dN/dx Reconstruction with Machine Learning for Drift Chamber

Guang Zhao, Zhefei Tian, Linghui Wu, Mingyi Dong, Franco Grancagnolo, Nicola De Filippis, Muhammad Anwar, Gang Li, Xu Gao, Zhenyu Zhang, Shengsen Sun

zhaog@ihep.ac.cn

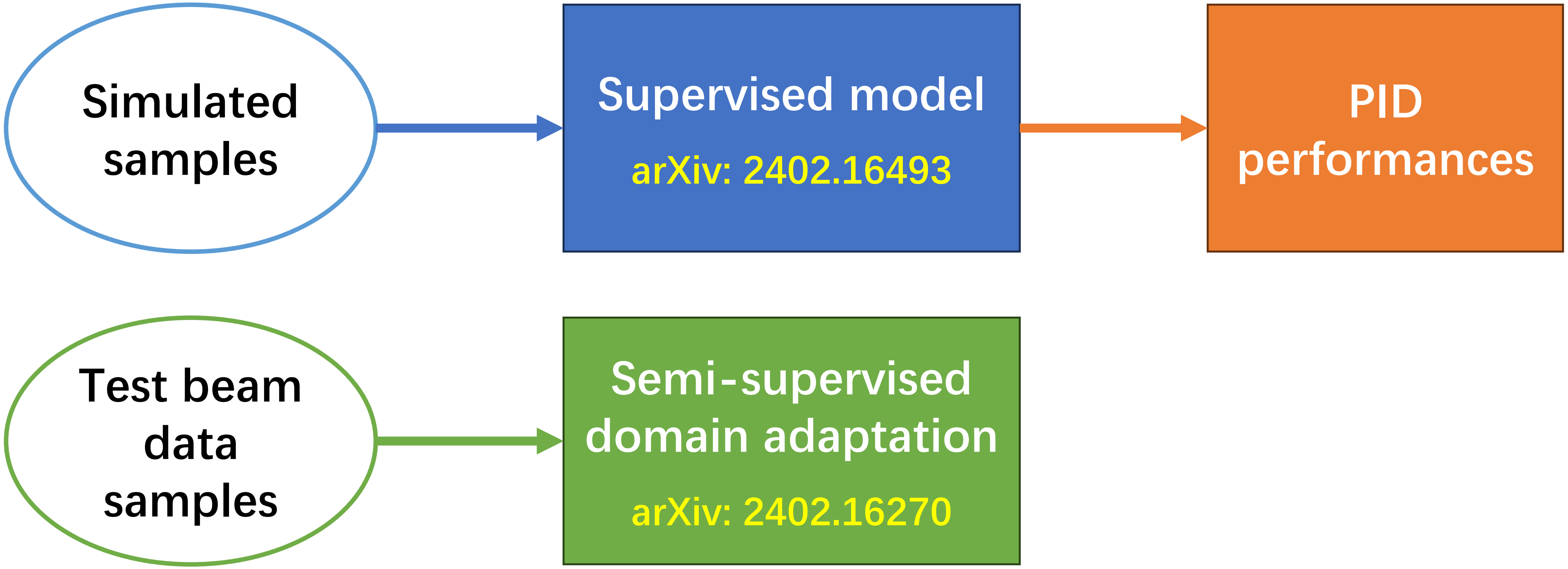
9 Apr 2024, Marseille, France

CEPC EU Workshop

中国科学院高能物理研究所
Institute of High Energy Physics
Chinese Academy of Sciences

Istituto Nazionale di Fisica Nucleare

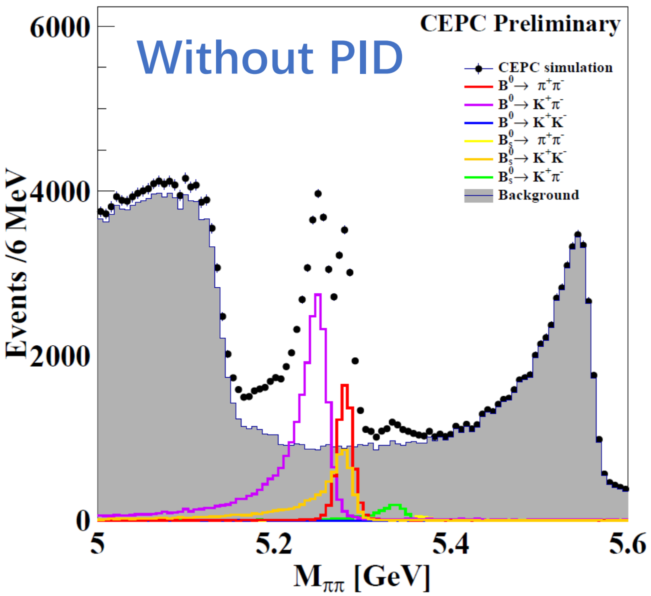
ML algorithms for dN/dx reconstruction



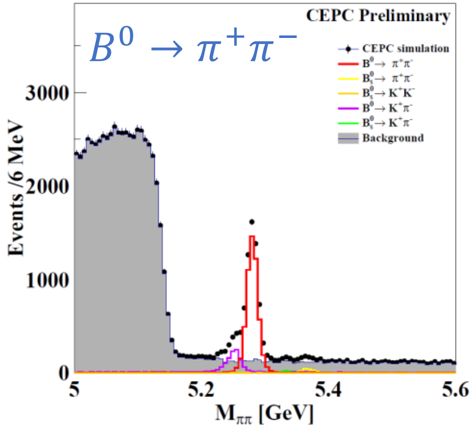
Motivation: Particle identification

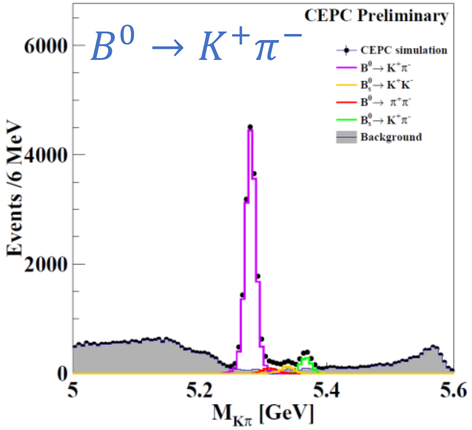
- **PID is essential for CEPC, especially for flavor physics**
  - Suppressing combinatorics
  - Distinguishing between same topology final-states
  - Adding valuable additional information for flavor tagging of jets
  - …

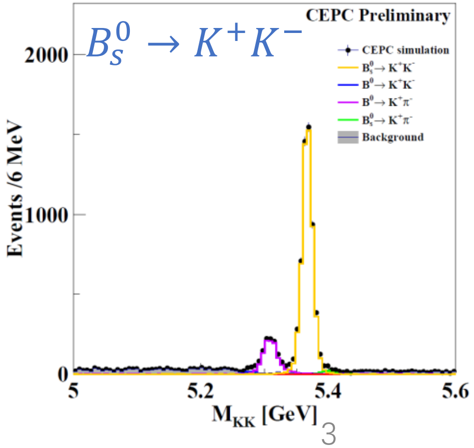
Benchmark channel:
$B^0_{(s)} \to h^+h'^-$



With PID

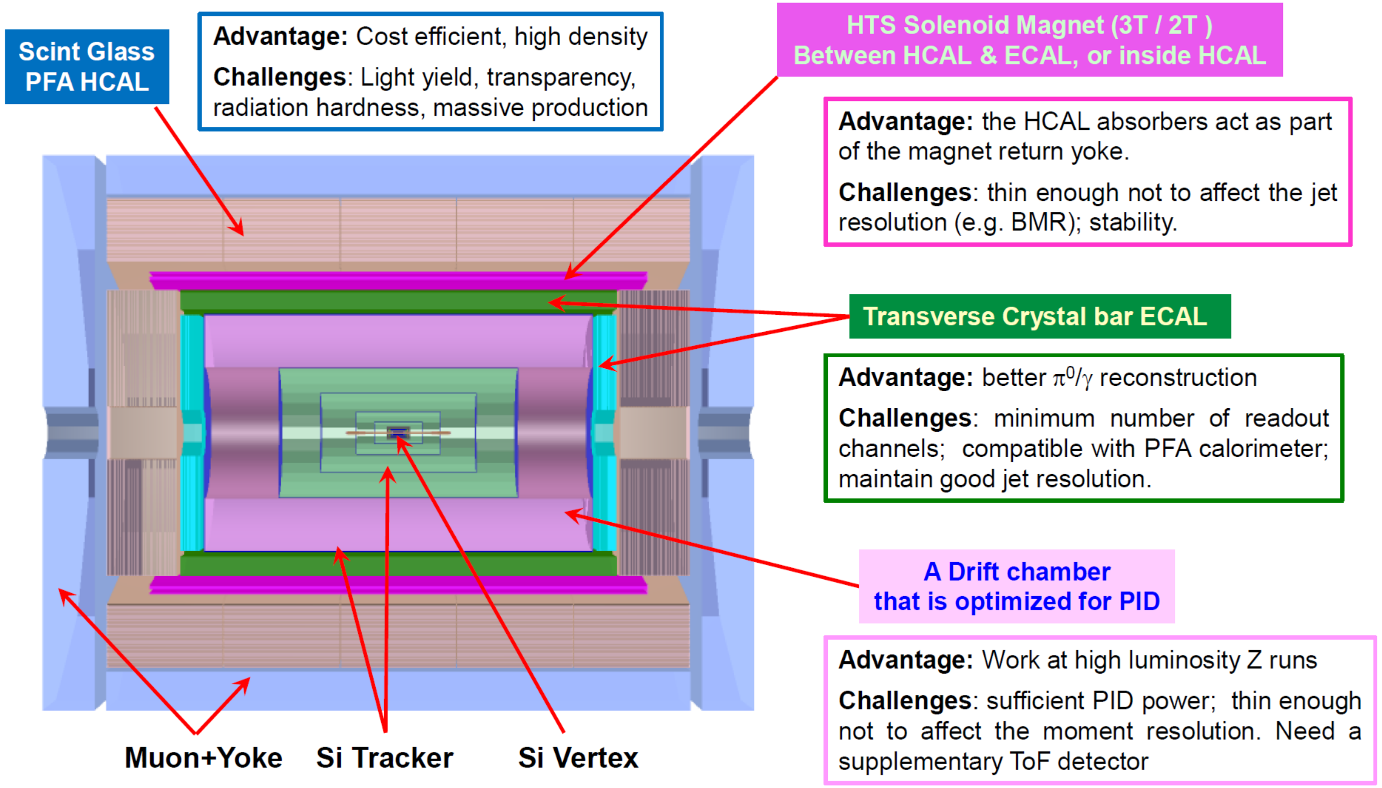






Drift chamber with PID capability

The CEPC 4th concept



A drift chamber with cluster counting (dN/dx) is one of the gaseous detector options

Key parameters:

- Full length: 5800 mm
- Barrel coverage: |cosθ| < 0.85
- Radius: 600 – 1800 mm
- Support: 8x8 carbon fiber frame
- Endcap: 20 mm Al plate
- Gas mixture: 90/10 He/iC_4H_{10}

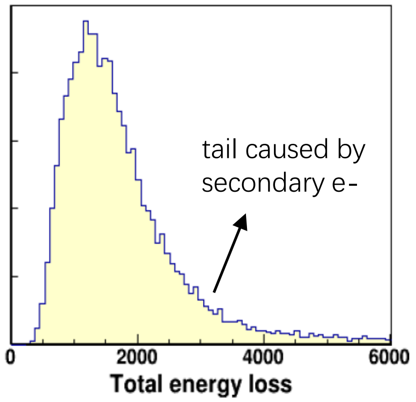
➢ See Mingyi's talk for more details on the drift chamber design

Cluster counting in drift chamber (dN/dx)

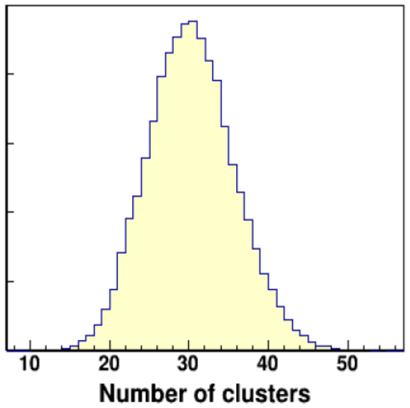
- **dE/dx: Measure the total energy loss**
  - Landau distributed
  - Large fluctuation from many sources

- **dN/dx: Measure the number of primary ionizations (breakthrough PID tech.)**
  - Poisson distributed
  - Small fluctuation; Potentially improve the resolution by a factor of 2

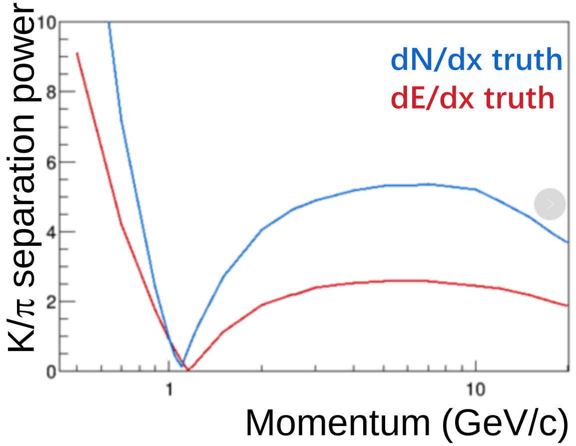
dE/dx



dN/dx

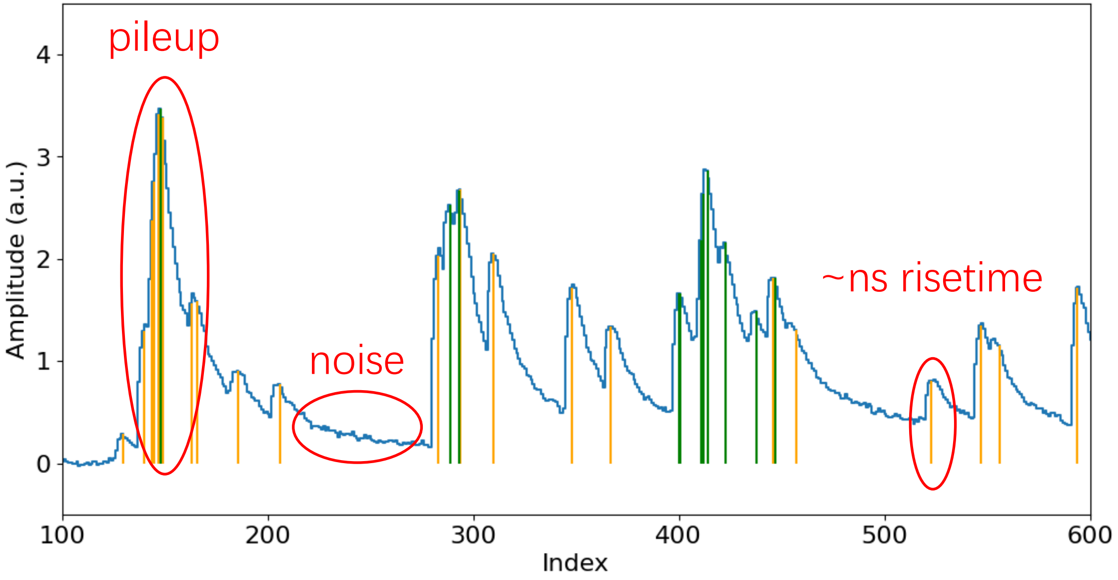


K/π separation power dN/dx vs dE/dx



Challenges of dN/dx measurement

Orange lines: Primary electrons (MC truth)
Green lines: Secondary electrons (MC truth)

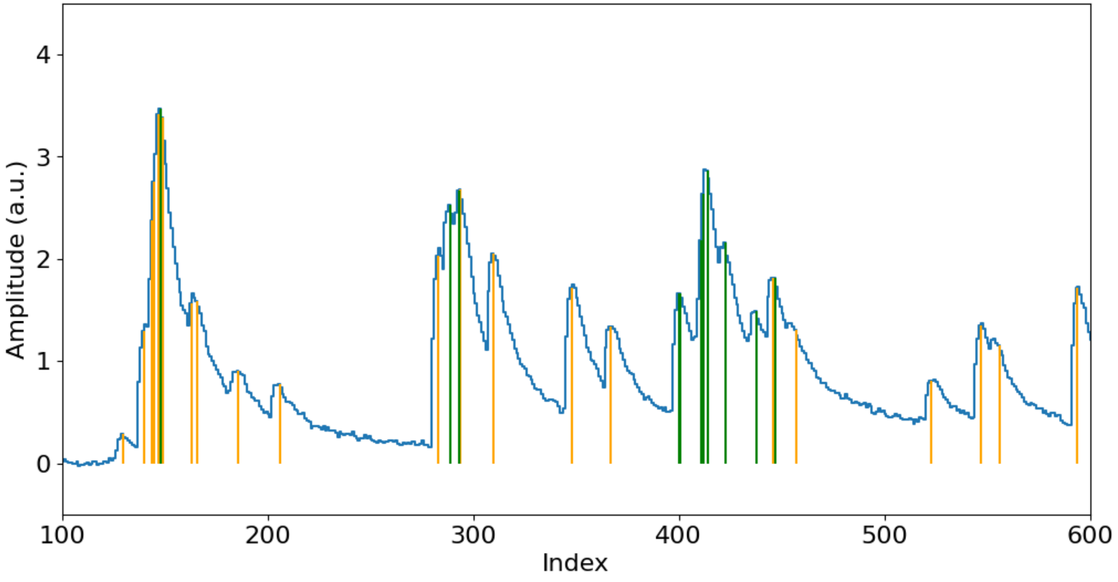


- **Single pulse risetime ~ns, require fast electronics**
  - Bandwidth > 1 GHz
  - Gain > 10
  - Sampling rate > 1.5 GS/s
  - Bit resolution > 12 bit

- **Signals are superimposed with noises and are heavily piled-up in some regions, require sophisticated reconstruction algorithm**

dN/dx reconstruction

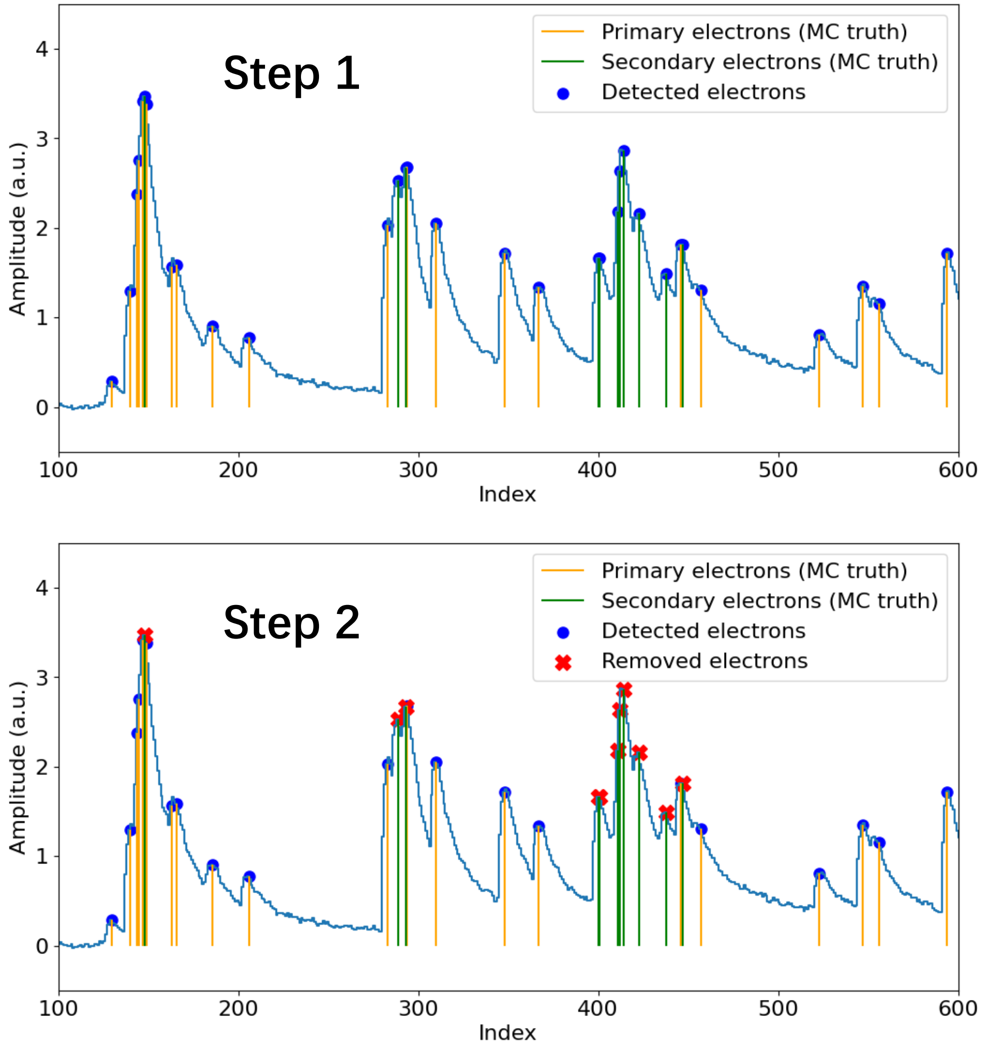
Orange lines: Primary electrons (MC truth)
Green lines: Secondary electrons (MC truth)



What is the dN/dx reconstruction?

- As implied by the name “cluster counting”, the dN/dx reconstruction is to determine the number of primary electrons in the waveform

dN/dx reconstruction (II)



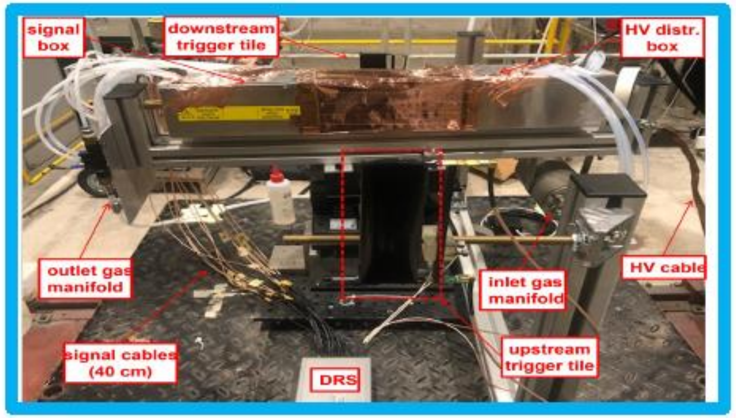
2-step algorithm

- **Peak finding:**
  - Detect peaks from both primary and secondary electrons
- **Clusterization:**
  - Remove secondary electrons from the detected peaks in step 1

Software package and data samples

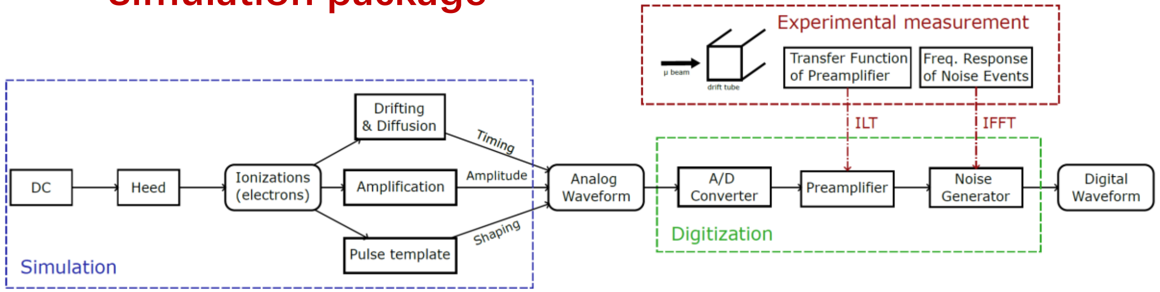
- **Simulation package**
  - Garfield++-based simulation + data-driven digitization
- **Data samples**
  - Simulated samples
    - 0-20 GeV/c pions and kaons
  - Experimental samples
    - 180 GeV/c muons from CERN/H8 beam

Test beam at CERN

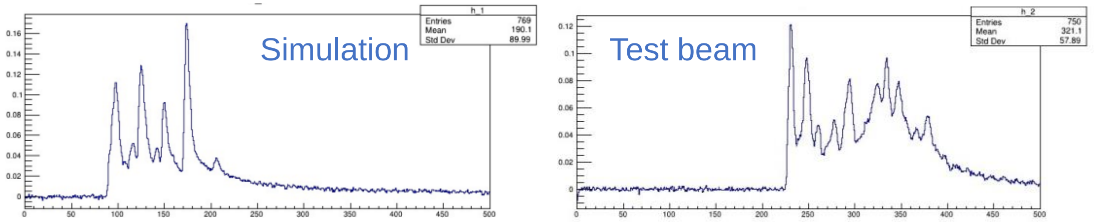


From INFN group leaded by Franco Grancagnolo and Nicola De Filippis

Simulation package

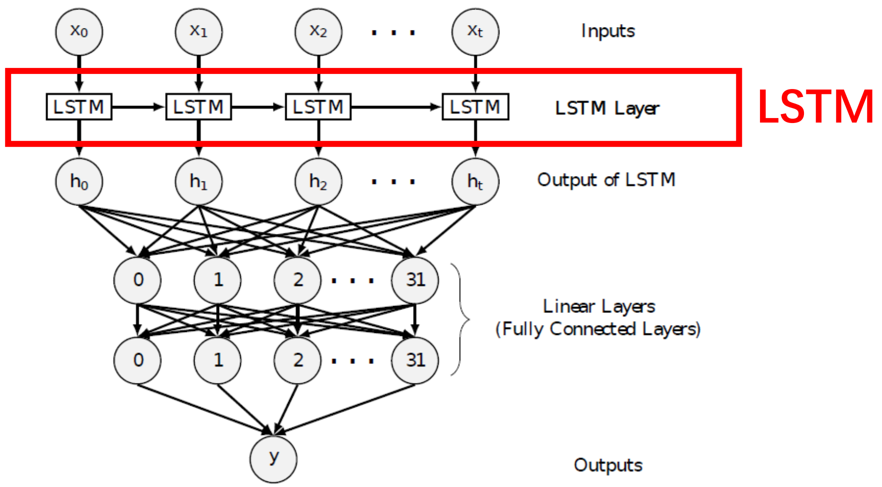


Tuned MC is comparable to data



Supervised model for simulated samples

Peak finding



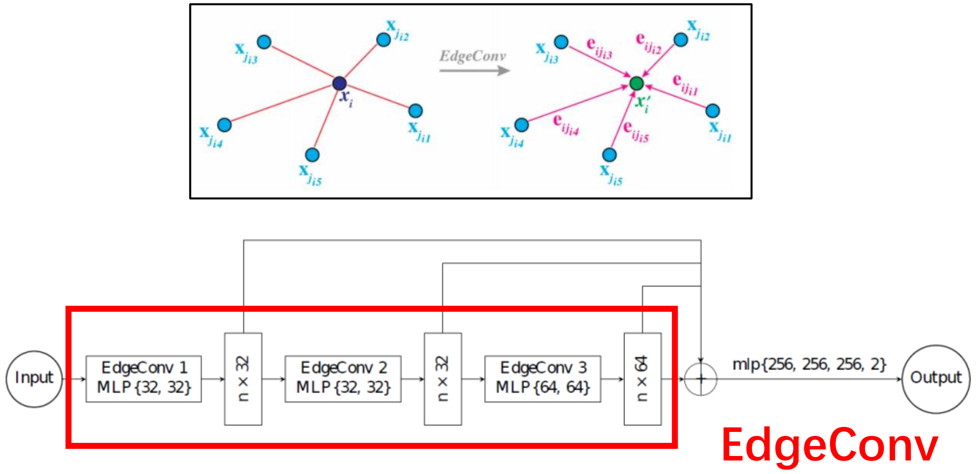
Long Short-Term Memory (LSTM)

- A specified recurrent neural network (RNN) that deals with the vanishing gradient problem
- Can handle long sequences efficiently

LSTM-based peak finding

- Waveform as sliding windows
- Binary classification of signals and noises

Clusterization



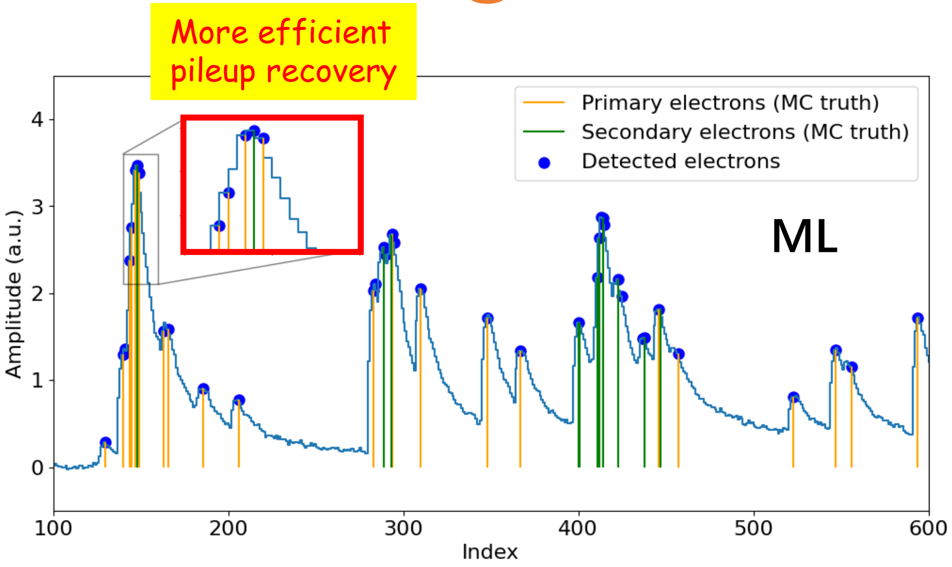
Dynamic Graph CNN (DGCNN)

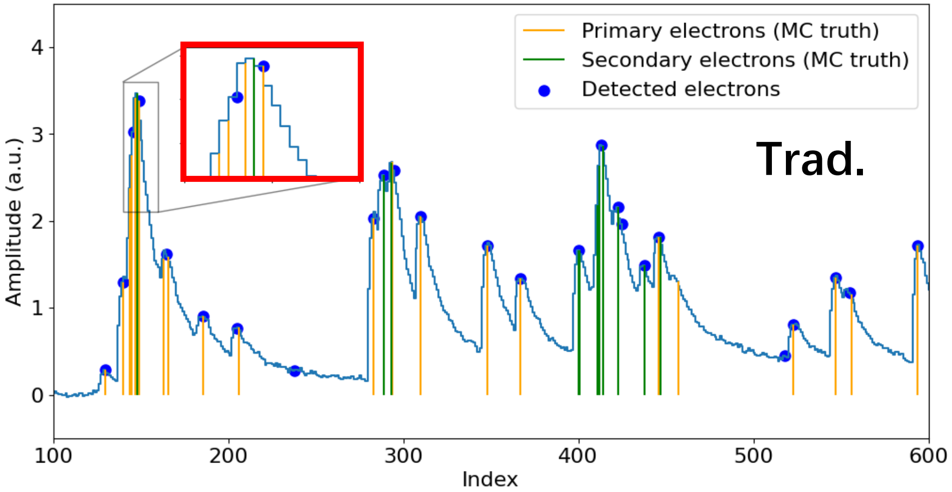
- A specified graph neural network (GNN) that incorporates local information and stacked to learn global properties, which is very suited for clusterization

DGCNN-based clusterization

- Peak timing as the node feature. Edges are initially connected by timing similarity.
- Binary classification of primary and secondary electrons

Peak finding results





Traditional peak-finding: second derivative

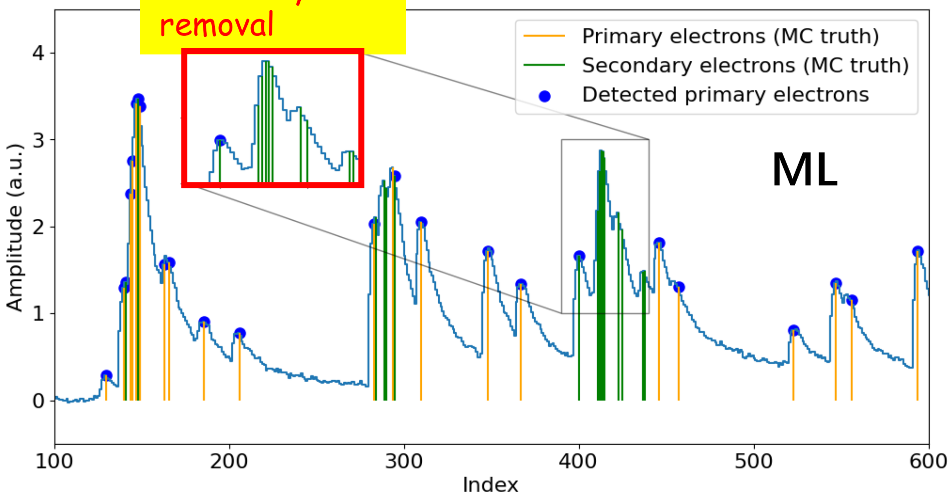
Table 2. The purity and efficiency comparison between LSTM-based algorithm and traditional D2 algorithm for peak-finding.

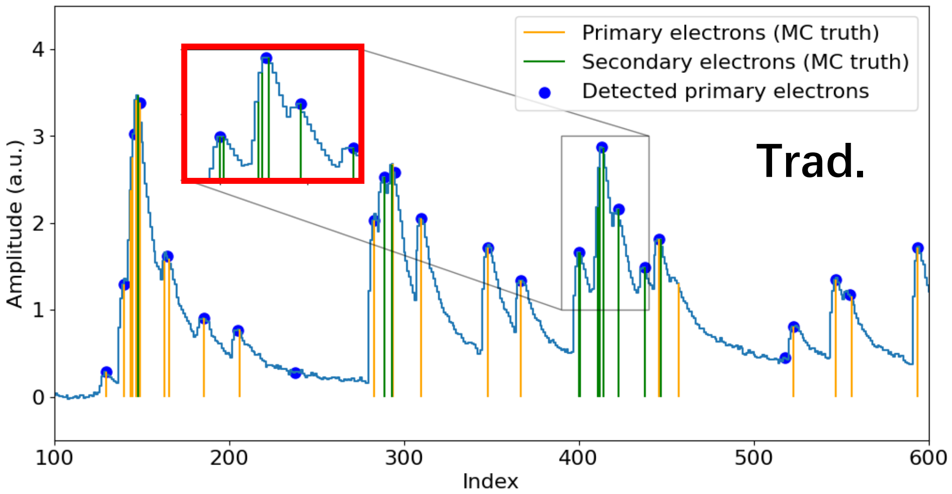
	Purity	Efficiency
LSTM algorithm	0.8986	0.8820
D2 algorithm	0.8986	0.6827

- The LSTM-based model is more powerful than the traditional derivative-based algorithm, especially for the pileup recovery

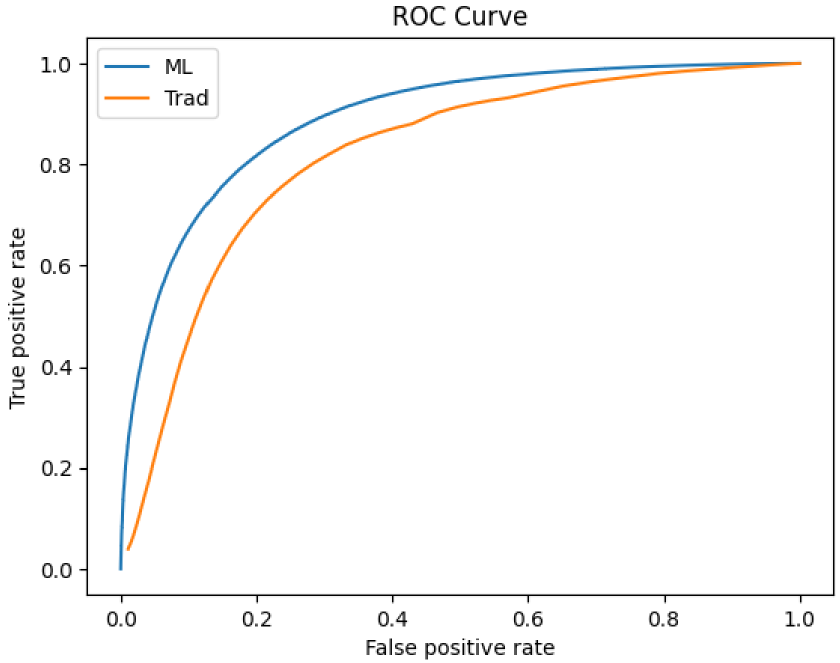
Clusterization results

More accurate secondary e- removal





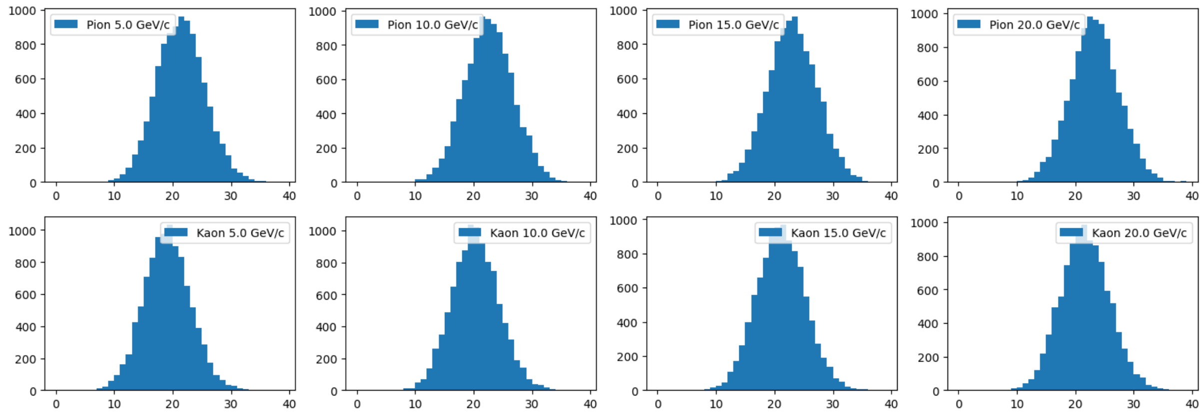
Traditional clusterization: adjacent-peak merge



- The DGCNN-based model is more powerful than the traditional peak-merge algorithm, as it can remove the secondary electrons more accurate

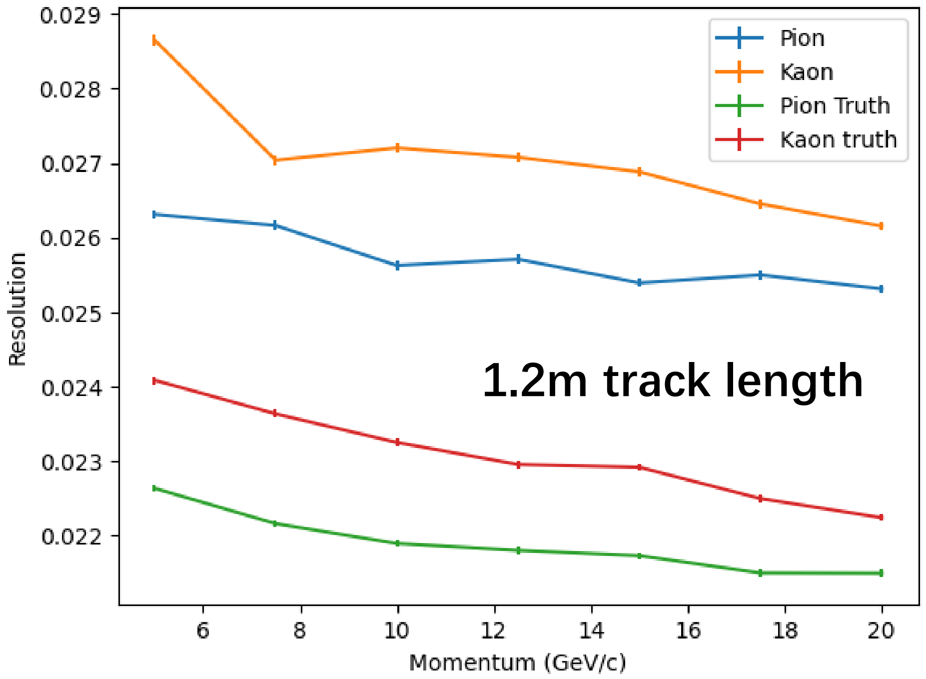
PID performances with supervised models

Reconstructed # of clusters distributions



The reconstructed n_{cls} distributions are very well Gaussian-like

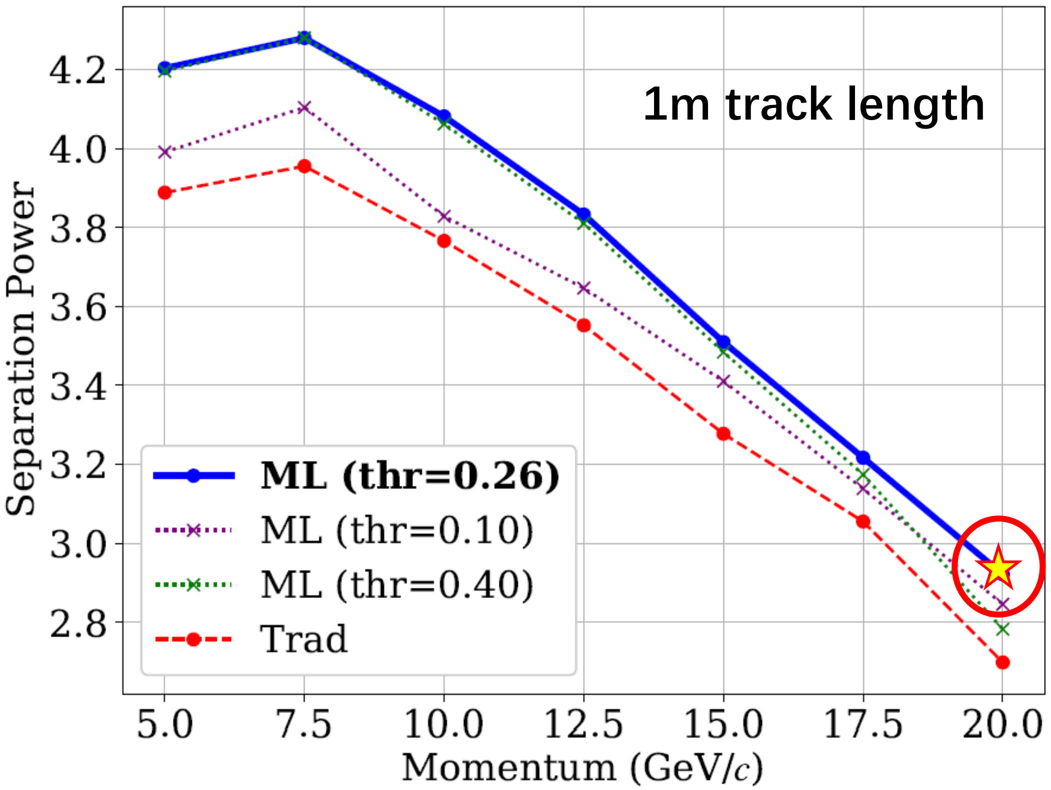
dN/dx resolution



dN/dx resolutions for high momenta pions/kaons are < 3%，which are much better than typical dE/dx ~5%

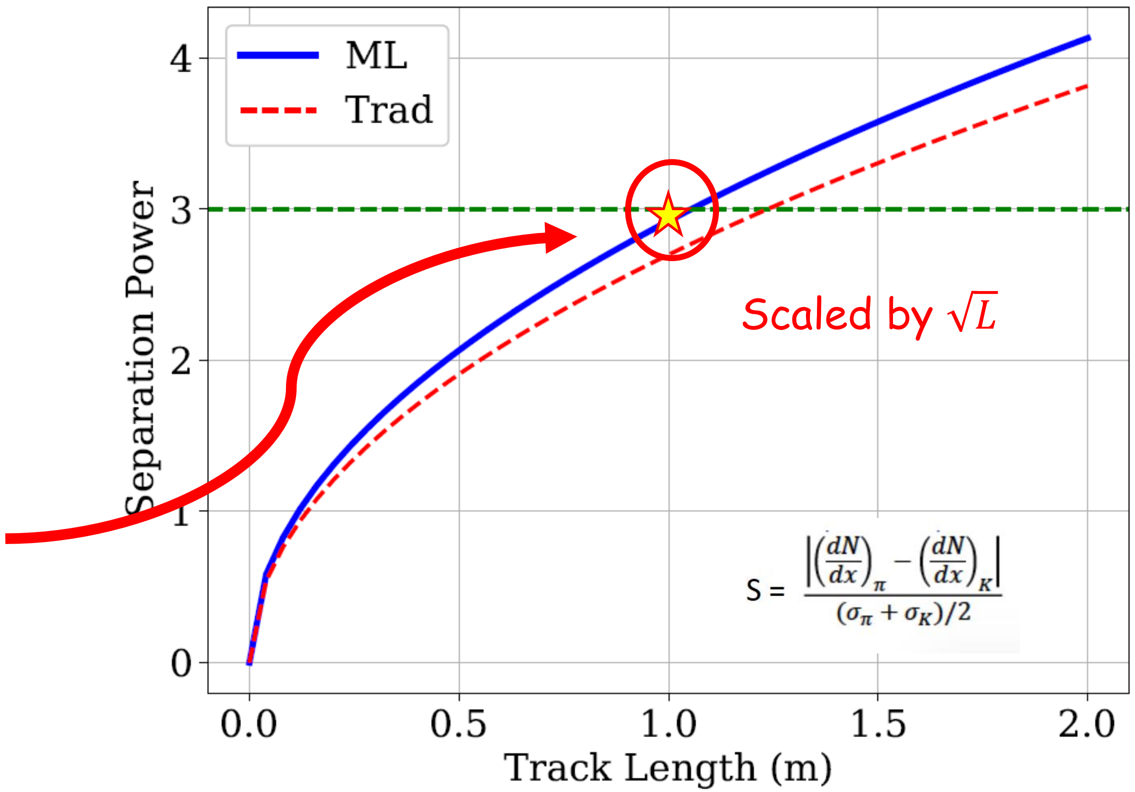
PID performances with supervised models (II)

K/π separation power vs. momentum



~10% improvement for ML (equivalent to a detector with 20% larger radius)

K/π separation power @ 20 GeV/c

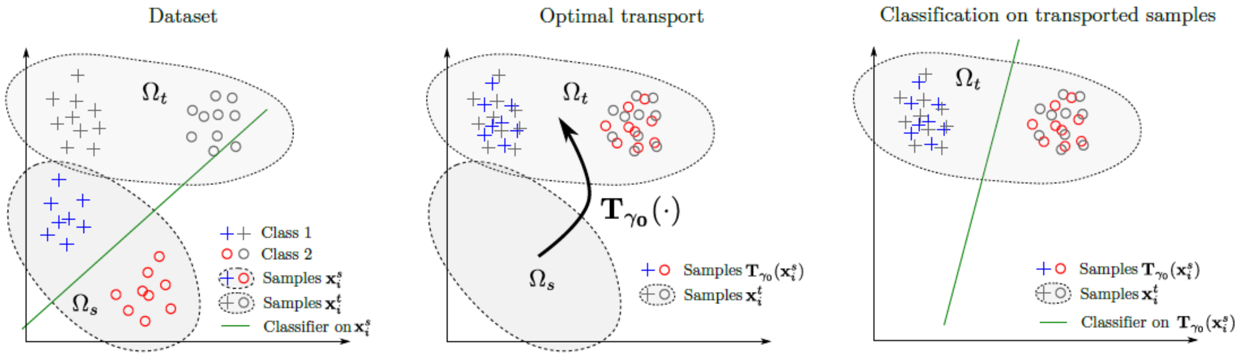


Could achieve 3σ for 1m track length. For 1.2m track length (current CEPC baseline), the separation is 3.2σ

Domain adaptation for test beam data

- **Challenges for real data**
  - Imperfect simulation
  - Incomplete labels in real data

- **Solution: Domain adaptation**
  - Transfer knowledge between simulation and real data



Align data/MC samples with Optimal Transport

Loss for labeled samples in source domain

$$\min_{f,g}\left[\sum_{i=1}^{m} L_s\left(y_i^s, f(g(x_i^s))\right) + \frac{1}{m_l}\sum_{i=1}^{m_l} L_t\left(y_i^{t,l}, f\left(g(x_i^{t,l})\right)\right) + \min_{\gamma\in\Delta}\sum_{i,j}^{m}\gamma_{ij}\left(\alpha\left\|g(x_i^s)-g(x_j^t)\right\|^2 + \lambda_t L_t\left(y_i^s, f\left(g(x_j^t)\right)\right)\right)\right]$$

Loss for labeled samples in target domain (THIS WORK)

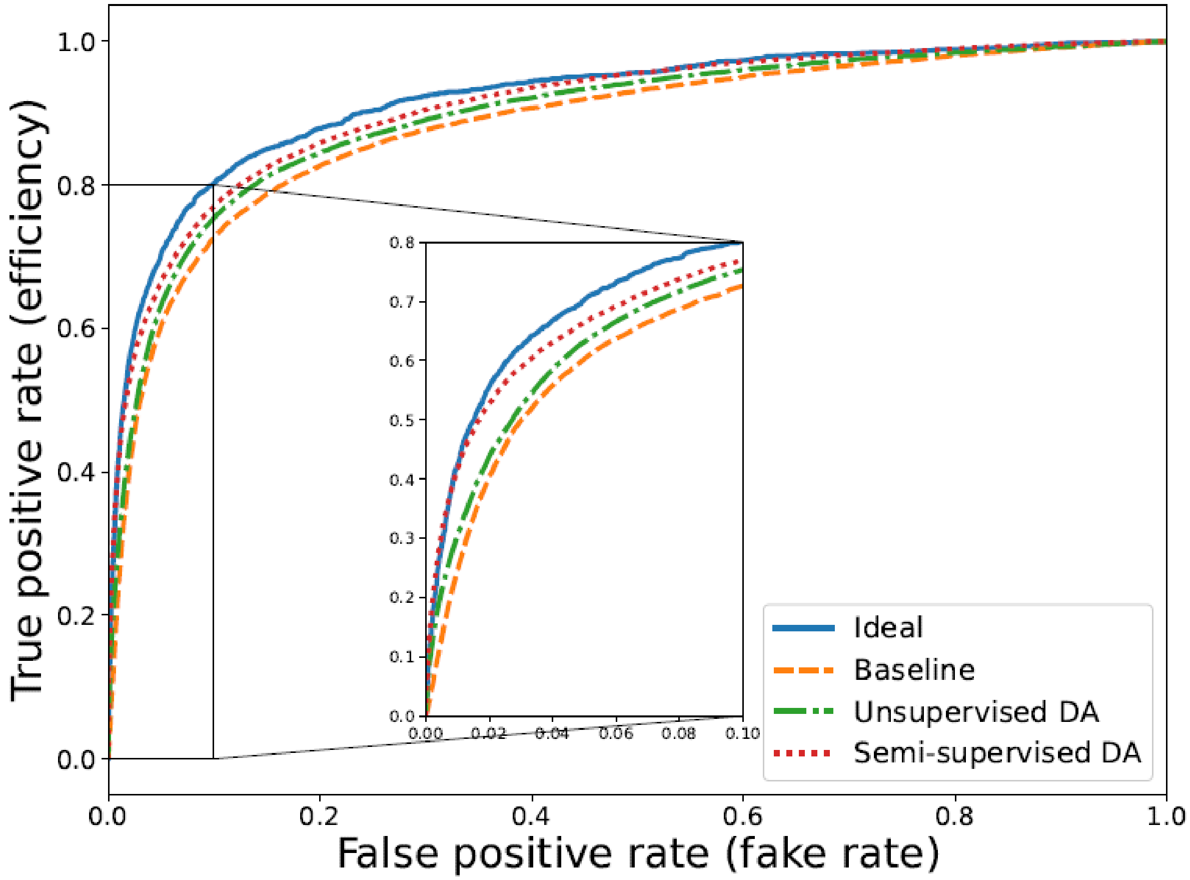
Cost of feature differences between source and target

Cost of 'label' differences between source and target

Cost of joint feature-label distribution for OT

Semi-supervised domain adaptation

Model validation by pseudo data



Numeric experiment with pseudo data in 2 domains (simulation domain & data domain)

Model	AUC	pAUC (FPR<0.1)
Ideal	0.926	0.812
Baseline	0.878	0.749
Unsupervised DA	0.895	0.769
Semi-supervised DA	0.912	0.793

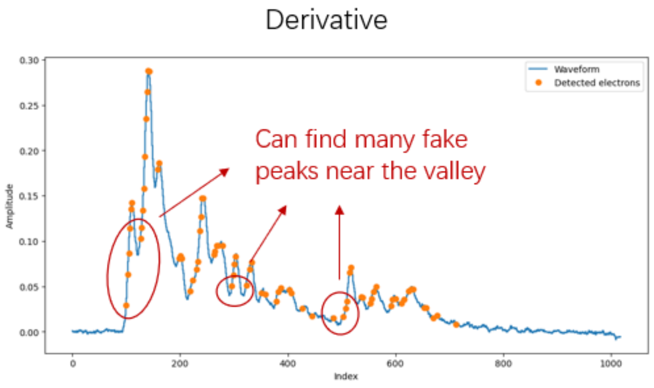
Improve

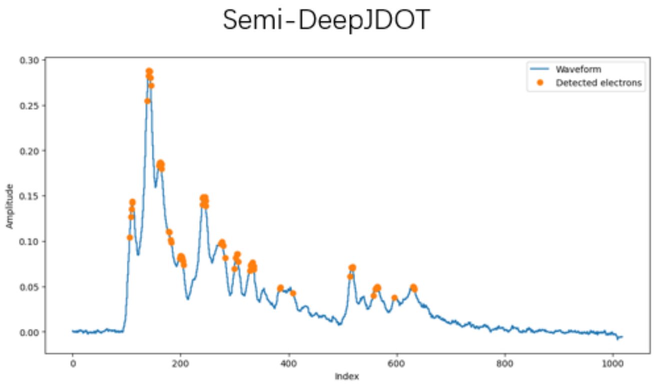
Improve

- **Note:**
  - Ideal = Supervised model in data domain
  - Baseline = Supervised model in sim. domain
  - Unsupervised DA = Baseline + OT
  - Semi-supervised DA = Baseline + OT + semi-supervised setup
- **The OT and the semi-supervised loss improve the results, and the performance of the semi-supervised DA model is very close to the ideal model**

Peak finding for test beam data

Single-waveform results between derivative alg. and DL alg.

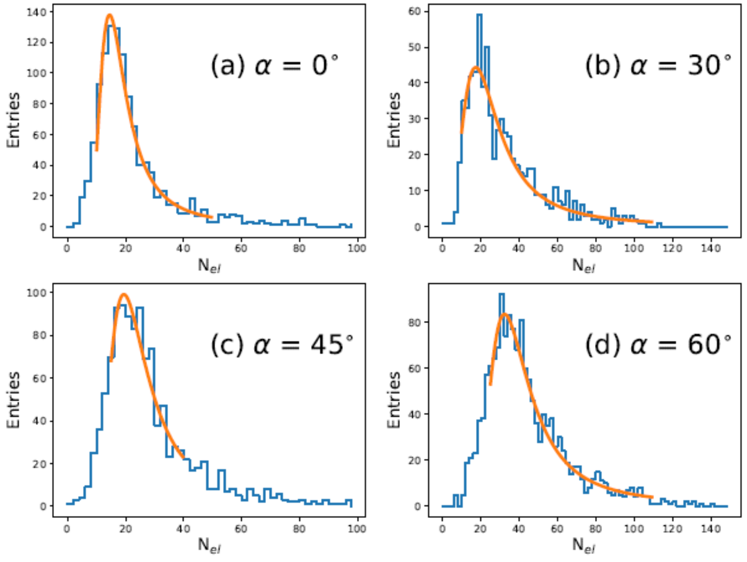


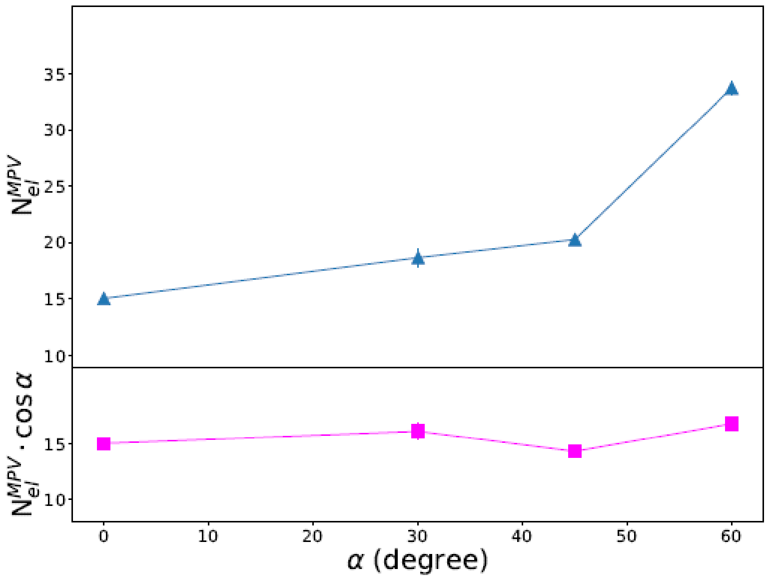


Note: Require similar efficiency for both cases

DL algorithm is more powerful to discriminate signals and noises

Multi-waveform results for samples in different angles





Scale w.r.t. track length

The algorithm is stable w.r.t. track length

Conclusion

- Two machine learning algorithms are developed for dN/dx reconstruction. In principle, the method can be applied to similar feature extraction tasks in signal processing.

- The supervised model has 10% improvement on K/pi separation w.r.t. traditional algorithm. The situation could be similar for the semi-supervised domain adaptation model.

- When studied with the full-simulation samples using a supervised model, the CEPC drift chamber achieves $< 3\%$ K/pi resolution and $> 3.2\sigma$ K/pi separation.

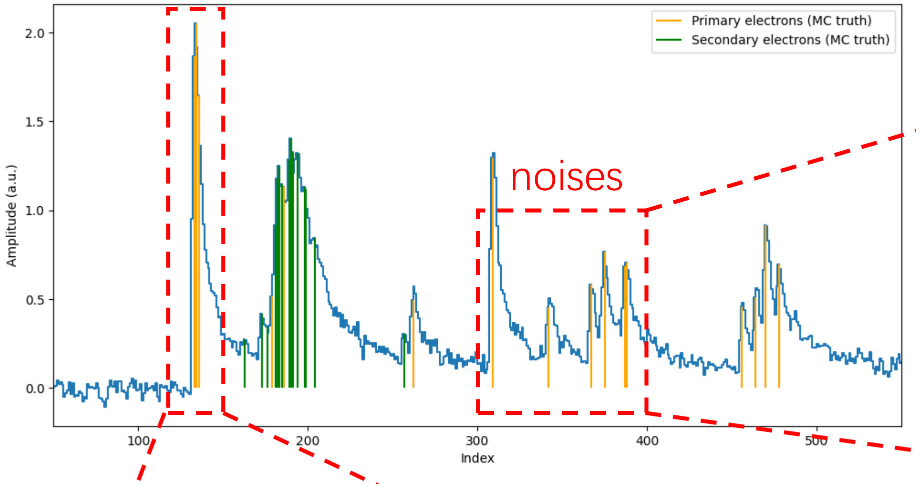
- When studied with the test beam samples, the semi-supervised domain adaptation model successfully transfer information from simulation and achieve stable performances.

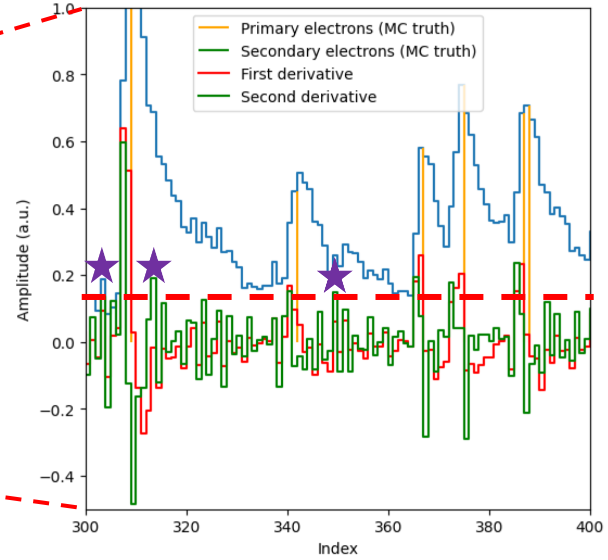
Thank you!

Backup

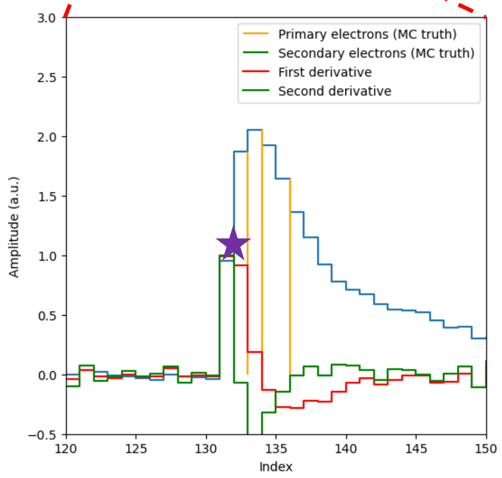
Traditional peak finding

pileup signals





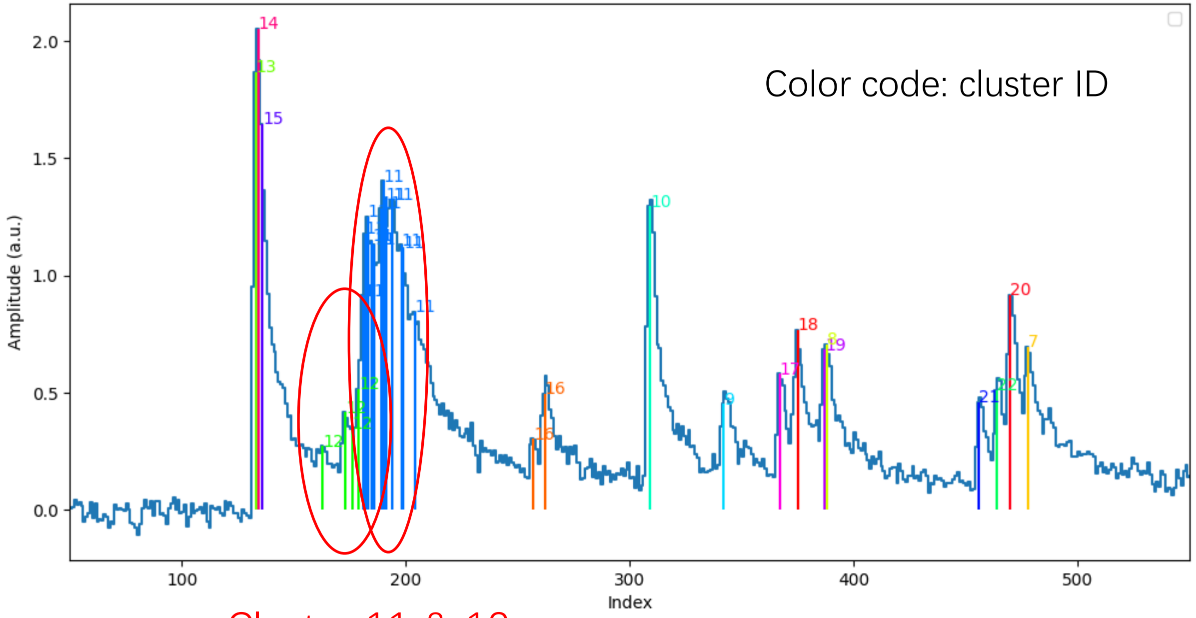
Some noises can also pass the threshold



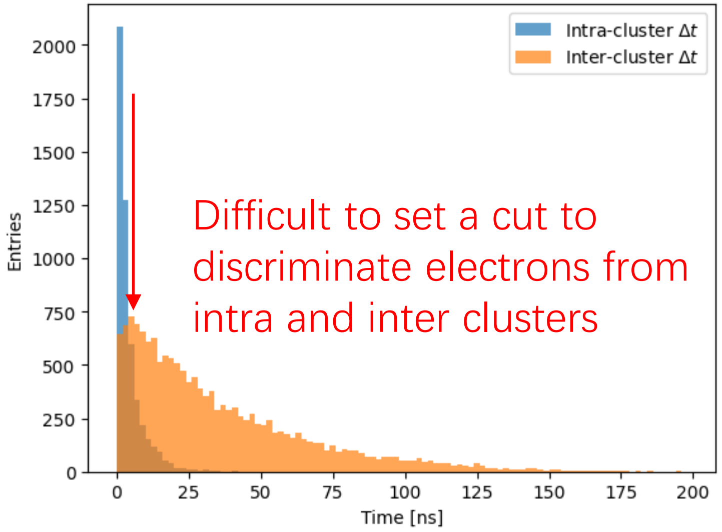
Only 1 out of 3 signals is detected

- **Derivative-based peak finding**
  - Take first and secondary derivatives
  - Require threshold passing
- **Challenges**
  - Noises can pollute the signal
  - Signals are highly piled up

Traditional clusterization



Cluster 11 & 12
are overlapped



- **Timing-based clusterization**
  - Merge adjacent peaks
- **Challenges**
  - Electrons from different clusters can overlap

Additional plots for domain adaptation

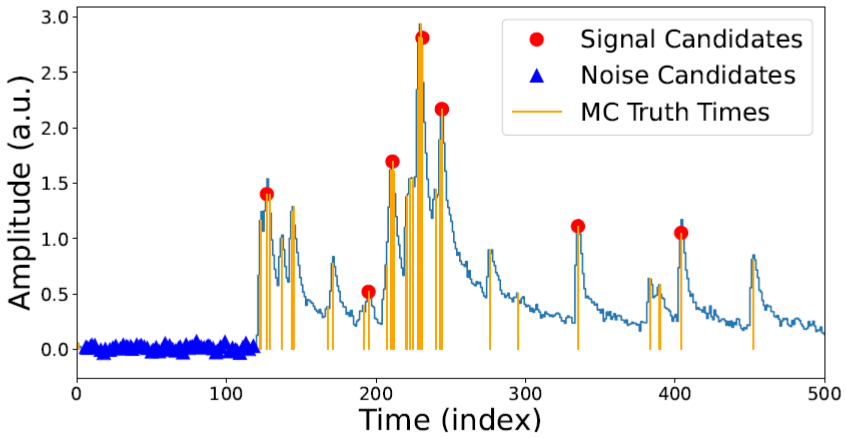


Figure 1: An example of simulated waveform. The blue histogram is the waveform. The red solid circles are the signal peaks selected by the CWT algorithm. The blue solid triangles are the noise peaks selected by requiring the 3 RMS requirement. The orange lines indicate the electron signal times from MC truth information.

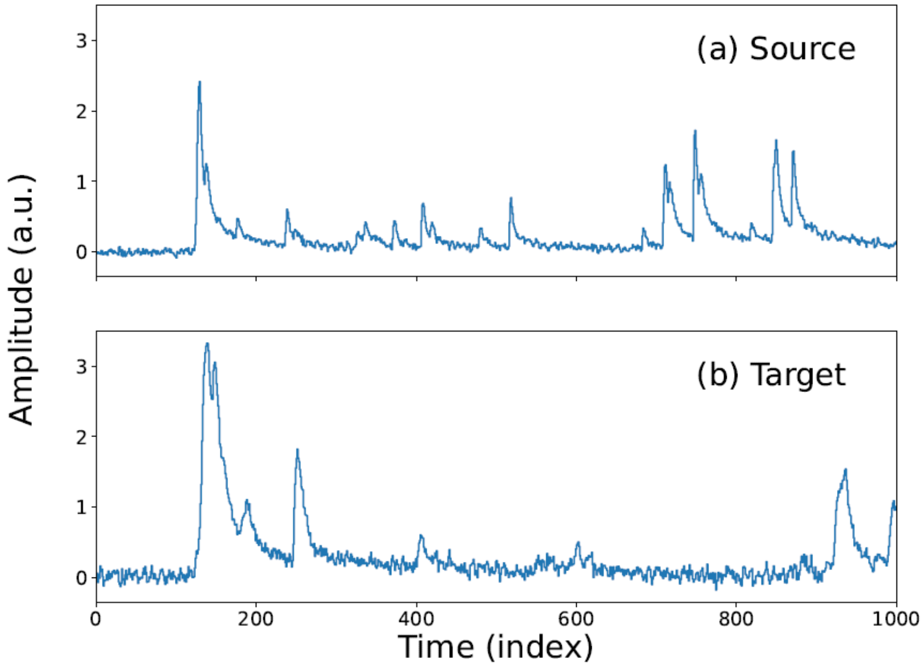


Figure 4: Waveform examples from the source sample (a) and the target sample (b). The source waveforms are generated with a noise level of 10% and a pulse risetime of 2 ns, while the target waveforms with a noise level of 20% and a pulse risetime of 4 ns.