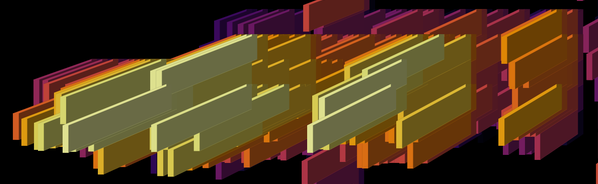
Neural network clusterization for the ALICE TPC online processing

Christian Sonnabend - ALICE PDP group
On behalf of the ALICE Collaboration



CHEP 2024, track 2: online processing - Krakow
23.10.2024

ALICE

CERN

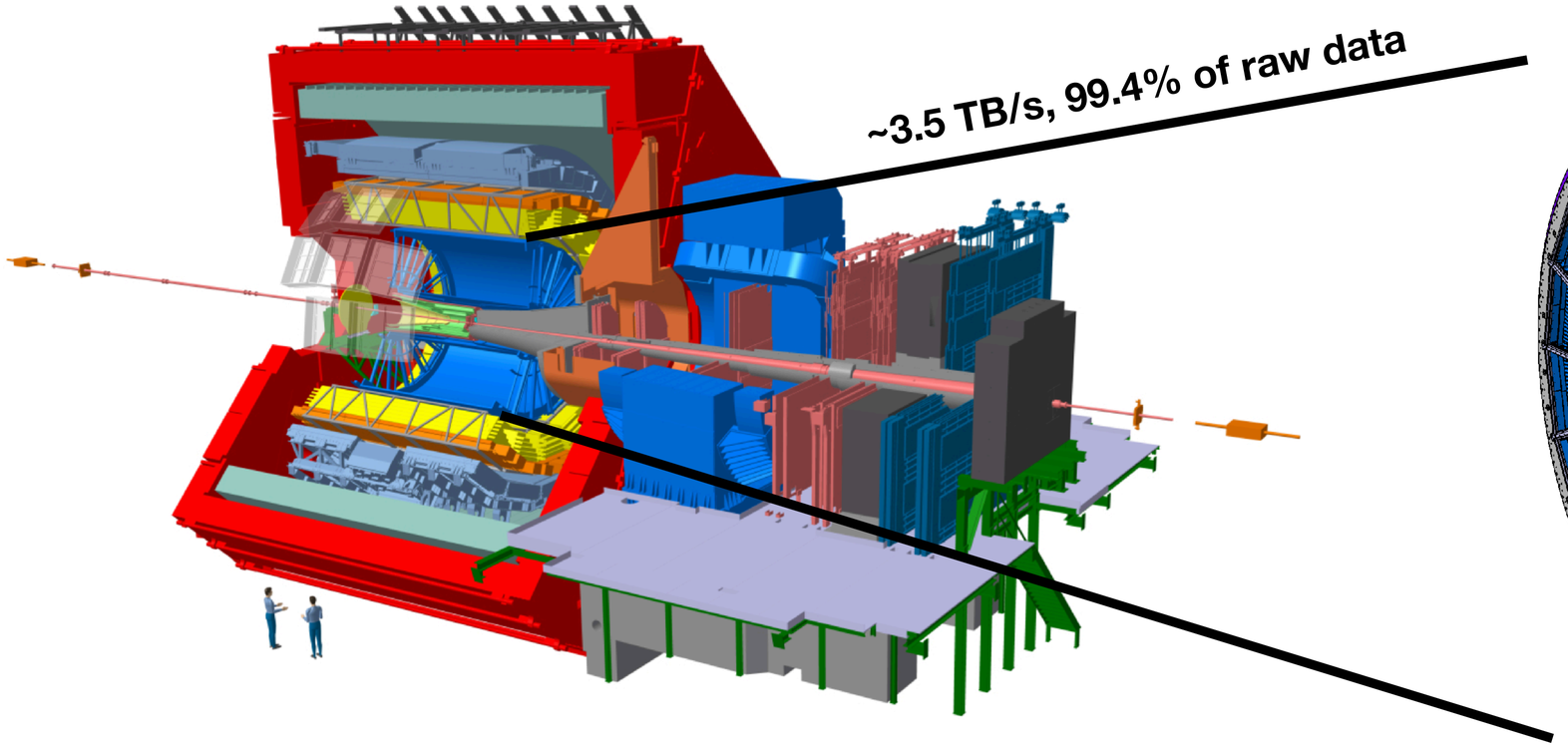
SPONSORED BY THE
Federal Ministry of Education and Research

UNIVERSITÄT HEIDELBERG
ZUKUNFT SEIT 1386

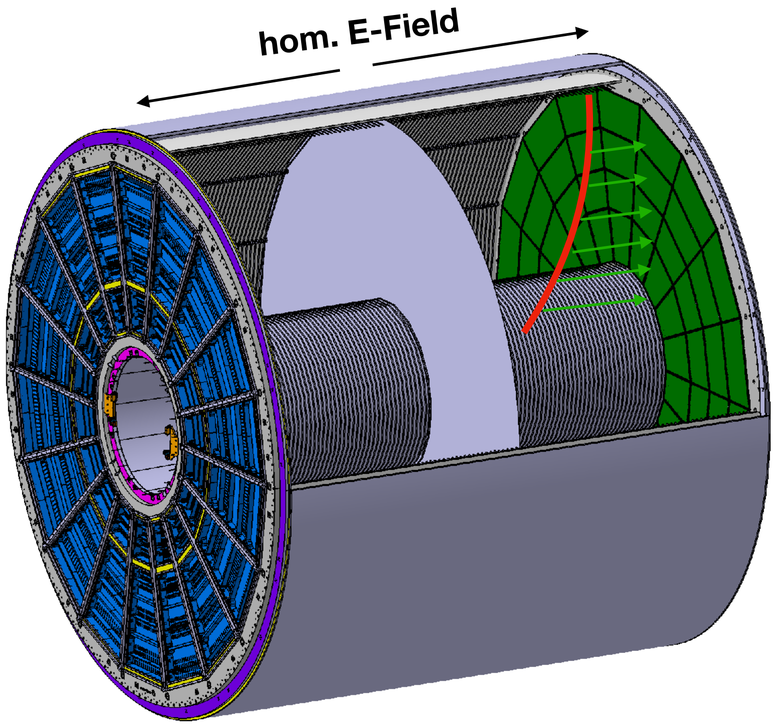
GSI Helmholtzzentrum für Schwerionenforschung

Introduction

ALICE



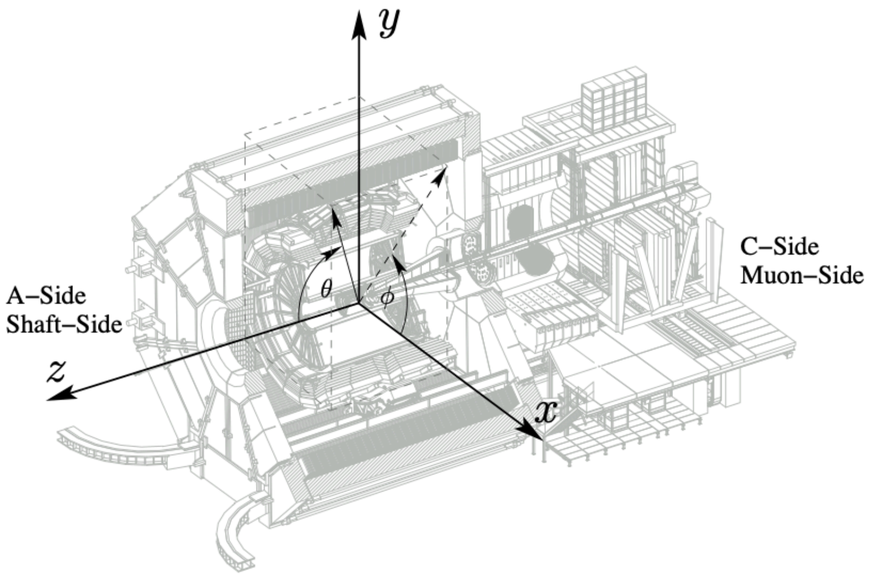
TPC - Time projection chamber

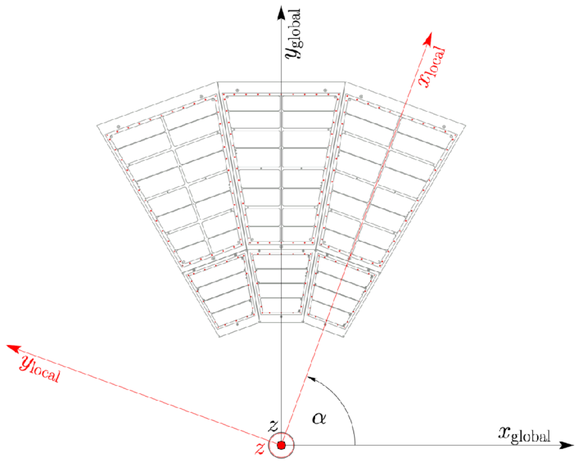


Central tracking and PID detector

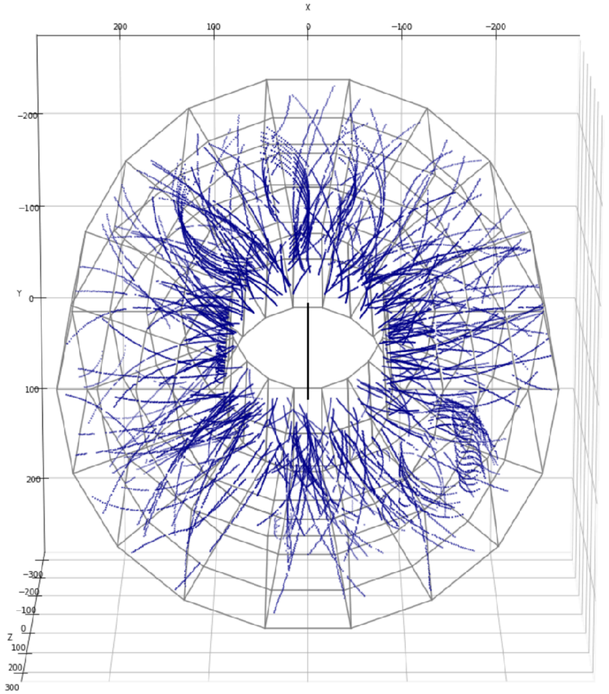
Introduction

The ALICE coordinate system





The ALICE TPC



Raw electrons from tracks → **GEM amplification stage** → **Readout** → **Digits with measured charge value** → **Clusters**

From clusters we can do e.g. tracking or PID via dE/dx

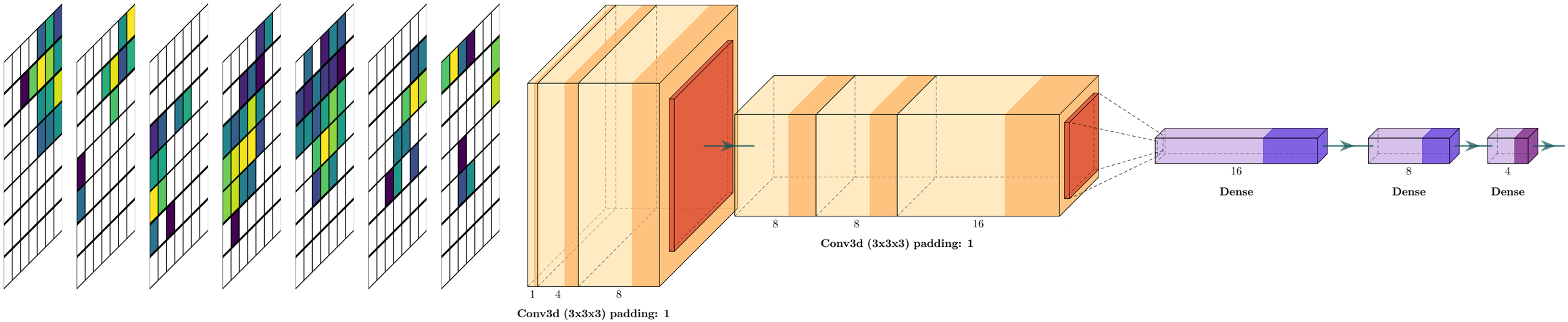
1) https://cds.cern.ch/record/1622286/files/ALICE-TDR-016.pdf

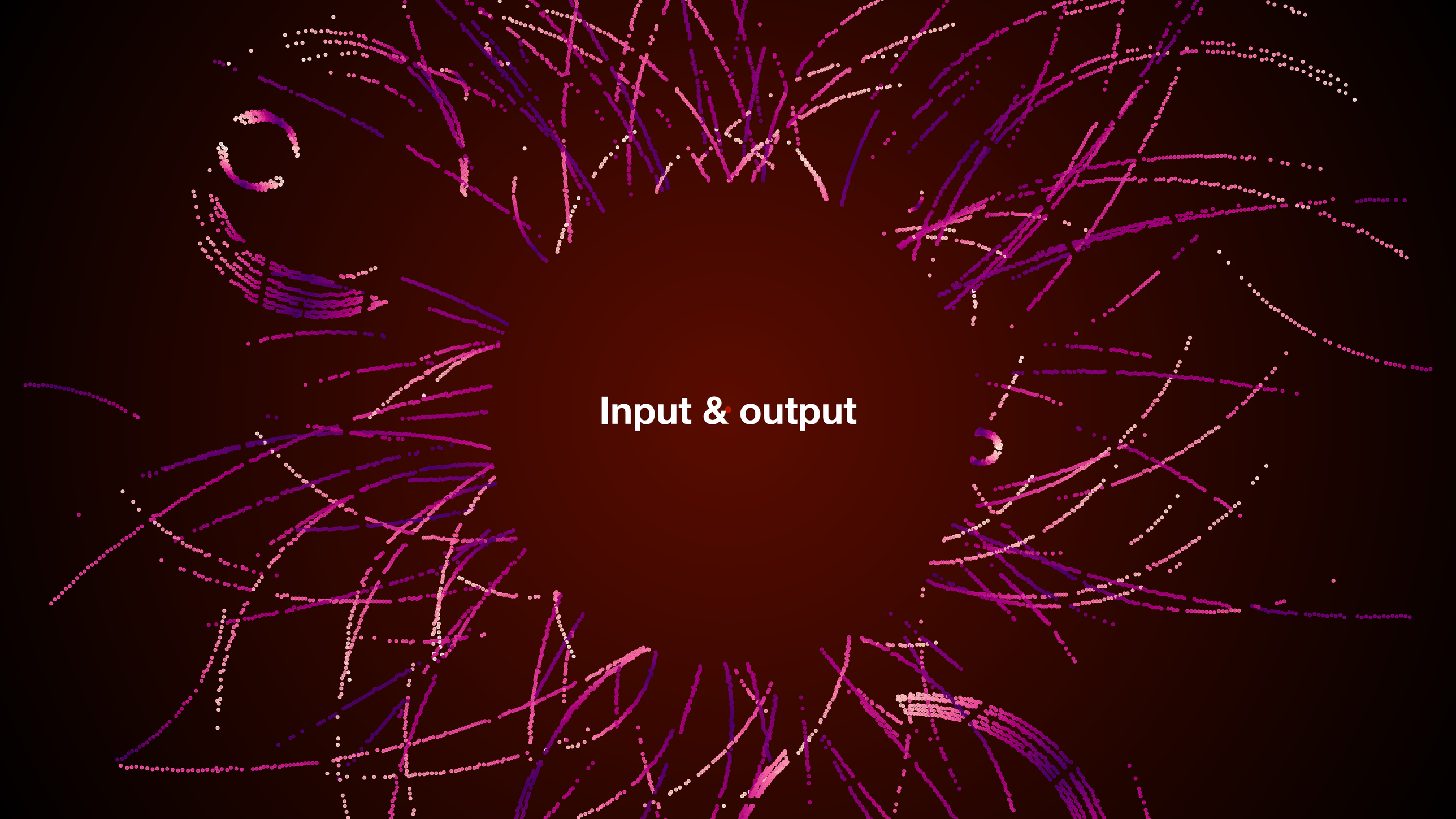
This project

Neural networks for online cluster classification and regression

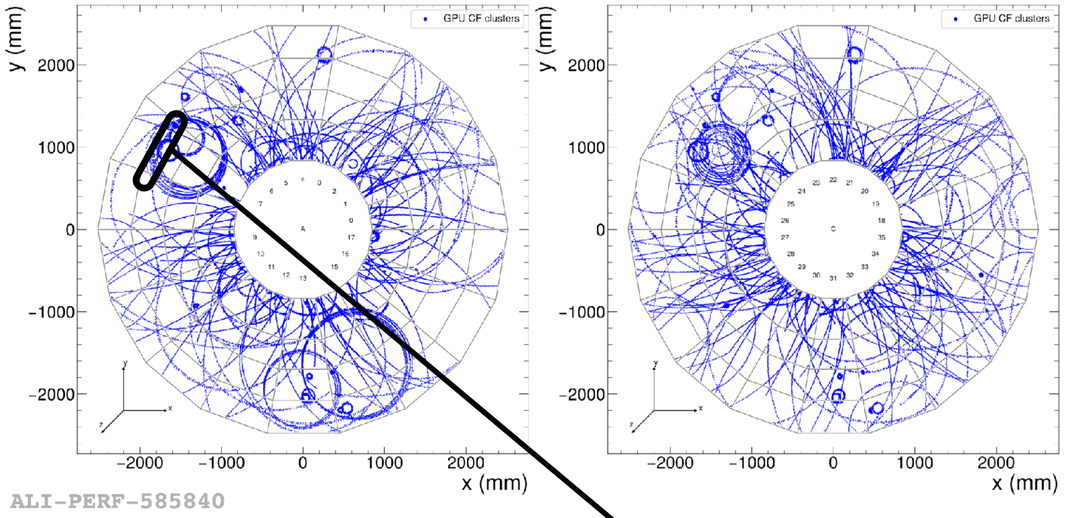
- Classification: Should digit maximum be converted to cluster or be rejected
- Regression: Predict cluster position, sigma, total charge and momentum vector
- Splitting: Should a cluster be split into two or more clusters

—> Make it fast enough for online processing



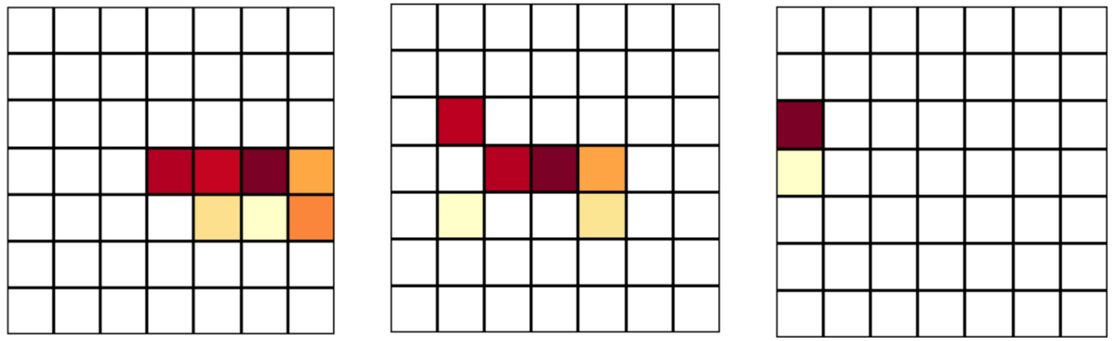

Input & output

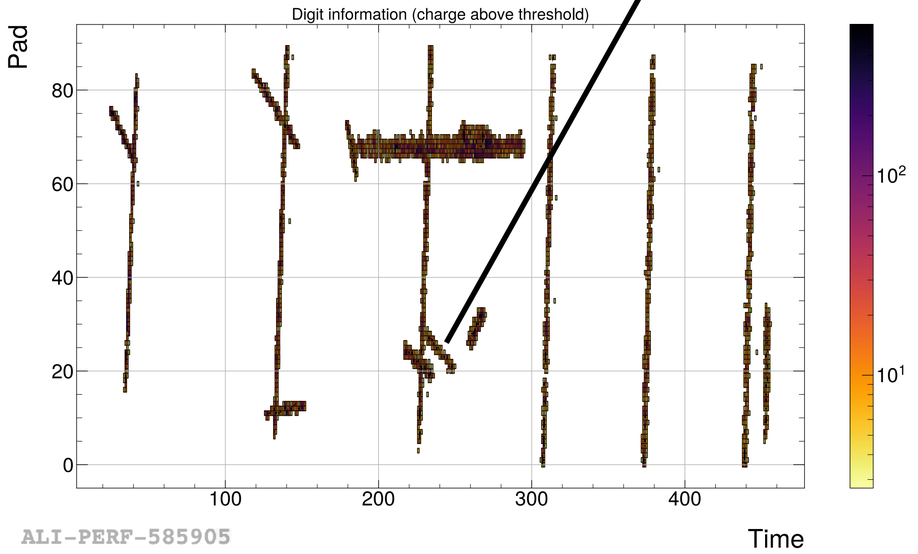
Data generation



Raw TPC digit data

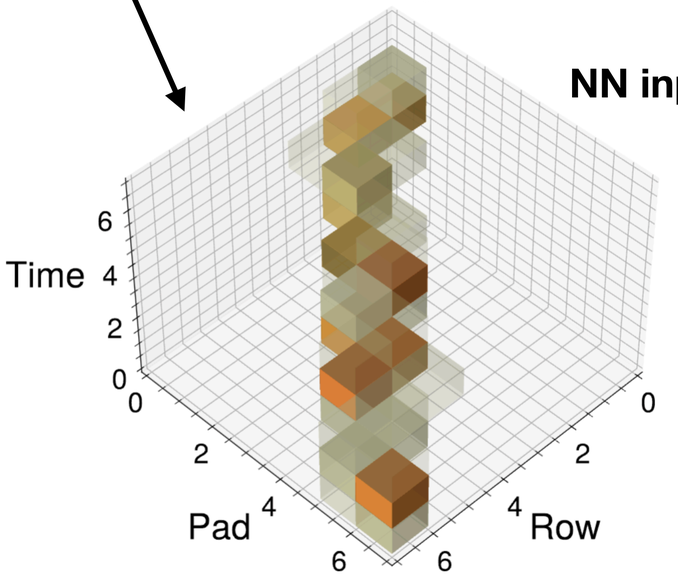
Adjacent rows





Network training data

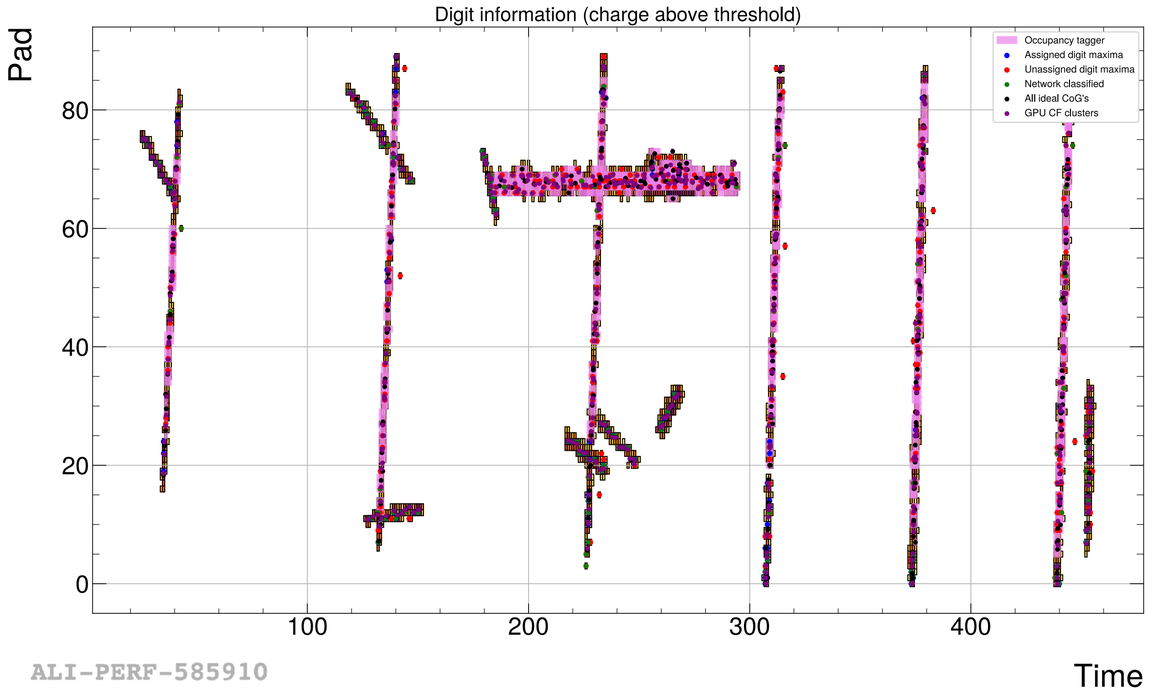
NN input



Assignment & training data selection

Simulation data

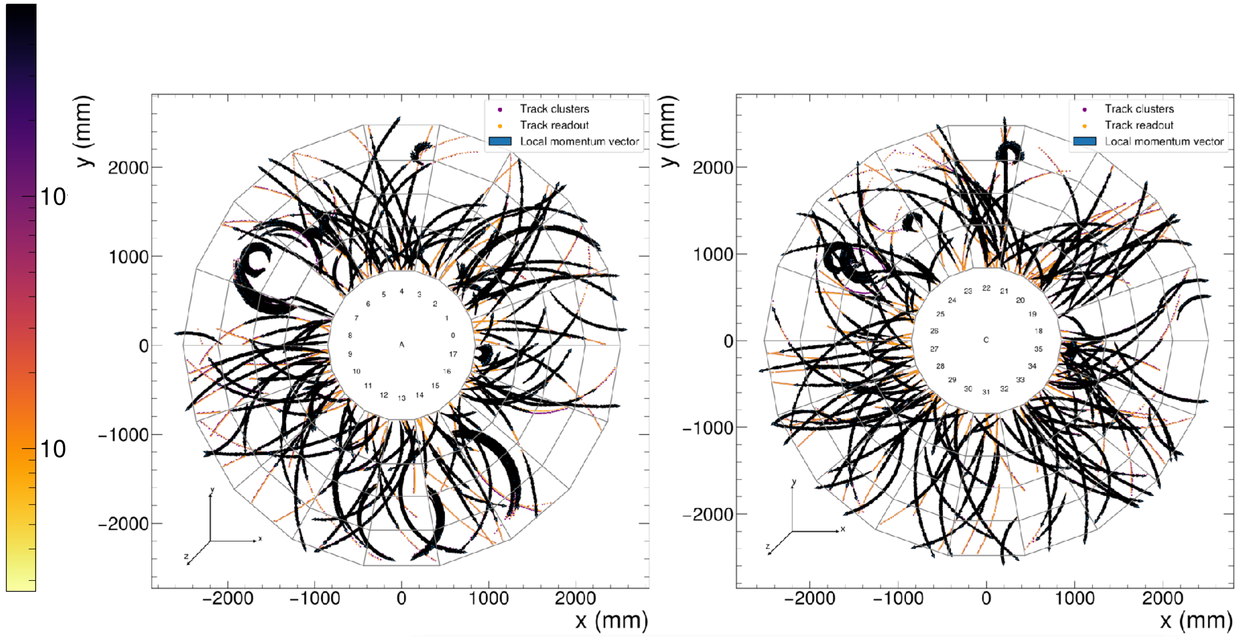
- Assign digit maxima with regular clusters
- Reject clusters if MC label occurs often in specified region of pad and time



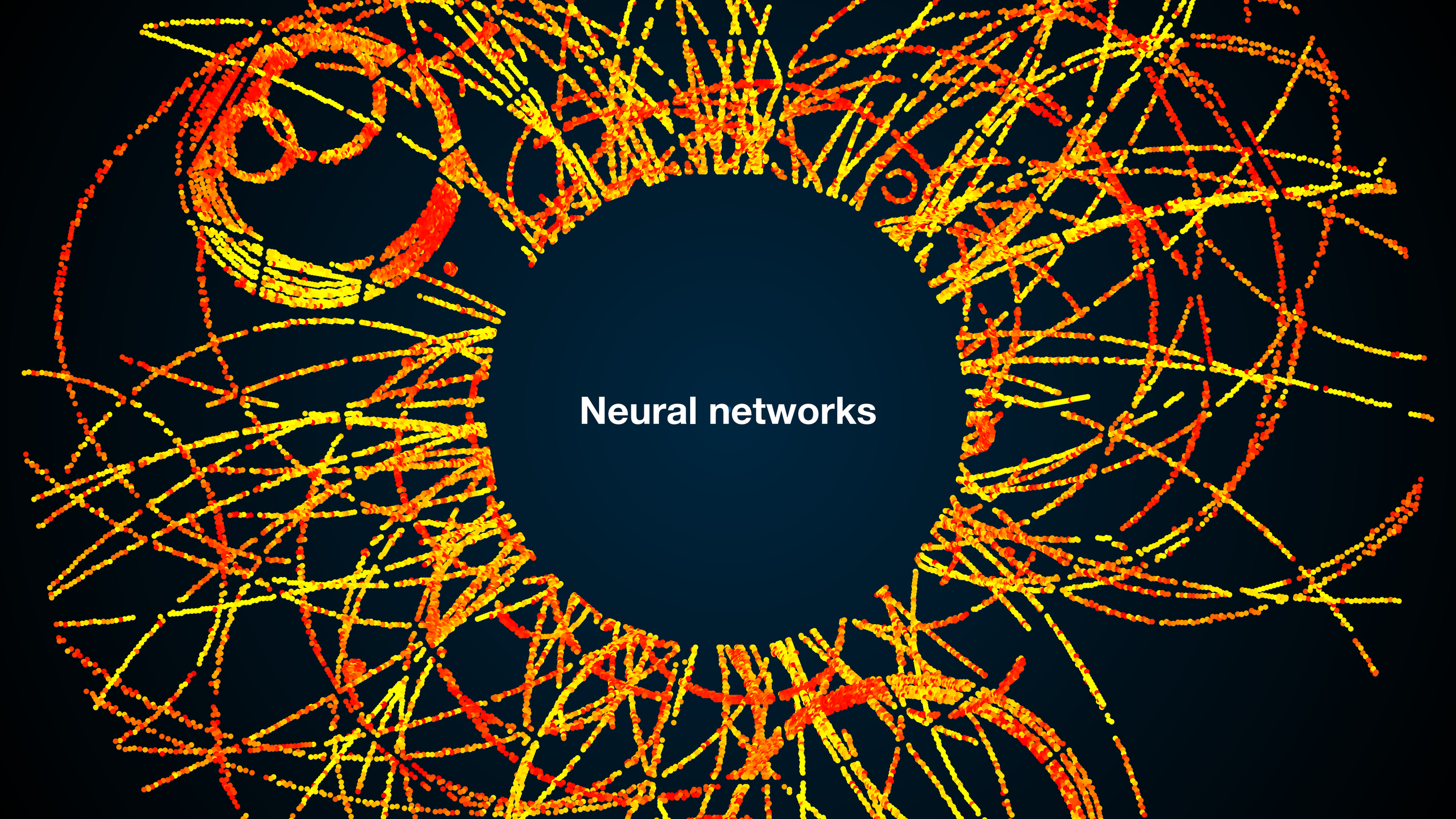
ALI-PERF-585910

Real data

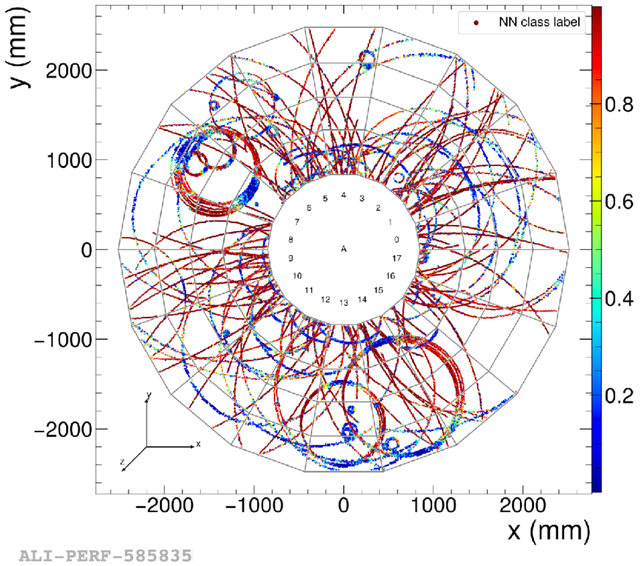
- Perform assignment between digit maxima and track paths
- Attach local momentum vector after reconstruction and reject clusters where loc. inclination angle is too high



ALI-PERF-585840

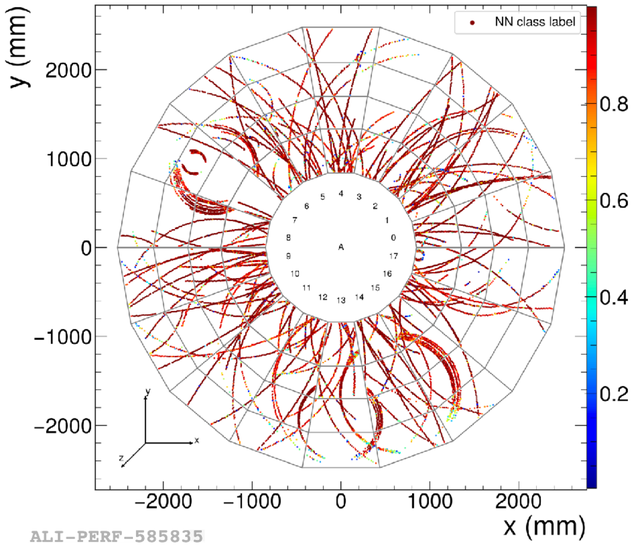

Neural networks

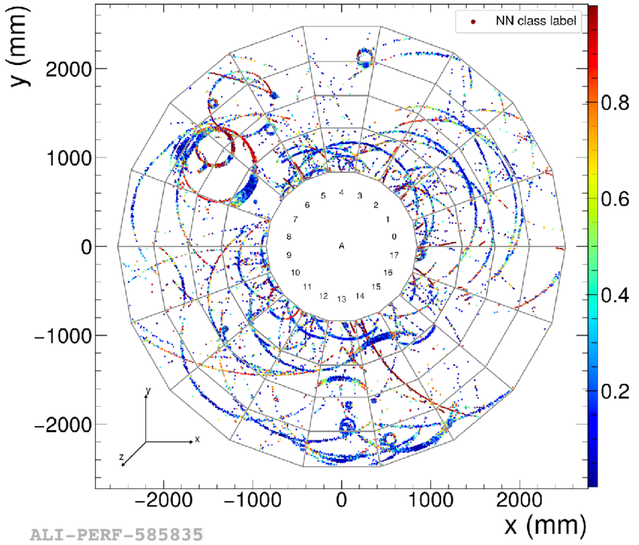
Classification network performance

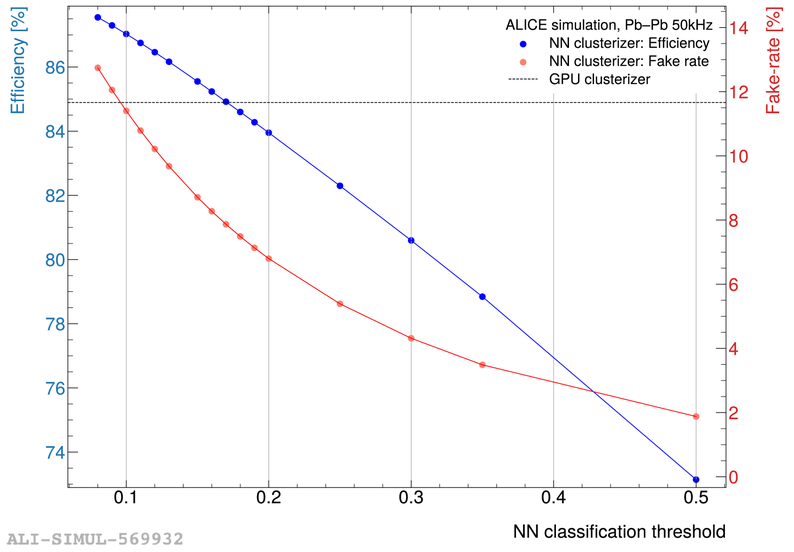


Should accept

Should reject



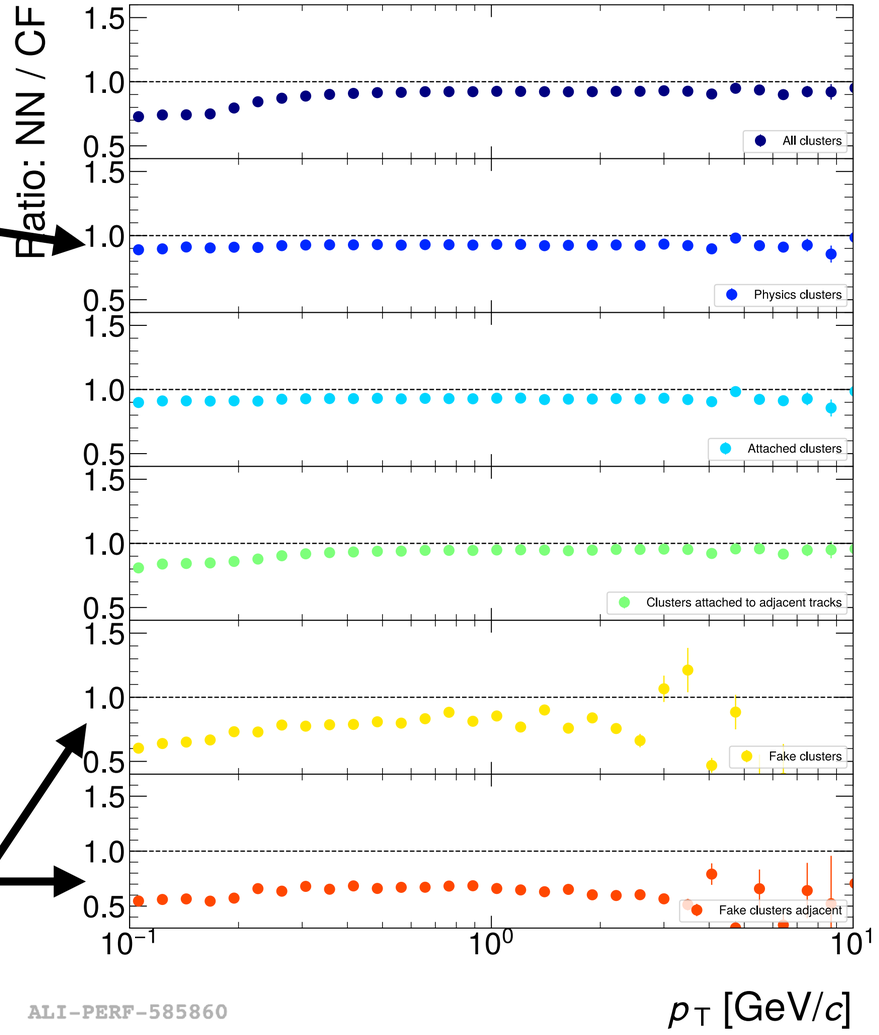




NN can reduce clusterization fake-rate by O(30%)!

Tracking and clusterization performance

More physics clusters by GPU cluster finder



BUT!

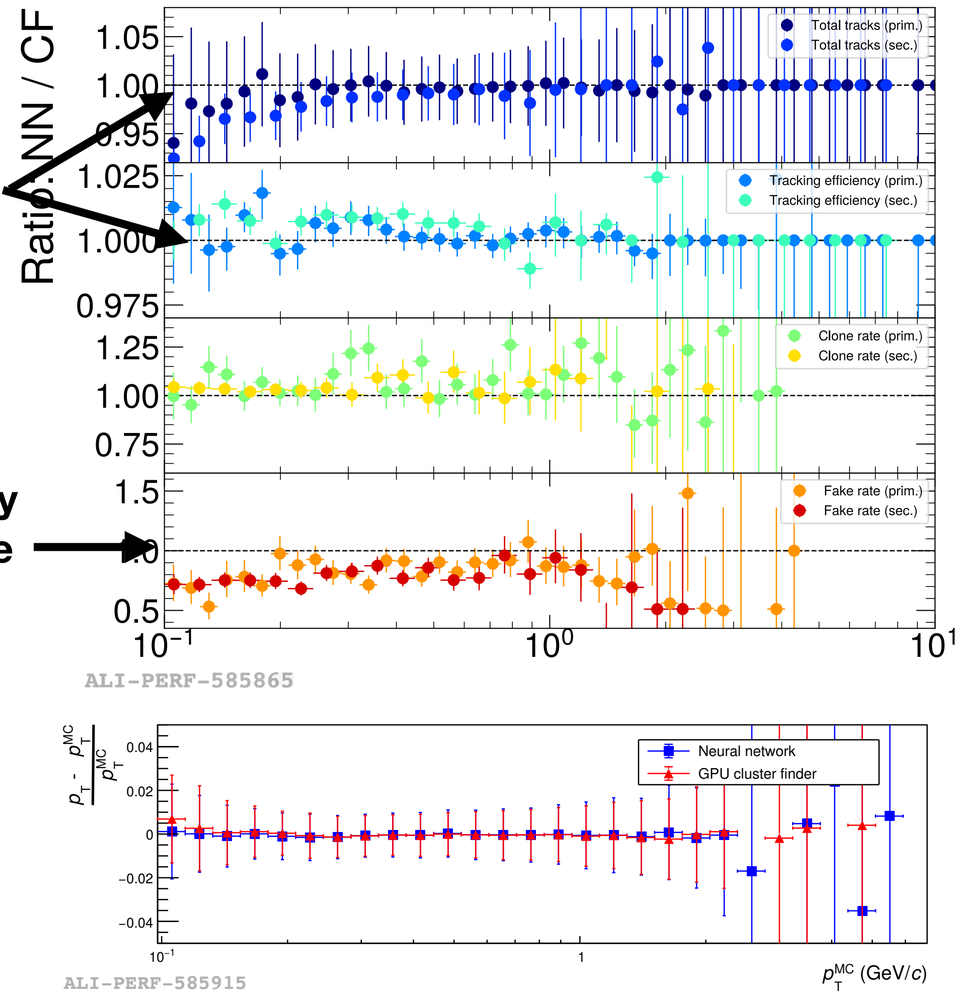
They are fakes!

Total clusters: 17.0 mio. (NN) vs. 21.4 mio. (GPU CF)

Maintain tracking efficiency

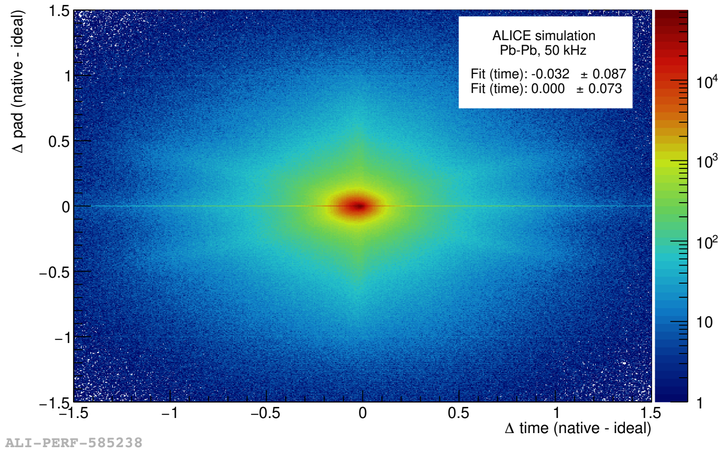
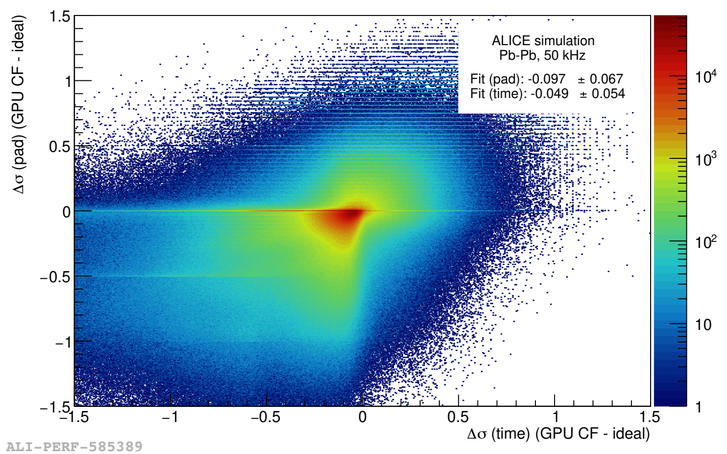
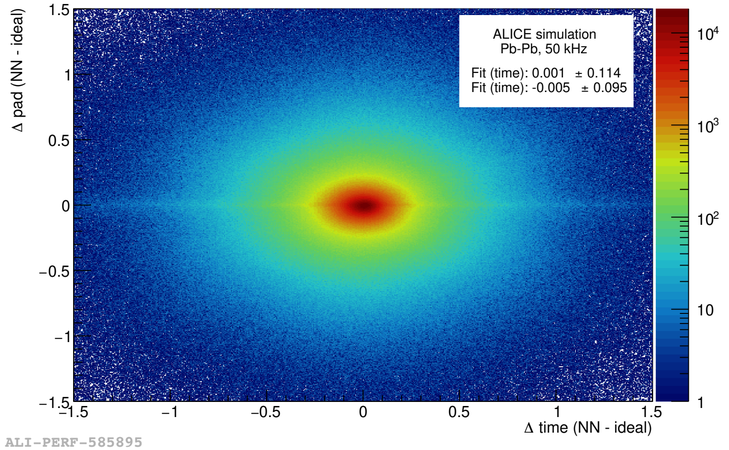
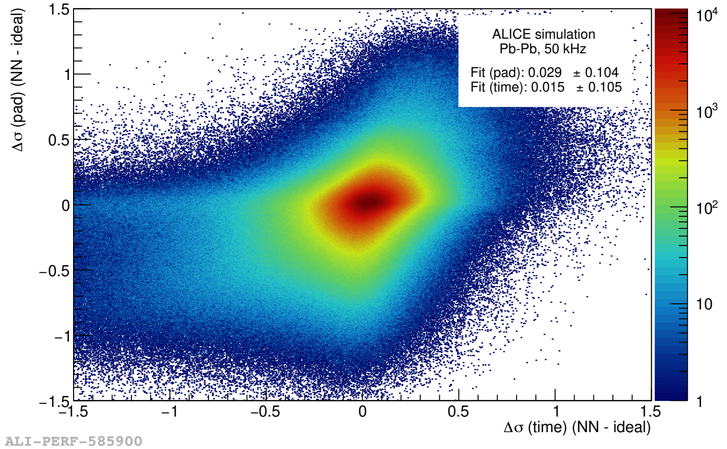
Significantly reduce fake track rate

Maintain tracking resolution



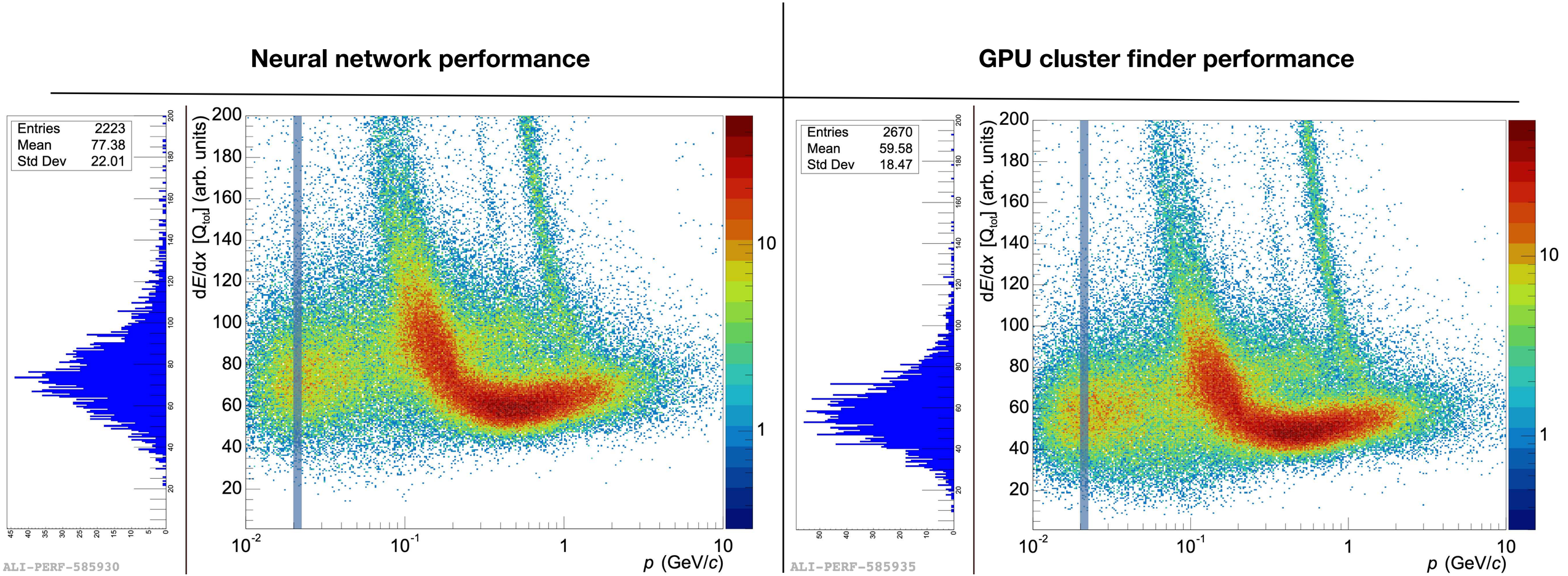
Total number of tracks: 180.4k (NN) vs. 195.5k (GPU CF)

Regression network performance

Pad vs. time	Center-of-gravity estimate	Cluster sigma estimate
GPU cluster finder		
NN cluster finder		

Comparable performance for centre-of-gravity and cluster sigma estimate

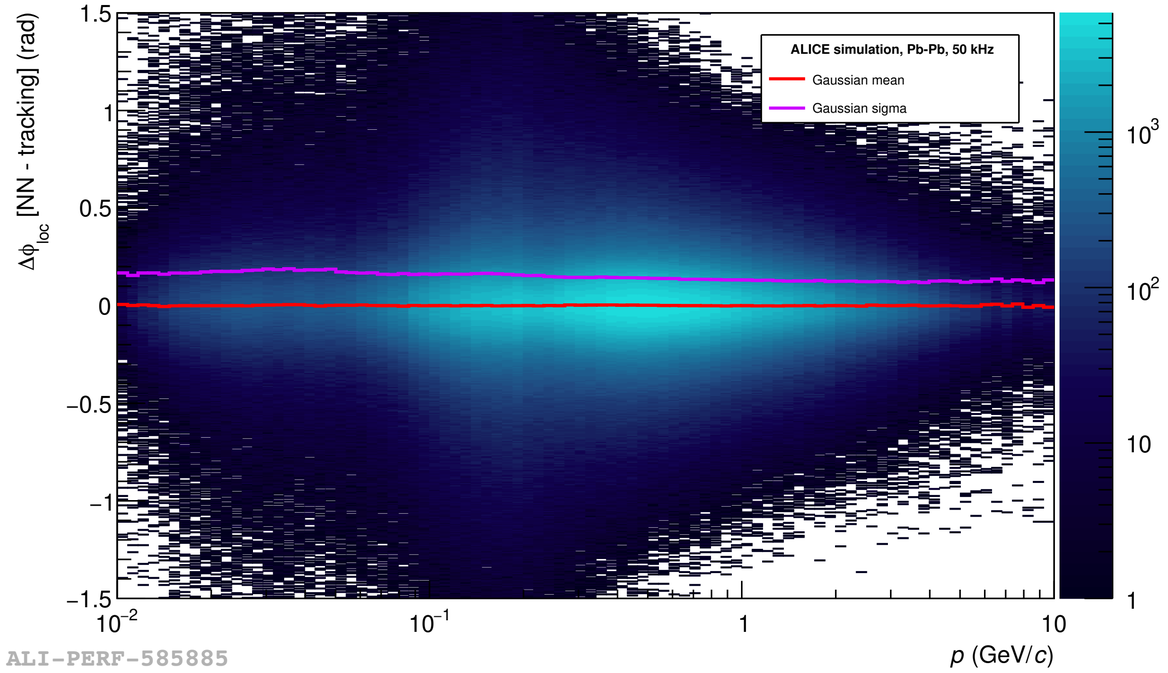
Neural network performance



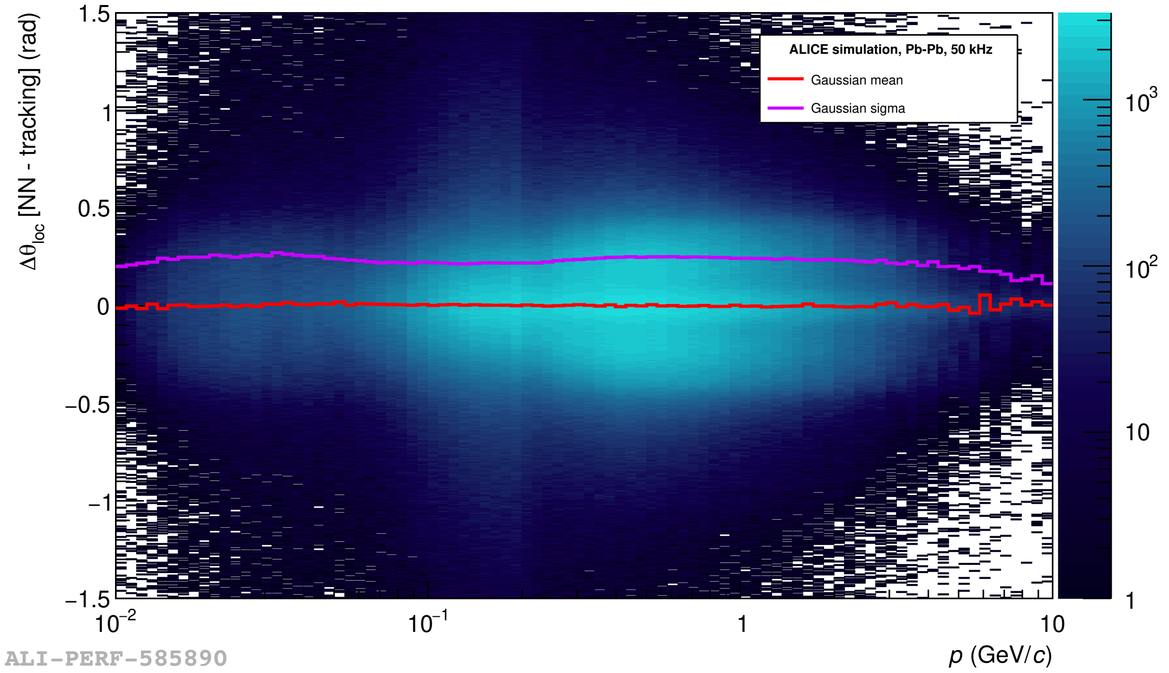
NN produces reliable dE/dx signal even in areas with very high cluster rejection

Neural network performance

NN: ϕ = arctan(p_Y/p_X) performance



NN: θ = arctan(p_Z/p_X) performance



Reasonably good estimate of local momentum vector estimate, useful for track seeding

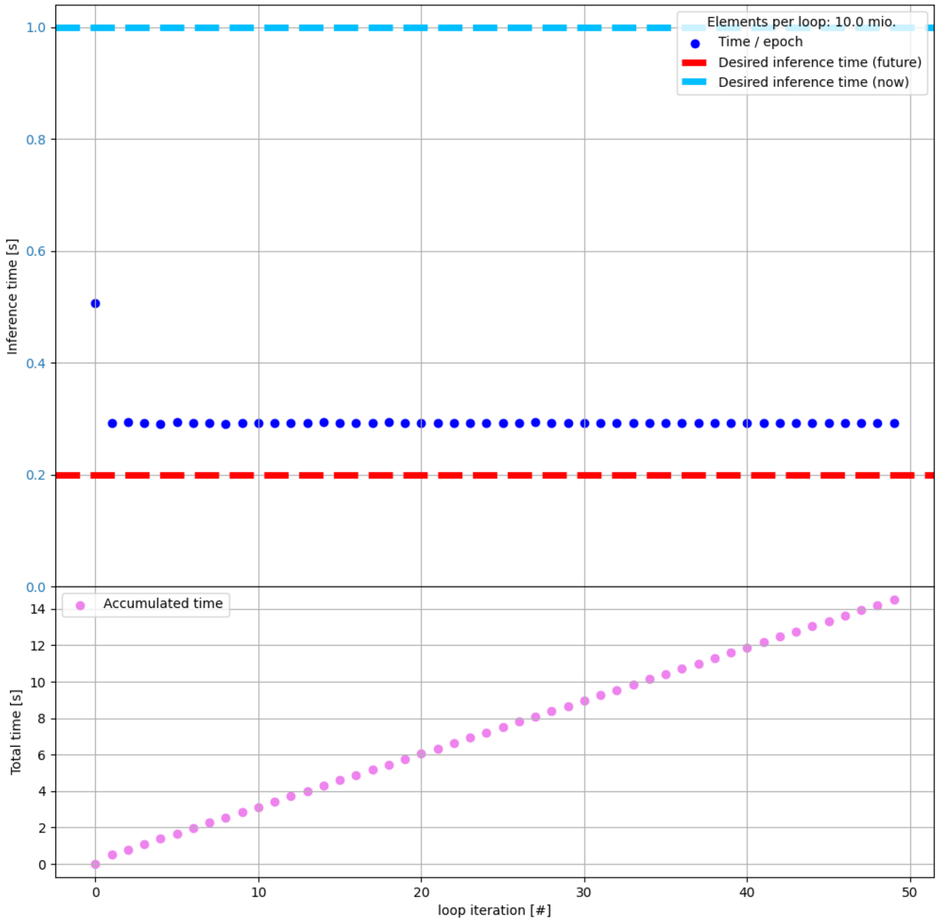
Processing speed & Design choices

Goal: Inference needs to be fast enough for online processing

- Trade-off: precision <-> speed -> Use Float16 implementation
- Measured in clusterization code: ~30 mio. clusters / s
  - Current GPU clusterizer: ~50 mio. clusters / s
  - Reduced number of clusters also leads to reduced combinatorics for tracking

Design choices

- NN design choices: Fully-connected or 2D convolutional layers are well optimised
- Inference framework: ONNX runtime with build options for MI50 & MI100 GPU's



Conclusion

Classification network

- Successfully rejects clusters that are not used in tracking
  - This could reduce effective data-size by ~20%!
- To-do: Predict cluster splitting -> Limited in training data

Regression network

- For single clusters: Comparable performance to current clusterizer
- Novel: Predict momentum of cluster (apparently with great success!)
- To-do: Can this be done well also for clusters that need to be split?

Thank you for listening!

Github: https://github.com/ChSonnabend/PhD, Email: christian.sonnabend@cern.ch

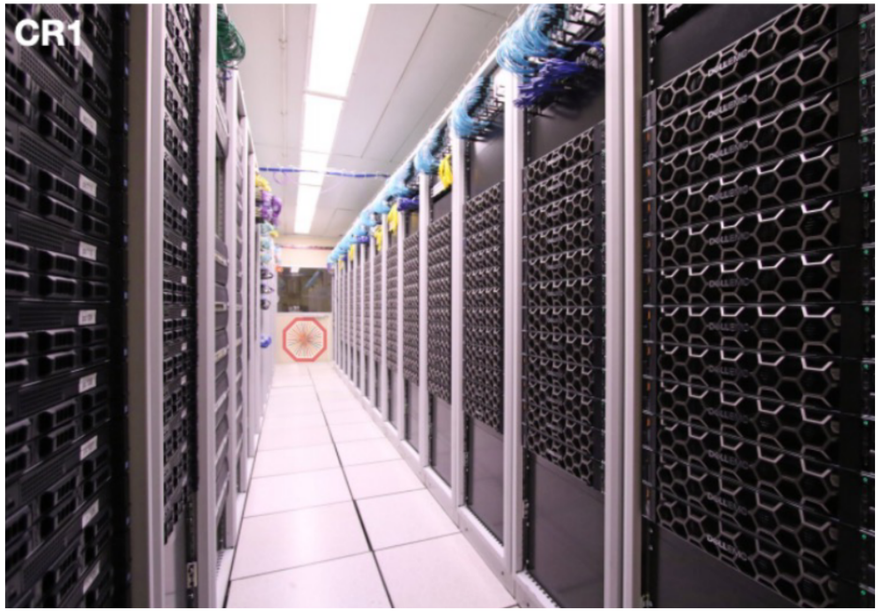
This work has been sponsored by the Wolfgang Gentner Programme of the German Federal Ministry of Education and Research (grant no. 13E18CHA)

BACKUP

Introduction

Hardware resources & constraints

- 350 EPNs (event processing nodes) for online reconstruction
- Each server: 8 MI50/MI100 GPUs, O(100) cores, O(1 TB) RAM
- Incoming data-rate: ~3.5 TB/s at peak load, ~50 mio. clusters/GPU/s



First level processors (FLP)

Event processing nodes (EPN)