.8,
Particle:
  Name: pi,
  Pos(cm): [0.2, -1.2, 0.0],
  Cos: [0.0, 1.0, 0.0],
  Momentum(GeV/c): [0.2, 20.0]
PreAmplifier:
  ScaleFactor: 1.0,
  ResponseSize: 50,
  Option: 0,
  Opt0-SimpleAmp-Tau: 2.0,
Noise:
  NoiseAmplitude: 0.02405,
  NoiseType: FFT noise

Simulation



Beam test data

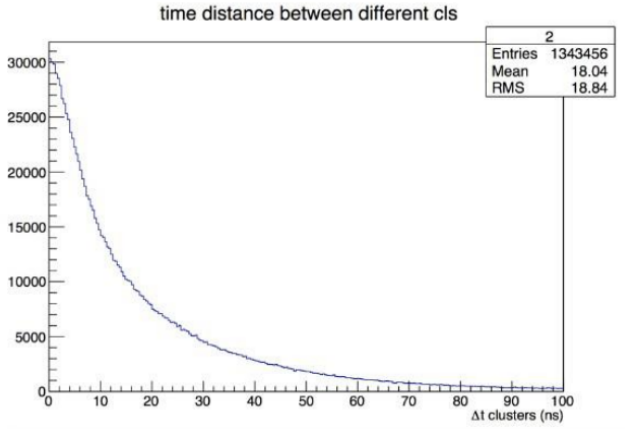


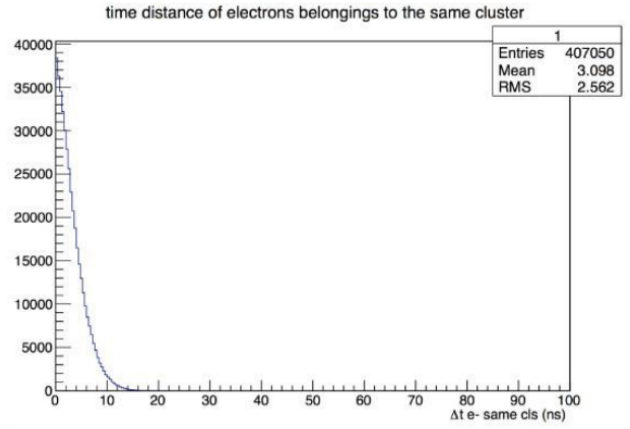
传统合并 (Clusterization) 算法

Convert peaks in cluster clusterization

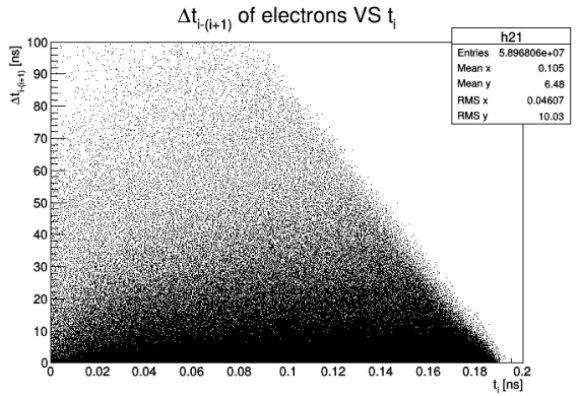
How to convert found peaks in clusters?

- Look to the time difference between electrons belonging to different clusters and those to the same cluster





- Event by event plot the difference $t_i - t_{i+1}$ Vs t_i



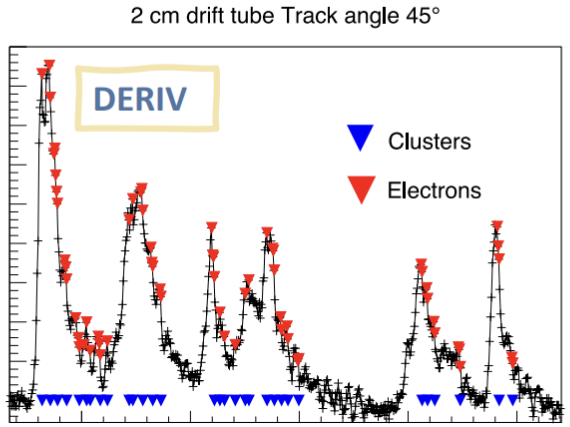
Source: [F. Cuna, Cluster counting techniques The beam test and the preliminary results, 2021]

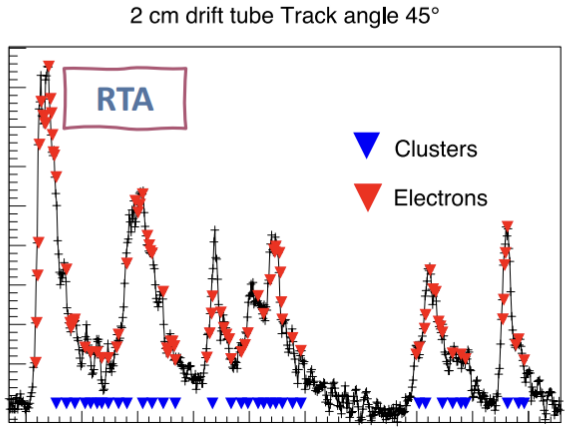
传统合并 (Clusterization) 算法

A single clusterization algorithm

Once find the electron peaks, clusterization of the electron peaks into ionization clusters has been implemented:

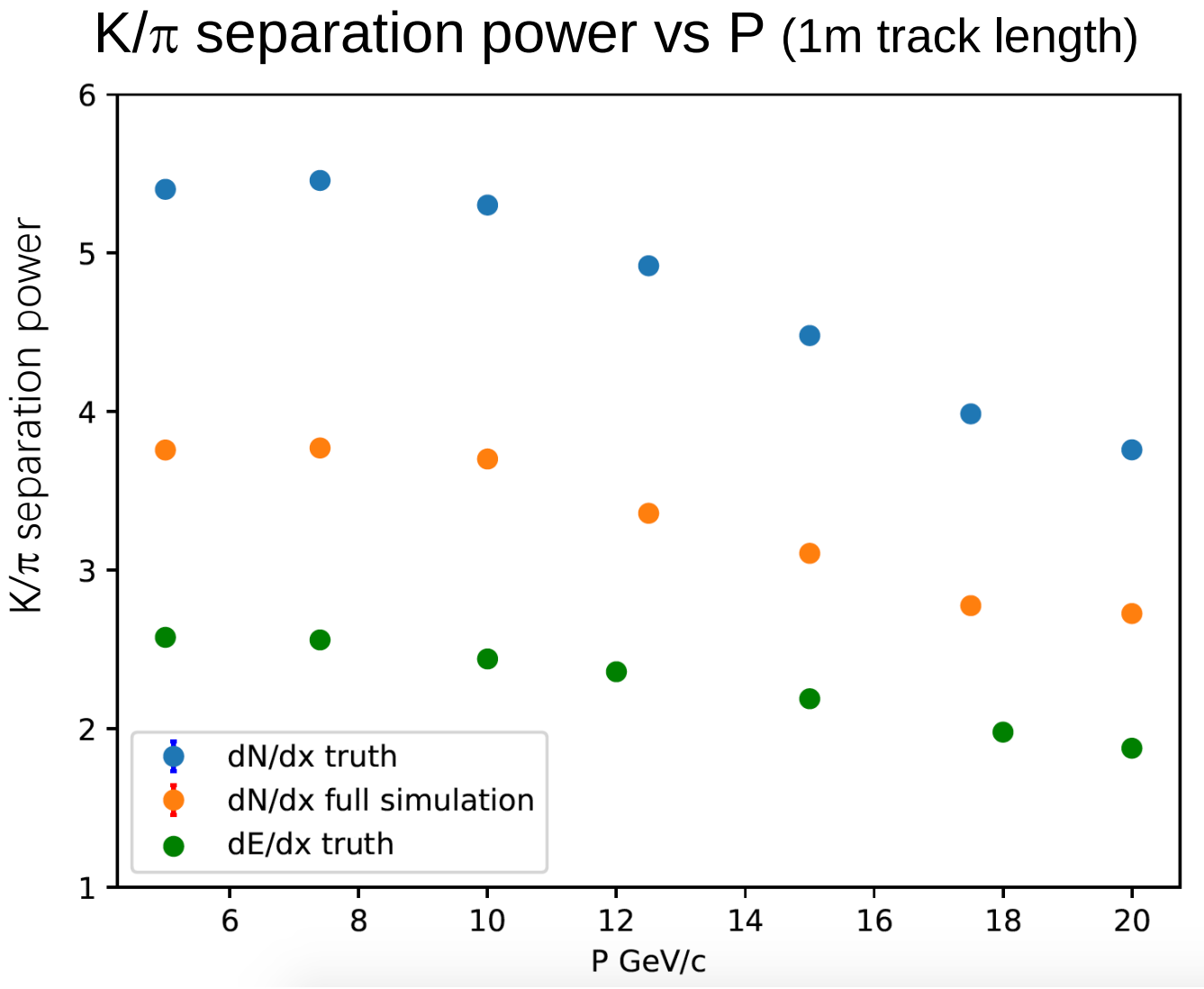
1) Association of electron peaks consisting in consecutive bins (difference in time == 1 bin) electrons to a single electron in order to eliminate fake electrons.
2) Contiguous electrons peaks which are compatible with the electrons diffusion time (2.5 ns or 3 bins) must be considered belonging to the same ionization cluster.
3) Position of the clusters is taken as the position of the last electron in the cluster.





Source: F. Cuna, Particle identification with the cluster counting technique in the IDEA drift chamber, 2022

基于传统ClusterCounting算法的初步结果



Separation power

$$S = \frac{\left|\left(\frac{dN}{dx}\right)_{\pi} - \left(\frac{dN}{dx}\right)_{K}\right|}{(\sigma_{\pi} + \sigma_{K})/2}$$