dN/dx reconstruction with supervised learning and transfer learning

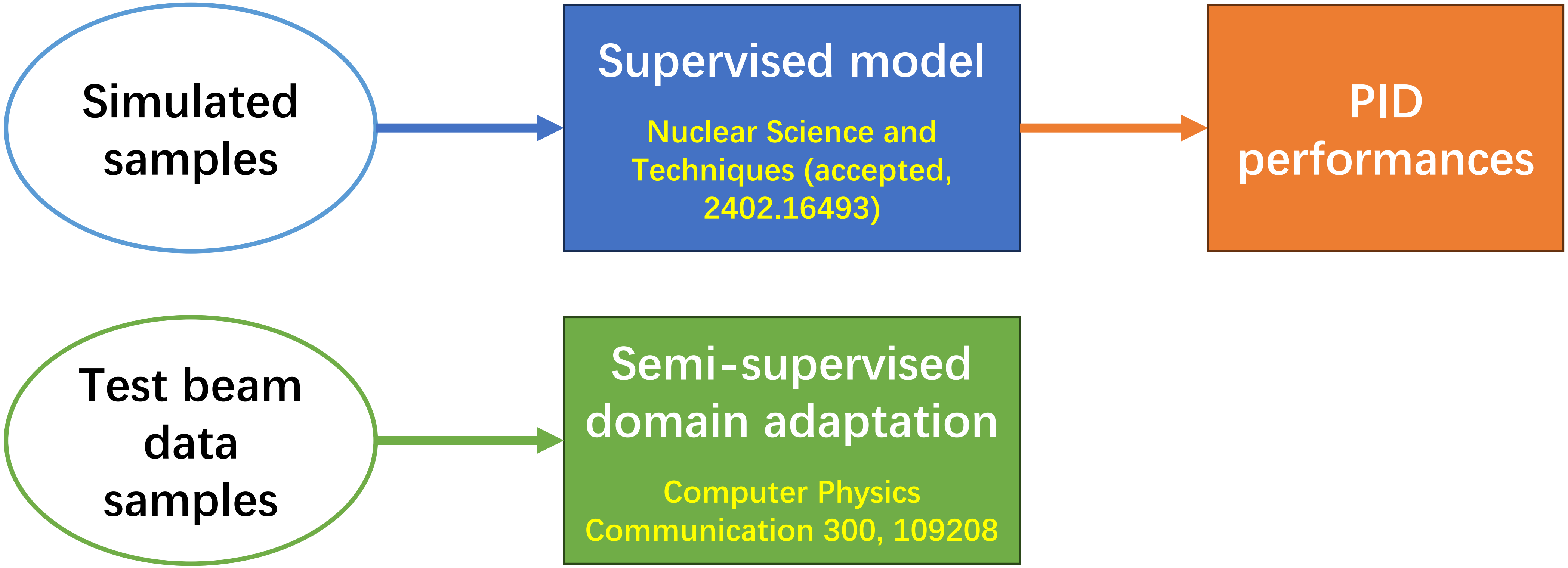
Guang Zhao (zhaog@ihep.ac.cn)

Institute of High Energy Physics

October 22, 2024

FTCF2024, Guangzhou

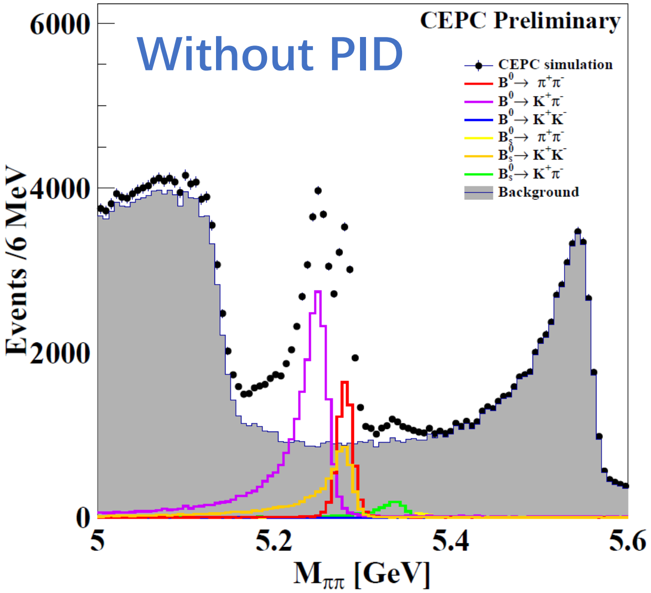
ML algorithms for dN/dx reconstruction



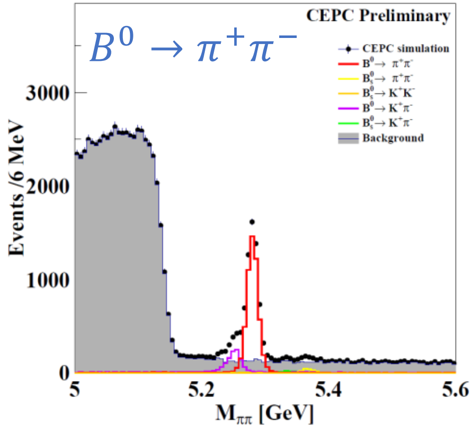
Motivation: Particle identification

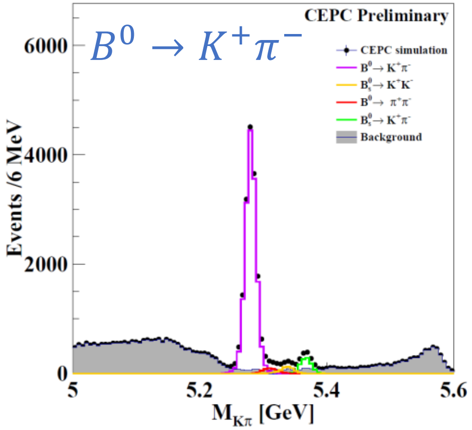
- **PID is essential for high energy physics experiments**
  - Suppressing combinatorics
  - Distinguishing between same topology final-states
  - Adding valuable additional information for flavor tagging of jets
  - …

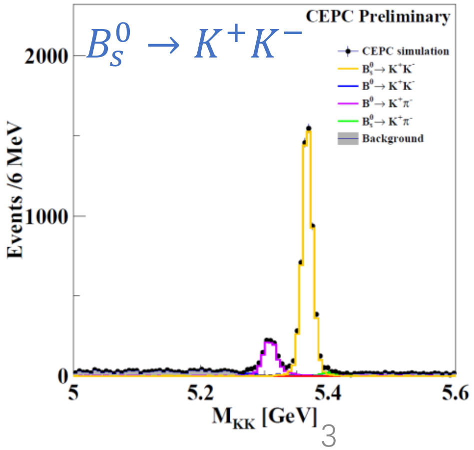
Benchmark channel:
$B^0_{(s)} \rightarrow h^+h'^-$



With PID

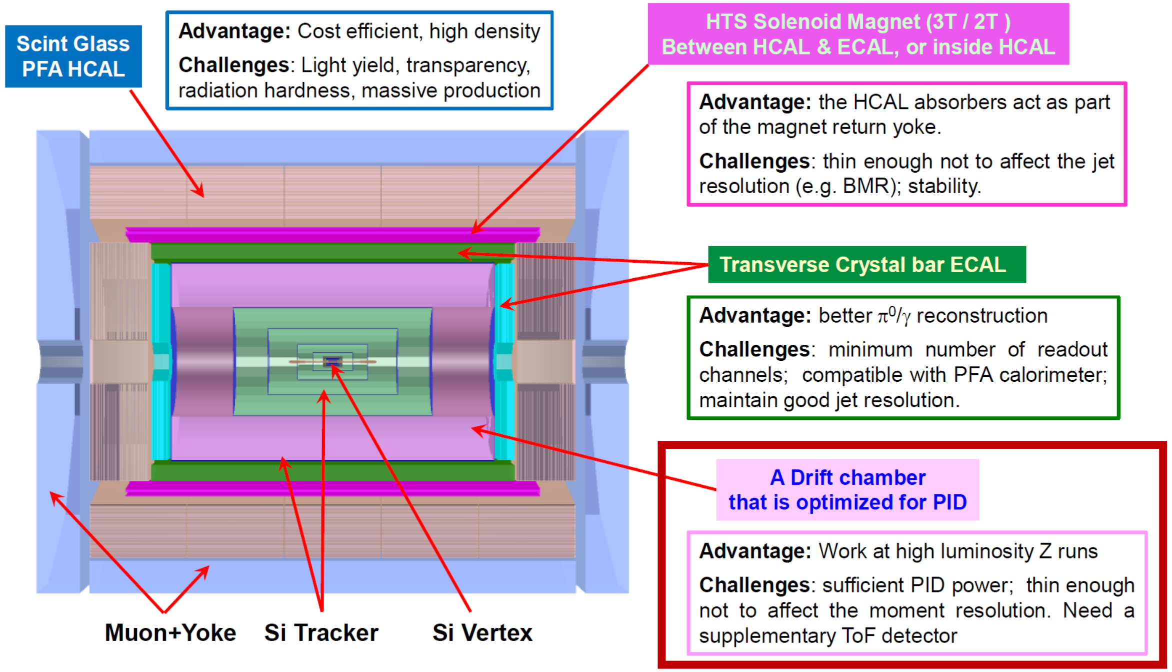






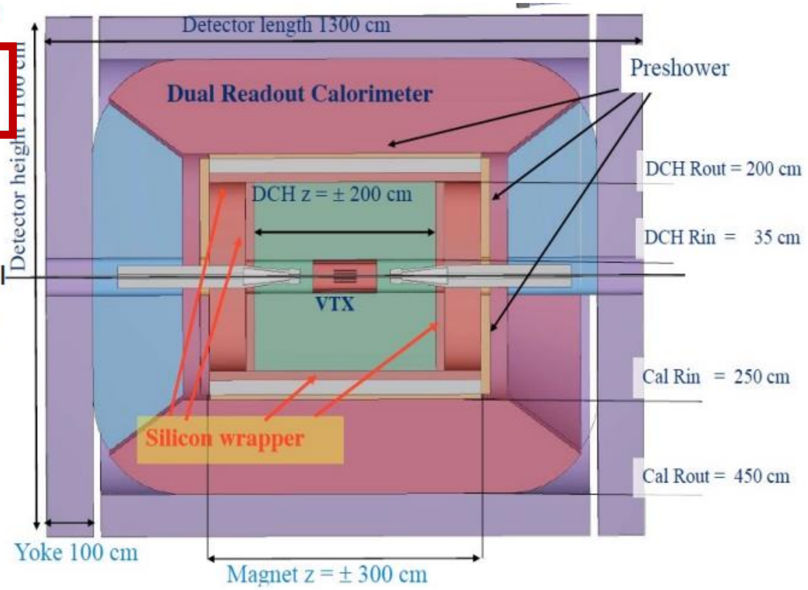
PID in next generation experiments

CEPC 4th Concept



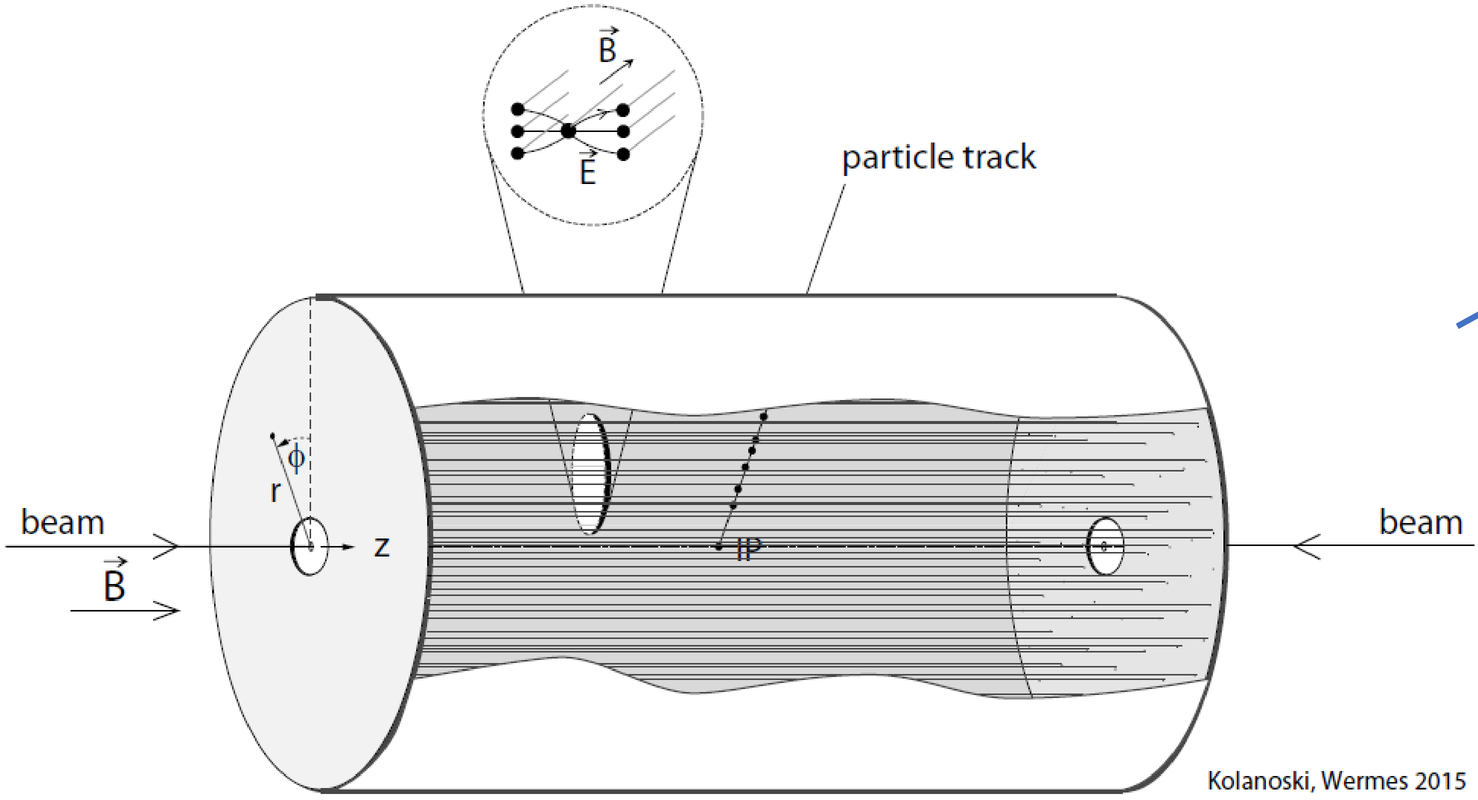
IDEA for FCC-ee

- a silicon pixel vertex detector
- a large-volume extremely-light **drift chamber**
- surrounded by a layer of silicon micro-strip detectors
- a thin low-mass superconducting solenoid coil
- a preshower detector based on **μ-WELL technology**
- a dual read-out calorimeter
- muon chambers inside the magnet return yoke, based on **μ-WELL technology**

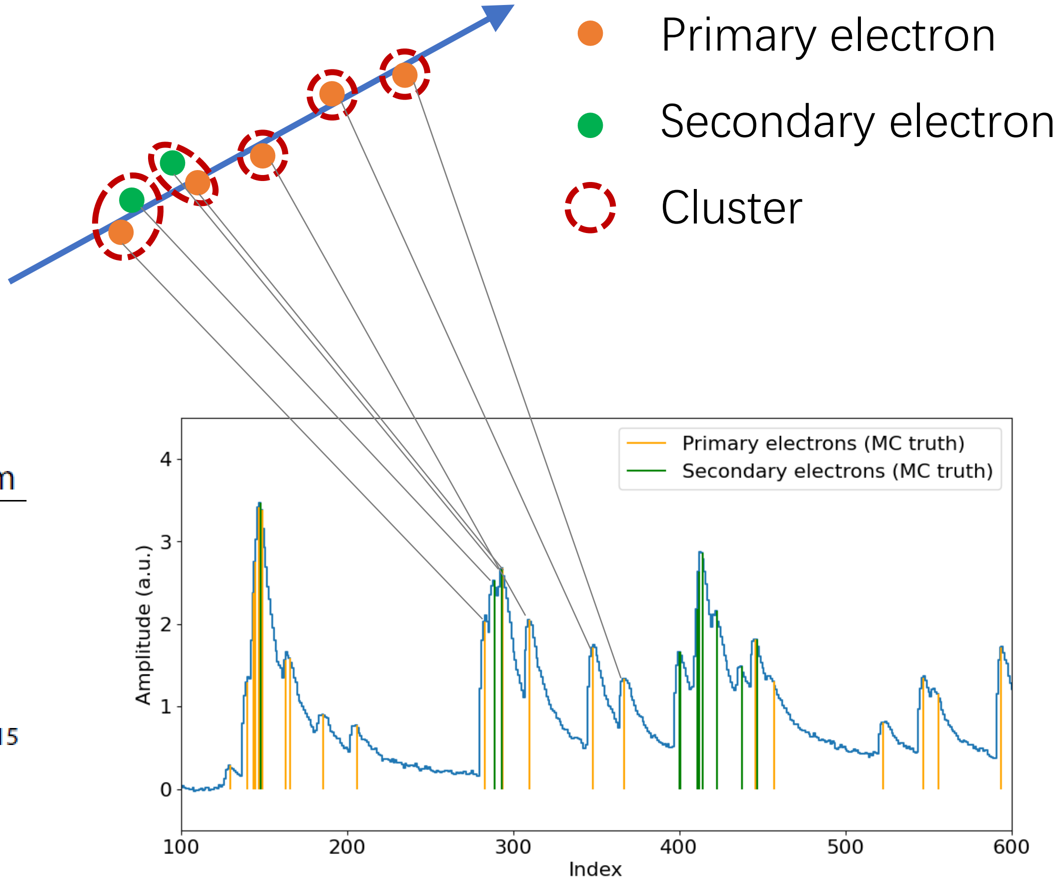


- In future high-energy and high-intensity frontier experiments (CEPC, FCC-ee, STCF), the requirements for PID are stringent
- In high-energy frontier case, flavor physics studies in high luminosity Z-pole run require high performance PID up to tens of GeV/c. Traditional techniques cannot meet such requirement
- Among the PID techniques, cluster counting (dN/dx) in drift chamber represents a breakthrough and is proposed in both CEPC and FCC-ee

Ionization measurement in drift chamber

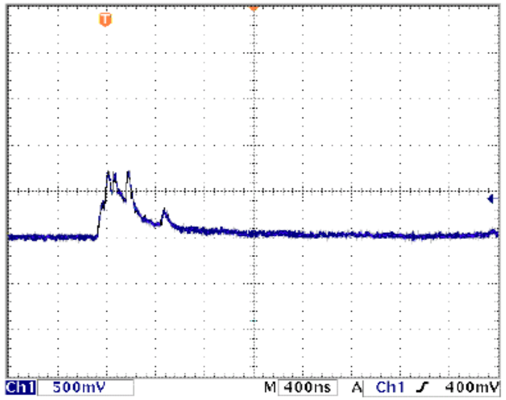


A track passing through the DC



A DC cell waveform

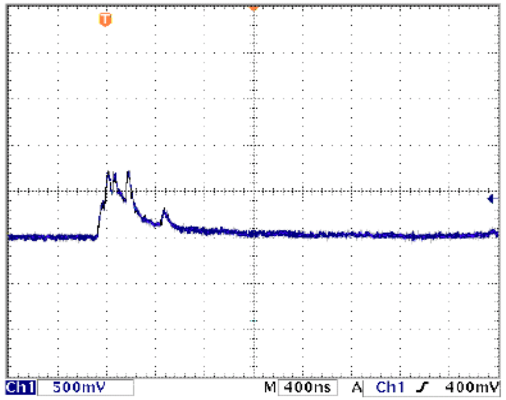
Ionization measurement in drift chamber

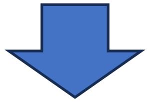


dE/dx (traditional method):

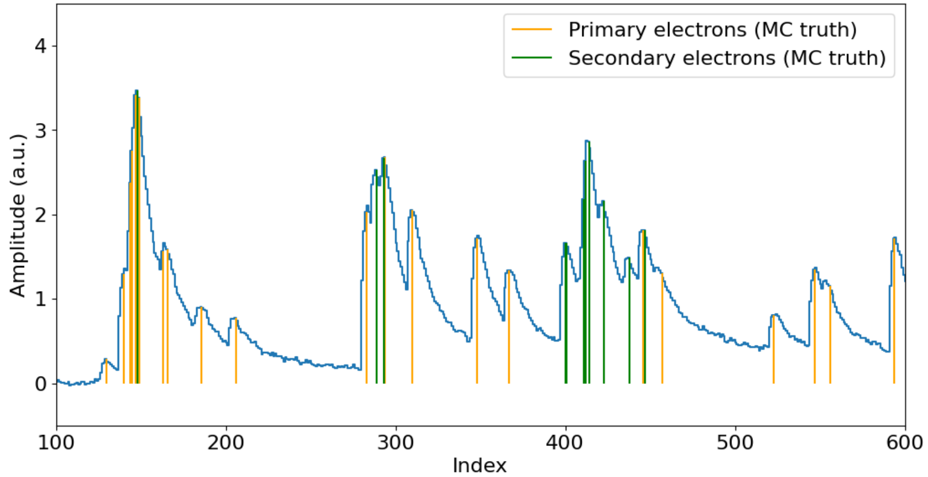
- **Method:** Total energy loss measurement by integrating the waveform
- **Characteristics:**
  - Landau distributed ➔ Loss ~30% statistics due to truncation
  - Large fluctuation from many sources

Ionization measurement in drift chamber





High bandwidth & sampling rate electronics



dE/dx (traditional method):

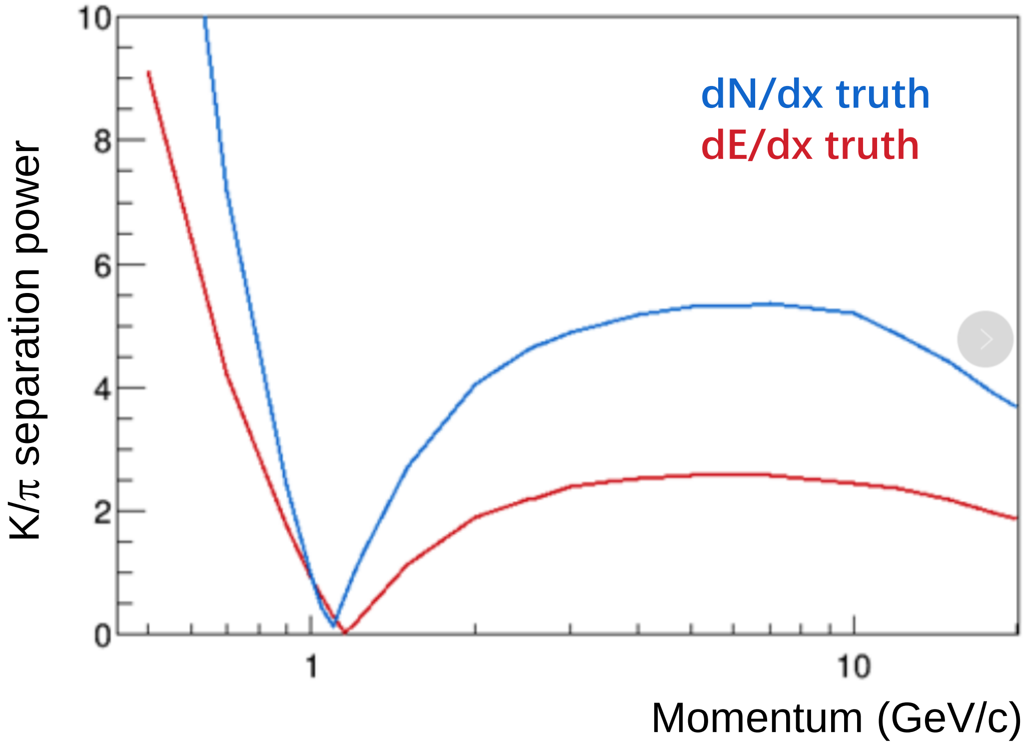
- **Method:** Total energy loss measurement by integrating the waveform
- **Characteristics:**
  - Landau distributed ➔ Loss ~30% statistics due to truncation
  - Large fluctuation from many sources

dN/dx or cluster counting ("ideal" method):

- **Method:** Number of primary ionization cluster measurement (require fast electronics)
- **Characteristics:**
  - Poisson distributed
  - Small fluctuation (resolution potentially improved by a factor of 2)

dN/dx vs. dE/dx

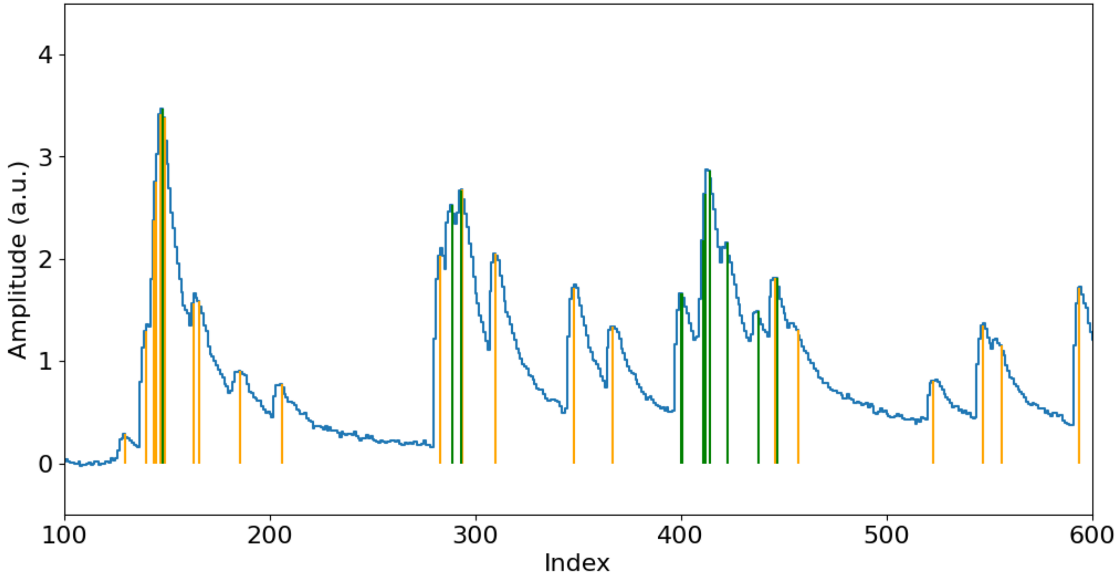
- **Particle separation power:**
  - Definition: $\frac{\text{separation}}{\text{resolution}} = \frac{|\mu_A - \mu_B|}{(\sigma_A + \sigma_B)/2}$
- **Typical K/π separation power:**
  - dE/dx: > 2σ up to 2…20 GeV/c
  - dN/dx: > 3σ up to 2…20 GeV/c



dN/dx has much better PID power than dE/dx
dN/dx is a breakthrough in PID

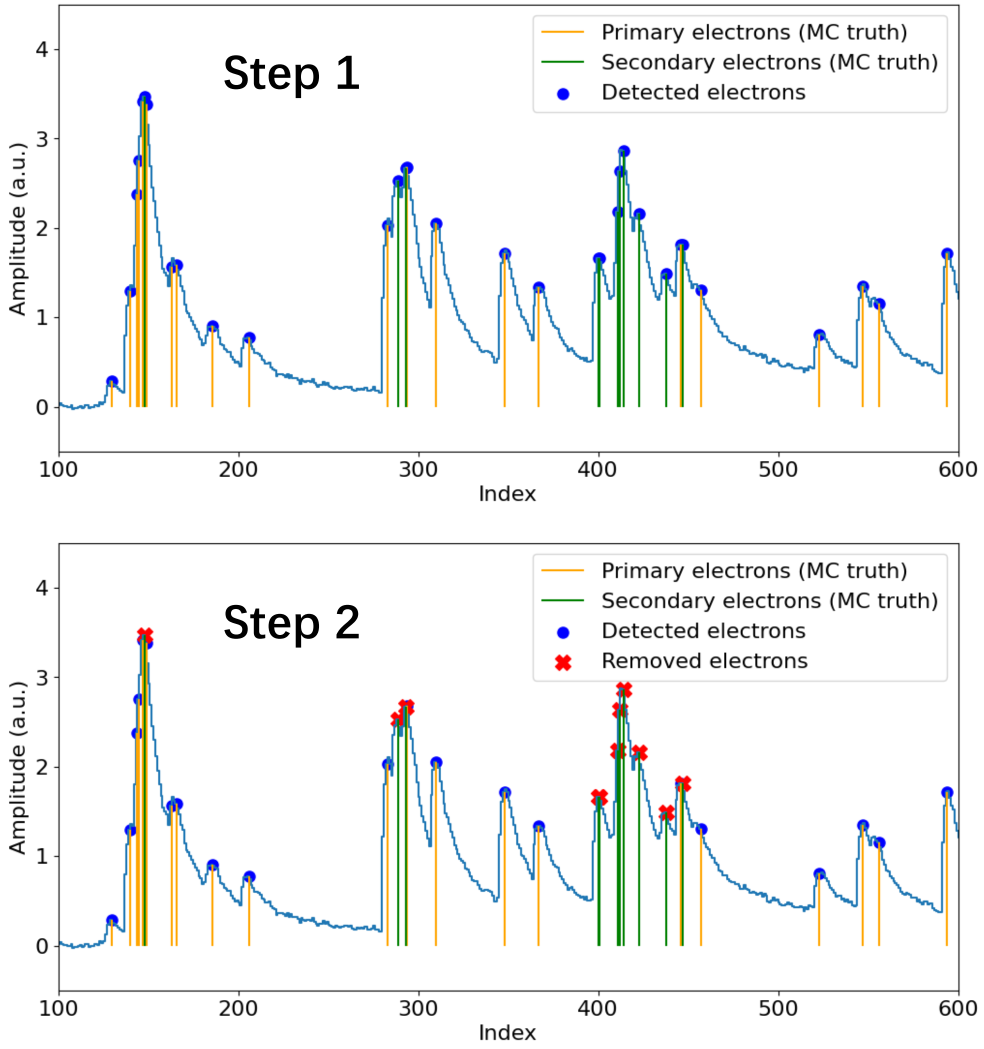
dN/dx reconstruction

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



As the name "**cluster counting**" implies, dN/dx reconstruction is to determine the number of **primary electrons** in the waveform

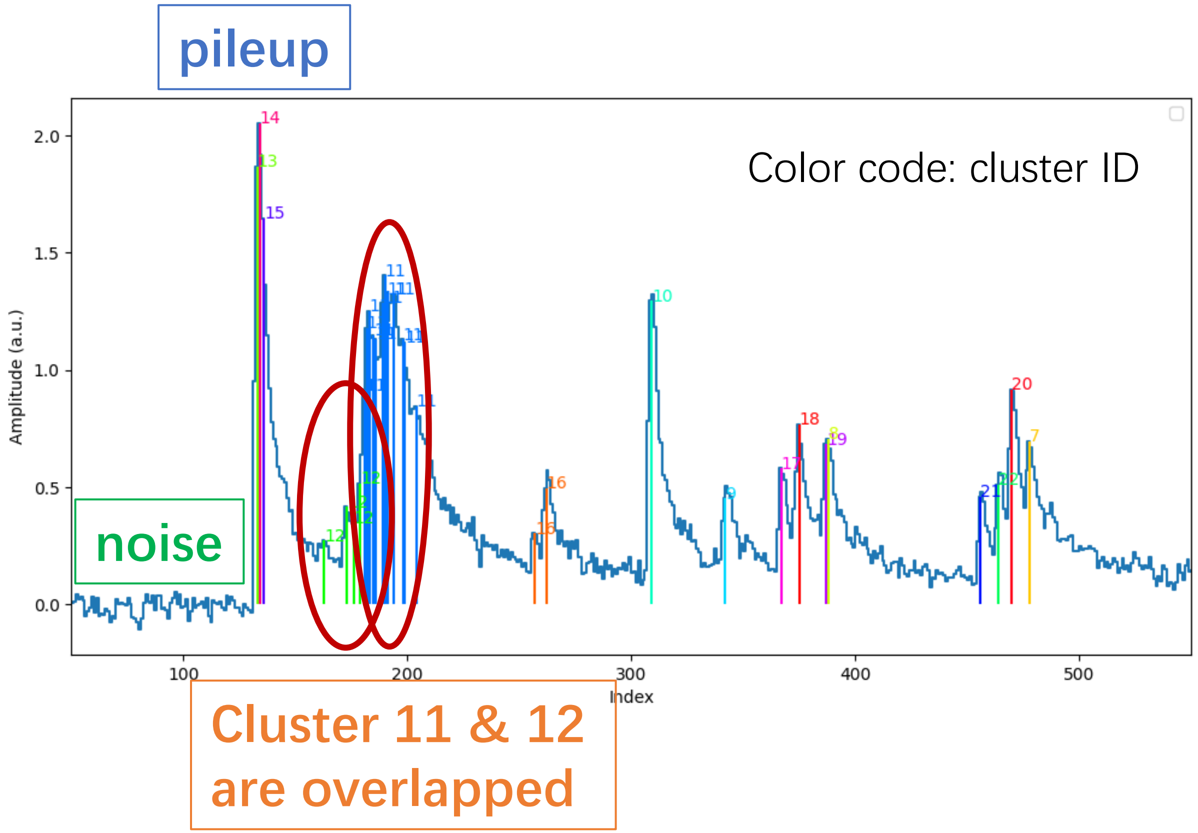
dN/dx reconstruction



2-step algorithm

- **Peak finding: find "electron" signals**
  - By detecting peaks from both primary and secondary electrons
- **Clusterization: find "cluster" signals**
  - By removing secondary electrons from the detected peaks in step 1

dN/dx reconstruction is challenging



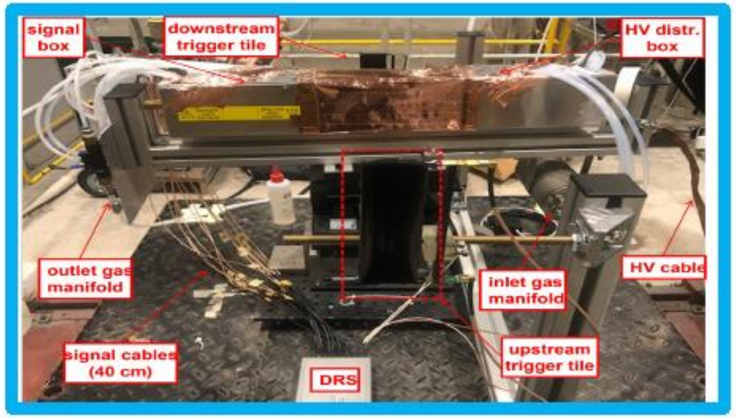
- Highly piled-up ➔ Difficult for an efficient peak-finding
- Noisy ➔ Filtering could (significantly) lose efficiency
- Overlapping between clusters ➔ Difficult for clusterization

Solution: Deep learning

Software package and data samples

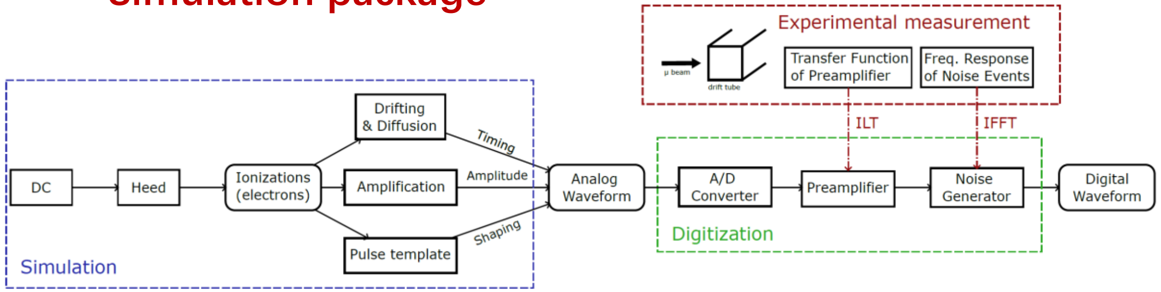
- **Simulation package**
  - Garfield++-based simulation + data-driven digitization
- **Data samples**
  - Simulated samples
    - 0-20 GeV/c pions and kaons
  - Experimental samples
    - 180 GeV/c muons from CERN/H8 beam

Test beam at CERN

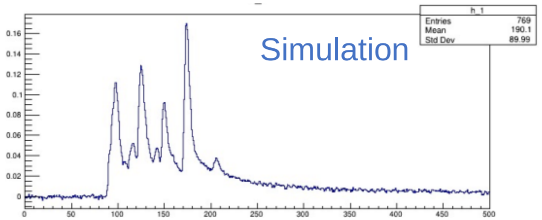


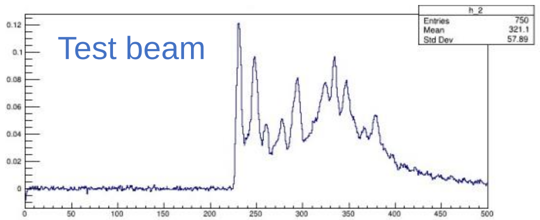
From INFN group led by Franco Grancagnolo and Nicola De Filippis

Simulation package



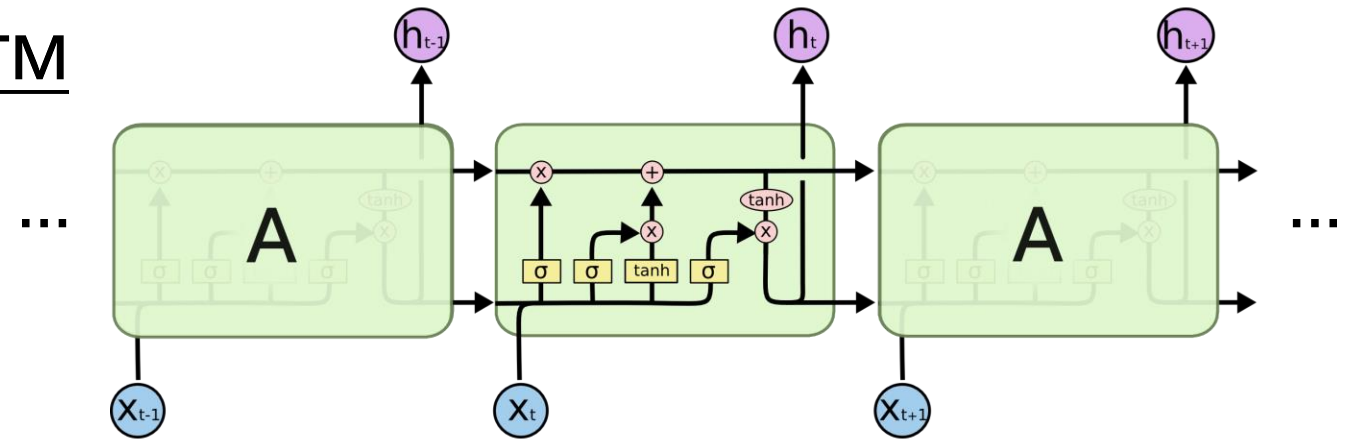
Tuned MC is comparable to data





Alg. 1: Supervised learning for simulated data

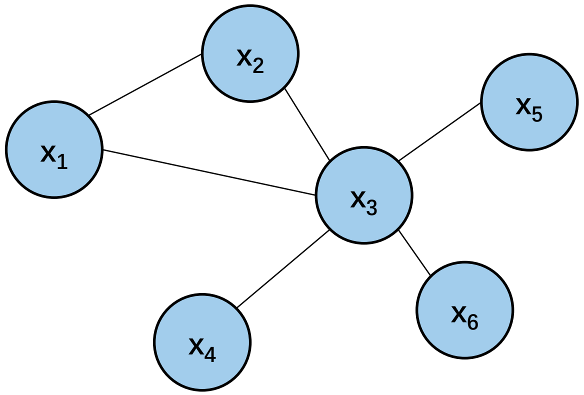
LSTM



LSTM-based peak finding:

- Can efficiently handle time-sequence
- Waveform slices as the LSTM input
- Binary classification of signals and noises

DGCNN



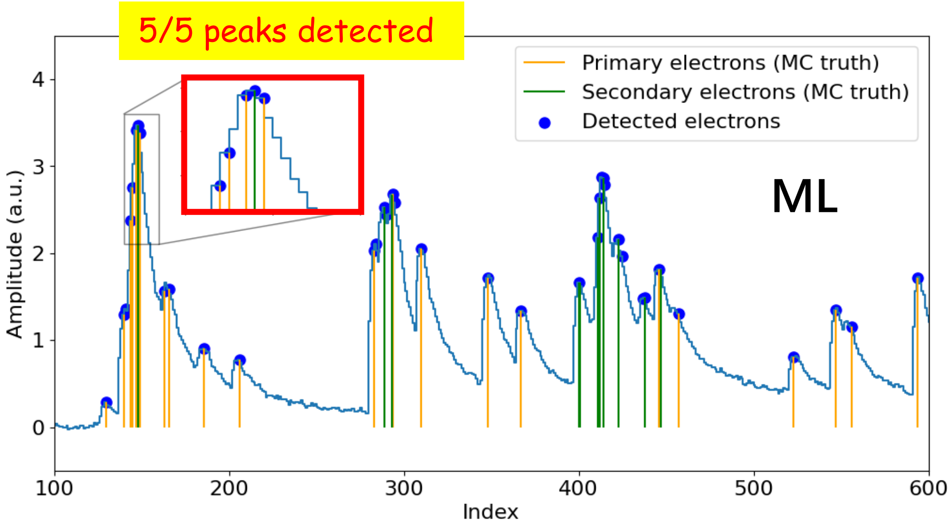
Dynamically connected Graph by k-NN

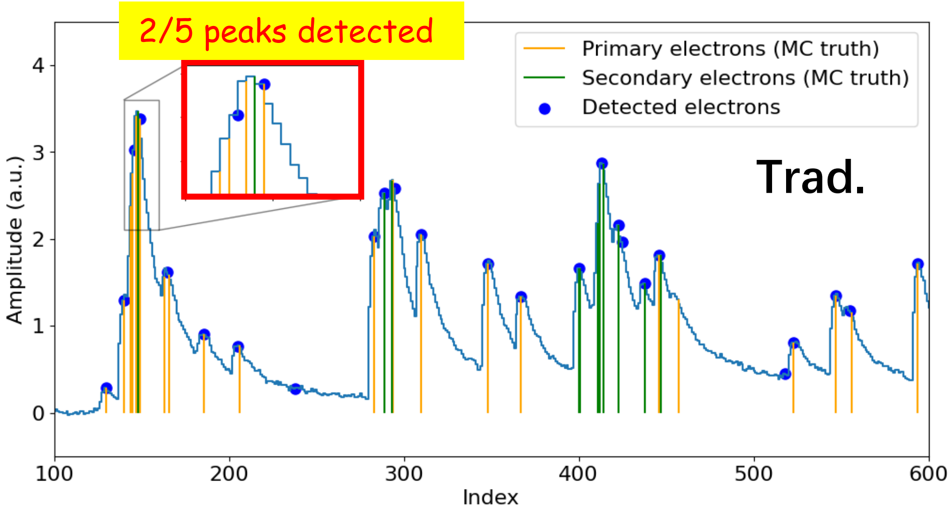
DGCNN-based clusterization:

- Incorporate local information to learn global properties
- Detected timings from the peak-finding as the DGCNN input
- Binary node classification of primary and secondary electrons

Nuclear Science and Techniques (accepted, 2402.16493)

Peak finding results





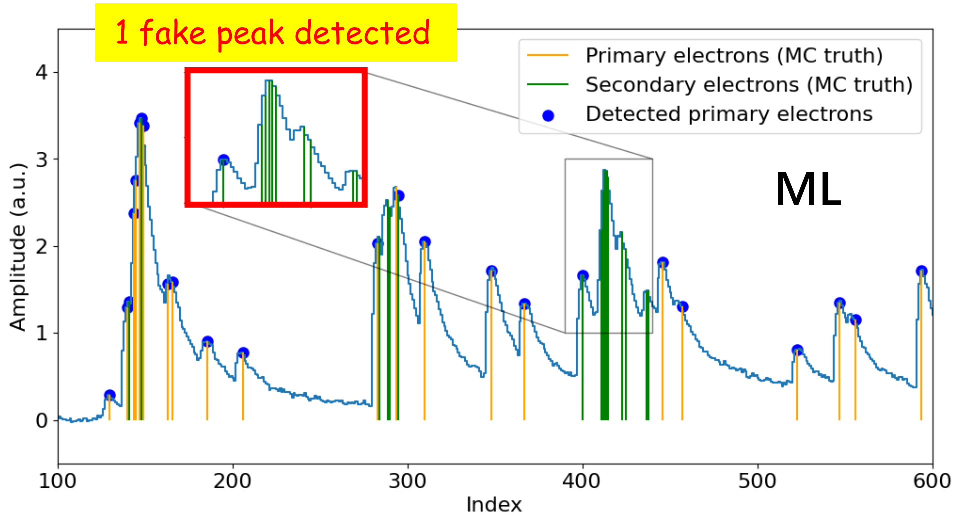
Traditional peak-finding: second derivative

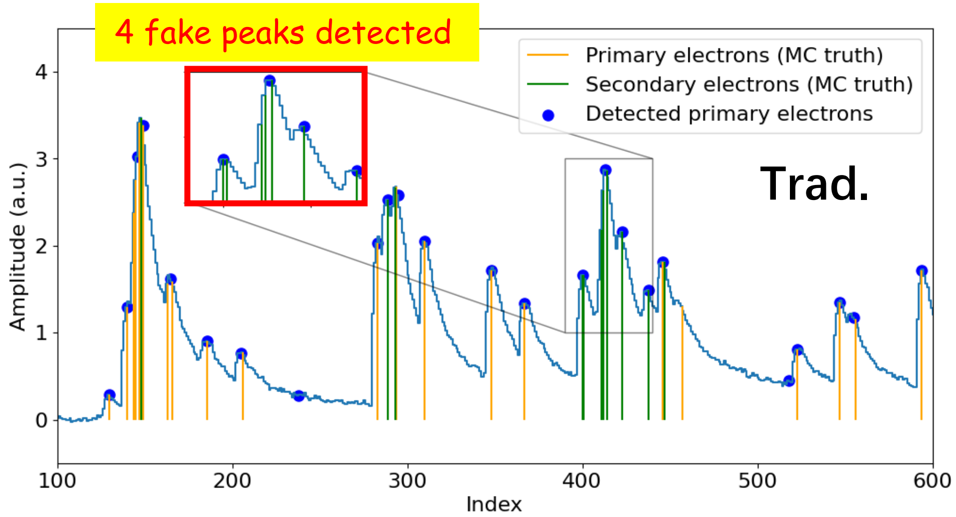
Table 2. The purity and efficiency comparison between LSTM-based algorithm and traditional D2 algorithm for peak-finding.

	Purity	Efficiency
LSTM algorithm	0.8986	0.8820
D2 algorithm	0.8986	0.6827

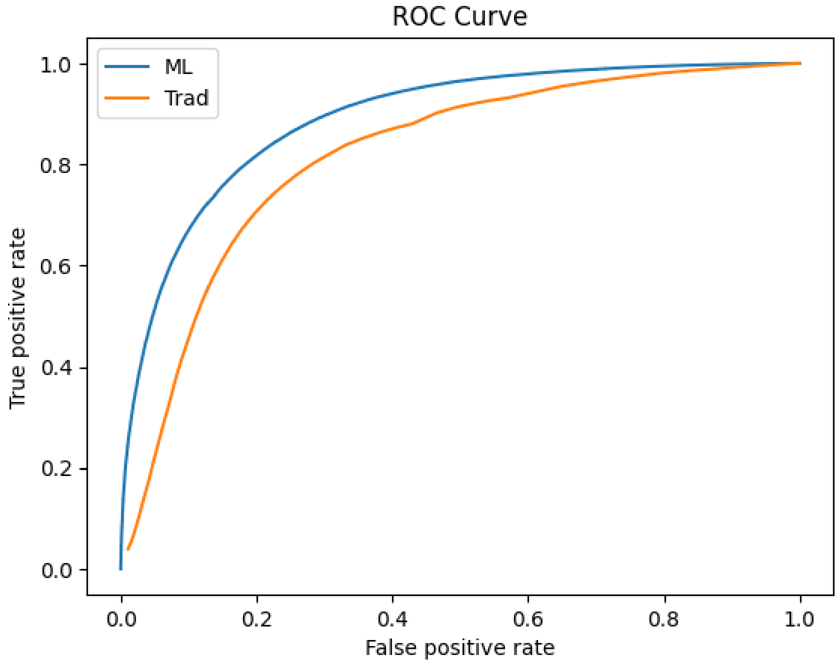
- The LSTM-based model is more powerful than the traditional derivative-based algorithm, especially for the pileup recovery

Clusterization results





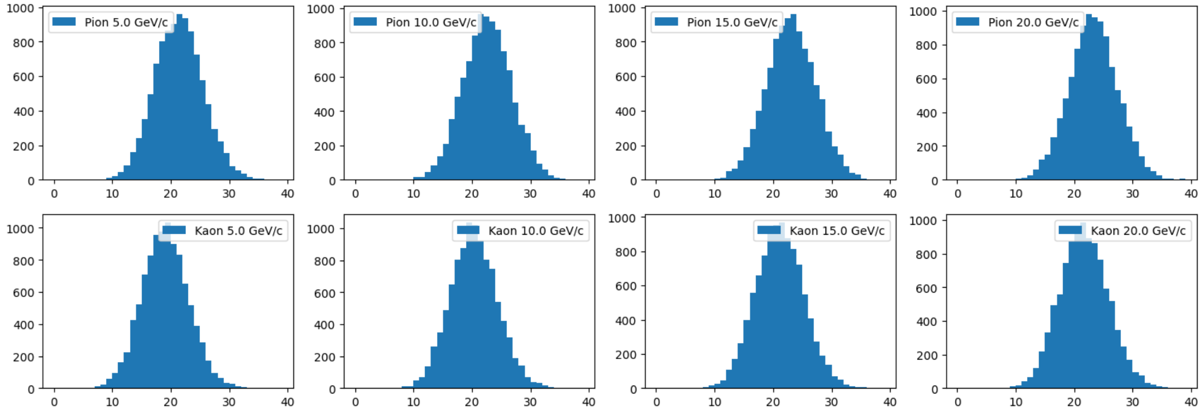
Traditional clusterization: adjacent-peak merge



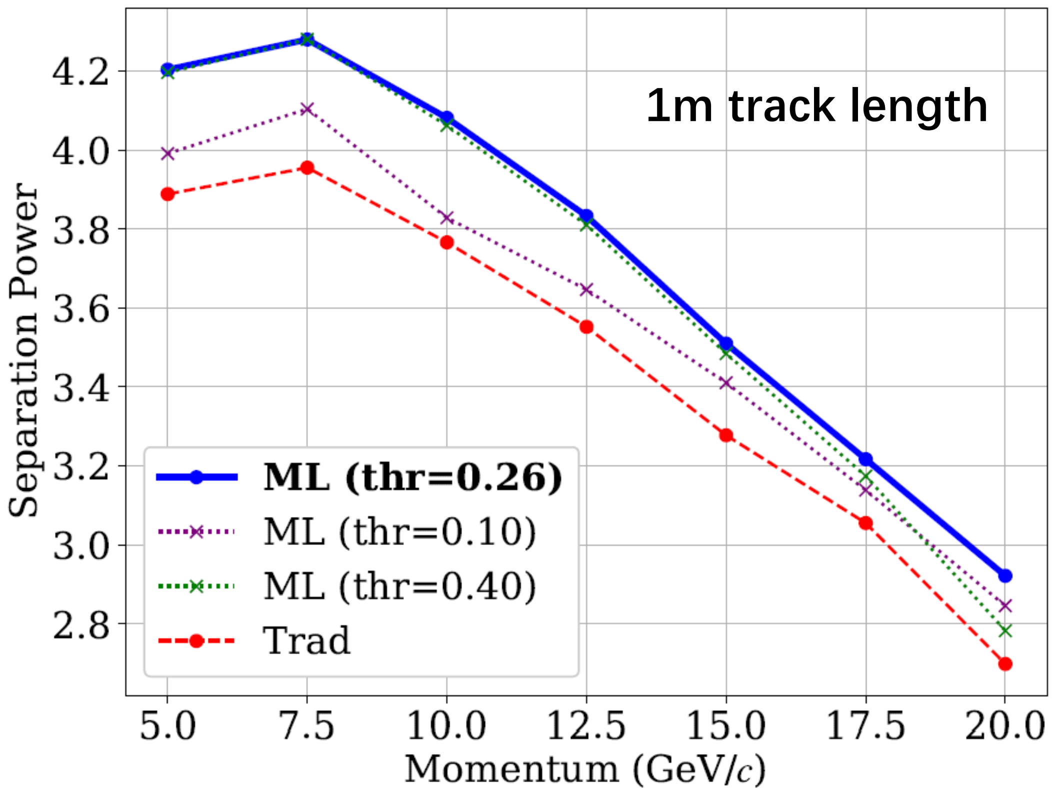
- The DGCNN-based model is more powerful than the traditional peak-merge algorithm, as it can remove the secondary electrons more accurate

PID performances with supervised models

Reconstructed # of clusters distributions



K/π separation power vs. momentum



- Very good Gaussians ➔ Efficient secondary electron removal
- For 1m track length, dN/dx resolution < 3%, typical ~5% for dE/dx

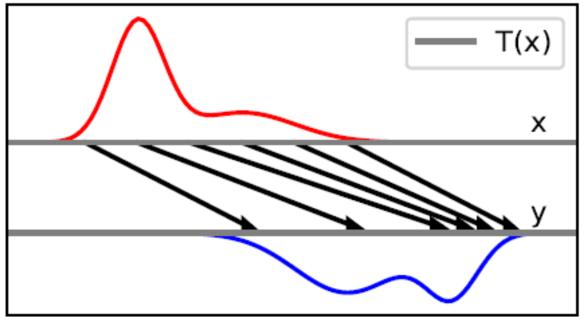
~10% improvement for ML (equivalent to a detector with 20% larger radius for trad. algorithm)

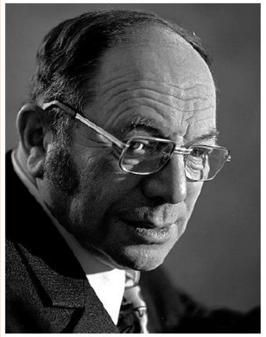
Alg. 2: Transfer learning for real data

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

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

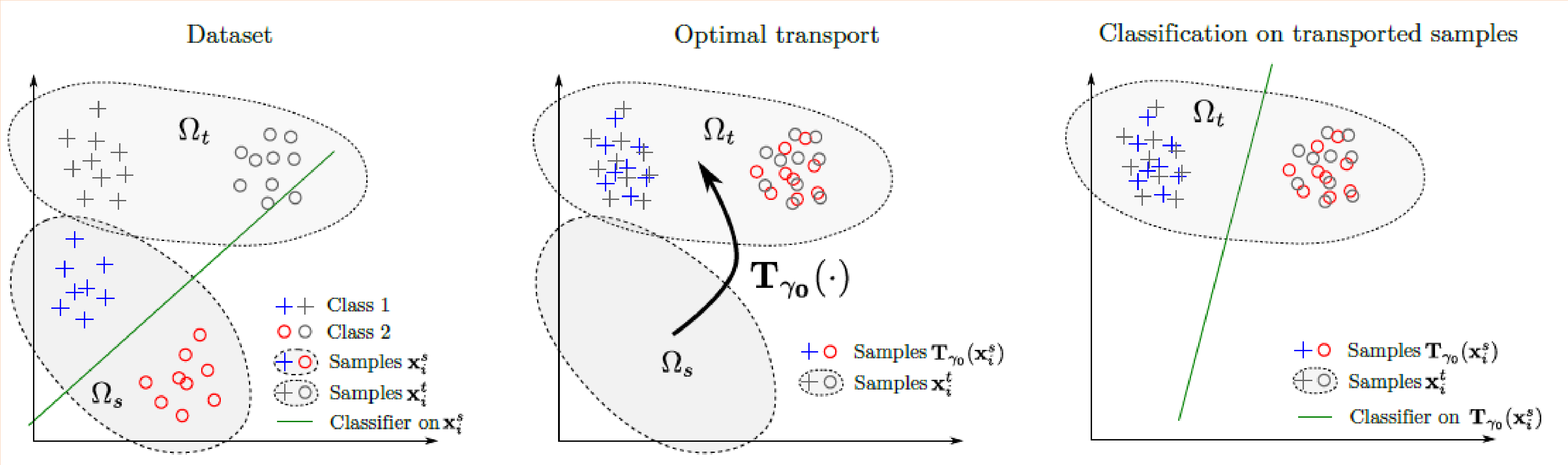
Optimal Transport





Kantorovich
(Economic Nobelist 1975)

Domain adaptation



Figures from Flamary's slides

Align data/MC samples with Optimal Transport

Semi-supervised domain adaptation

Loss for labeled samples
in source domain

* **Based on Deep-JDOT (1803.10081)**

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

Loss for labeled samples in target domain (THIS WORK)

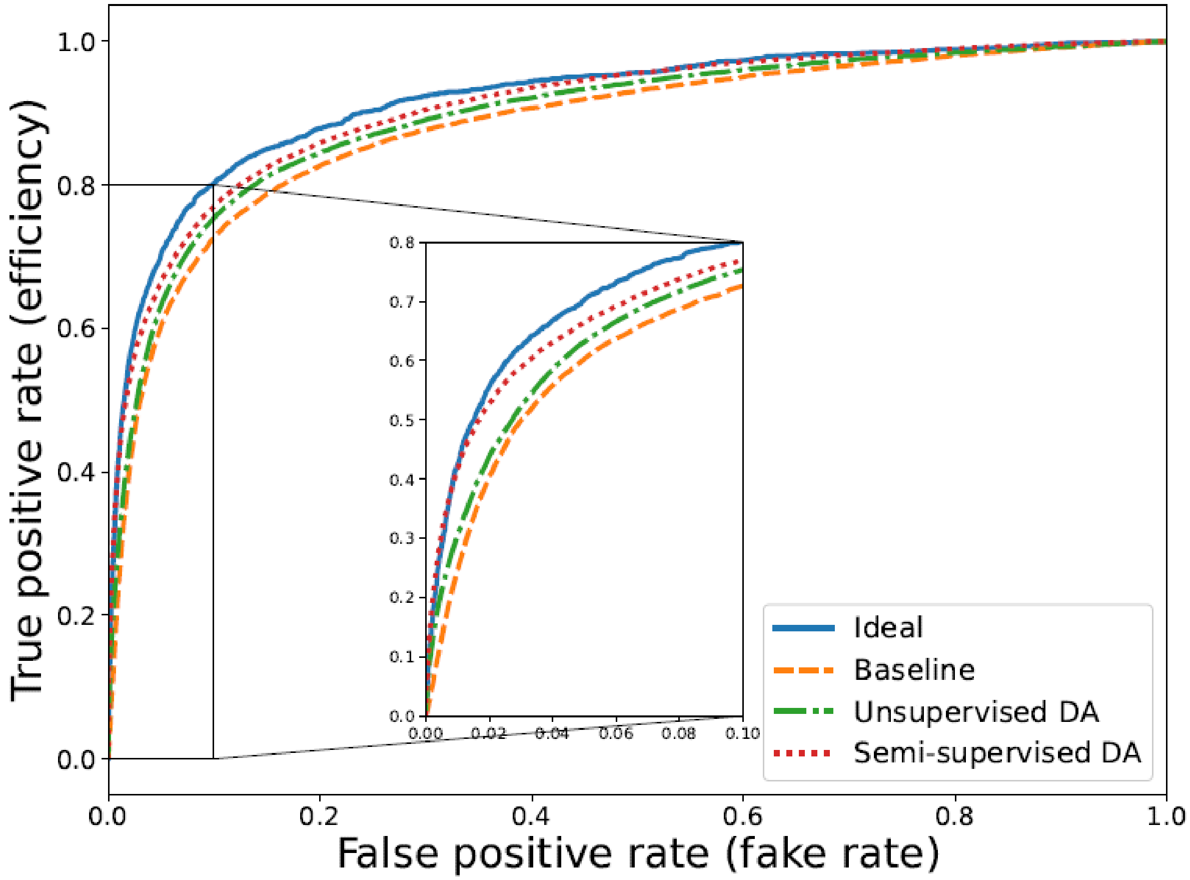
Cost of feature differences between source and target

Cost of 'label' differences between source and target

Cost of joint feature-label distribution for OT

Computer Physics Communication 300, 109208

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

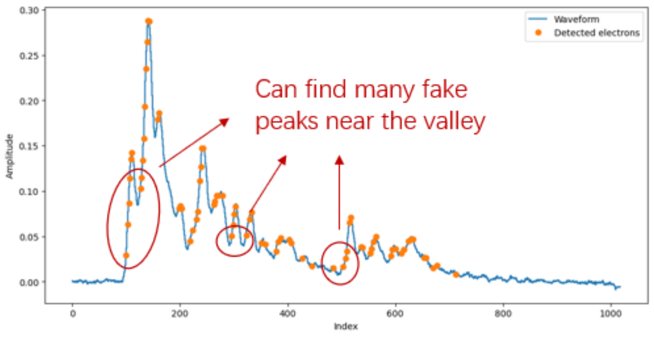
Improve

- **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**

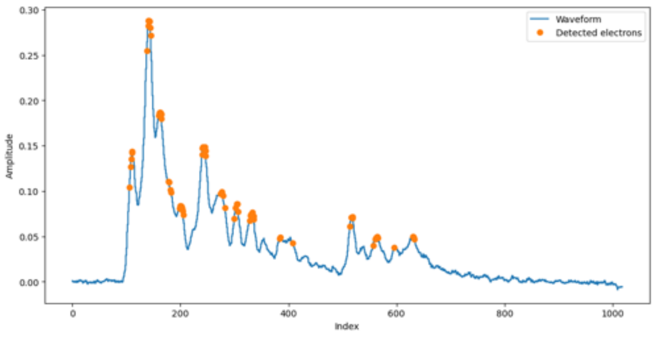
Peak finding for test beam data

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

Derivative

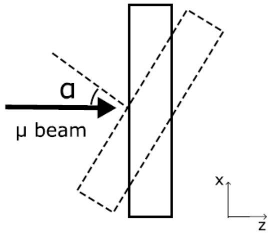


Semi-DeepJDOT

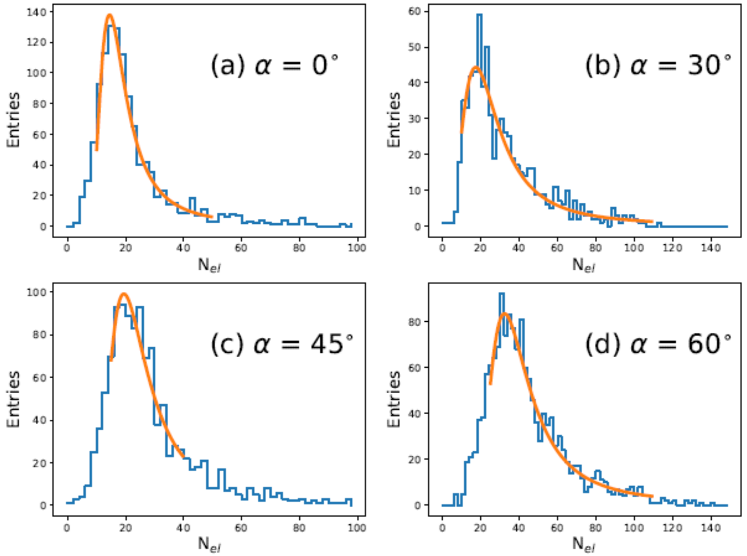


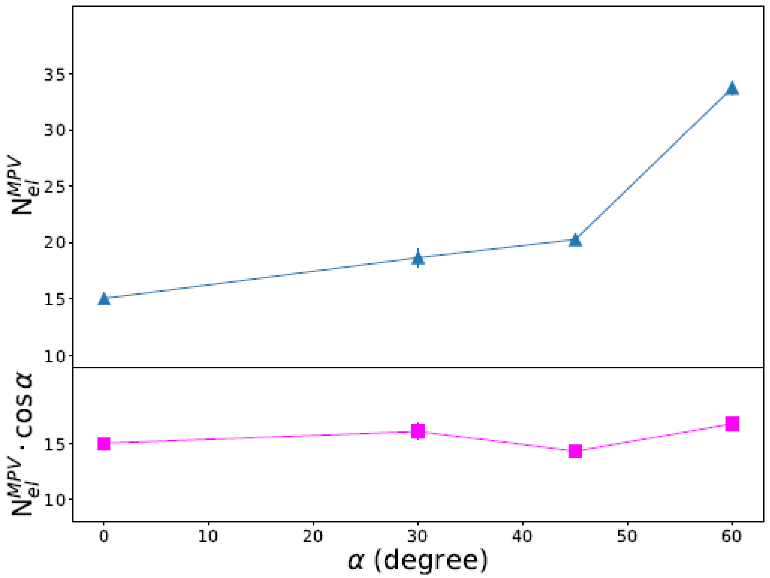
Note: Require similar efficiency for both cases

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

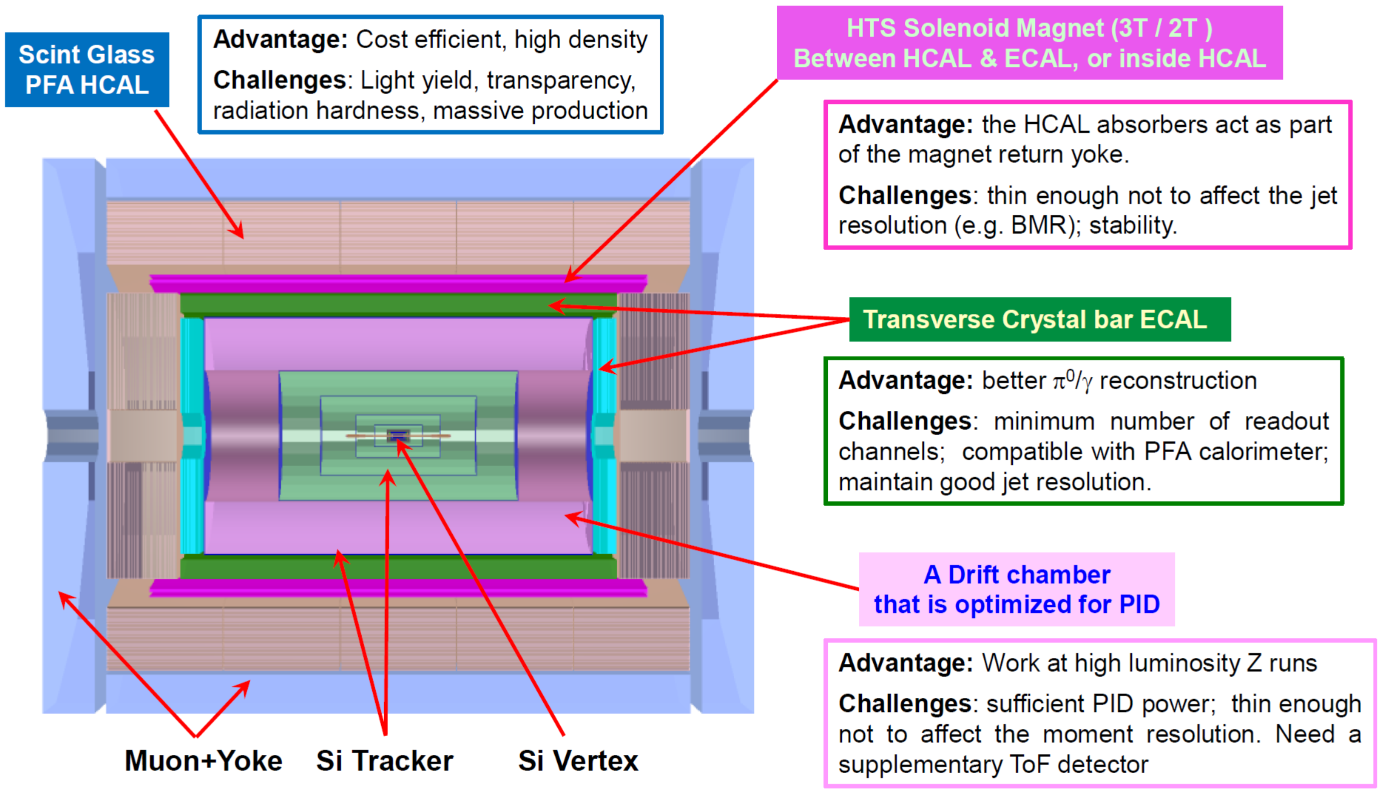
- dN/dx is the breaking through PID technique and its reconstruction is challenging. Two machine learning algorithms are developed for dN/dx reconstruction.
- The supervised model has 10% improvement on K/pi separation w.r.t. traditional algorithm. The situation could be similar for the semi-supervised domain adaptation model.
- When studied with the full-simulation samples using a supervised model, the PID performance achieves < 3% K/pi resolution and ~3σ K/pi separation for 1m track length.
- 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

Drift chamber with PID capability

The CEPC 4th concept



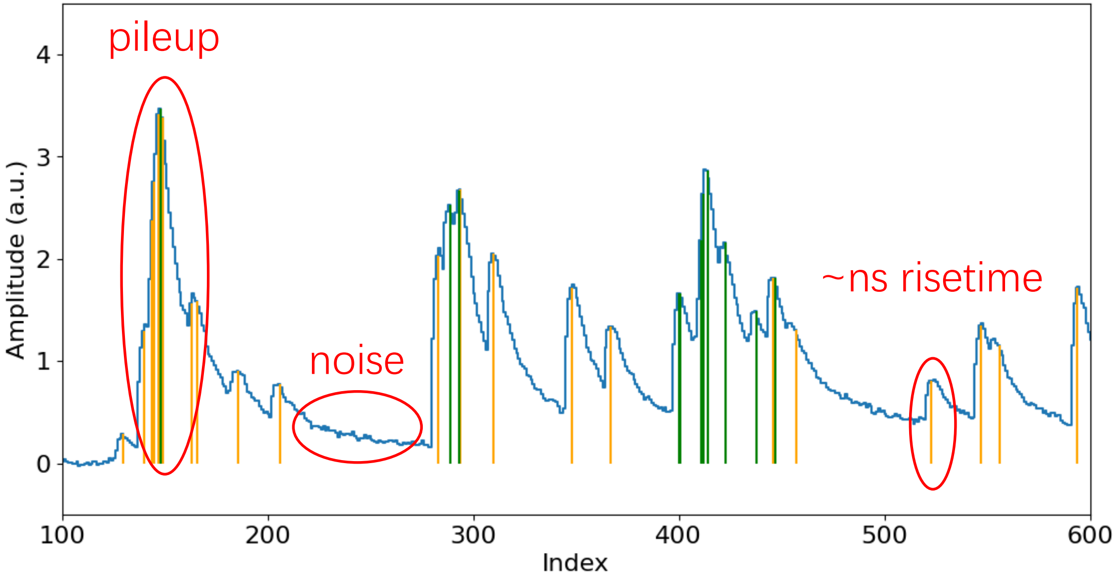
A drift chamber with cluster counting (dN/dx) is one of the gaseous detector options

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}

Challenges of dN/dx measurement

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

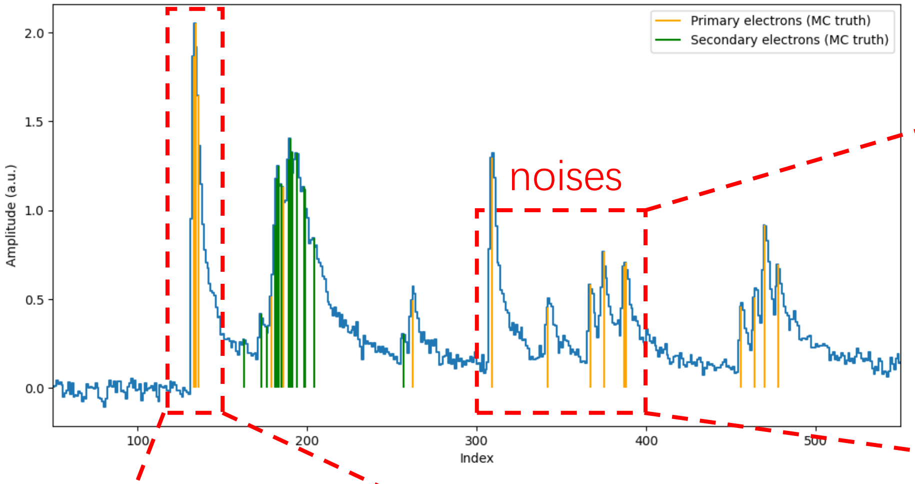


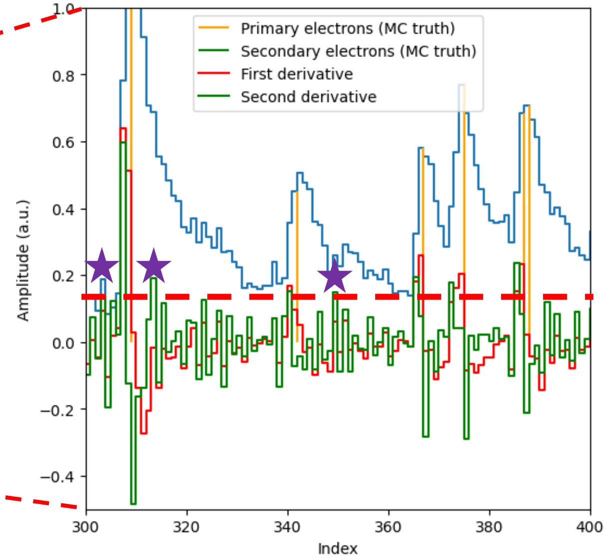
- **Single pulse risetime ~ns, require fast electronics**
  - Bandwidth > 1 GHz
  - Gain > 10
  - Sampling rate > 1.5 GS/s
  - Bit resolution > 12 bit

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

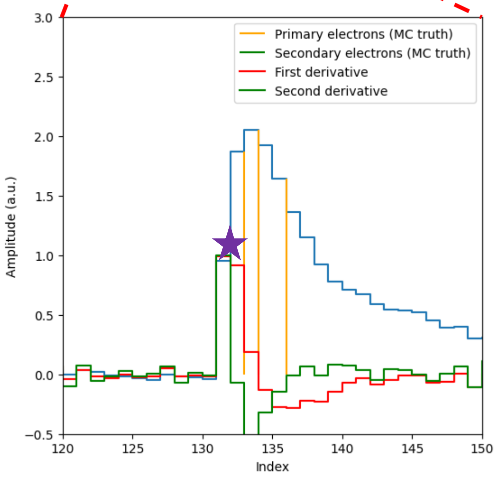
Traditional peak finding

pileup signals





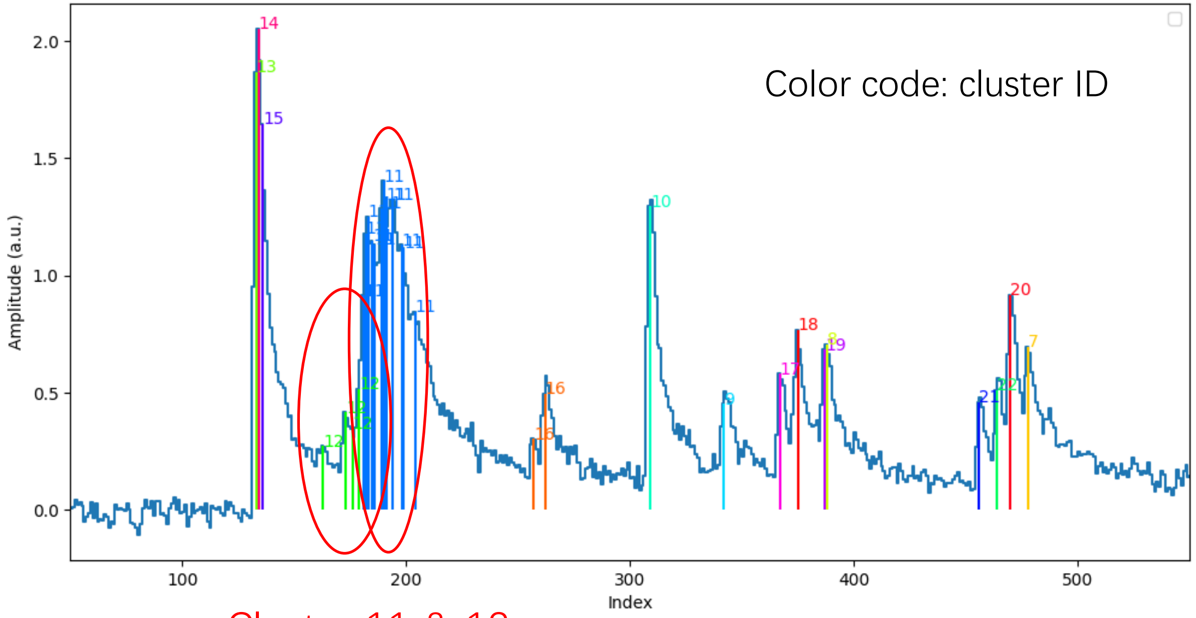
Some noises can also pass the threshold



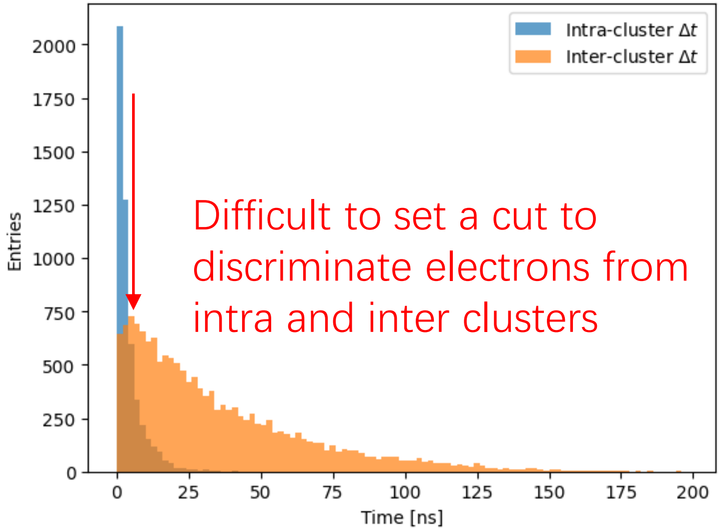
Only 1 out of 3 signals is detected

- **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

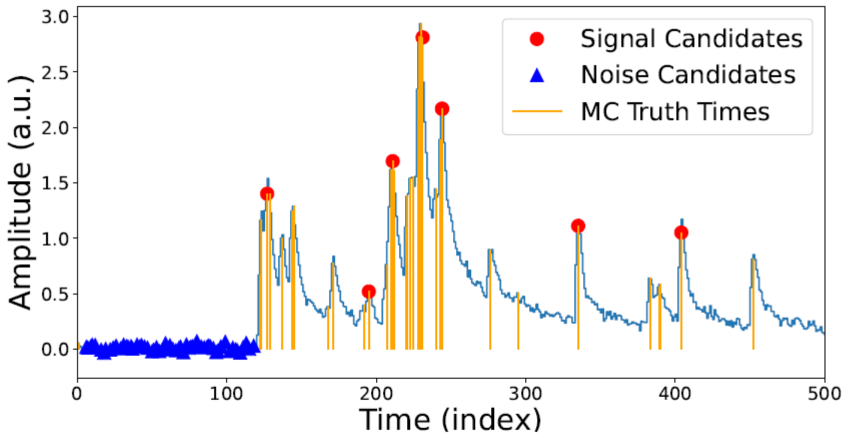


Figure 1: An example of simulated waveform. The blue histogram is the waveform. The red solid circles are the signal peaks selected by the CWT algorithm. The blue solid triangles are the noise peaks selected by requiring the 3 RMS requirement. The orange lines indicate the electron signal times from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

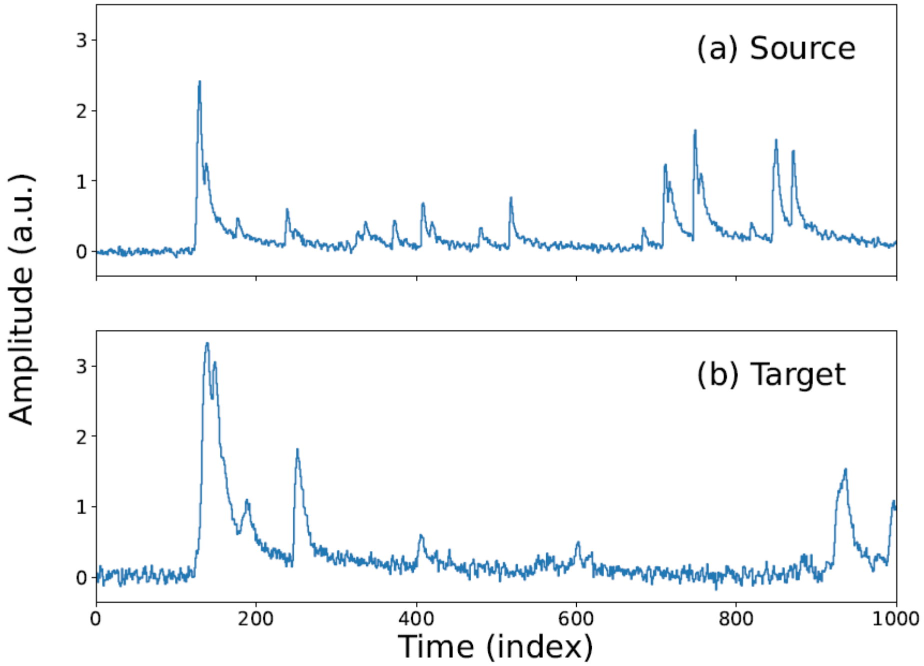


Figure 4: Waveform examples from the source sample (a) and the target sample (b). The source waveforms are generated with a noise level of 10% and a pulse risetime of 2 ns, while the target waveforms with a noise level of 20% and a pulse risetime of 4 ns.