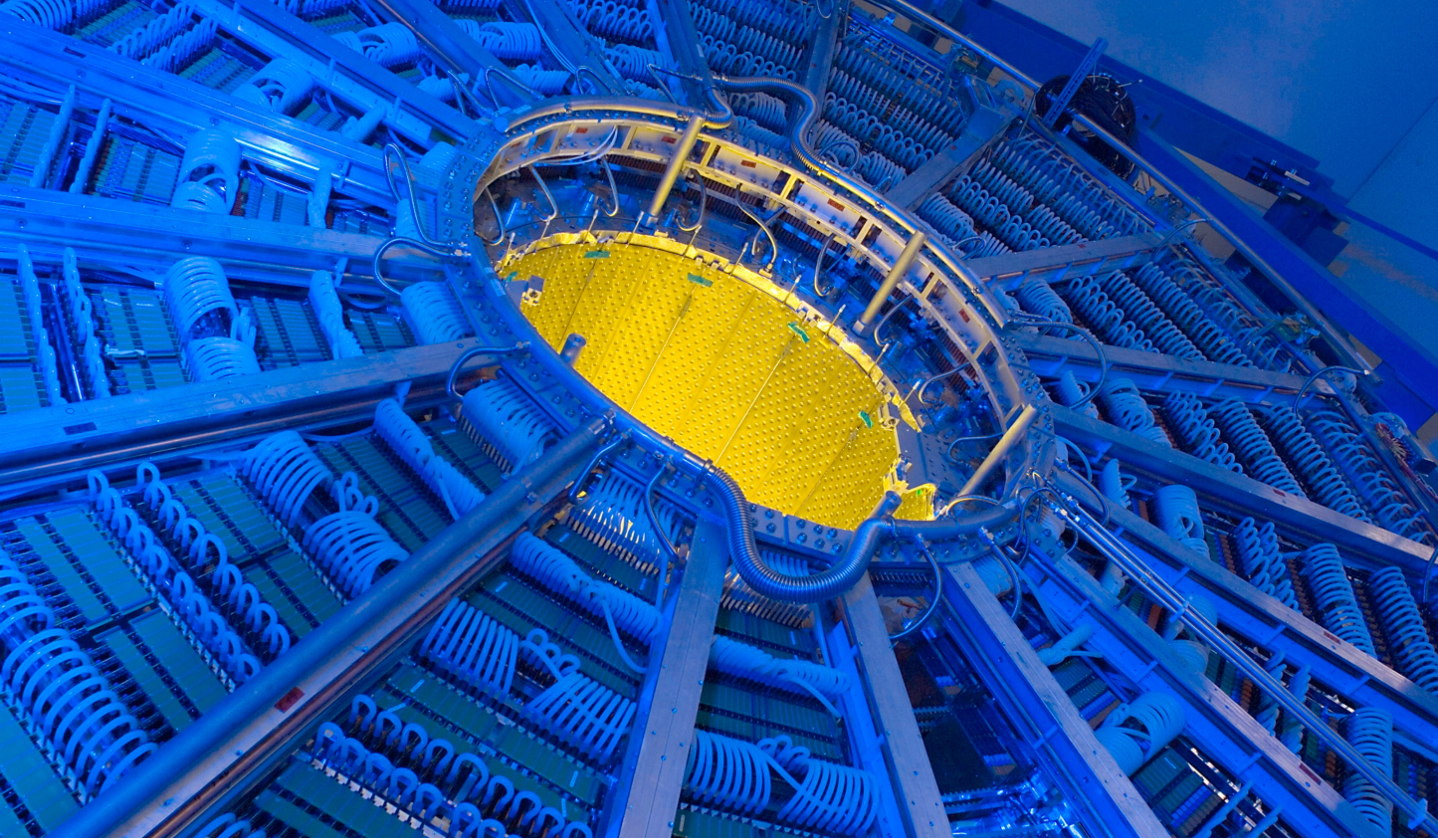


ALICE: ML and DA Challenges

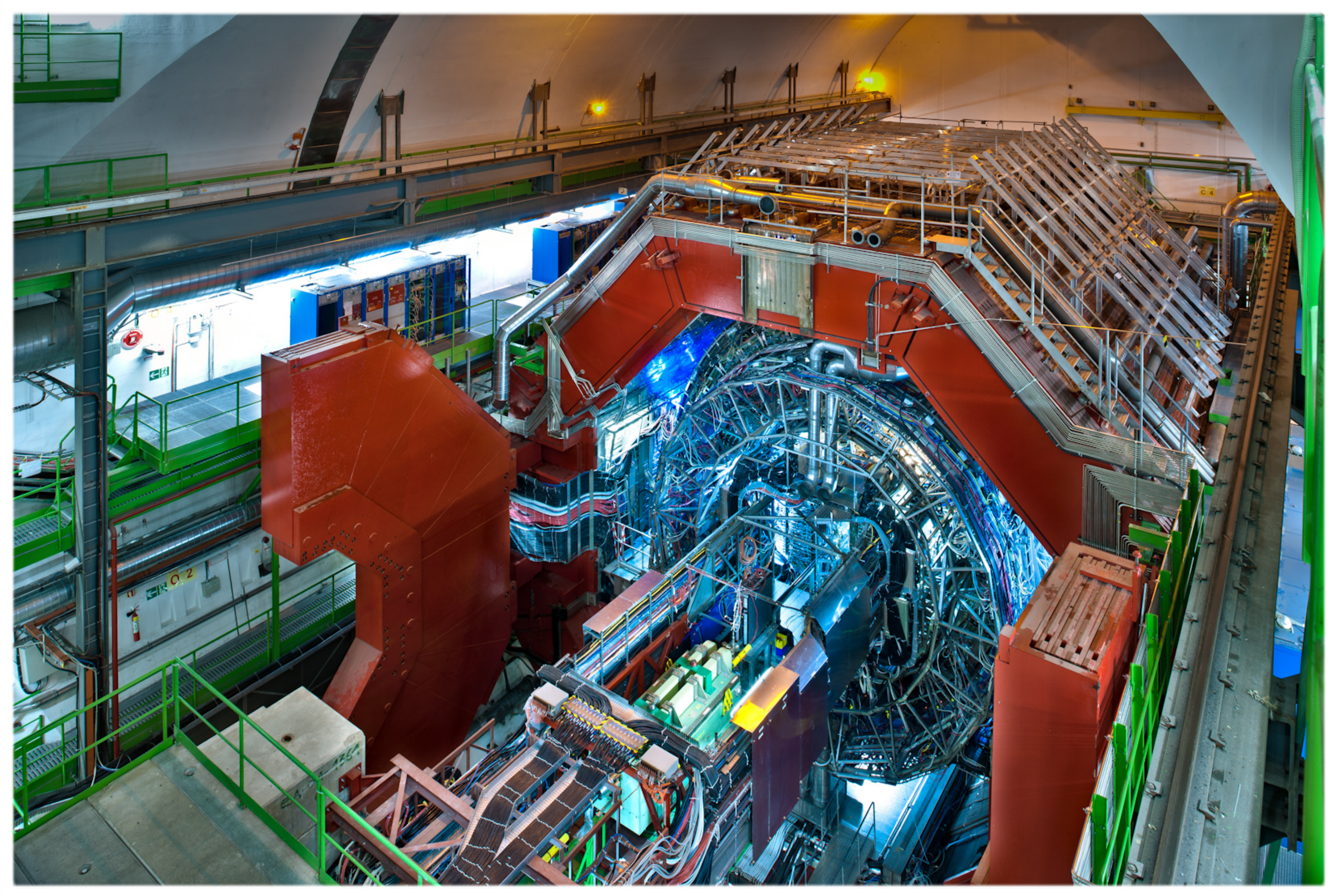
Michele Floris (CERN)
for the ALICE collaboration
CERN Openlab Workshop

Introduction

- Machine learning is in its infancy in ALICE
- Run I analysis mostly based on traditional methods:
  I will also show non-ML approaches
- Some attempts ongoing to apply ML and advanced "data science"
- In general, increasing interests in these tools

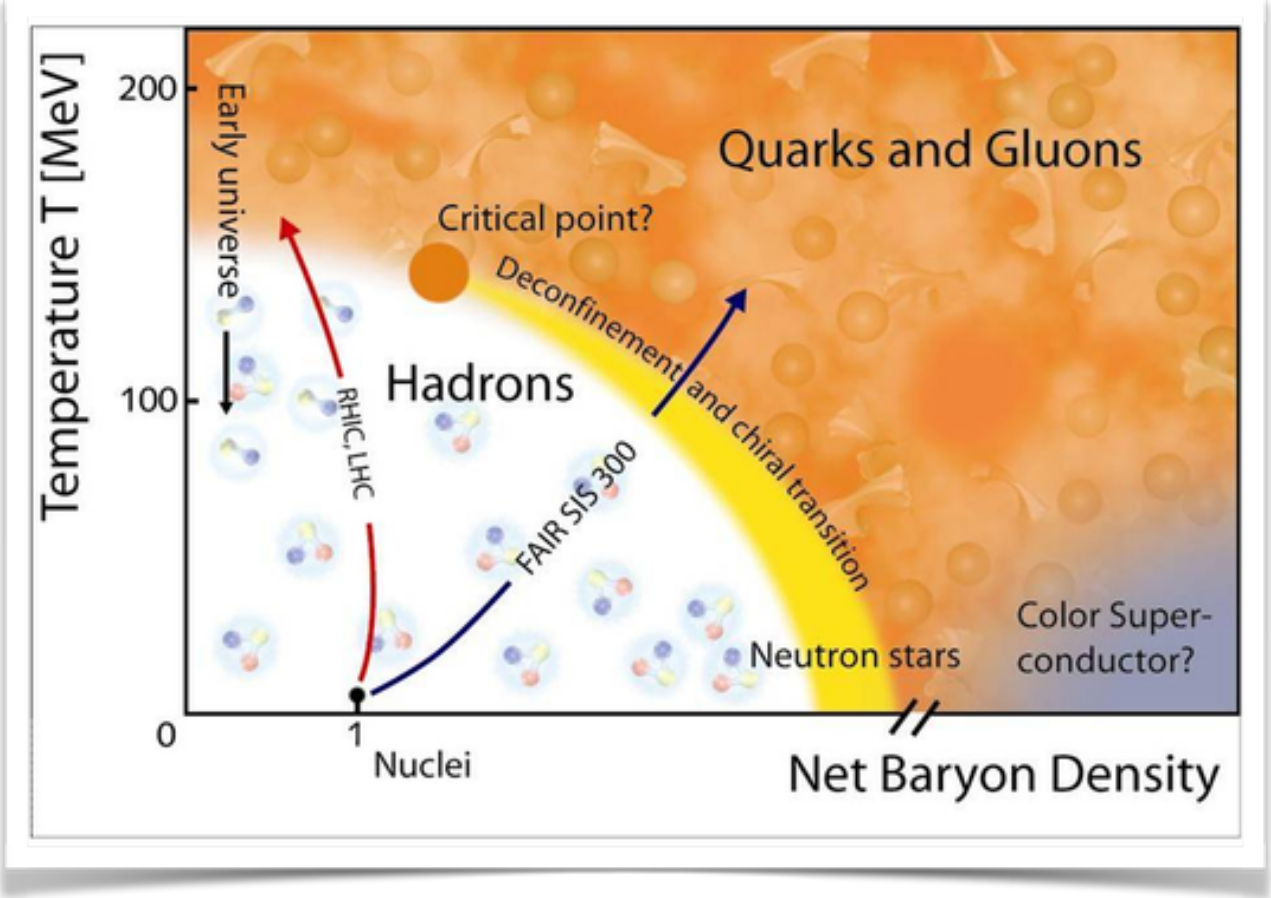
Outline

- Heavy Ion Physics and the ALICE Experiment
- Application at "detector level" (tracking and PID)
- Applications to Physics Analysis
- Applications to Computing
- Summary



Heavy-ion physics in a nutshell

- “Condensed matter” studies of QCD
  - Explore the **phase diagram** of QCD
  - Characterize the **deconfined phase** of QCD matter (quark gluon plasma)
- Understand **hadronization** and hadro-chemistry
  - How hadrons are produced from QGP
  - Hadron mass generation in QCD

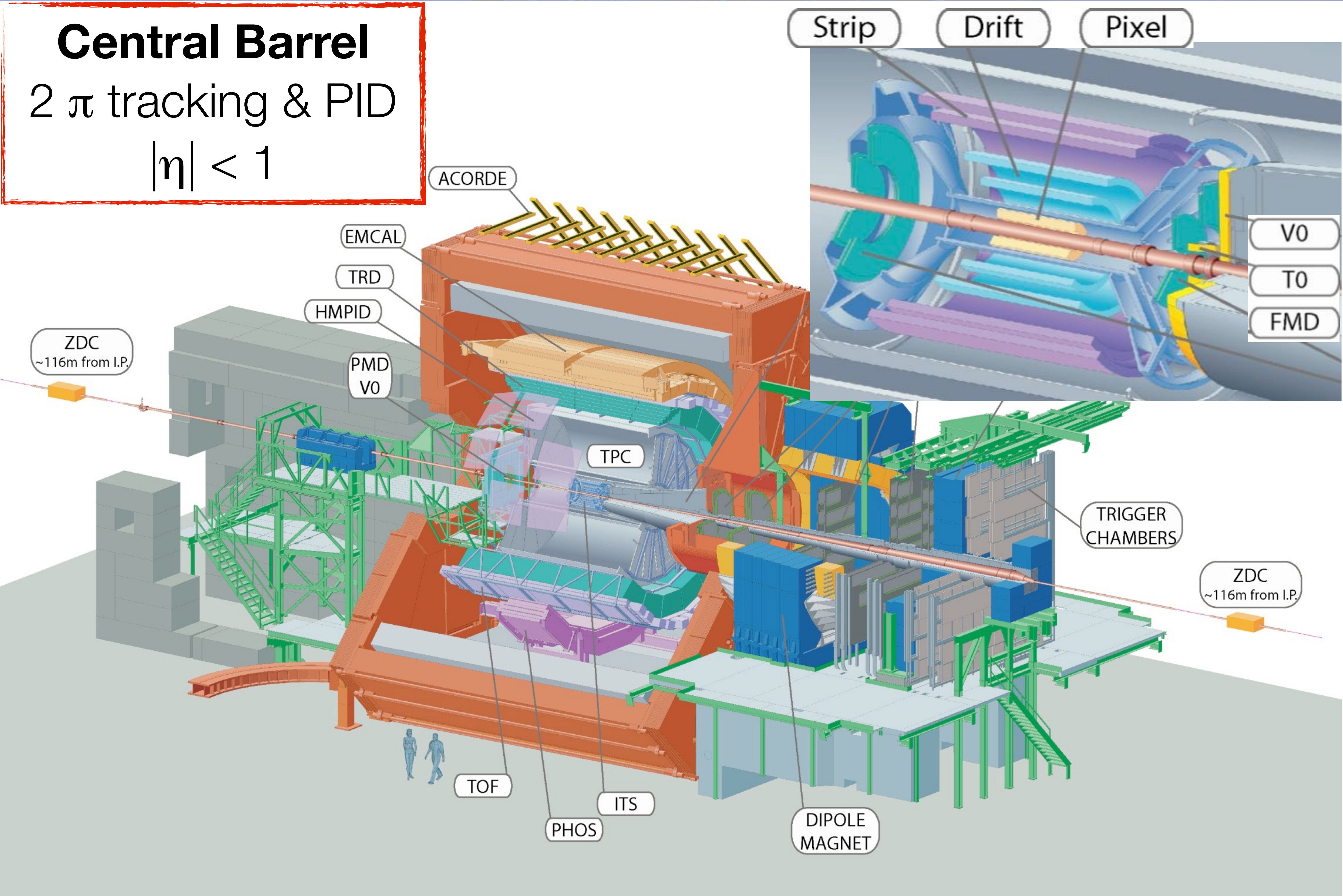


- Experimental needs: **low p_T** tracks, **particle identification** and **flavor tagging**
  - Extensive particle identification over broad momentum range
  - Low p_T tracking (“bulk” particle production and low p_T heavy flavor)
- **Colliding systems**
  - Pb-Pb: “create” the QGP
  - p-Pb, pp: control experiments, system size studies
    - and many surprises at the LHC!

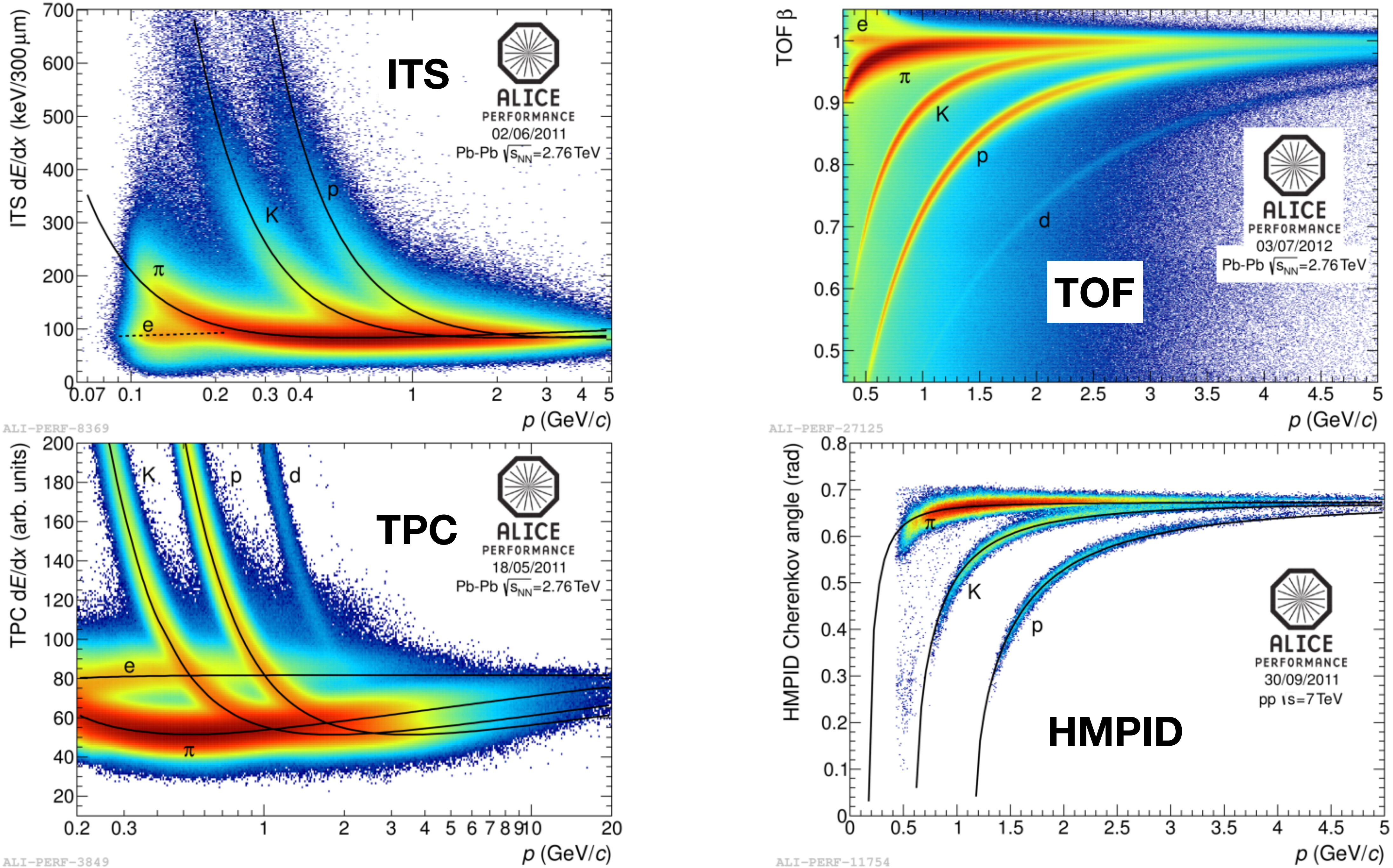
Central Barrel

$2\,\pi$ tracking & PID

$|\eta| < 1$



Particle Identification and Tracking



Particle identification (PID, many different techniques)

Extremely low-mass tracker ~ 10% of X_0

Excellent vertexing capability

Efficient low-momentum tracking – down to ~ 100 MeV/c

Challenges

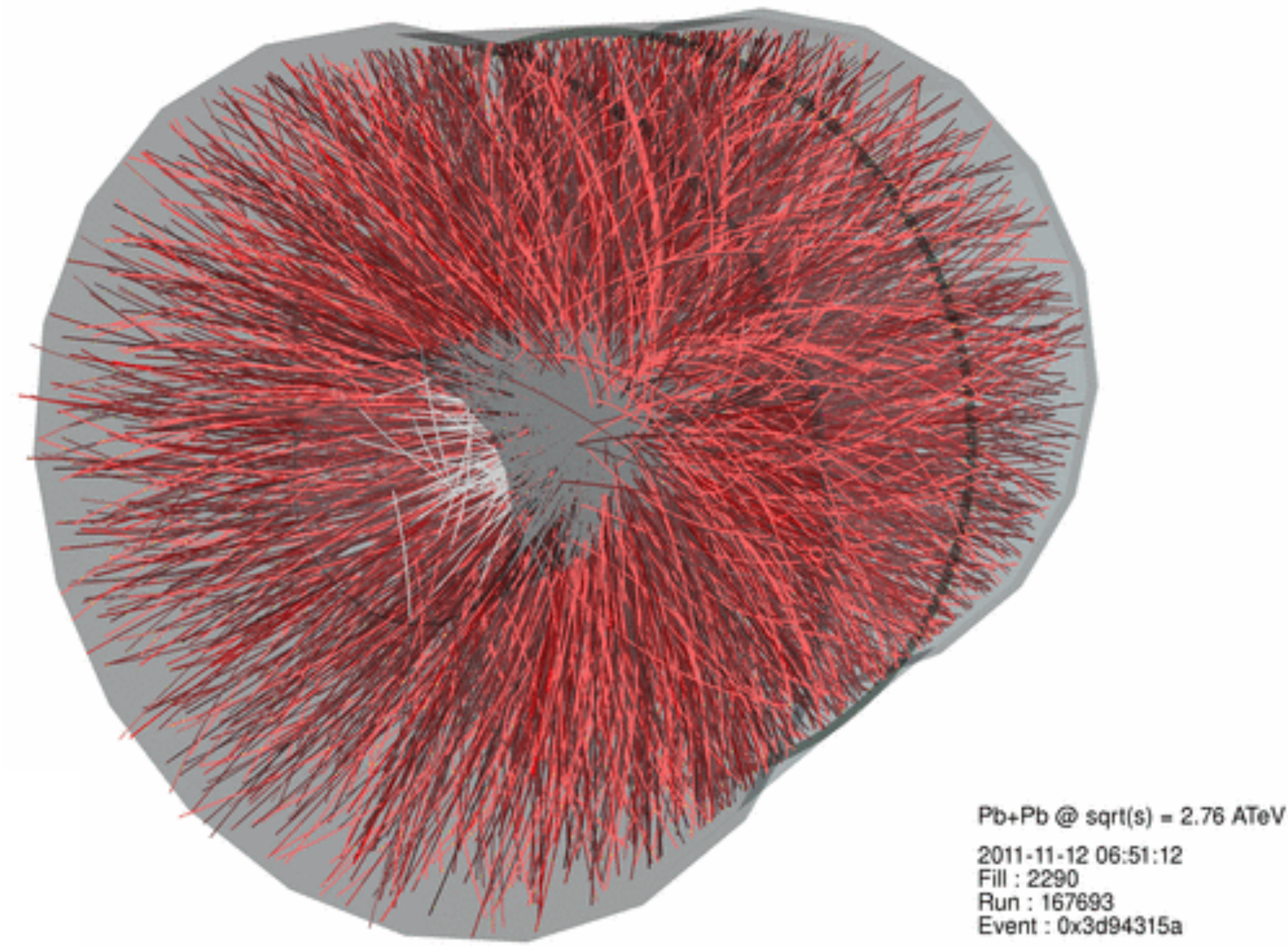
Very large charged **tracks multiplicity**:
several thousand tracks in TPC in a head-on Pb–Pb collision at the LHC

Data volume: ~10 PB of data so far, (~3 PB Pb-Pb 2015) almost twice that in MC

Complex detector **calibration**

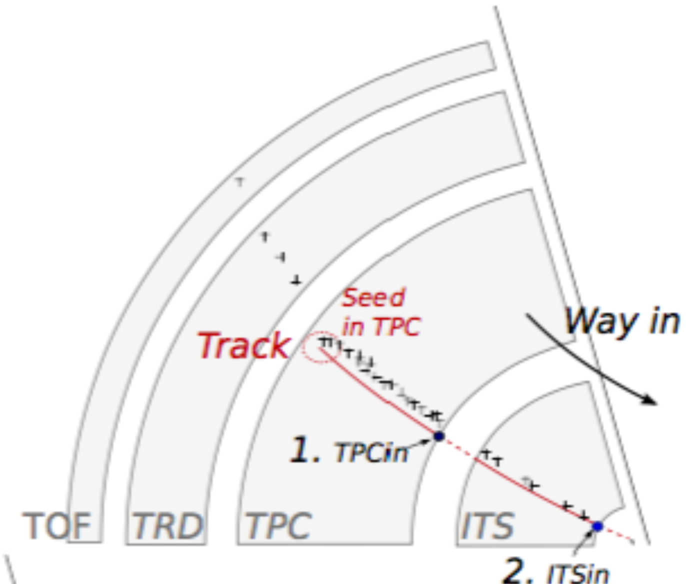
Combine **PID** in broad momentum region (0.1–20 GeV/*c*)

Key channels: very **low signal-to-background**

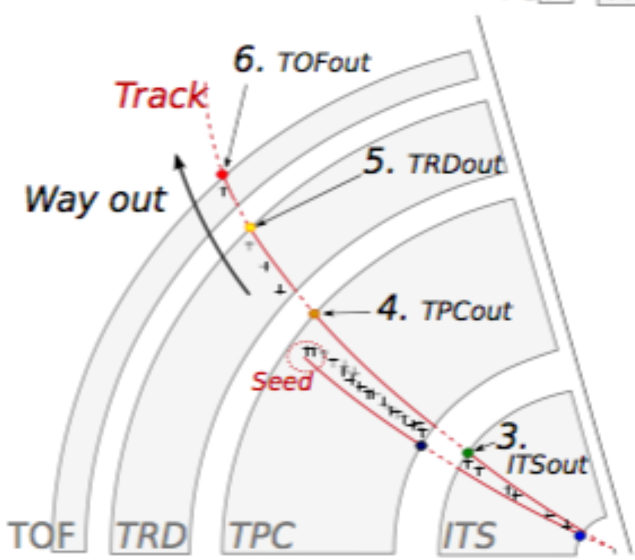


Detector: Track Reconstruction

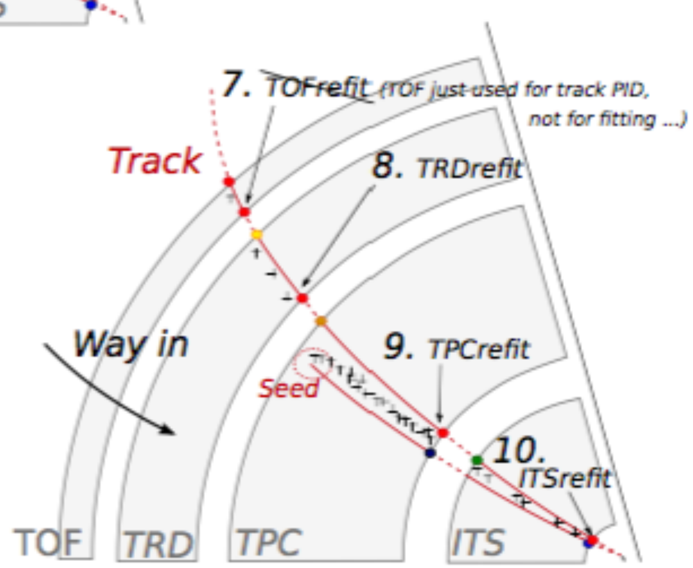
1st pass



2nd pass



3rd pass



Inward-outward-inward procedure to reduce combinatorics

Standard Kalman Filter

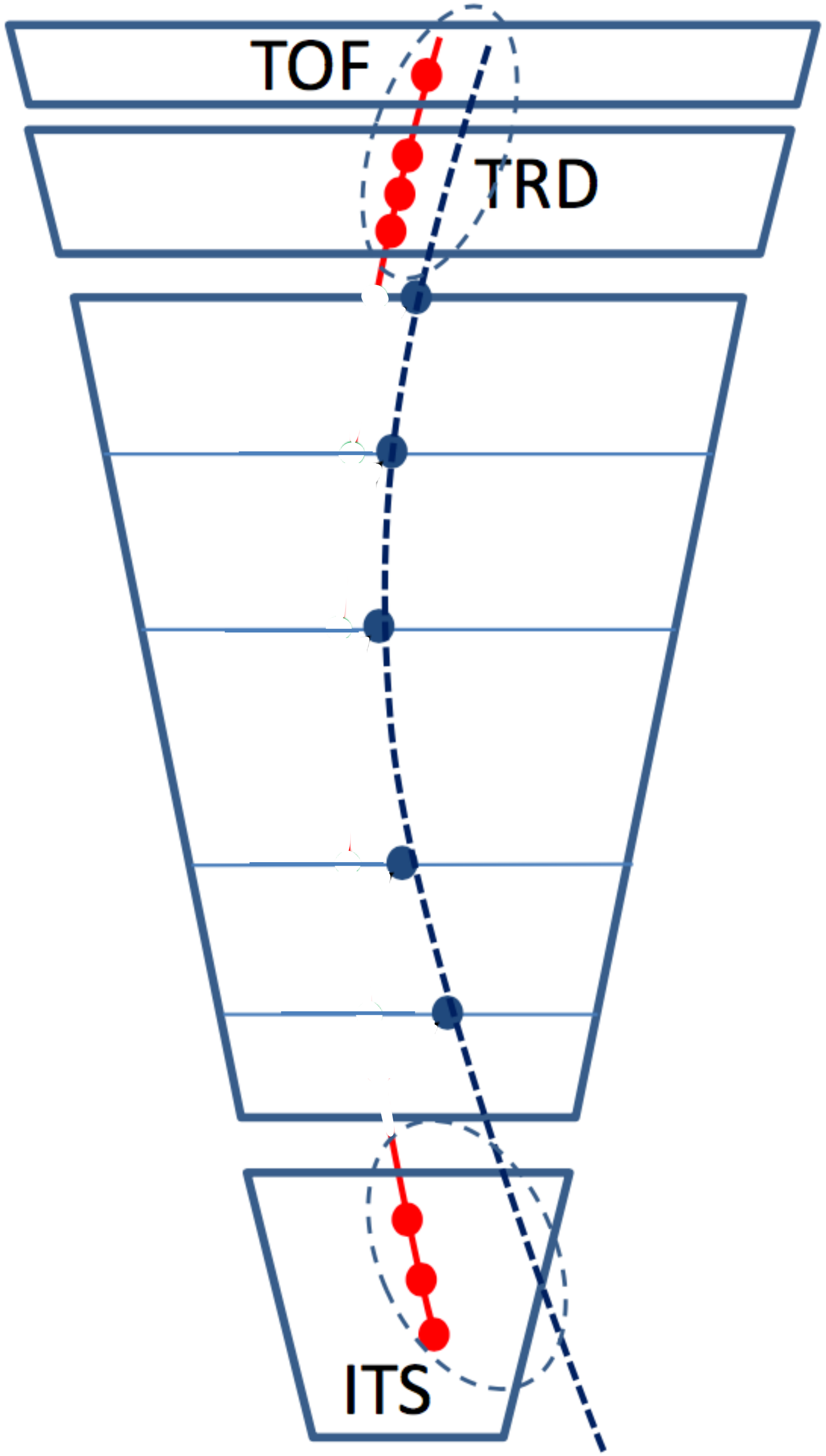
Bulk of data produced by TPC (80% of volume)

Calibration is also a major challenge

Drawing from: CERN-THESIS-2011-263

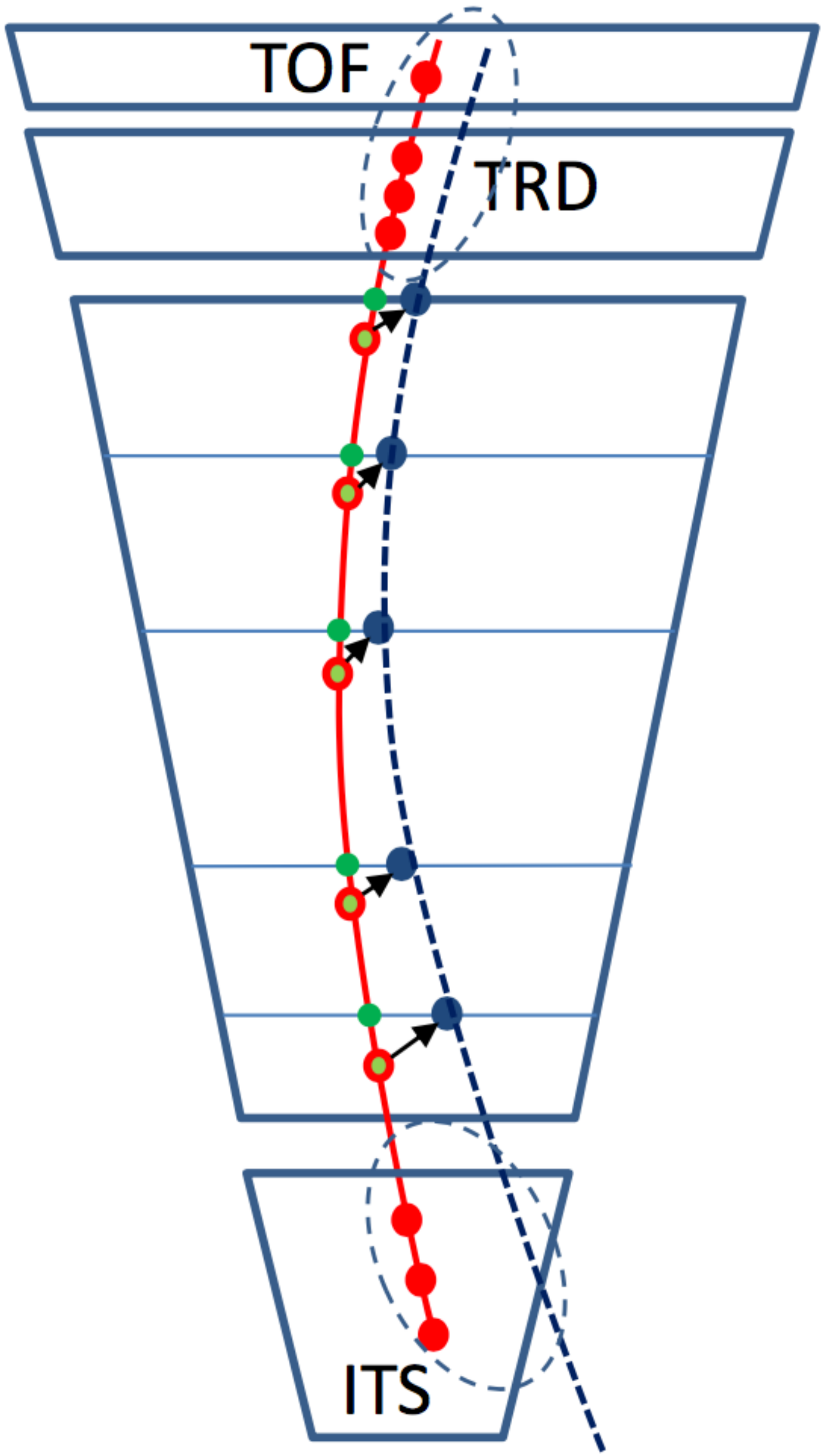
Int.J.Mod.Phys. A29 (2014) 1430044, arXiv:1402.4476

Space charge distortions



- **Charge** accumulated in the TPC may **distort electric field**
- **Clusters** (and reconstructed track) are **distorted**
- **Calibrate** cluster positions using inner and outer detectors
- **Challenges**:
  - Initial reconstruction with very large tolerances → **outliers**
  - Need **smooth parametrization** of corrections (currently: kernel smoother + Chebyschev polynomials)
  - **Time dependence** (need ~20-40 mins bins)
  - **Fluctuations**
  - Number of voxels (~850 K) + fits for pre-processing → **computational time**

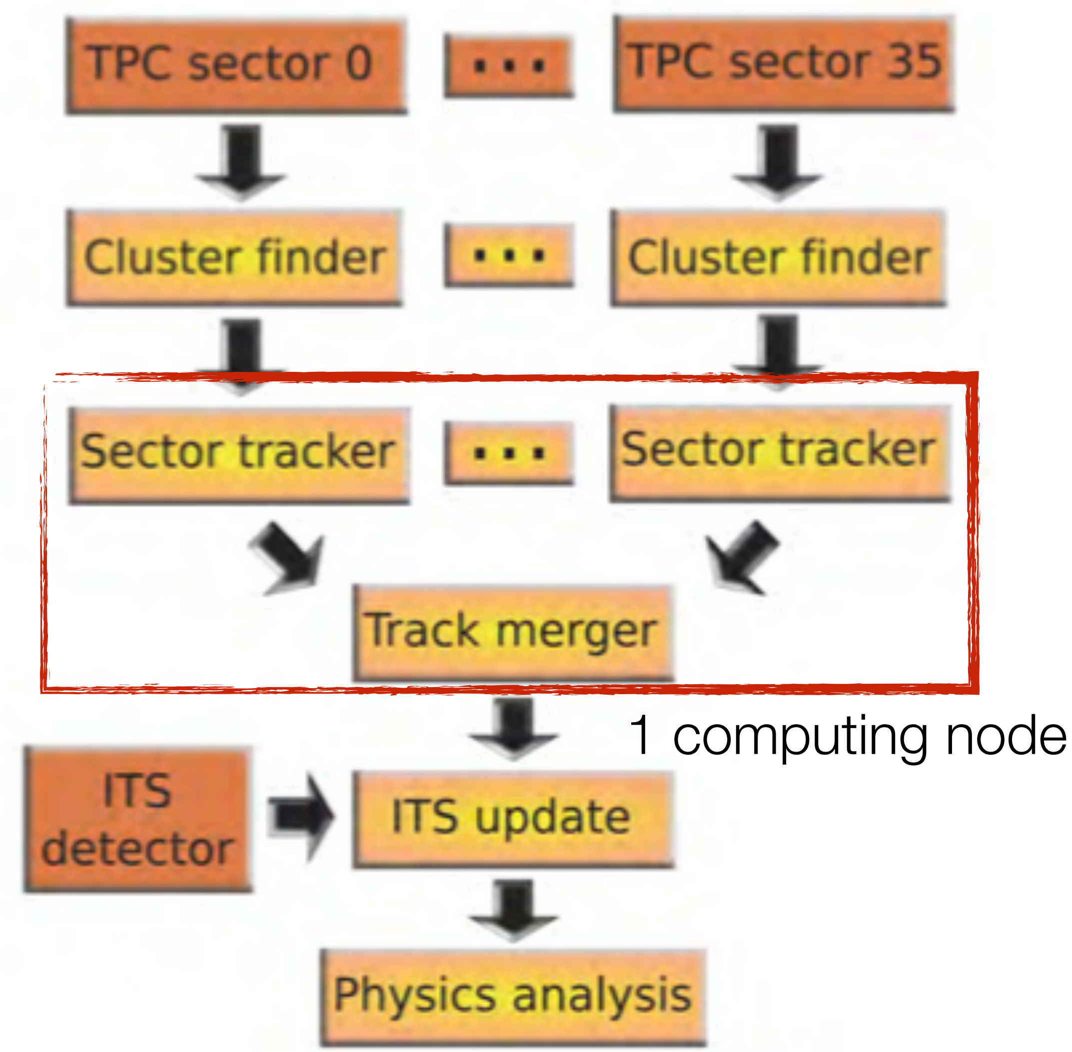
Space charge distortions



- **Charge** accumulated in the TPC may **distort electric field**
- **Clusters** (and reconstructed track) are **distorted**
- **Calibrate** cluster positions using inner and outer detectors
- **Challenges**:
  - Initial reconstruction with very large tolerances → **outliers**
  - Need **smooth parametrization** of corrections (currently: kernel smoother + Chebyschev polynomials)
  - **Time dependence** (need ~20-40 mins bins)
  - **Fluctuations**
  - Number of voxels (~850 K) + fits for pre-processing → **computational time**

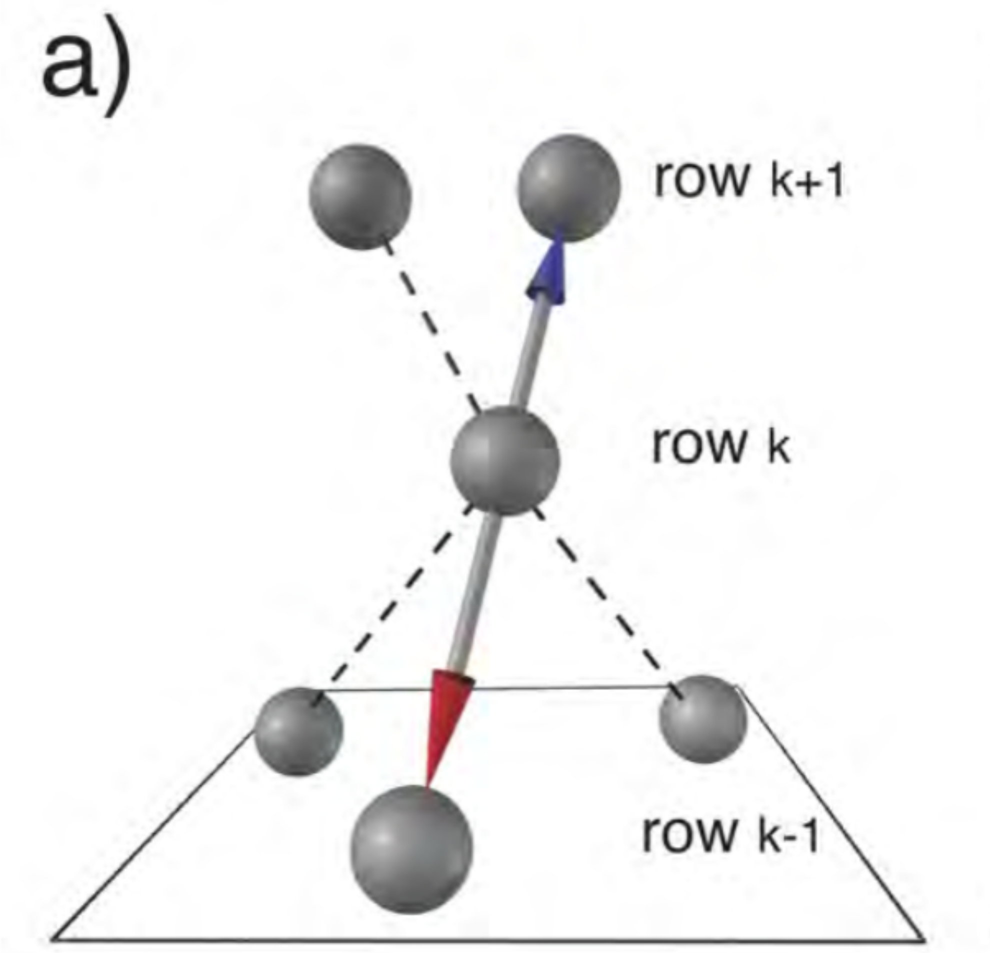
Track reconstruction in the HLT

- Need for **online cluster and track** reconstruction in the High Level Trigger
  - Data compression (factor ~4)
  - Quality Assurance
- **Parallelization** and **hardware** acceleration
  - FPGA-based cluster finder
  - Parallel tracking
    - Seeding based on “Cellular Automaton”
    - Track following based on Kalman filter
  - GPU-based algorithms
  - HLT farm: 180 nodes, 4320 CPU cores

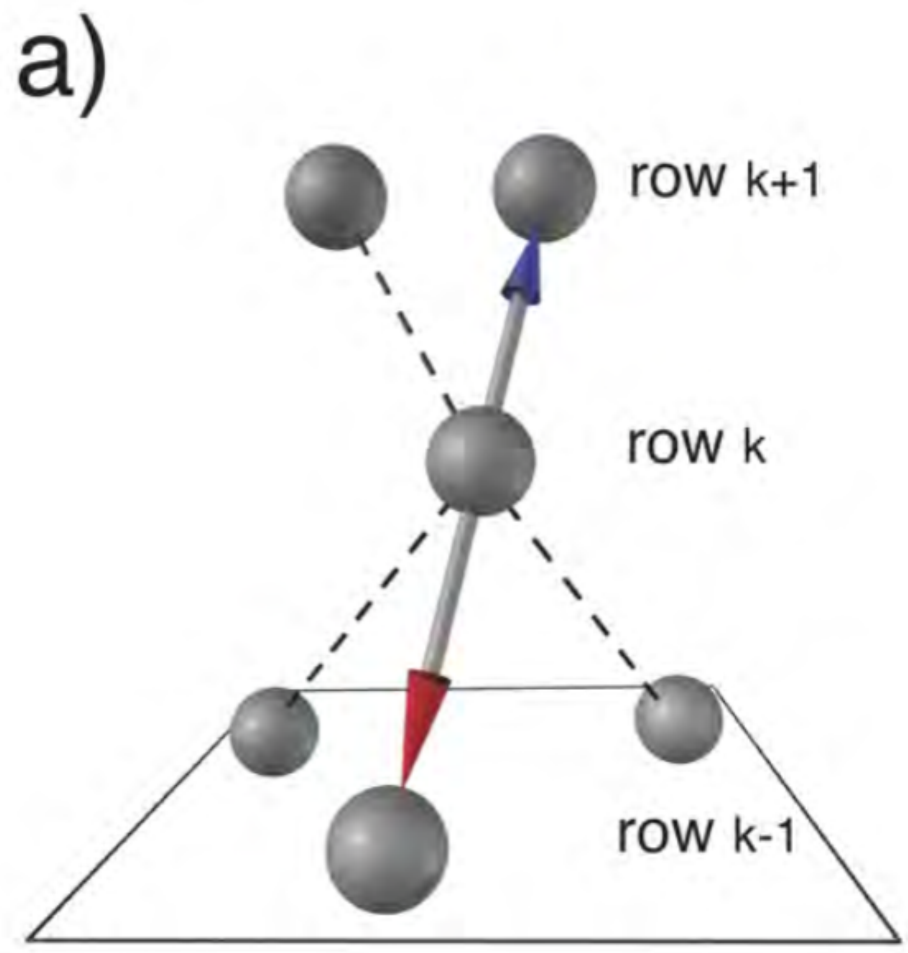


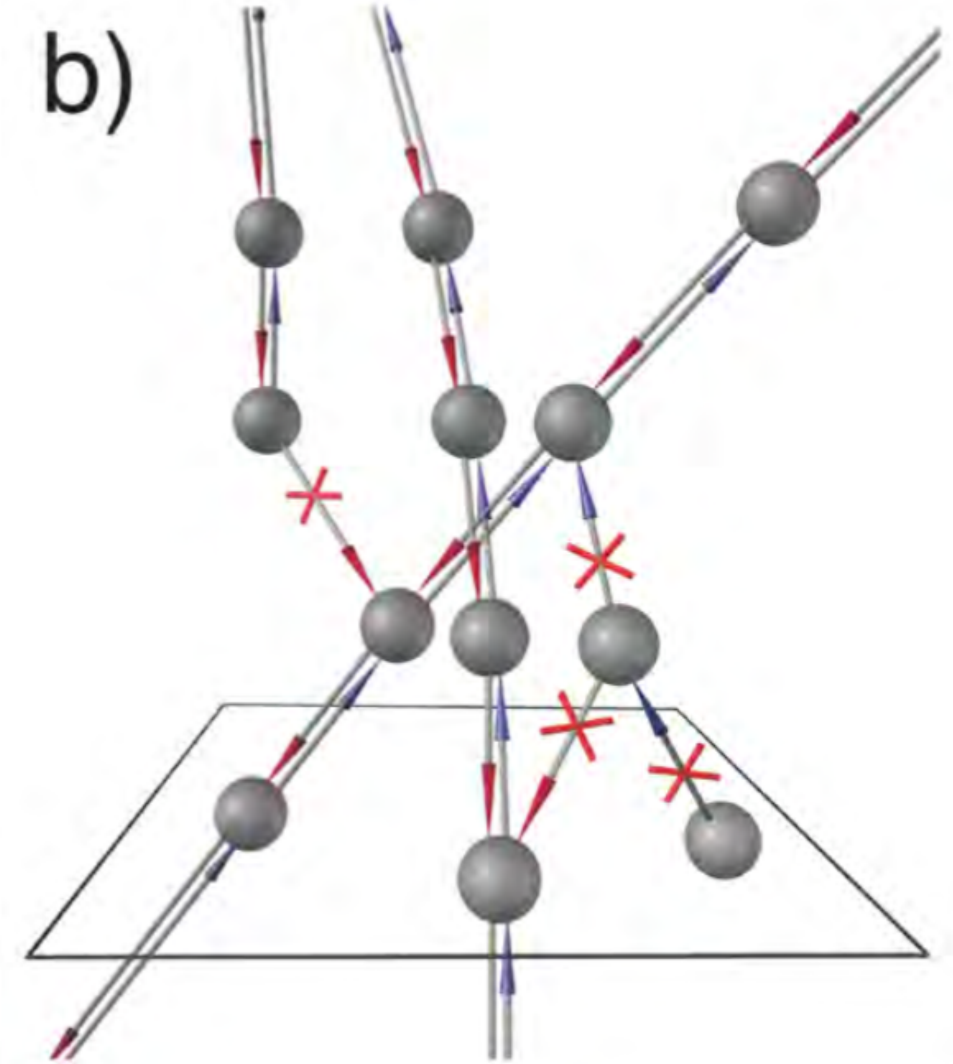
IEEE TNS, 58(4), 1845–1851, 10.1109/TNS.2011.2157702
CNNA 2012 proceedings, 10.1109/CNNA.2012.6331460

Cellular Automaton



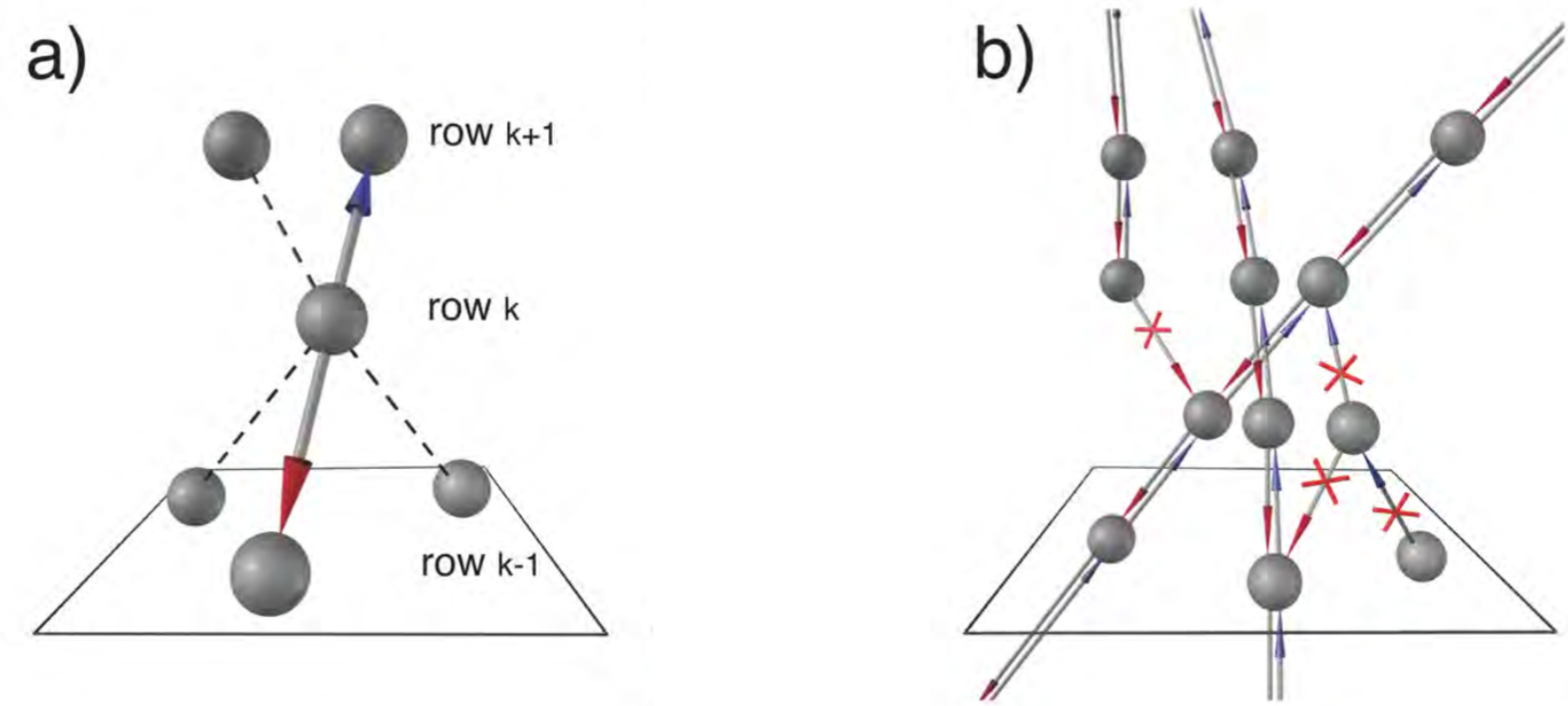
Neighbors finder:
segments of 3 clusters
forming a straight line (“link”)





Neighbors finder:
segments of 3 clusters
forming a straight line ("link")

Evolution step:
Only reciprocal links are kept



Neighbors finder:
segments of 3 clusters
forming a straight line ("link")

Evolution step:
Only reciprocal links are kept

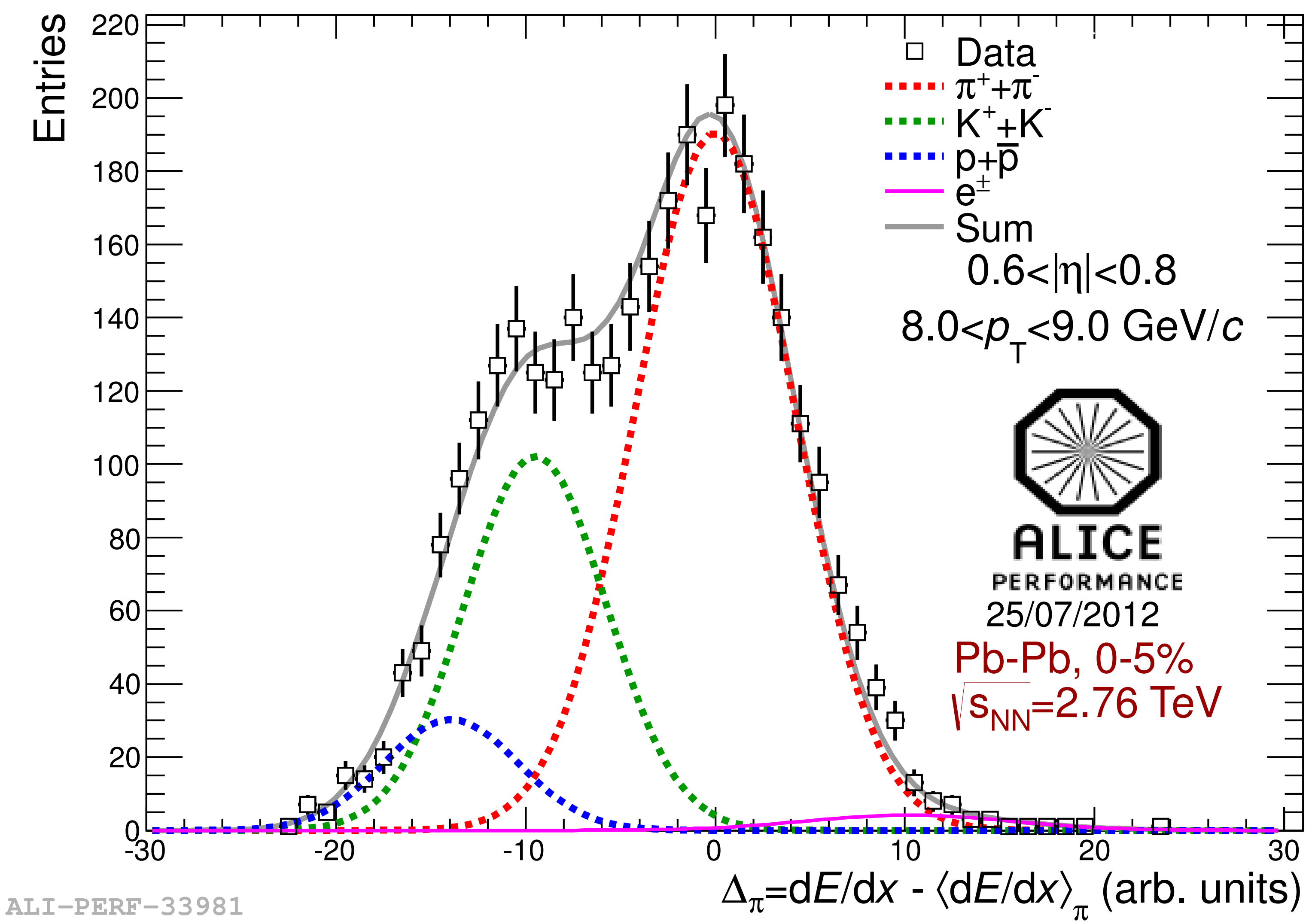
Chain of links for the track candidates → Kalman Filter

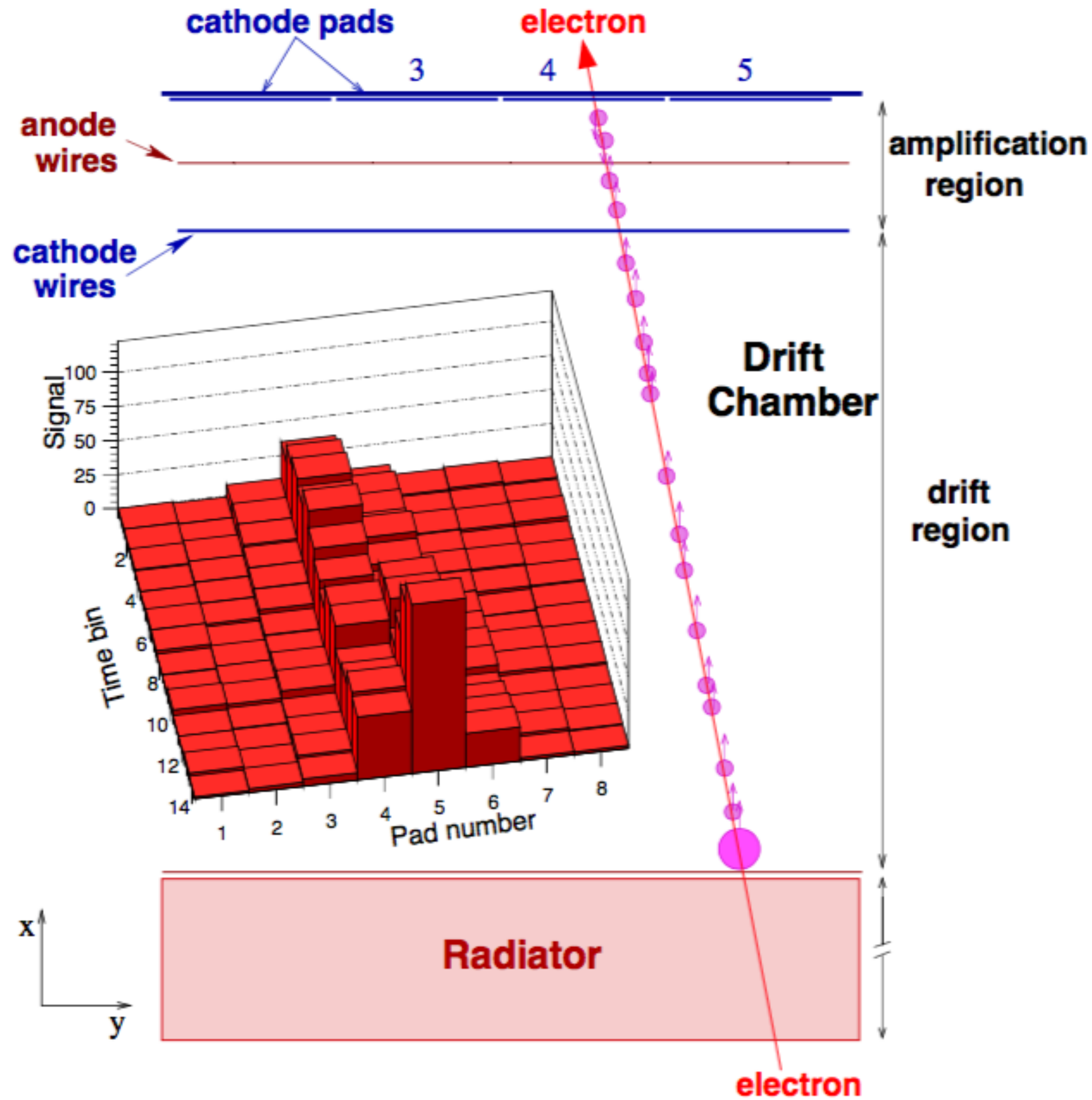
2 orders of magnitude faster than offline tracker

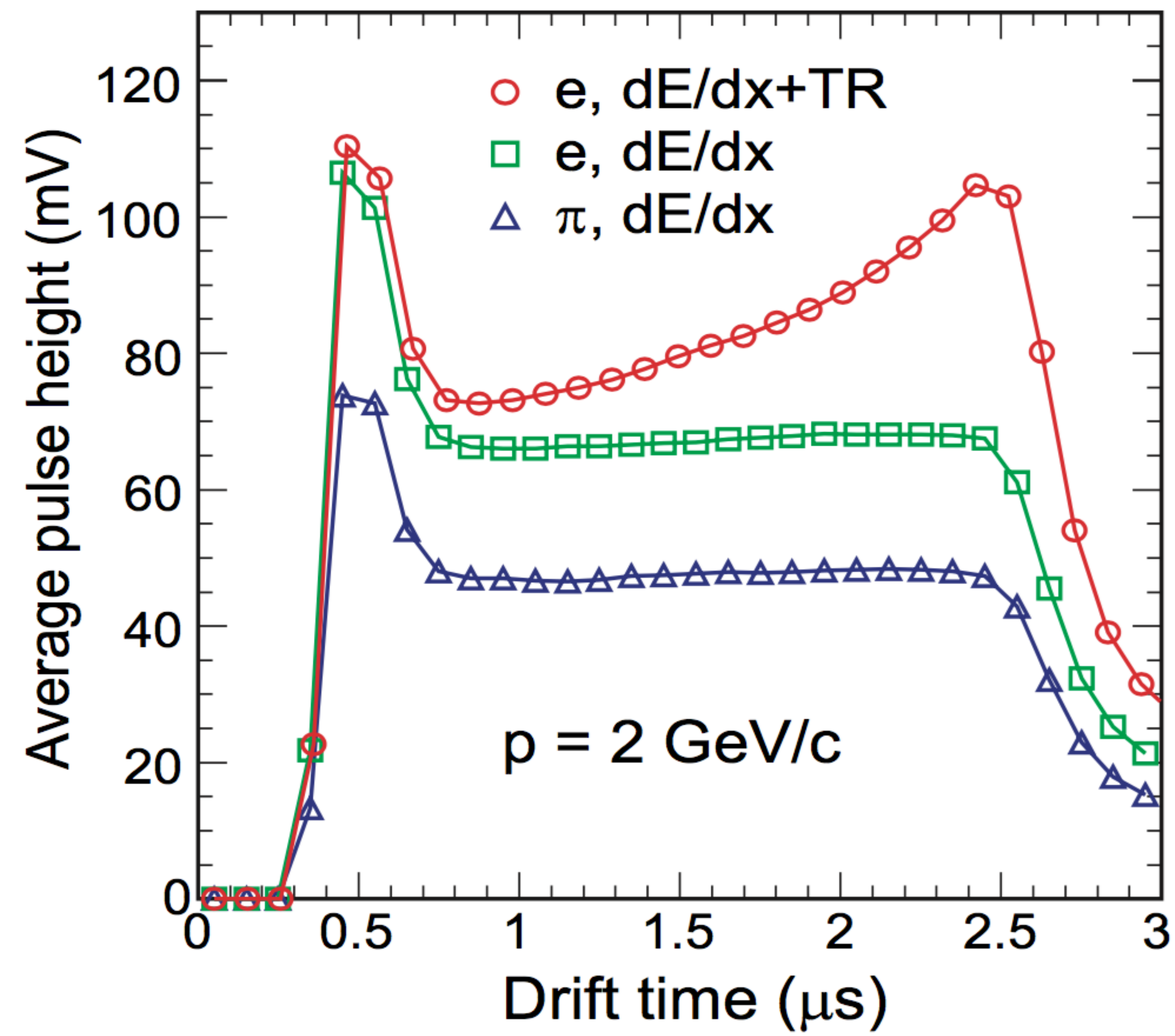
Detector: processing of (PID) signals

Particle Identification

- Can use **statistical identification**
- **Track-by-track** needed for some studies
- **Multidimensional "classification"** problems:
  - Extracting information for a single detector
  - Combining information from many detectors





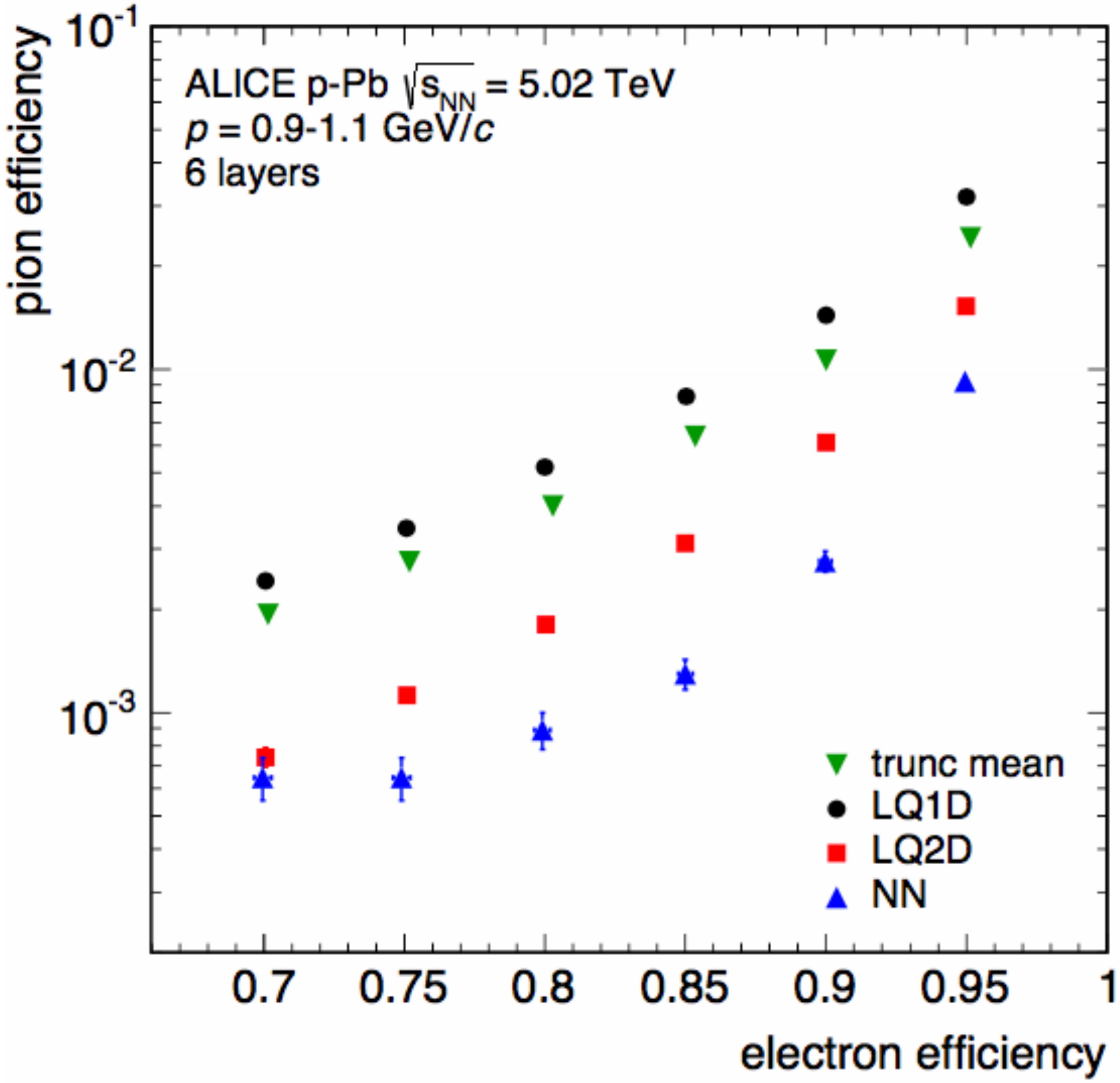


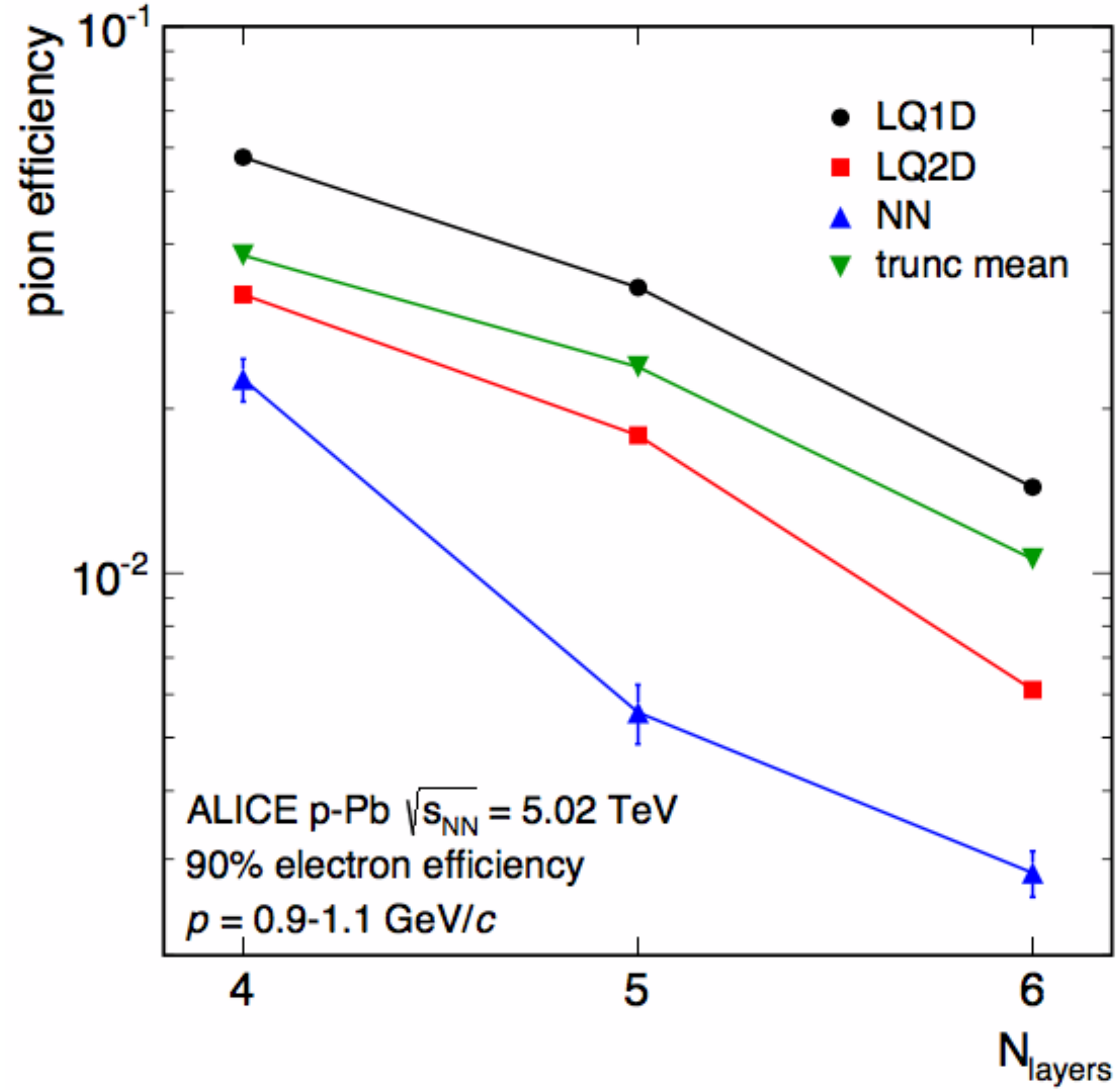
ALICE TRD: stack of 6 identical layers

Electrons: larger signal and different time dependence

2008 JINST 3 S08002

Comparison of e/π distrimination methods



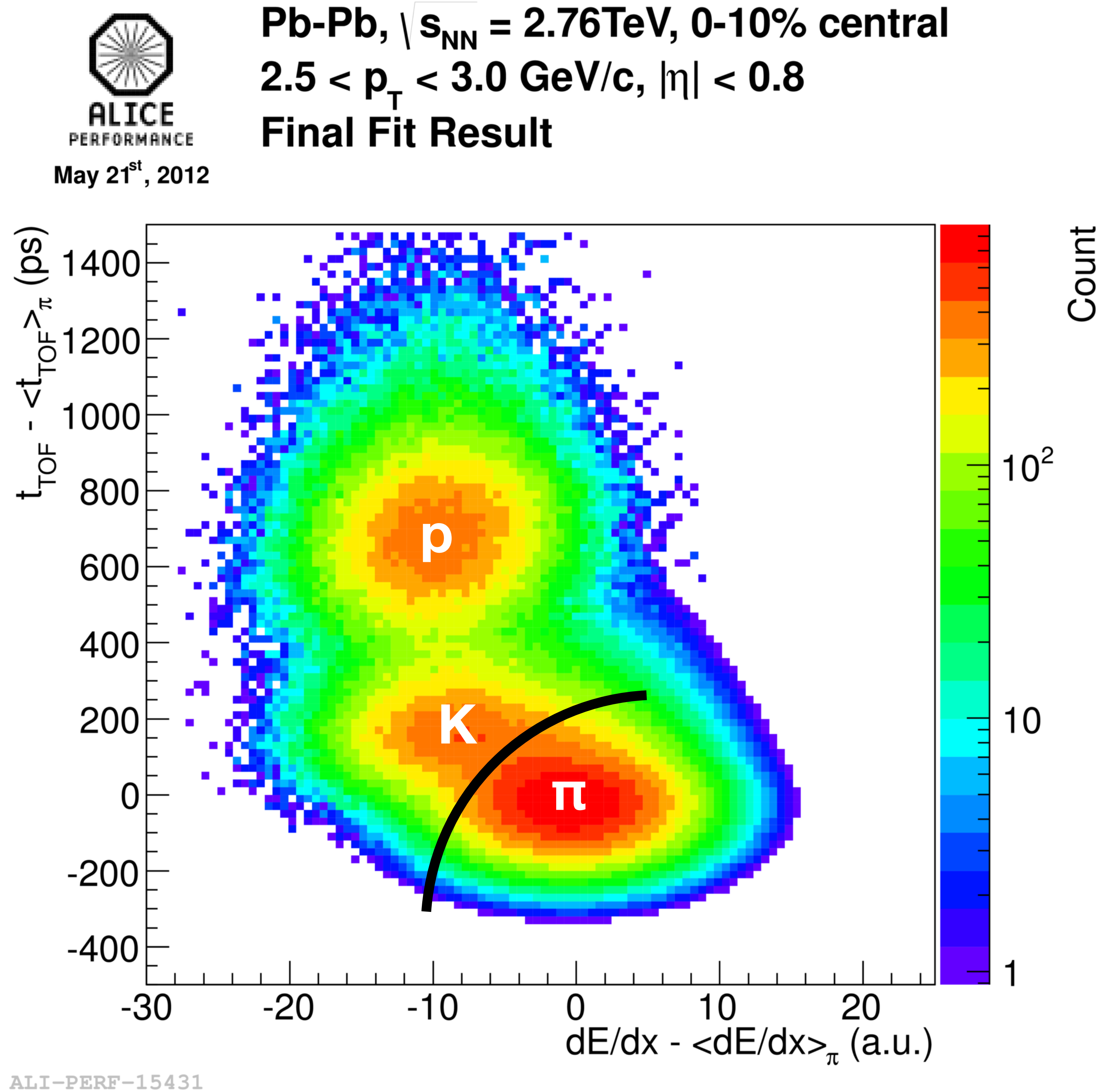


FF Neural Network (NN) works better than other methods,
but uses more information.
Next: include track properties

Int.J.Mod.Phys. A29 (2014) 1430044, arXiv:1402.4476

Combining Detectors: Bayesian PID

- **Many PID detectors** in ALICE: combination?
- **Basic approach**: rectangular cuts on PID variables (or nσ)
  - Sub-optimal:
    - Contamination depends on particle species abundances
    - Non-gaussian features in the signal distributions
- **Bayesian approach**:
  - Use knowledge of detector response and prior species abundances
  - Determine priors iteratively
- Early attempts to use **multivariate methods**

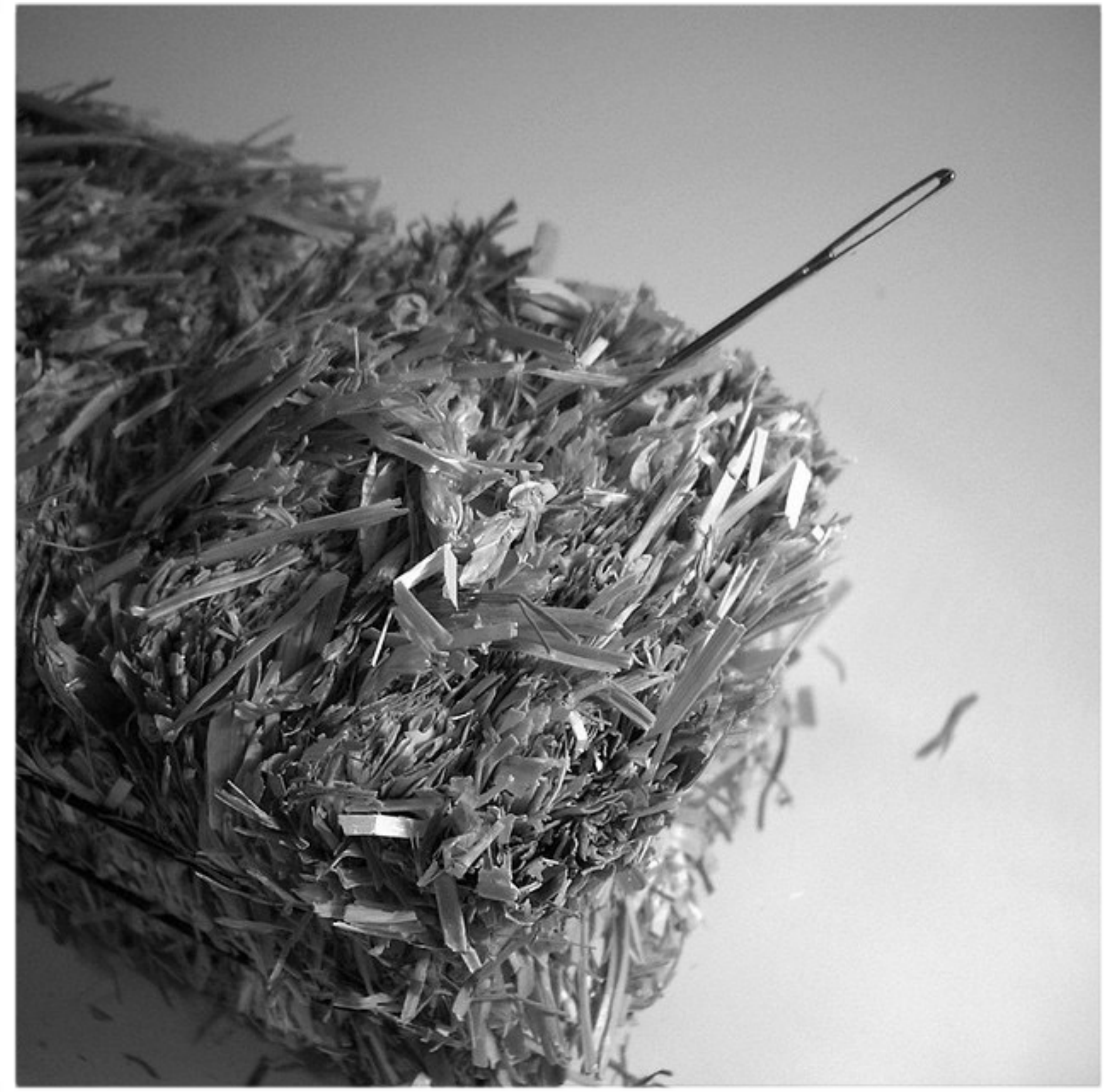


TPC+TOF

Signal extraction

Signal extraction in low S/B environment

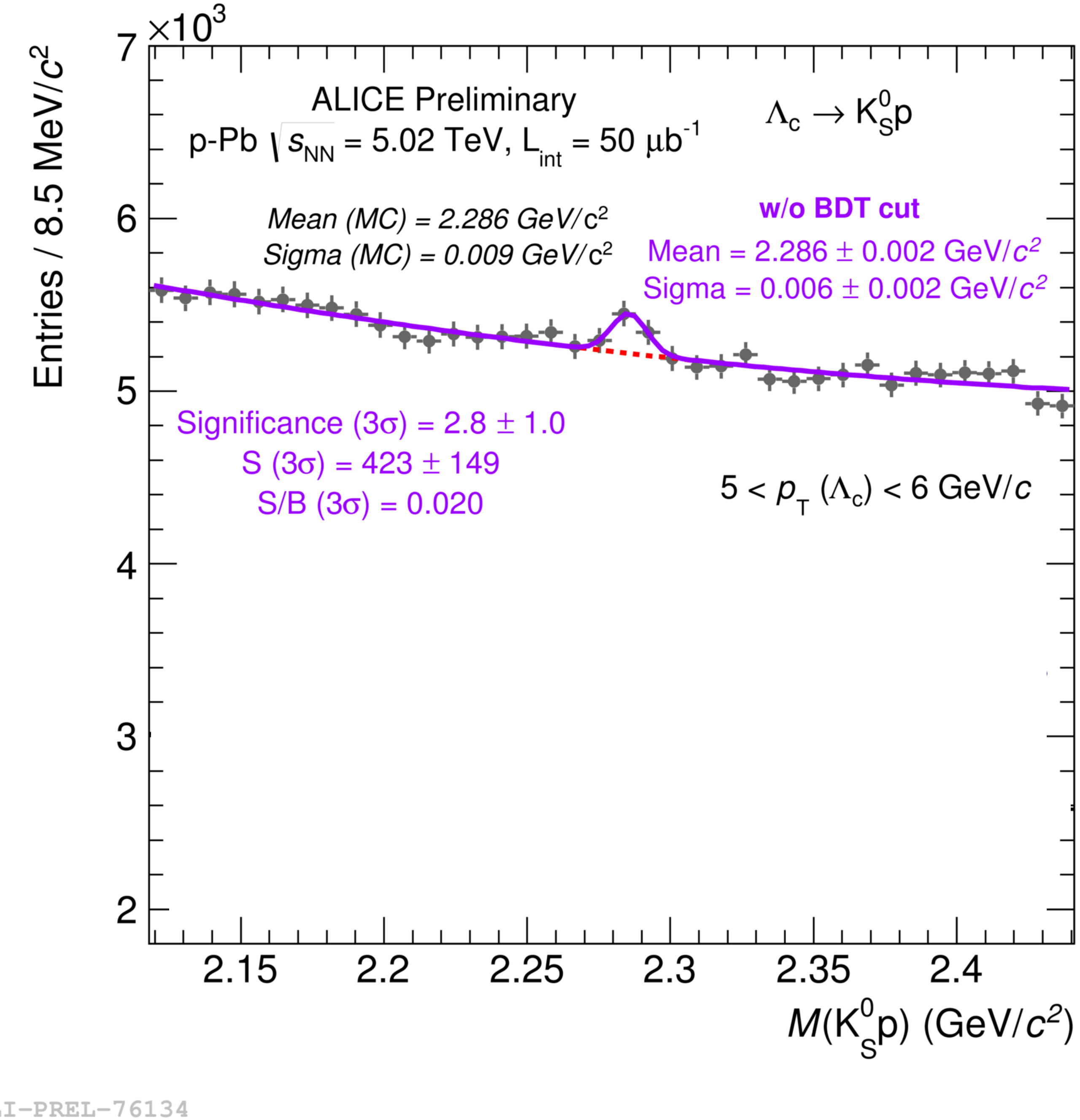
- Reconstruction of **2- and 3-prong decays** in heavy ion collisions is challenging: **large combinatorics**
    - (remember: several thousand particles/event)
- Many (topological, PID, …) **cut variables** available, often complex correlations: ideal playground for multivariate methods
- Limited “real-life” application so far:
    - Methods involved: hidden systematics?
    - Need excellent control over training sample (typically MC)
    - Not always clear gain with respect to traditional cuts analysis

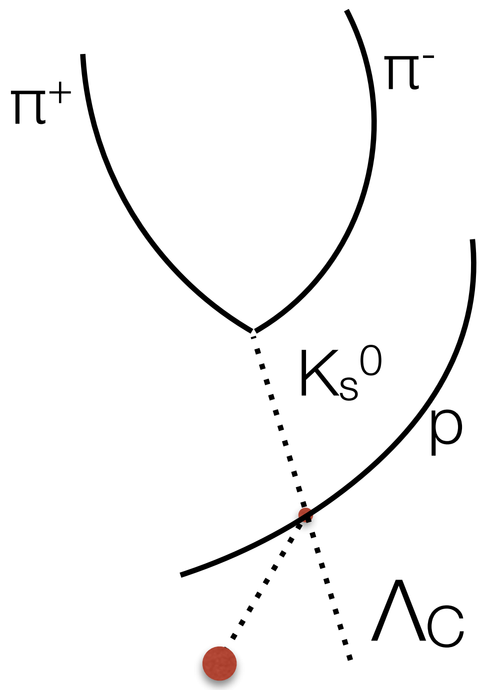


https://www.flickr.com/photos/mayaevening/138372058

$\Lambda_C \rightarrow K_S^0 p$ in p-Pb collisions

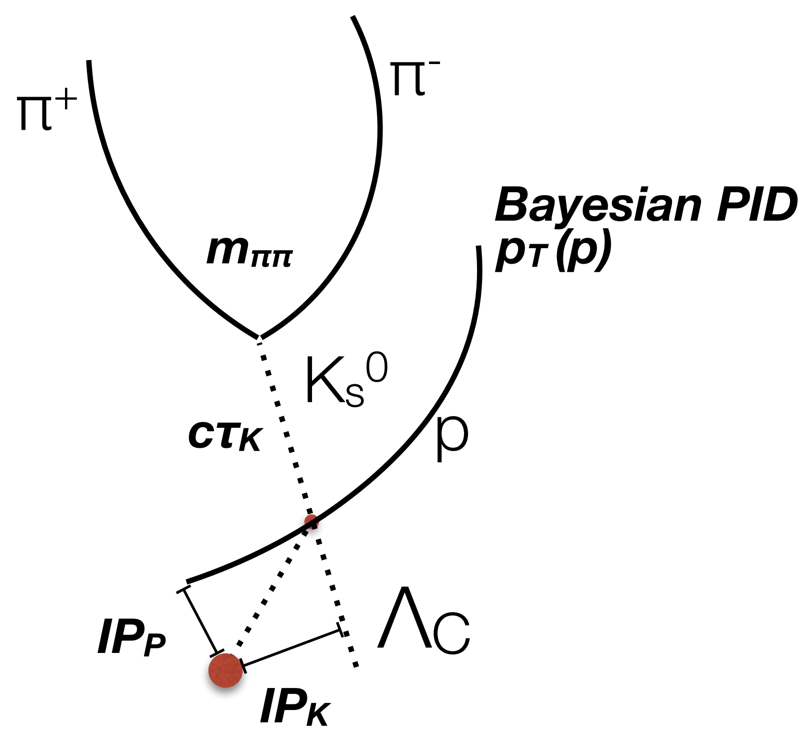
- **Recent attempts based on TMVA,** mostly **BDTs**
- Several channels studied:
  - $\Lambda \rightarrow p\pi$, $K_S^0 \rightarrow \pi\pi$, $\Lambda_C \rightarrow \pi K p$, …
- Example discussed here: $\boldsymbol{\Lambda_C \rightarrow K_S^0 p}$
- 3-prong decay: large combinatorial BG

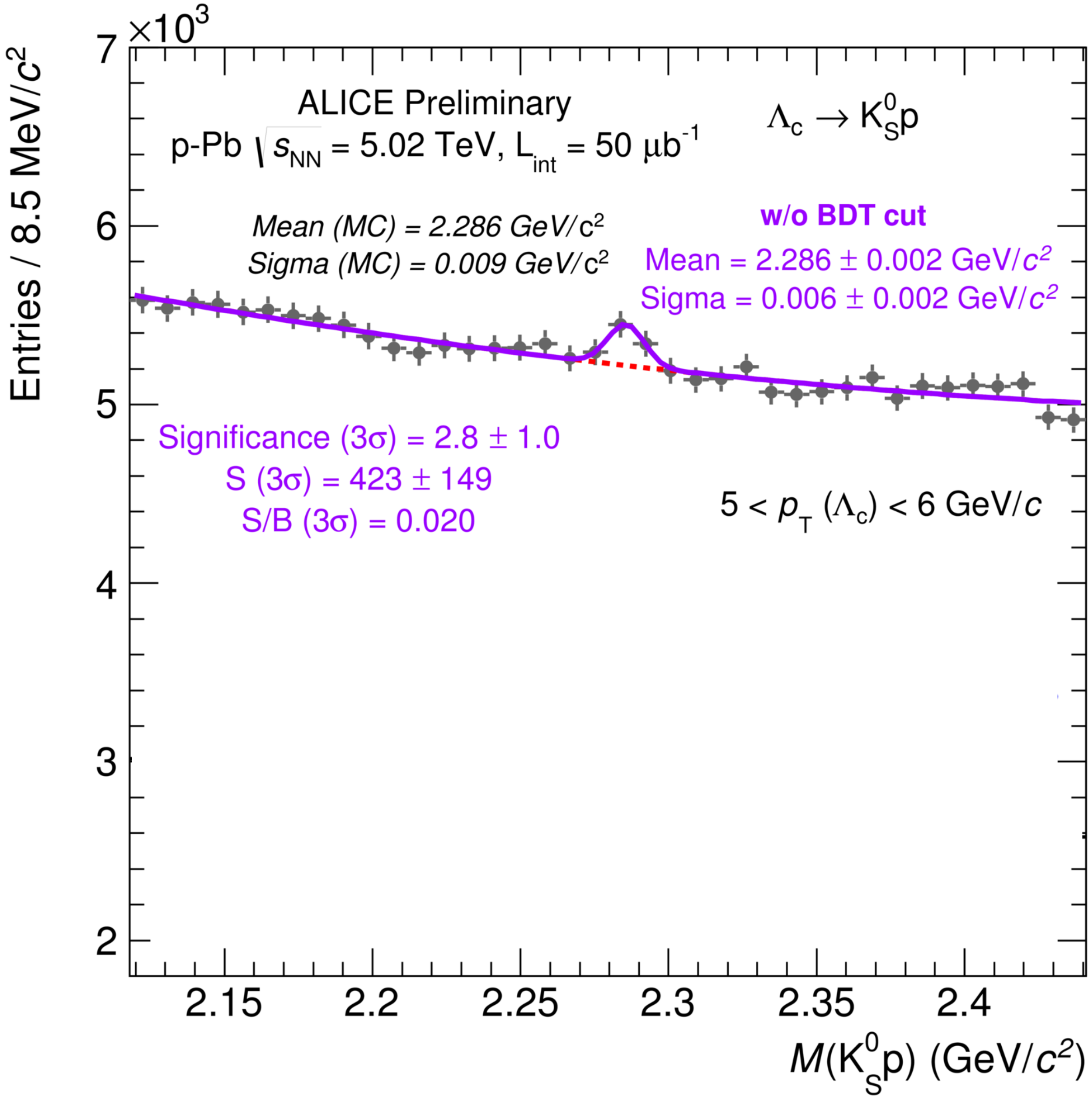




$\Lambda_C \rightarrow K_S^0 p$ in p-Pb collisions

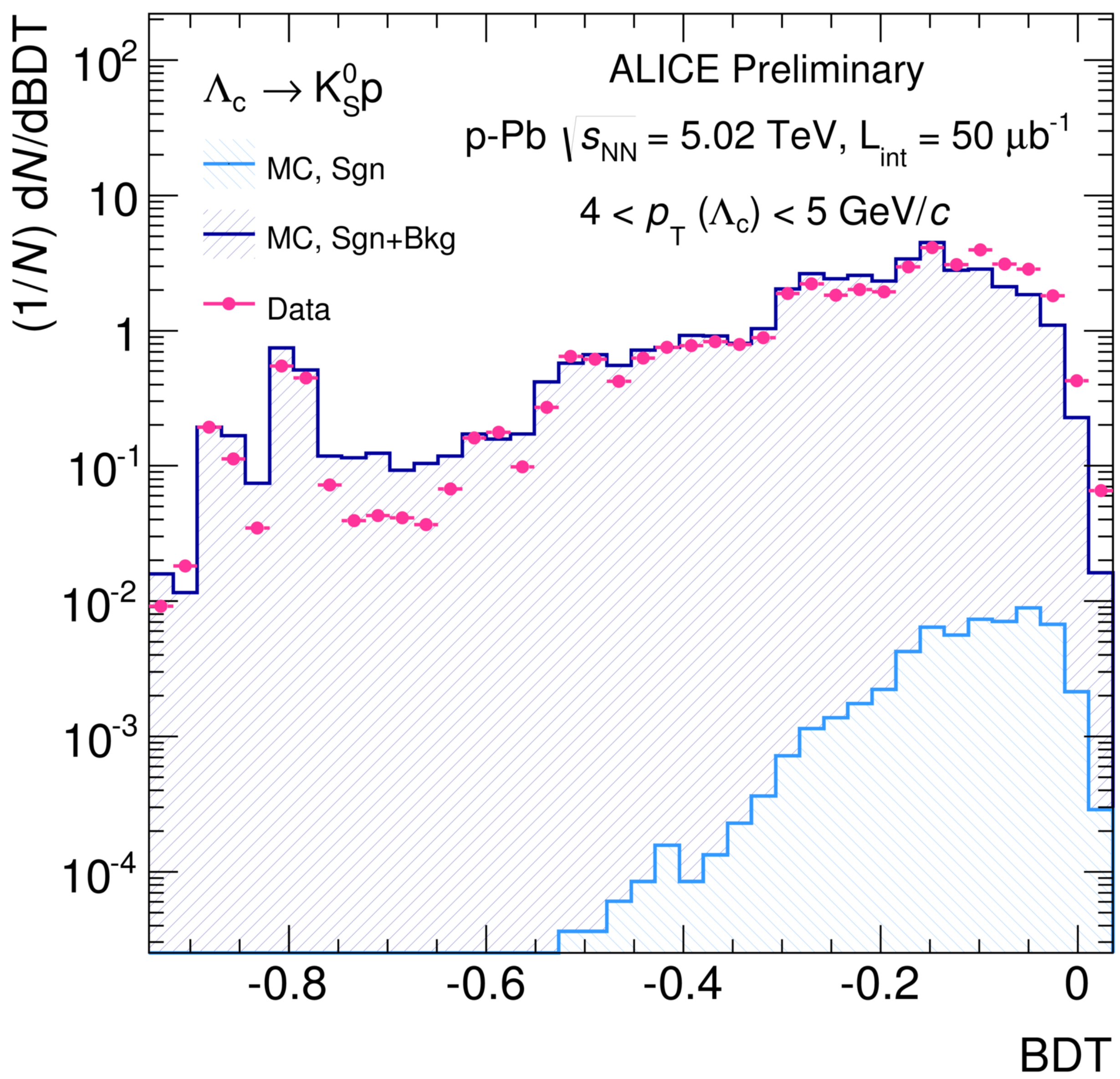
- **Recent attempts based on TMVA,** mostly **BDTs**
- Several channels studied:
  - $\Lambda \rightarrow p\pi$, $K_S^0 \rightarrow \pi\pi$, $\Lambda_C \rightarrow \pi K p$, …
- Example discussed here: $\boldsymbol{\Lambda_C \rightarrow K_S^0 p}$
- 3-prong decay: large combinatorial BG





ALI-PREL-76134

BDT output



ALI-PREL-76146

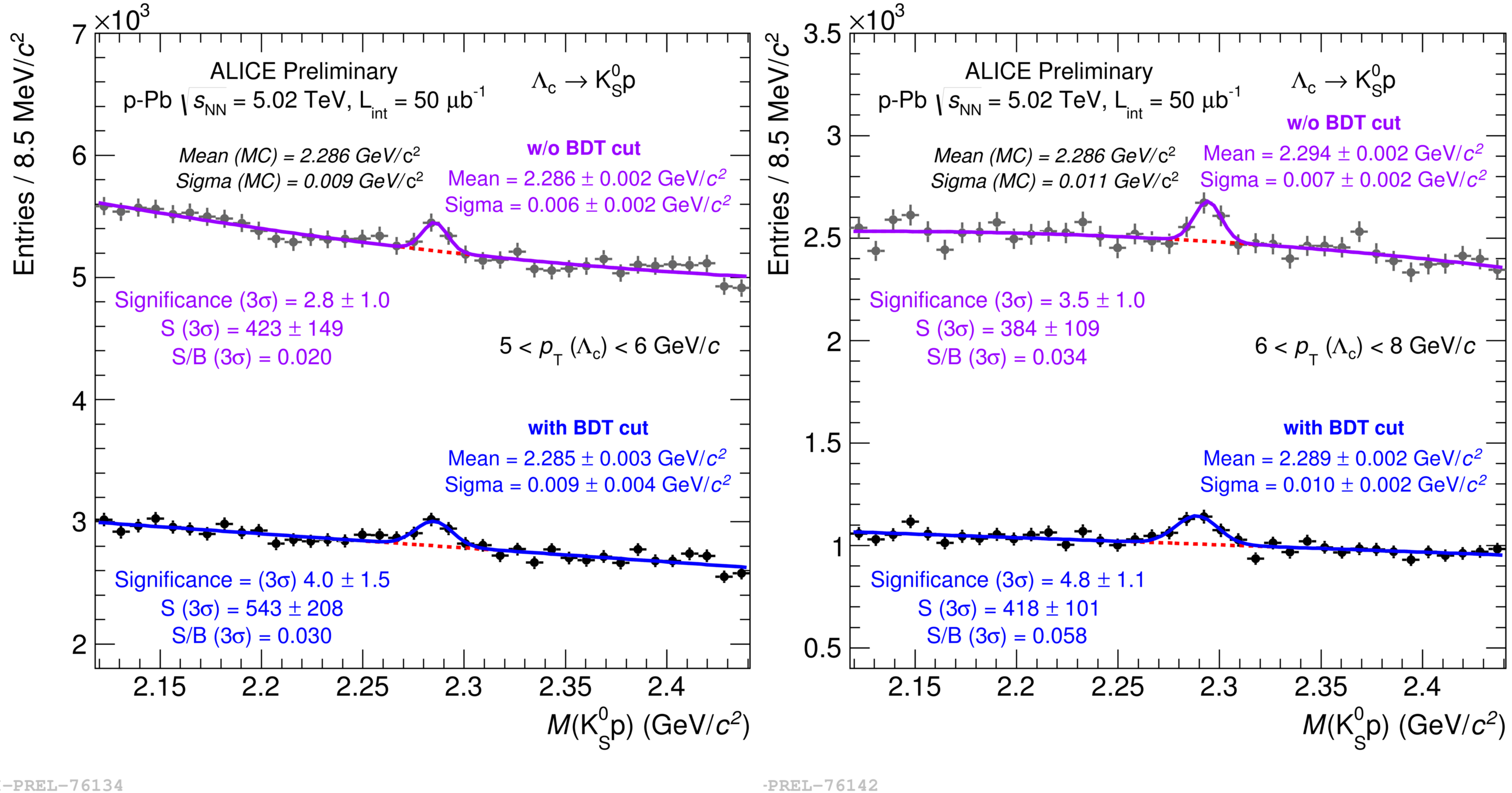
BDT output distribution in data and MC reasonably similar

Tuning repeated with BG from data (side bands)

Separation not perfect, tail at low BDT values for the signal

Optimization of BDT parameters in progress

$\Lambda_C \rightarrow Ks^0$: Results



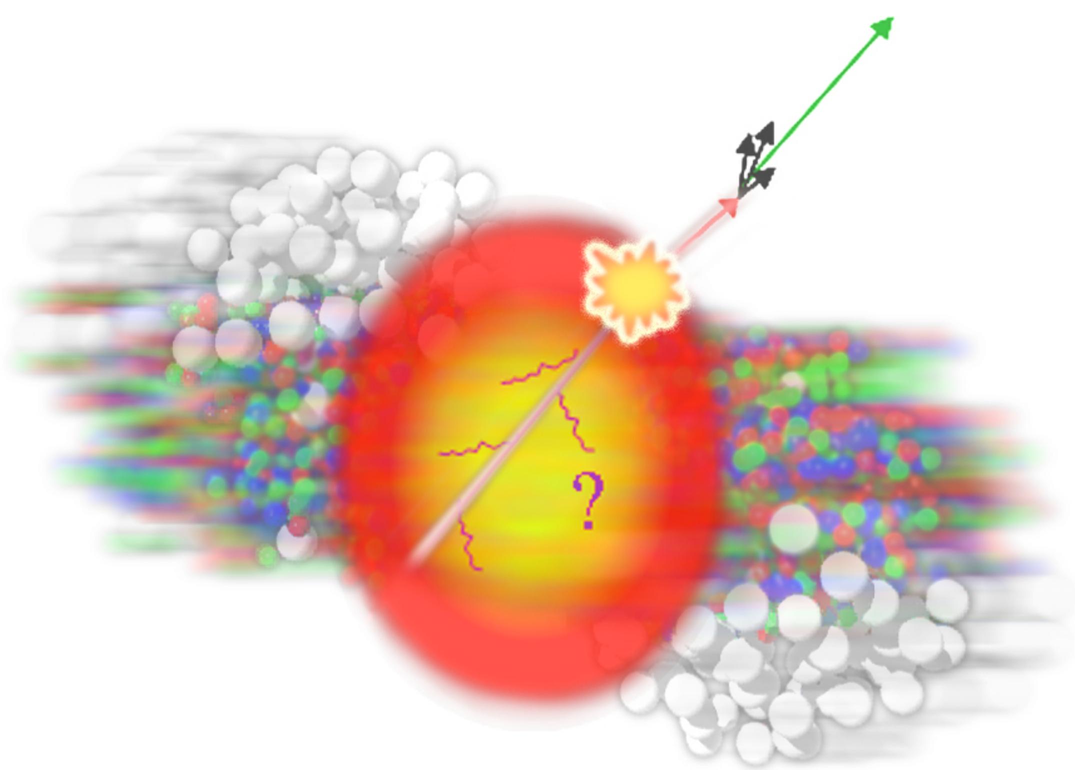
Significance improved by BDT

Multi-dimensional selection criteria simplified

Additional BDT systematics not dominant (large statistical error)

Quark vs Gluon Jet Discrimination

A primer on jet quenching



Recoil **jet loses energy** when traversing the medium "Radiative" and "Collisional" energy loss

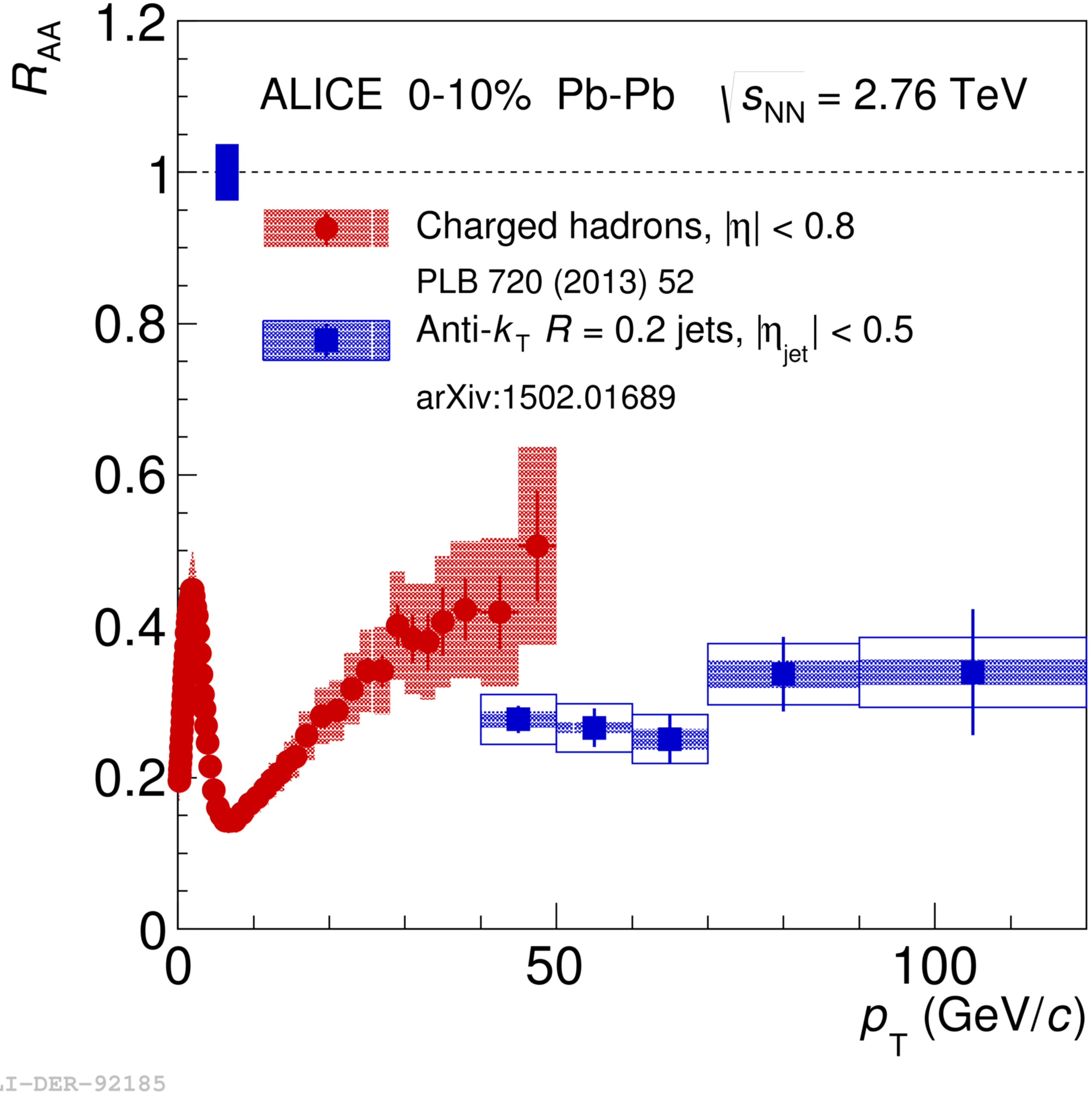
$\mathbf{\Delta} \boldsymbol{E}_{\mathbf{g}} \mathbf{>} \mathbf{\Delta} \boldsymbol{E}_{\mathbf{u,d,s}}$ (Color factors)

Distinguishing Quark and Gluon jets would allow to study **microsopic process** of energy loss in detail

"$\boldsymbol{R}_{\mathbf{AA}}$" is the simplest way of studying this modification

A primer on jet quenching

$$R_{AA} = \frac{AA}{\text{rescaled pp}}$$



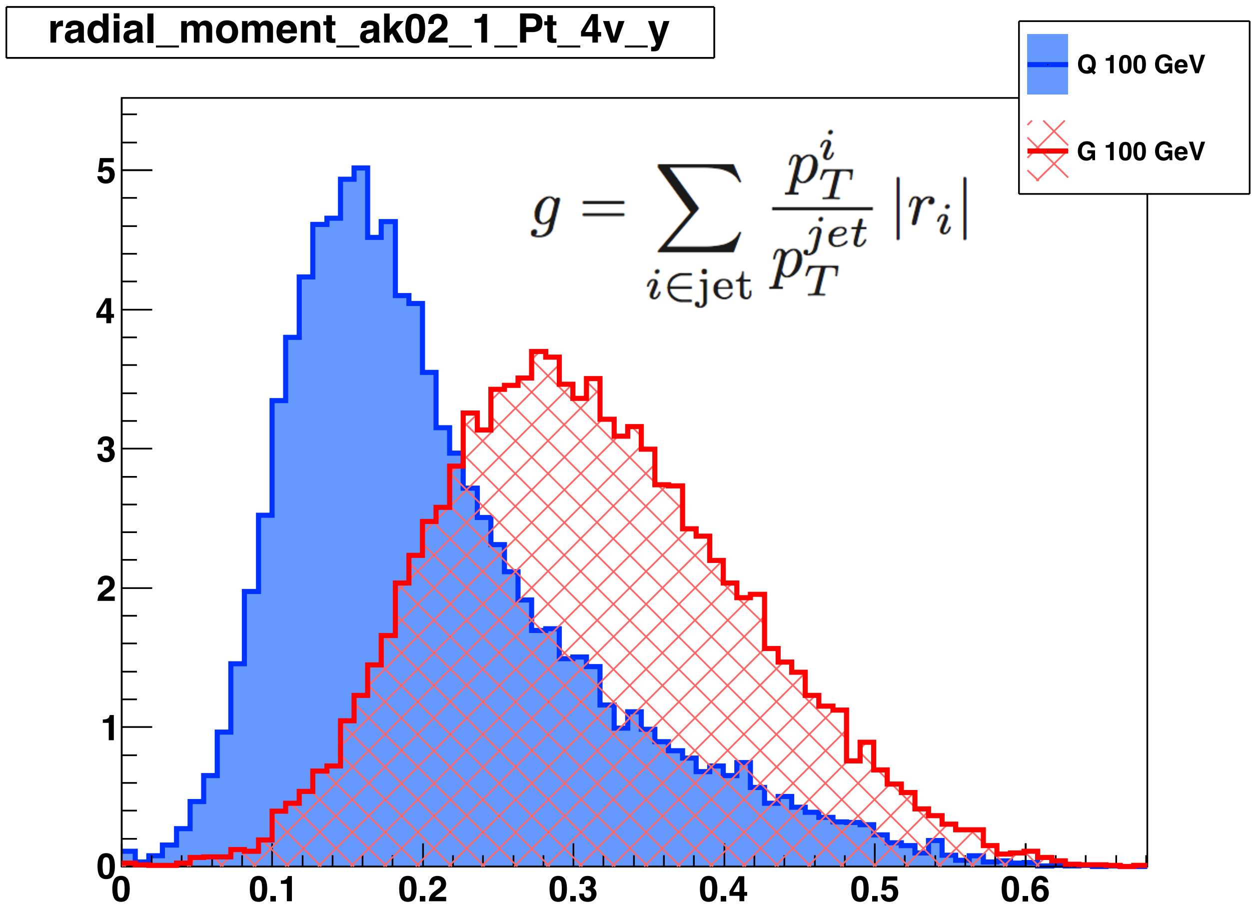
Recoil **jet loses energy** when traversing the medium "Radiative" and "Collisional" energy loss

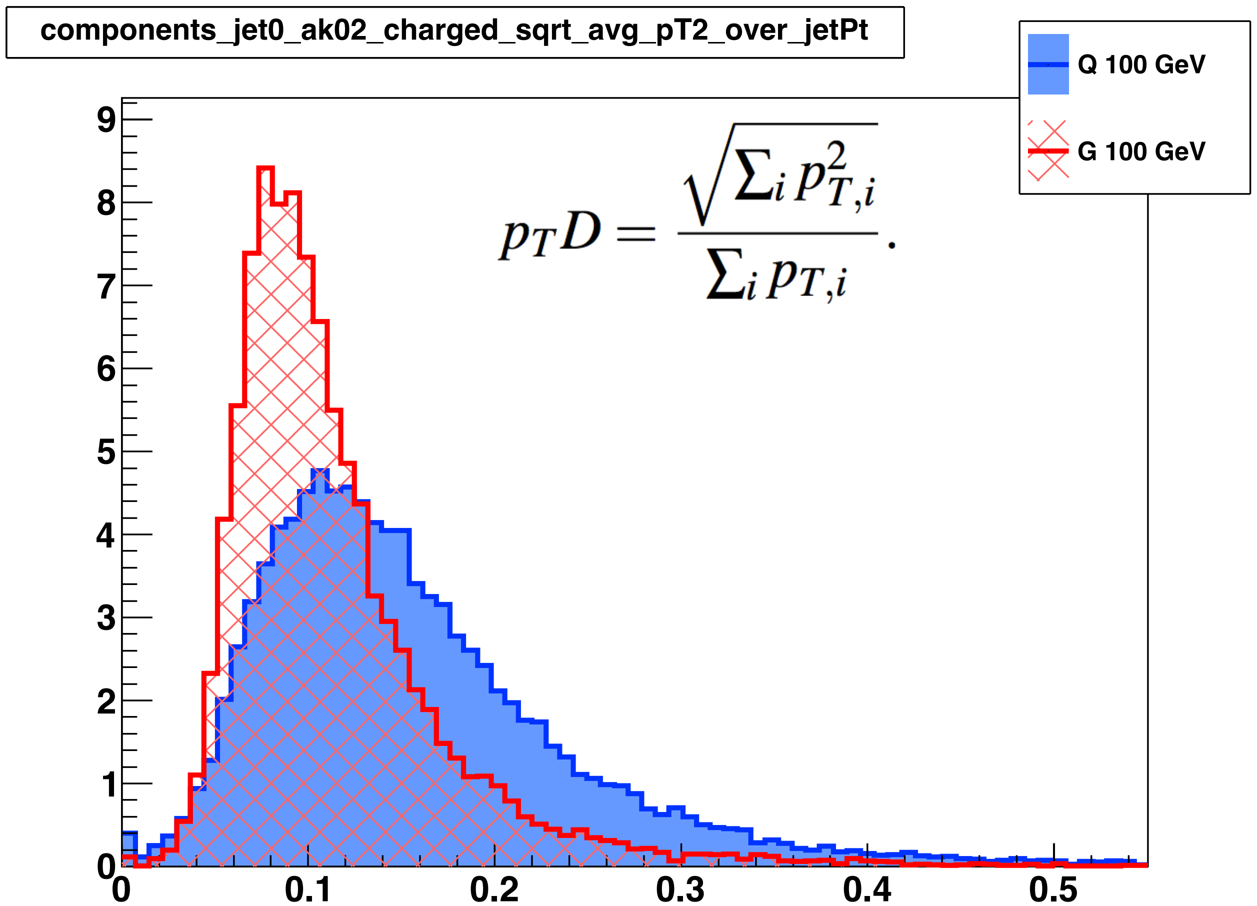
$\boldsymbol{\Delta E_g > \Delta E_{u,d,s}}$ (Color factors)

Distinguishing Quark and Gluon jets would allow to study **microsopic process** of energy loss in detail

"$\boldsymbol{R_{AA}}$" is the simplest way of studying this modification

Quark-Gluon Jet Discrimination





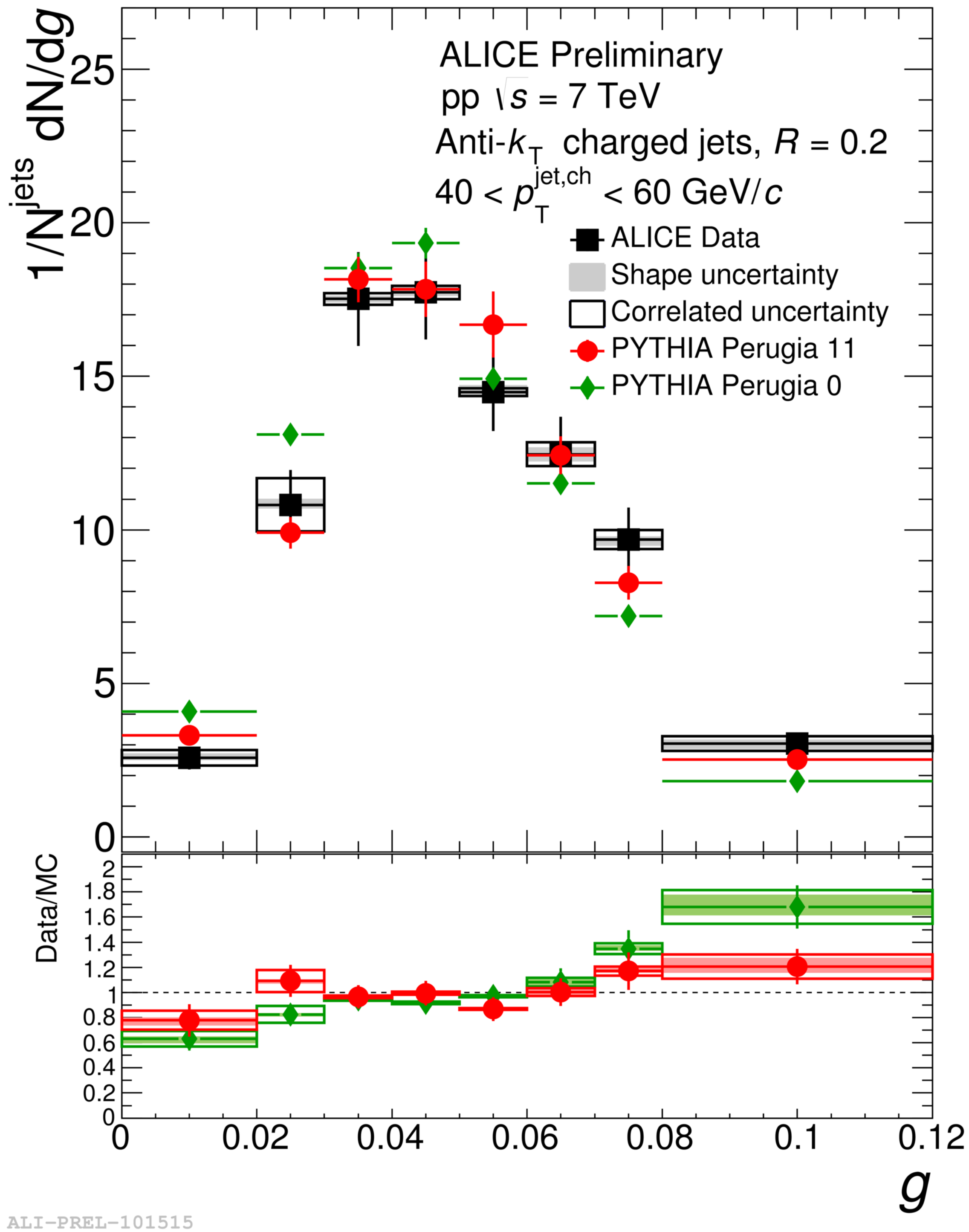
Jet shapes like angularities, radial moment or p_TD show sensitivity to differences between **quark and gluon** fragmentation

(Plots from: http://jets.physics.harvard.edu/qvg/)

Used as **input to ML** methods to **tag** jets as *q* or *g*

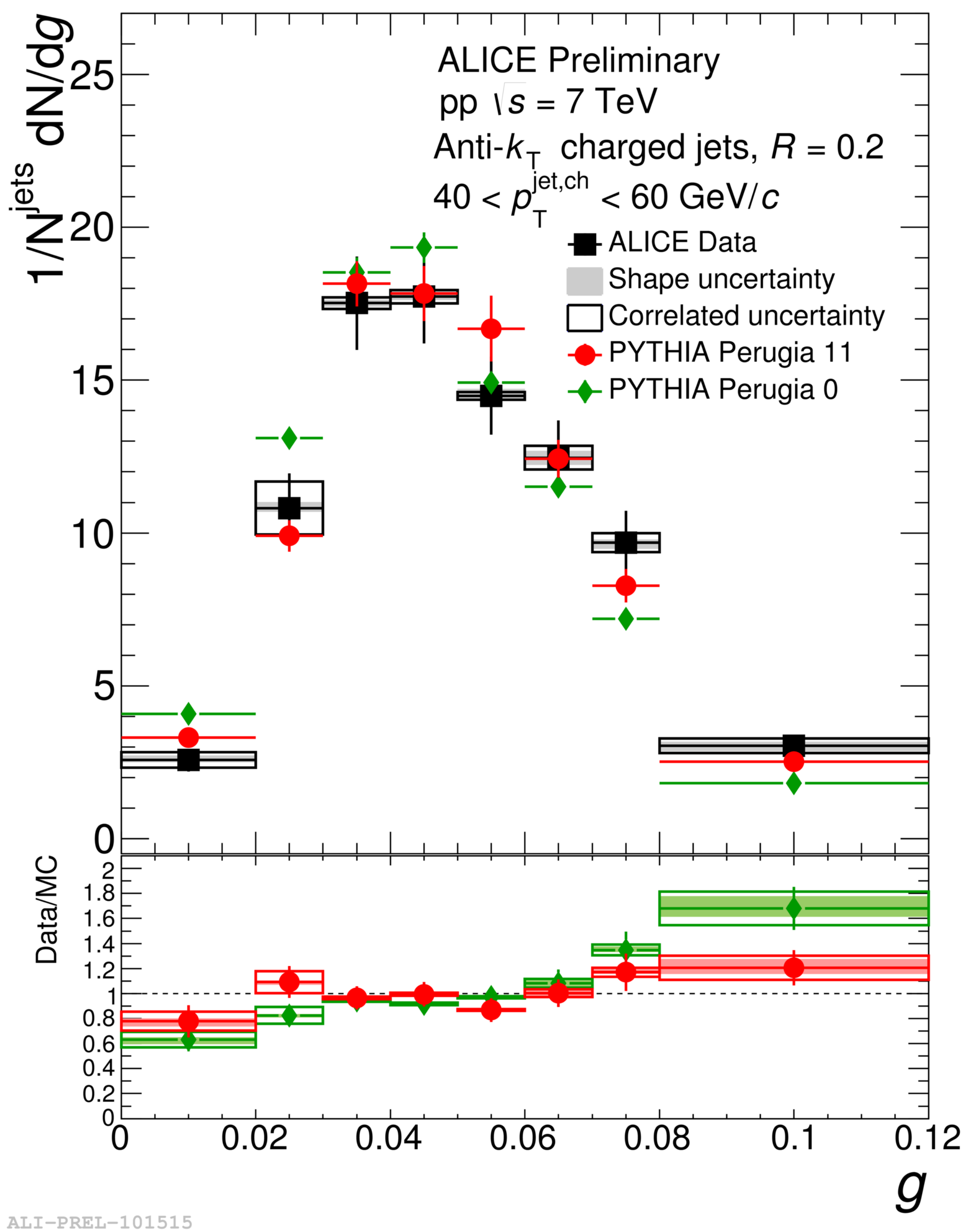
Other potential **areas of applications**: fake jets, jet energy estimation, heavy-flavor tagging, …

Jet Shapes, results

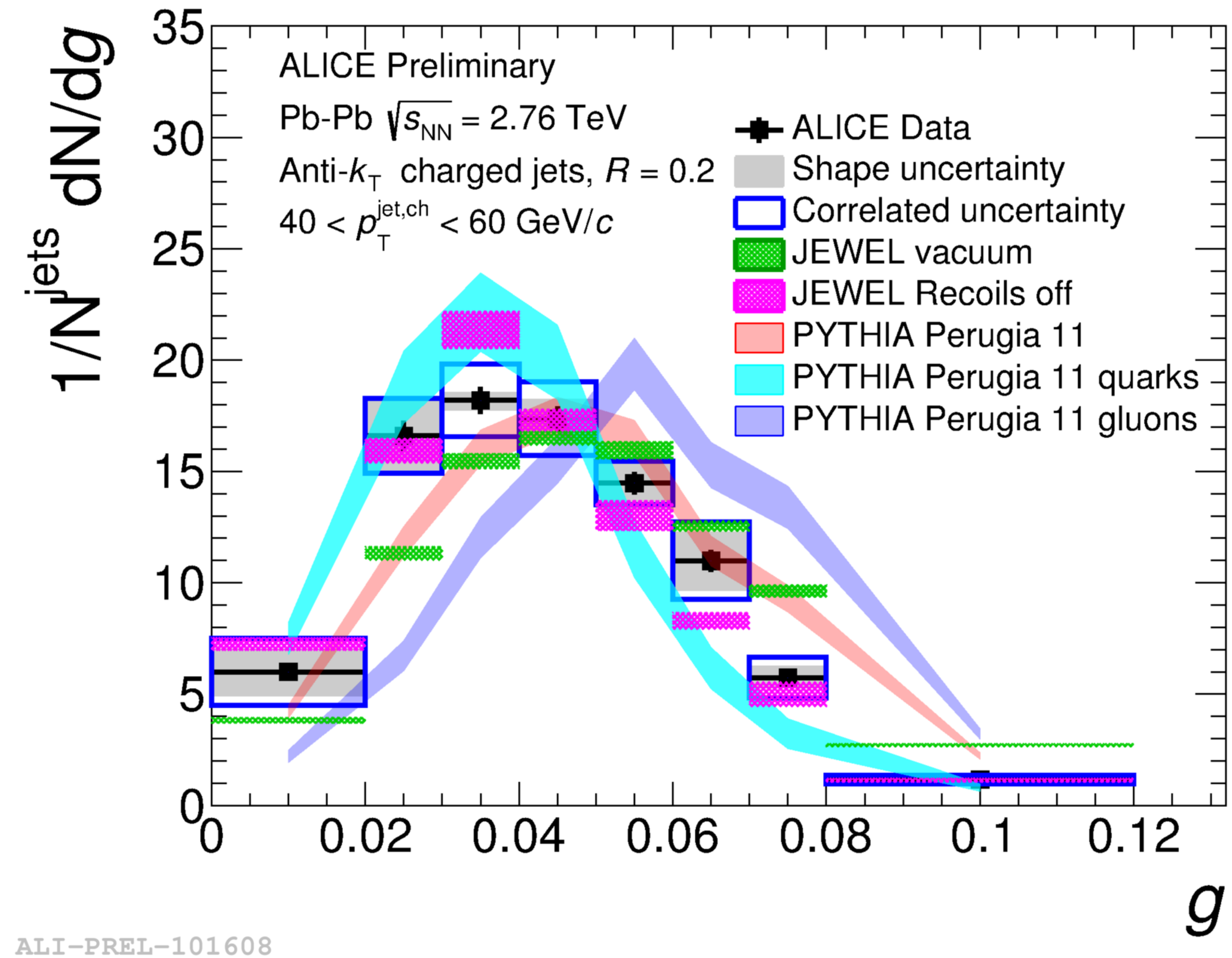


Pythia reproduces jet **shapes** (e.g. girth) in pp collisions

Jet Shapes, results



Pythia reproduces jet **shapes** (e.g. girth) in pp collisions



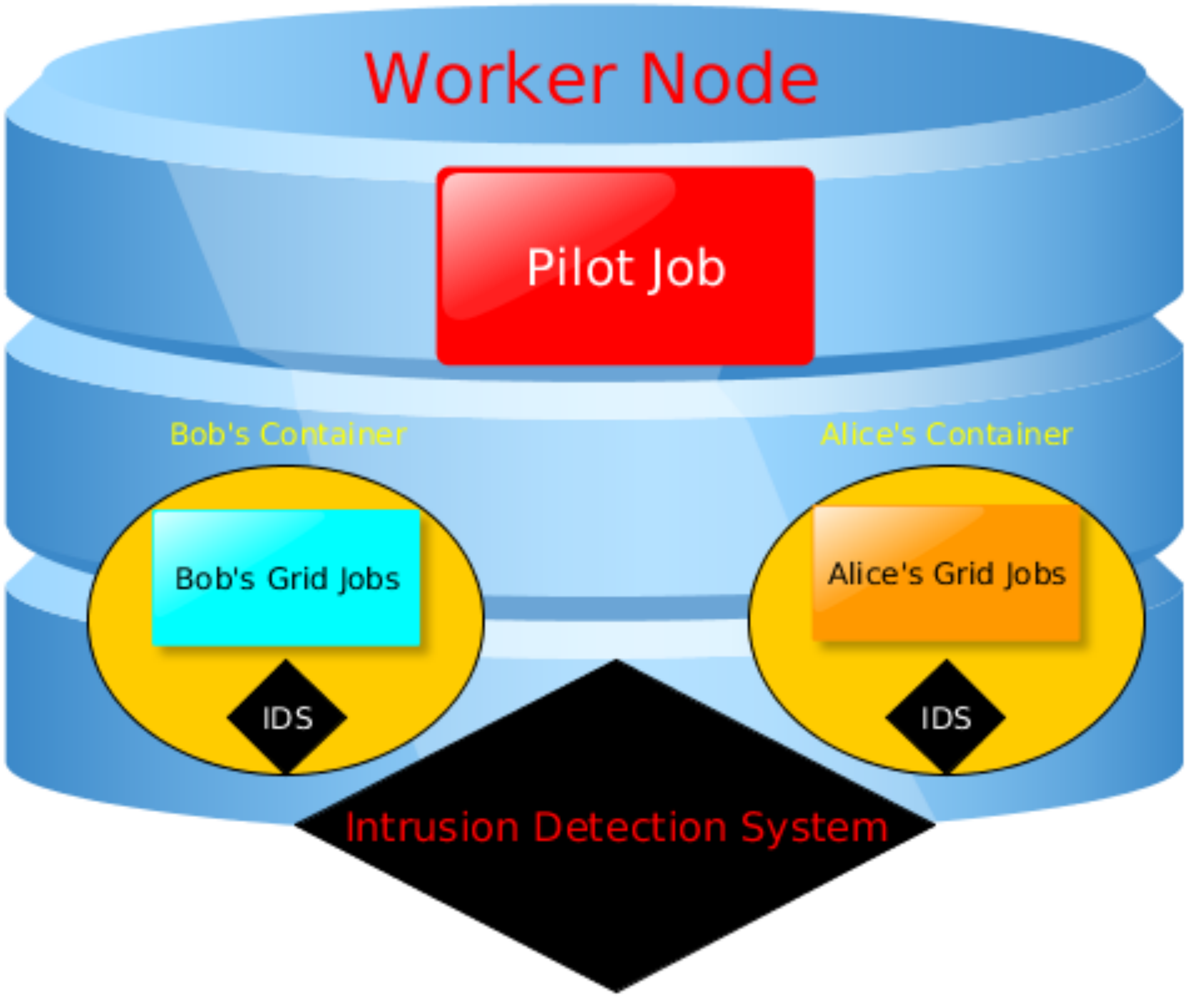
Shapes change in Pb-Pb, more “**quark like**”

Different suppression of q and g?

Modification of fragmentation?

“Exotic” Application: Grid Security

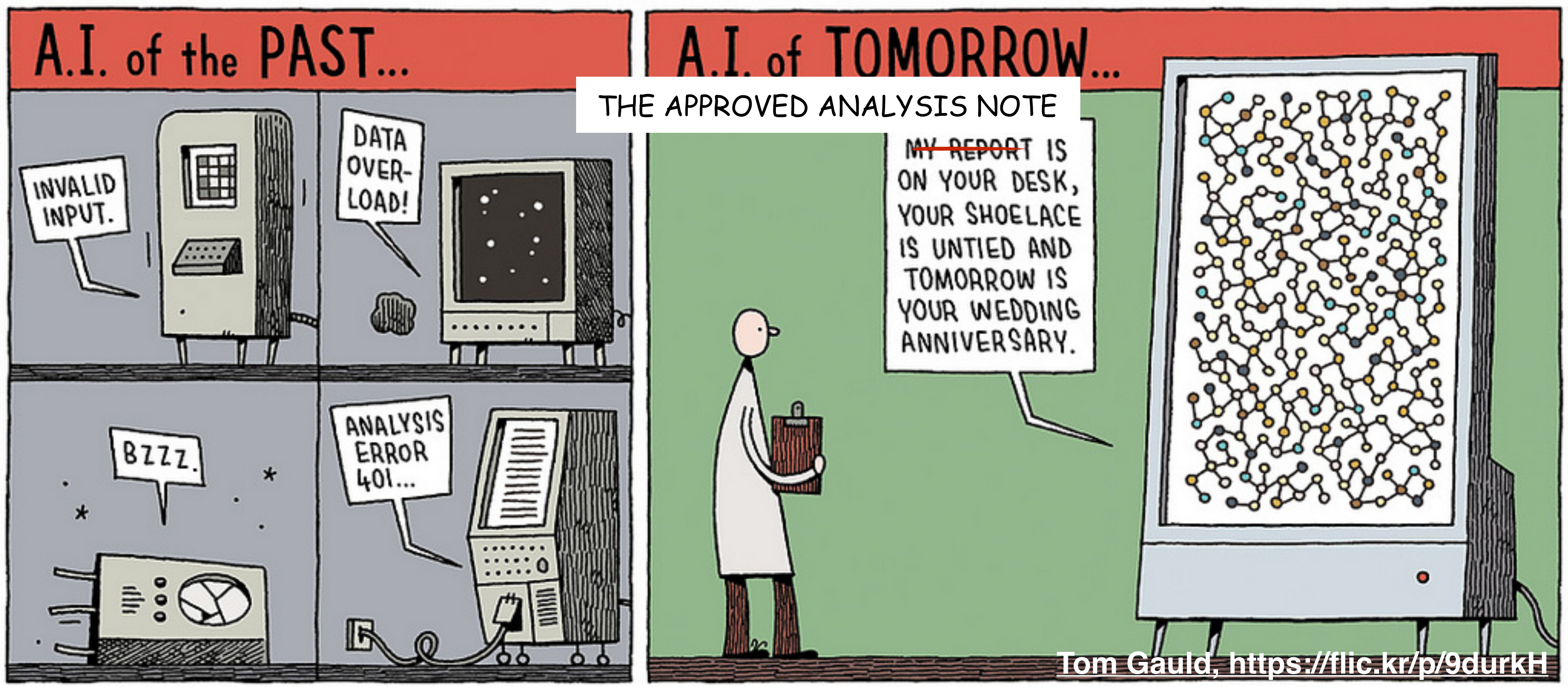
- **Feature space**: monitoring metrics
  - **Resource** consumption (Like CPU/ Memory)
  - **Connection** information (TCP/IP)
  - **System calls**
- **Machine Learning** Method:
  - Recurrent Artificial **Neural Network**
  - A cascade of **several algorithms?**
- **Malicious samples**:
  - Run test Jobs → DoS, Bitcoin mining, botnet, malware, ...
  - Capture metrics



CHEP2015, https://indico.cern.ch/event/304944/contribution/14

Summary

- Several potential applications for machine learning techniques in ALICE
  - Detector, reconstruction, physics analysis, computing
- Early attempts, no widespread use yet
- Increasing interest and expertise



Tom Gauld, https://flic.kr/p/9durkH

Thanks! Andrea Alici, Andres Gomez, Andrew Lowe, Chiara Zampolli, David Rohr, Davide Caffarri, Georgios Krintiras, Jaime Norman, Julien Faivre, Leticia Cunqueiro, Mike Sas, Michael Weber, Yvonne Pachmayer, Zaida Conesa Del Valle

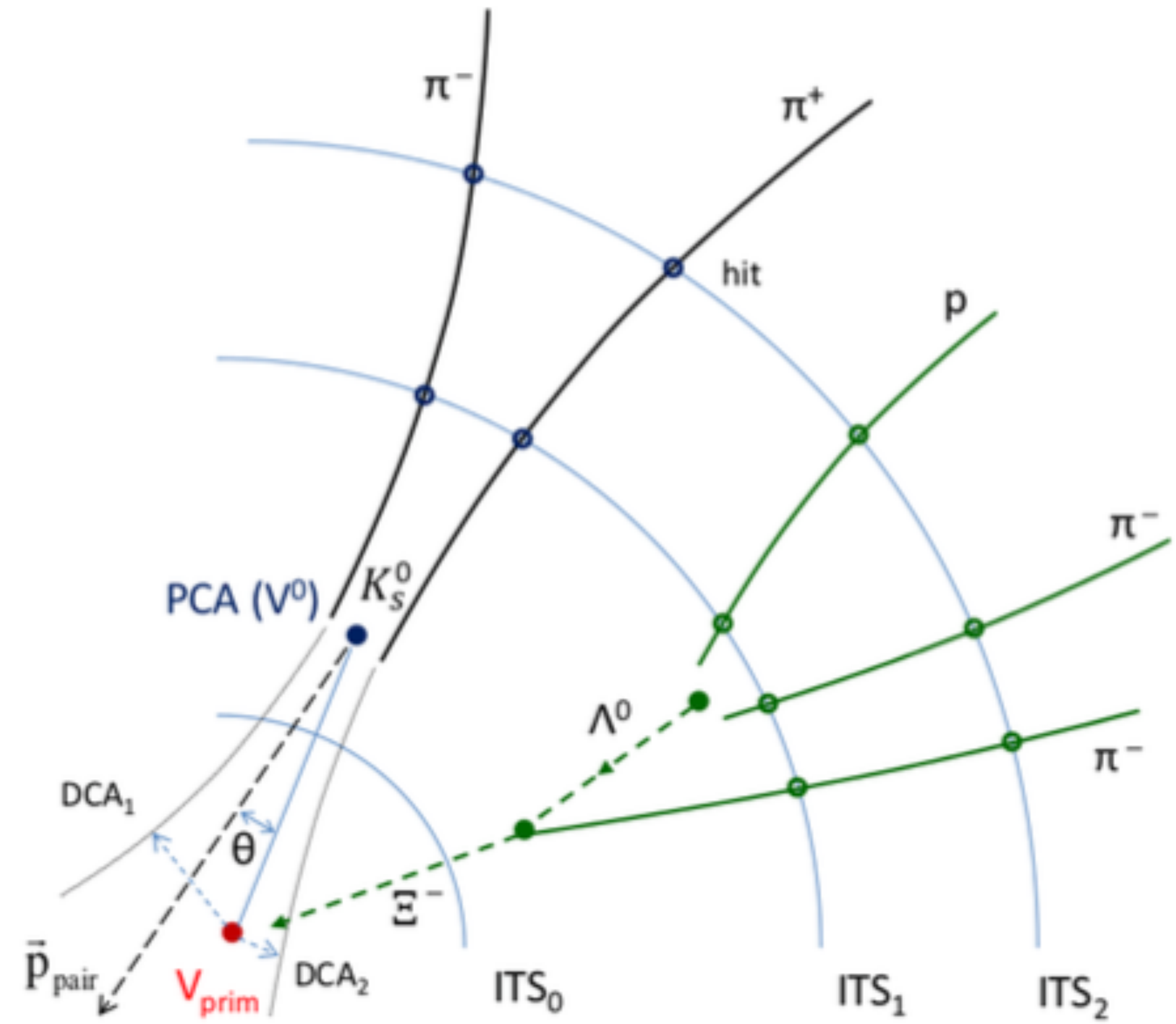
Backup

Invariant mass reconstruction

Particle identification cuts can be based on several sub-detectors (ITS, TPC, TOF…)

Topological reconstruction of weakly decaying particles ("**high level features**"):

- Decay radius
- cos(θ) – pointing angle
- Distance of their closest approach (DCA1 and DCA2) to V_{prim}
- Distance of daughters at the point of closest approach (PCA)
- Armenteros-Podolansky variables

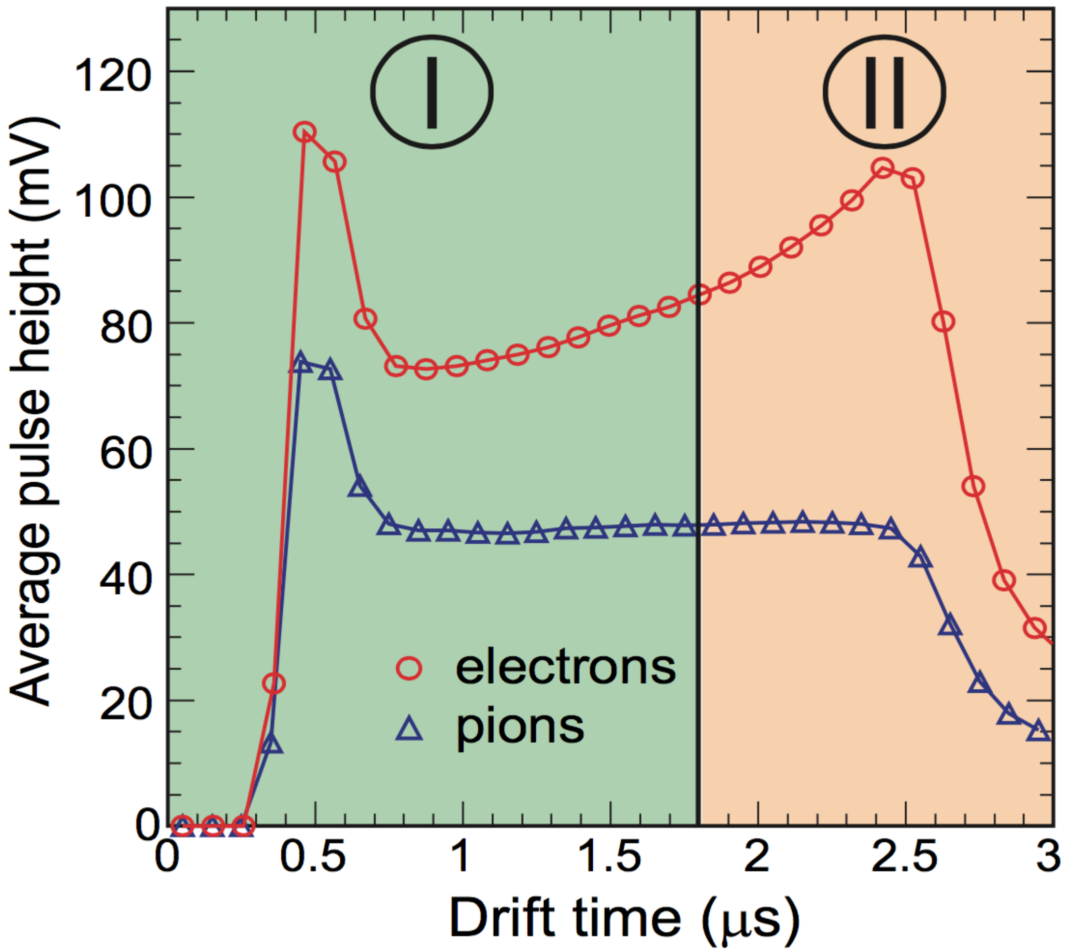


Correlations among the cut variables

1D Likelihood: start probability that a particle *k* deposits a charge Q

$$L\left(e|\overline{Q}\right) = \frac{P\left(\overline{Q}|e\right)}{\sum_k P\left(\overline{Q}|k\right)}$$ k= e, π, k, p, …

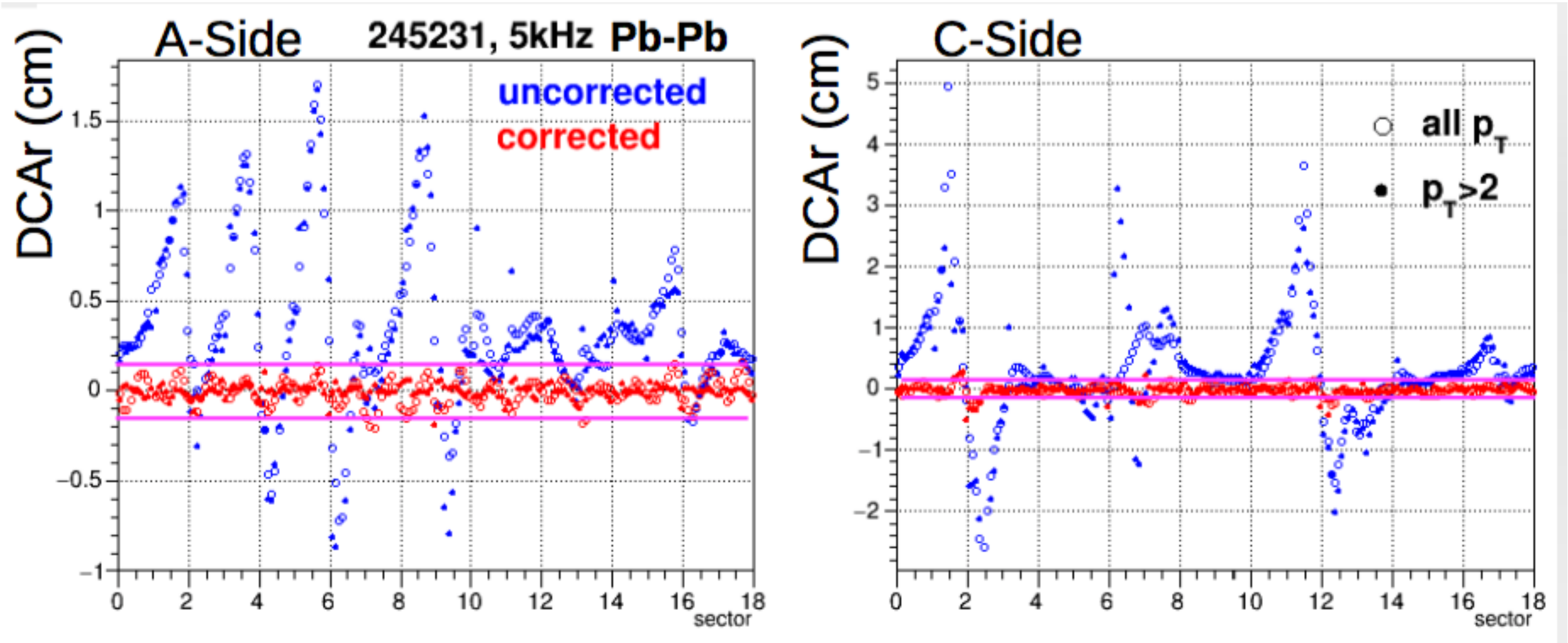
$$P\left(\overline{Q}|e\right) = \prod_{j=1}^{n} P^j\left(Q_j|e\right) = \prod_{j=1}^{n} P\left(Q_j|e\right).$$ j=layer



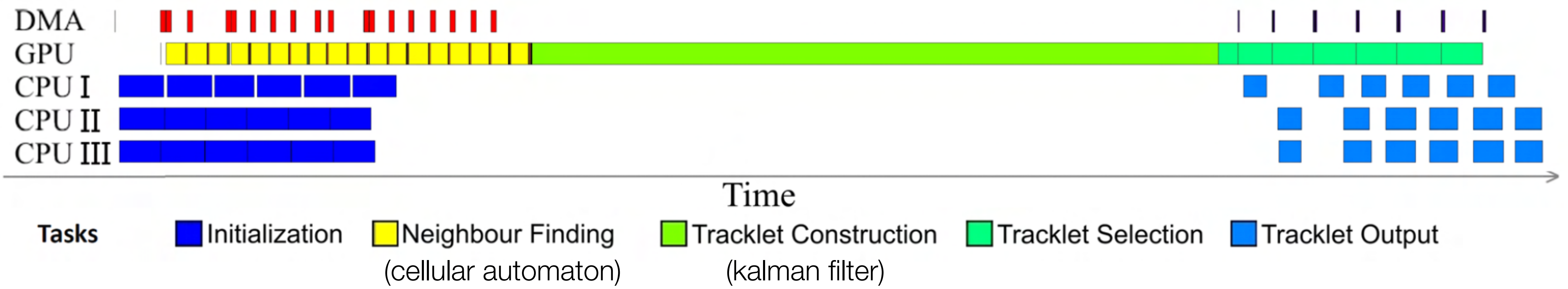
2D Likelihood: charged deposition in 2 time bins

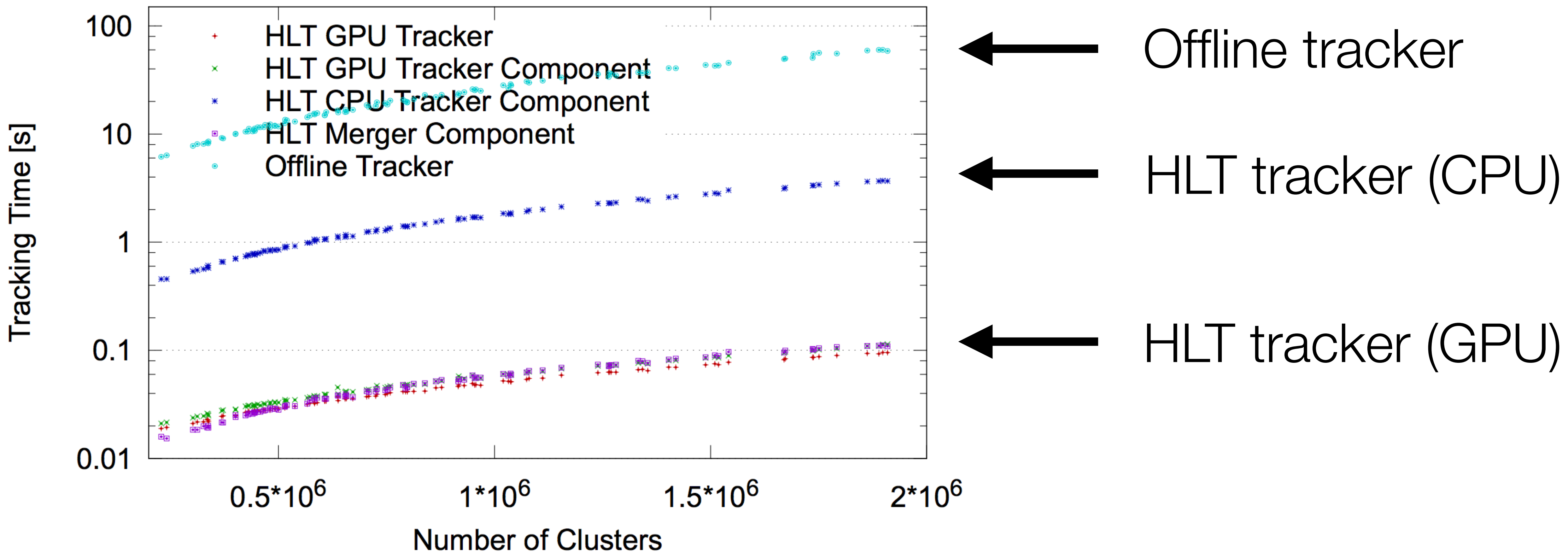
$$P\left(\overline{Q1},\overline{Q2}|e\right) = \prod_{j=1}^{6} P\left(Q1_j, Q2_j|e\right).$$

Alternative: NN (**MLP)** with charge deposited in *n* time bins (TMVA based)


A-Side
245231, 5kHz Pb-Pb
uncorrected
corrected
DCAr (cm)
sector
C-Side
all p_T
p_T>2
DCAr (cm)
sector

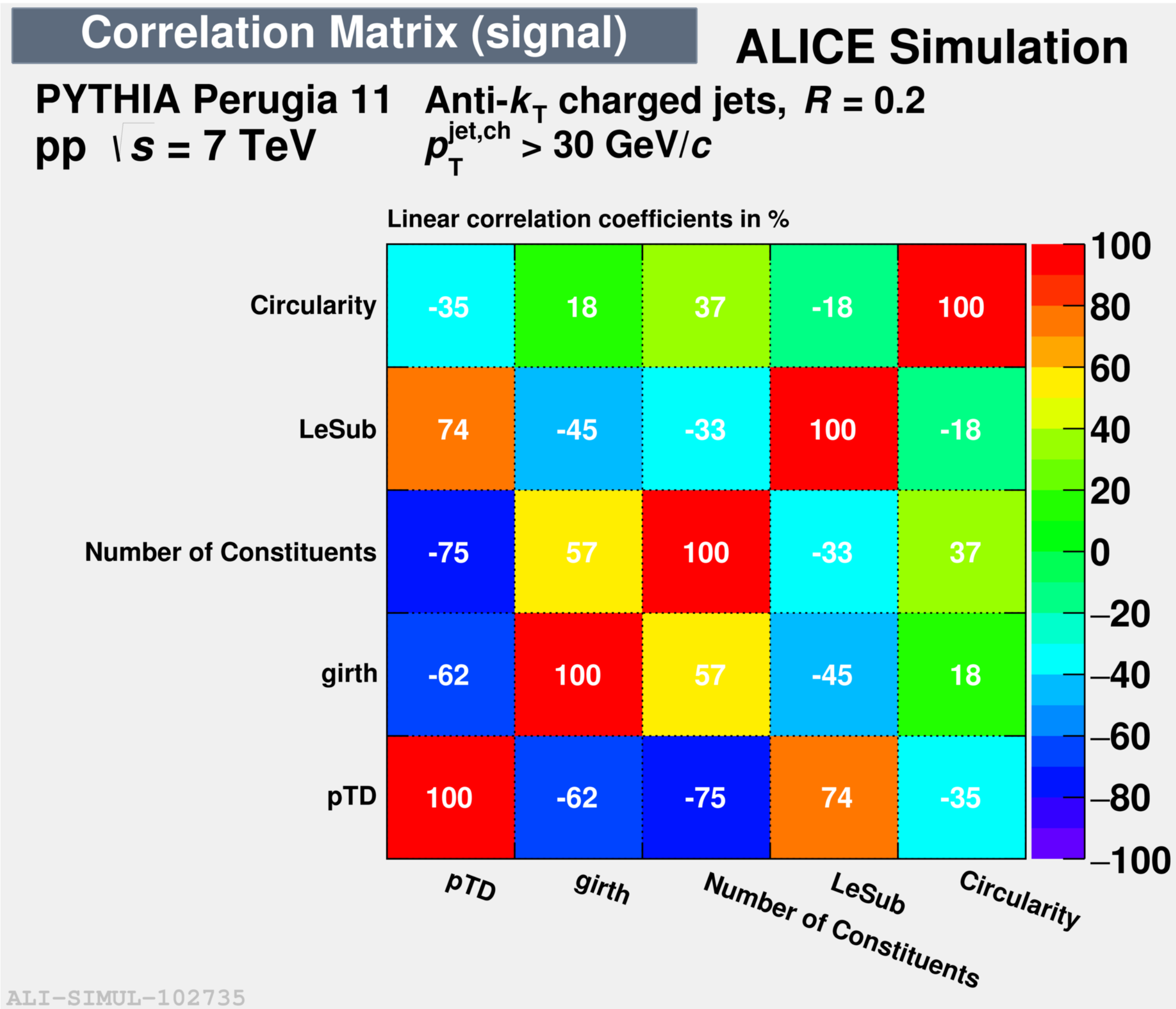
Performance

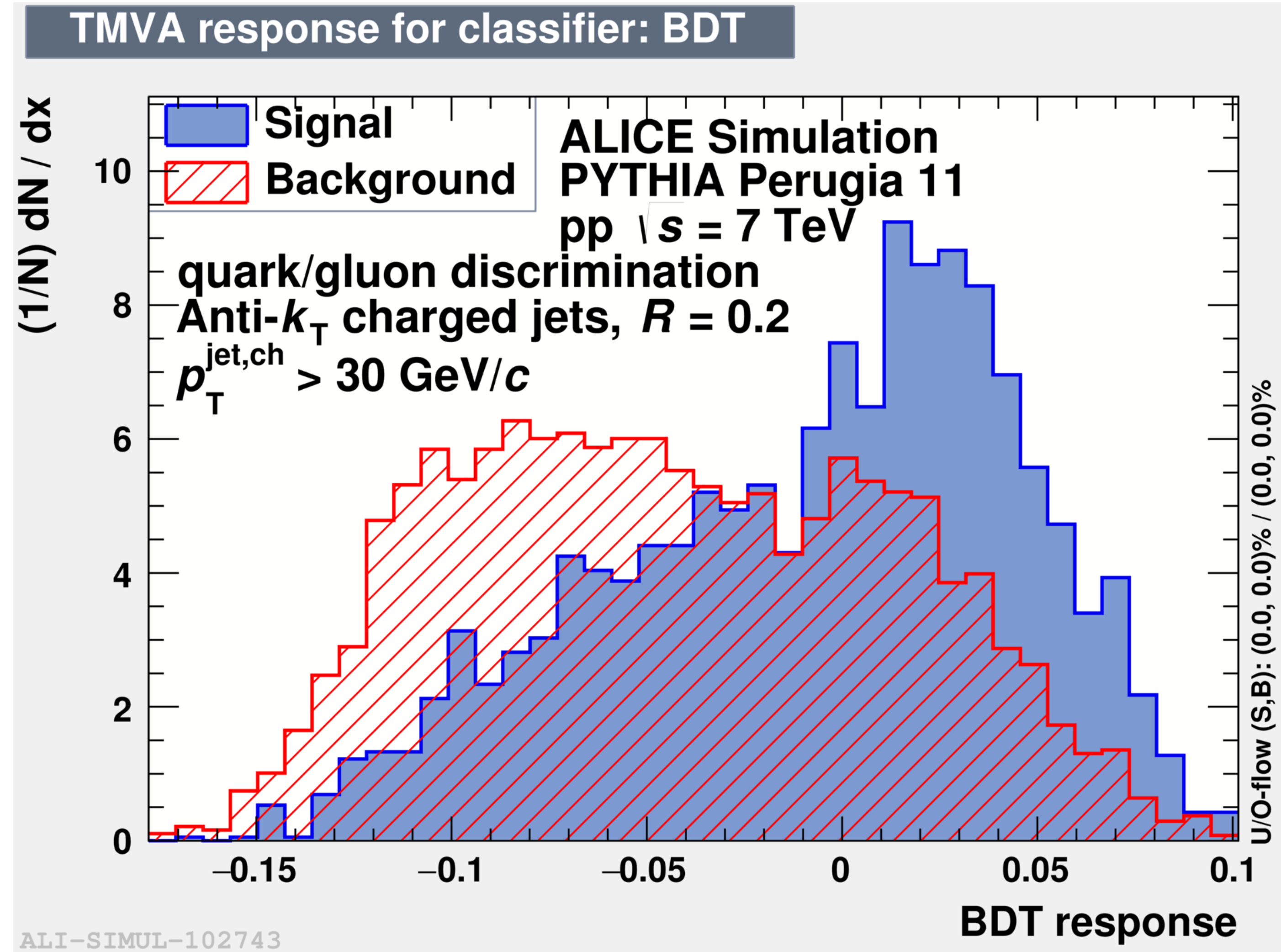




IEEE TNS, 58(4), 1845–1851, 10.1109/TNS.2011.2157702
CNNA 2012 proceedings, 10.1109/CNNA.2012.6331460

Tagging Jets with BDT



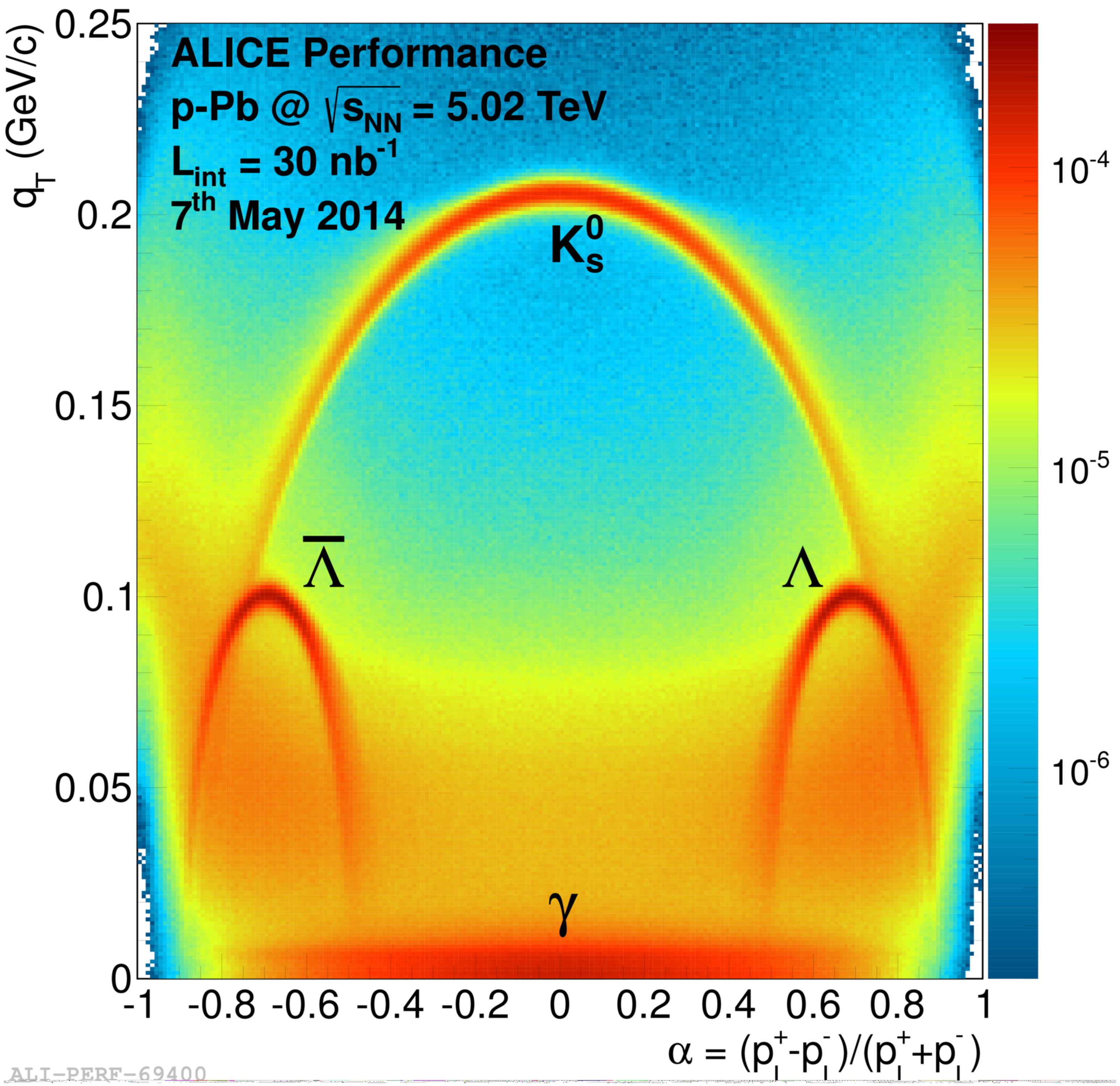


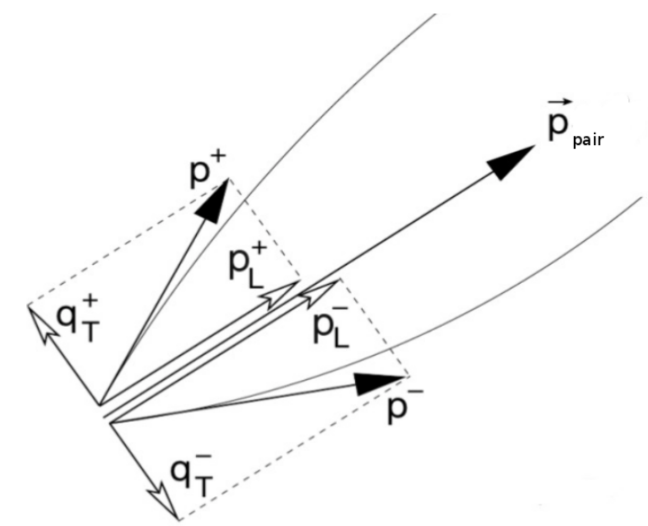
Pythia Perugia 2011, particle level

Anti-kT, R=0.2

Variables input to BDT: p_TD, girth, constituents, LeSub, Circularity

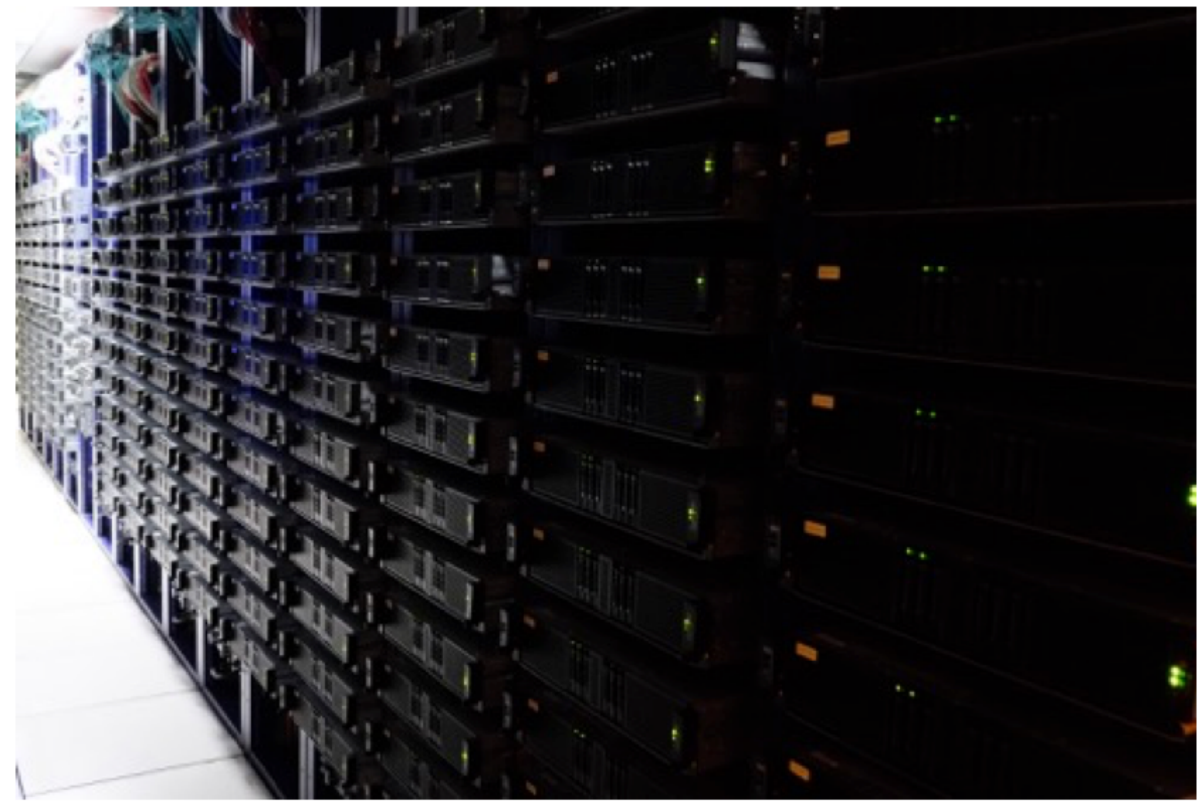
Armenteros-Podolanski





The ALICE High Level Trigger

- 180 nodes - 4320 CPU cores:
  - 2x Intel Xeon E5-2697 CPUs (2.7 GHz, 12 Cores each).
  - 128 GB RAM.
  - 2x 240 GB SSD (used in Raid 1 - Mirroring).
  - 1 AMD FirePro S9000 GPU.
  - 1 C-RORC board (installed in 74 nodes).



- 6+ Infrastructure Nodes:
  - 2x Intel Xeon E5-2690, 3.0 GHz 10 Cores.
  - 128 GB RAM.
  - 2x 240 GB SSD (Raid 1 - mirroring).
- Network:
  - <u>Data</u>: Infiniband in IPoIB Mode (FDR with 56Gb/s, full bisection bandwidth).
  - <u>Management</u>: gigabit ethernet with sideband IPMI - one physical ethernet port per node.
    - 10Gbit backbone.