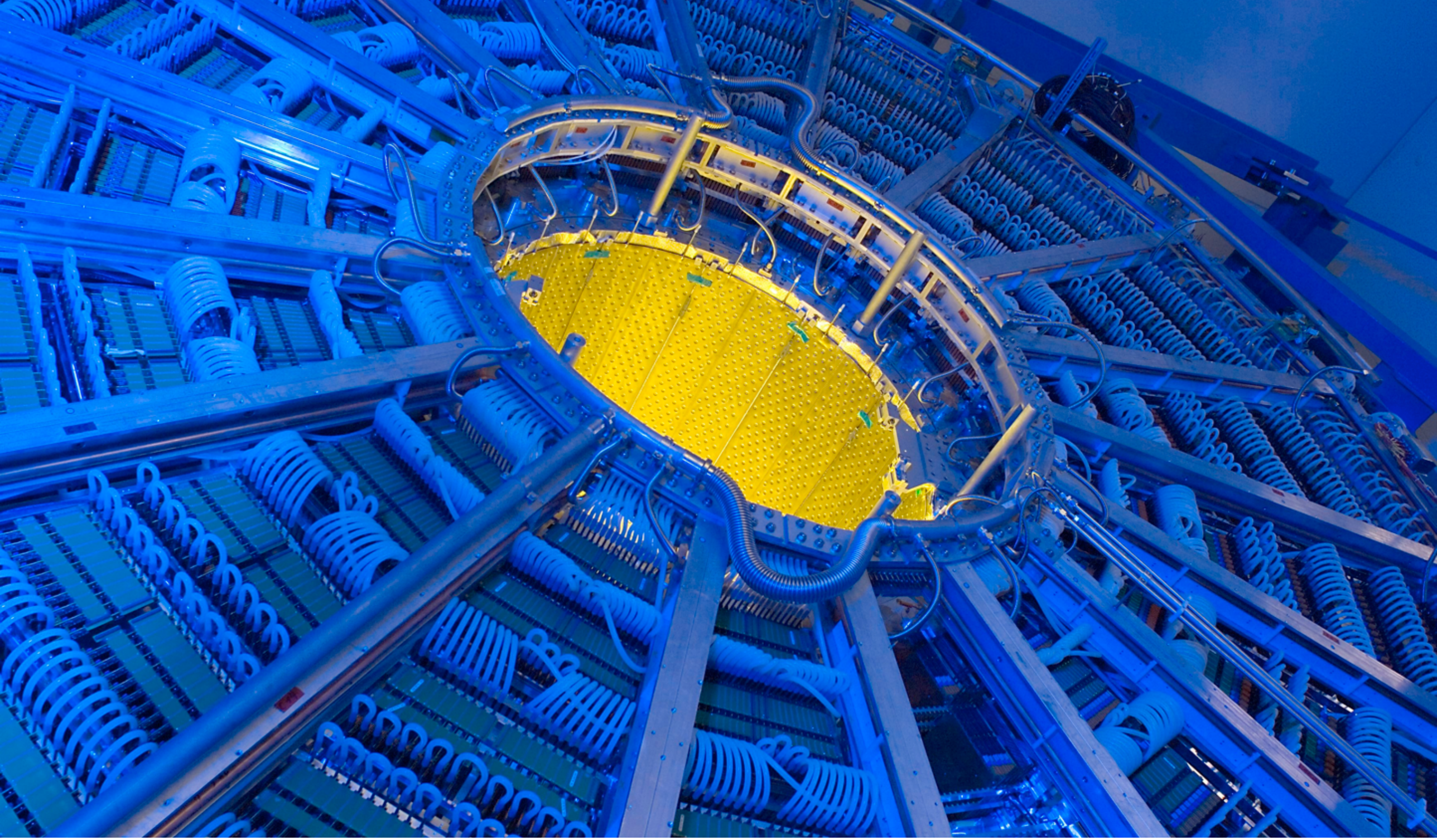


Data Science in ALICE

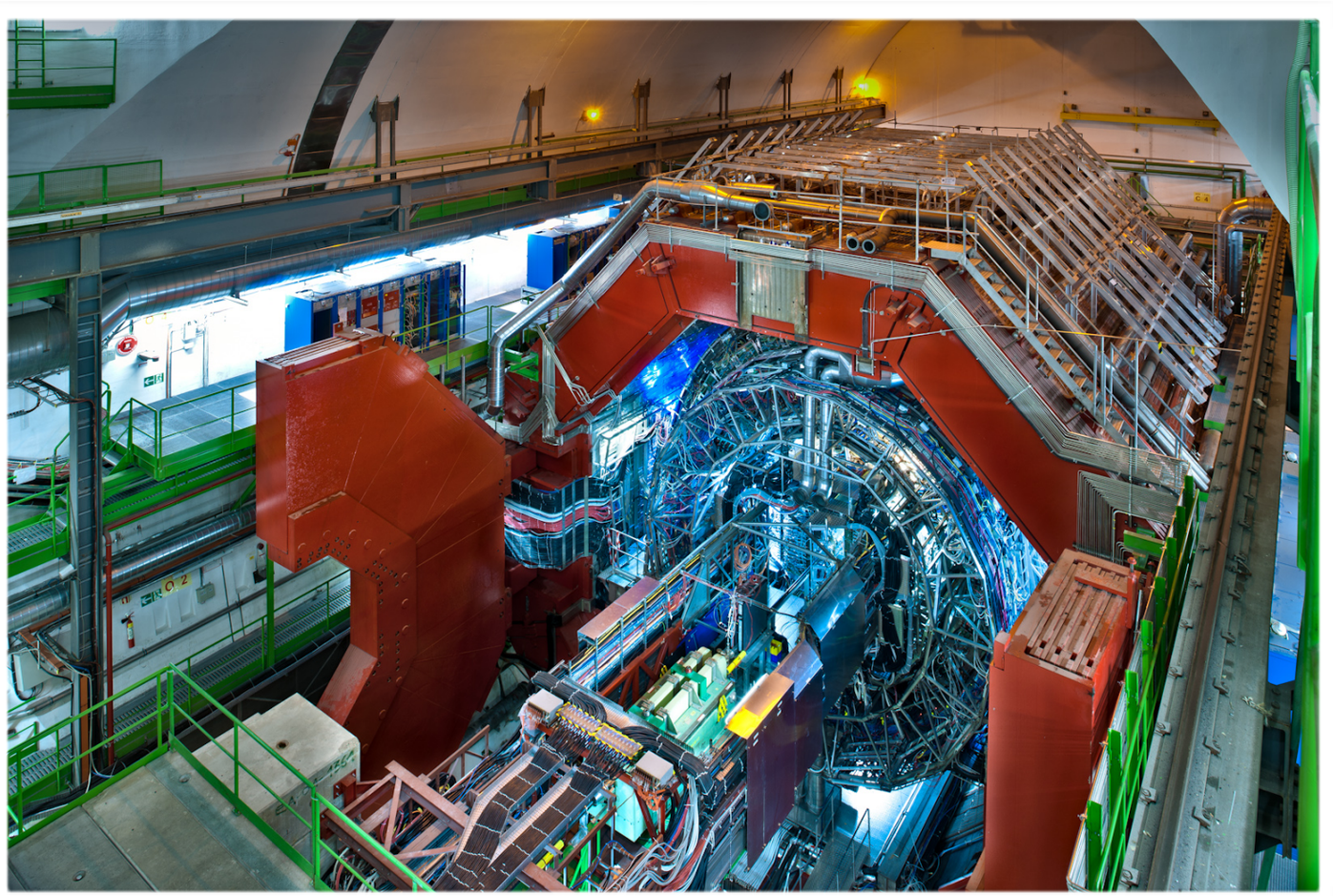
Michele Floris (CERN)
for the ALICE collaboration

Introduction

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

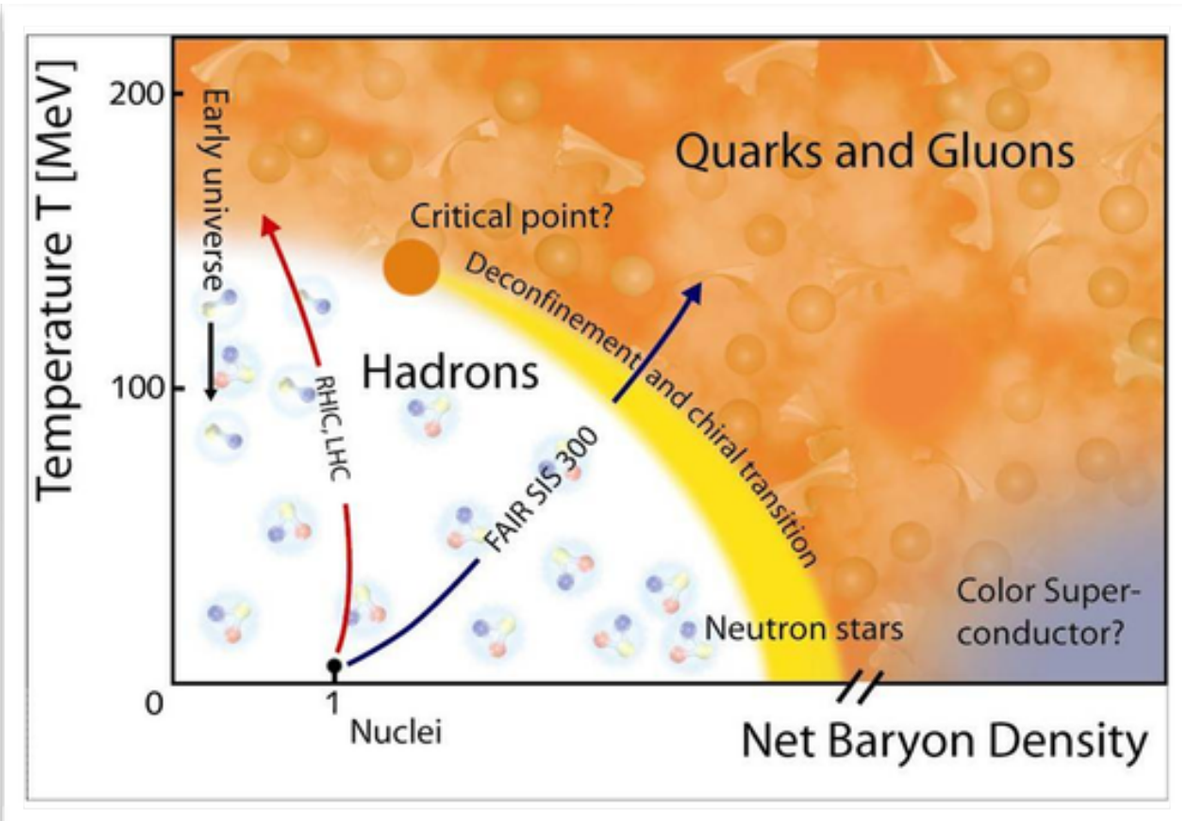
Outline

- 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



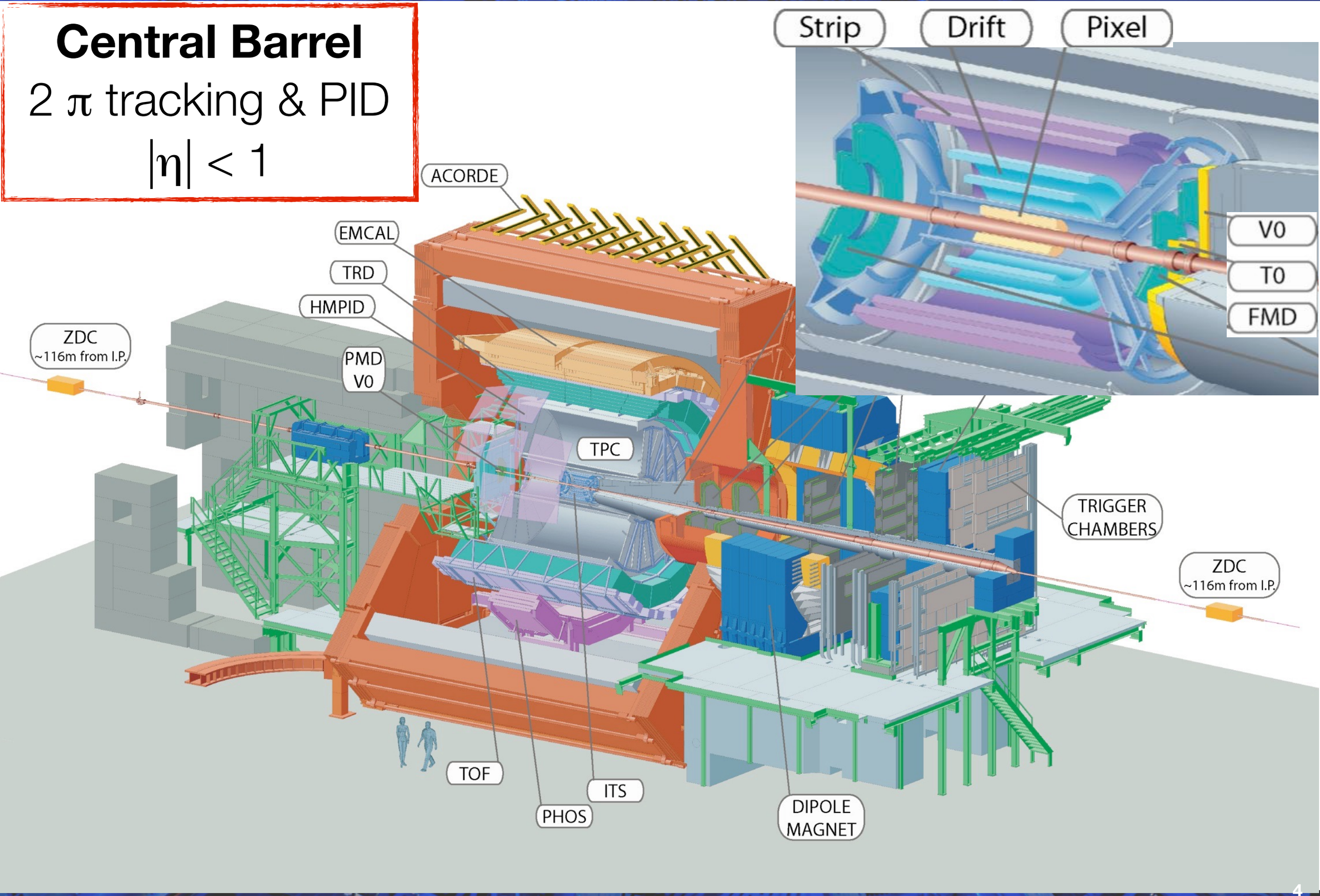
- Experimental needs: **low p_T** tracks, **particle identification** and **flavor taging**
  - Extensive particle identification over broad momentum range
  - Low p_T tracking ("bulk" particle production and low p_T heavy flavor)
- **Colliding systems**
  - Pb-Pb: "create" the QGP
  - p-Pb, pp: control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

The ALICE detector

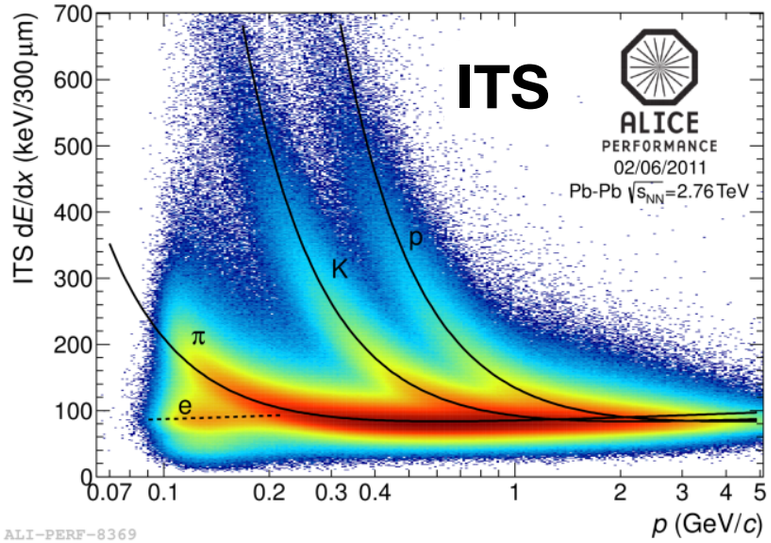
Central Barrel

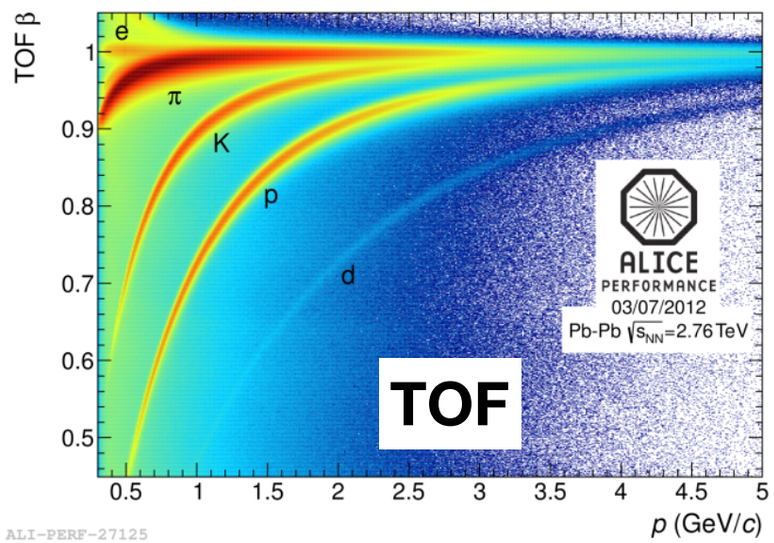
2 π tracking & PID

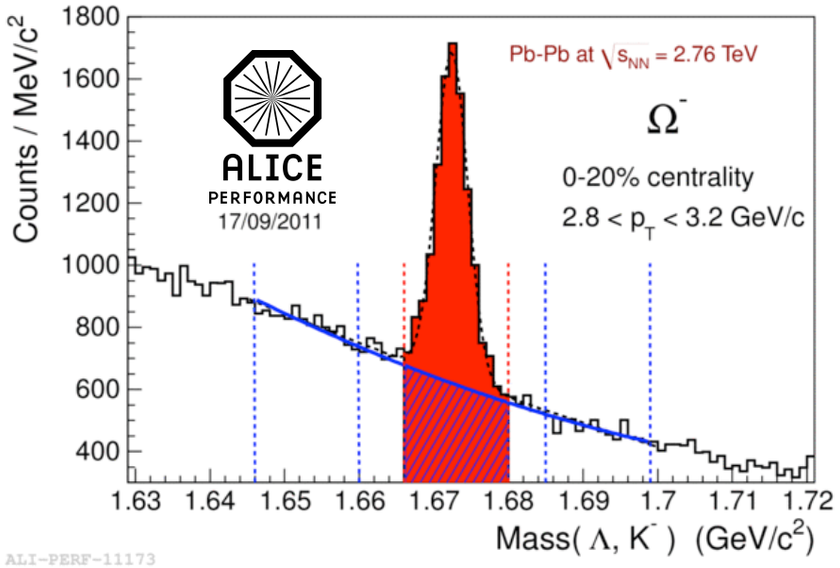
$|\eta| < 1$

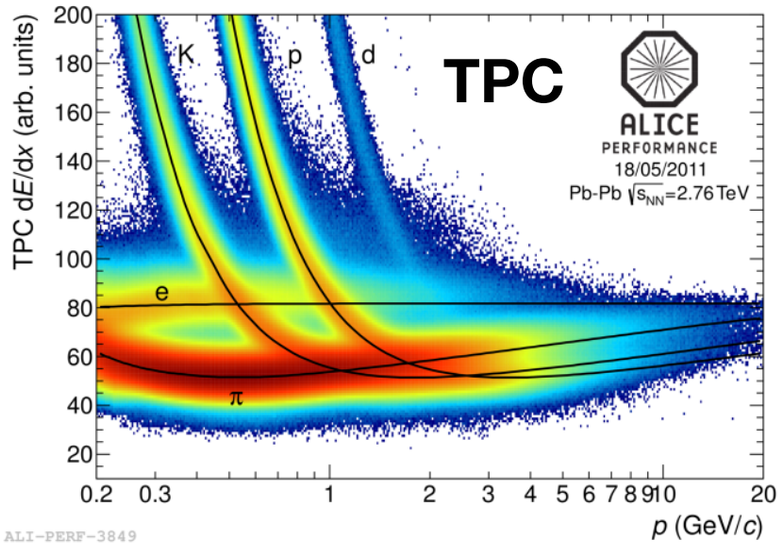


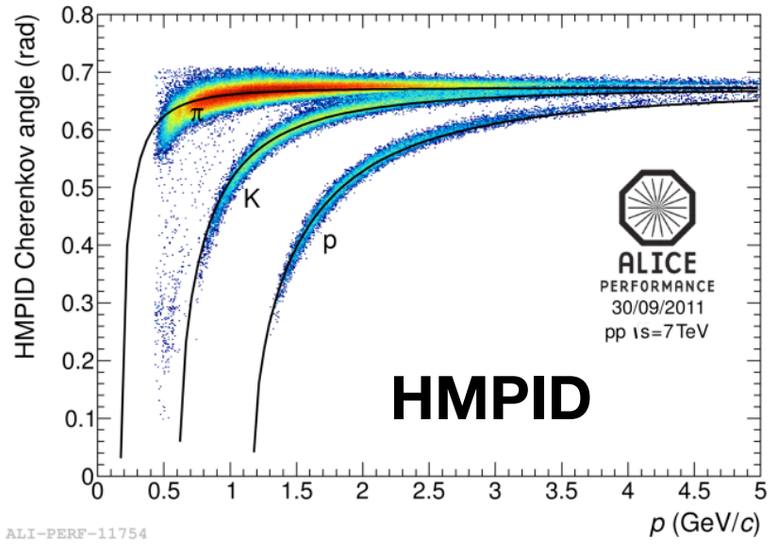
Tracking and Particle Identification

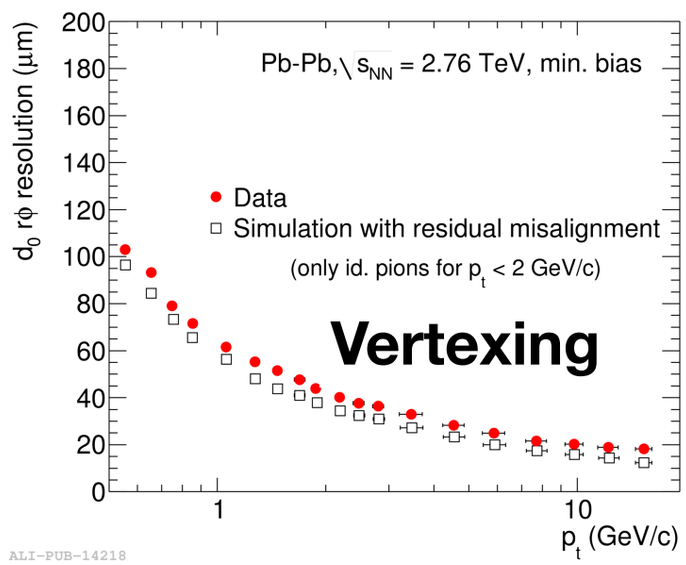












Particle identification (PID, many different techniques)

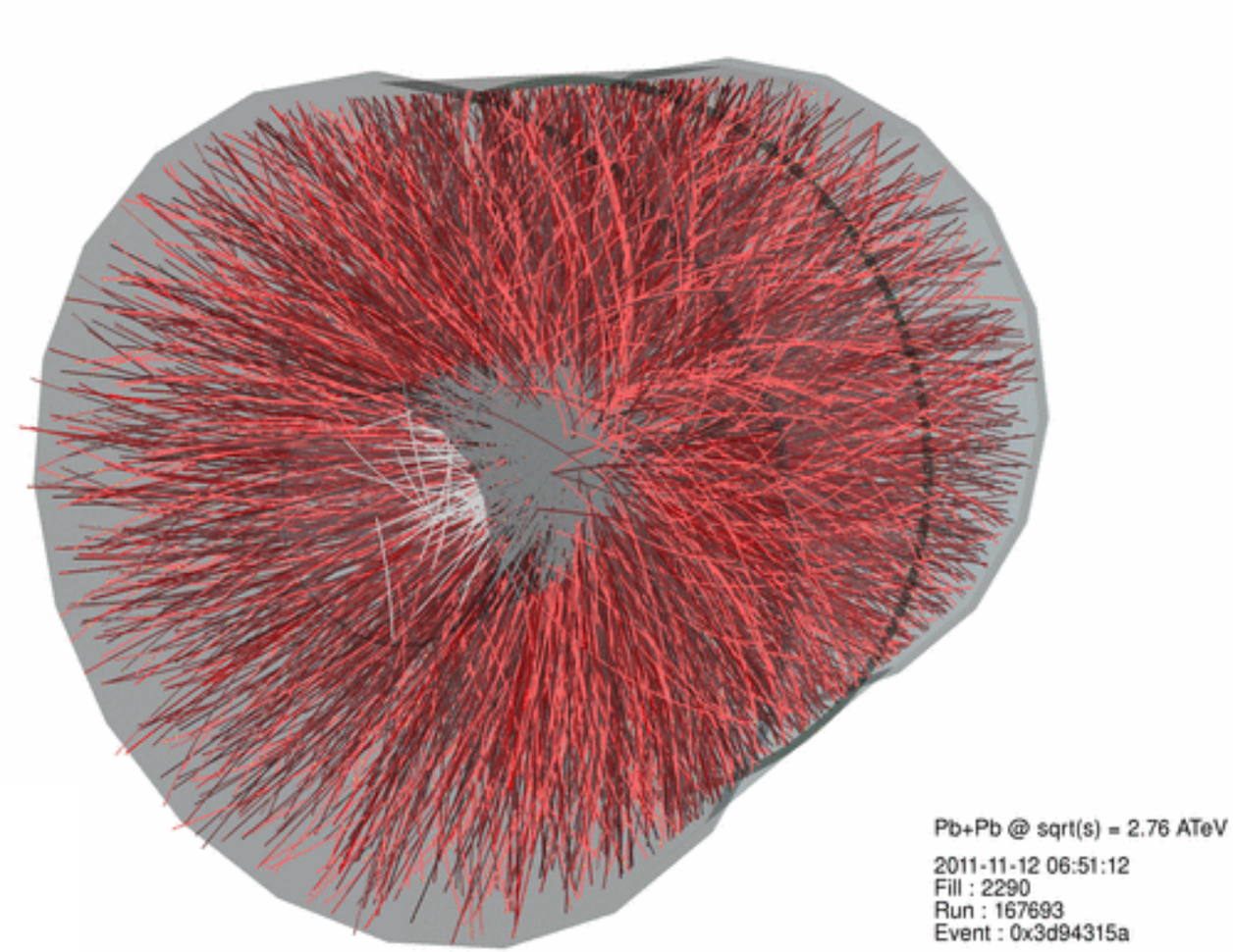
Extremely low-mass tracker ~ 10% of X_0

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: 7 PB of data so far, twice that in MC, 3 PB Pb-Pb 2015 expected

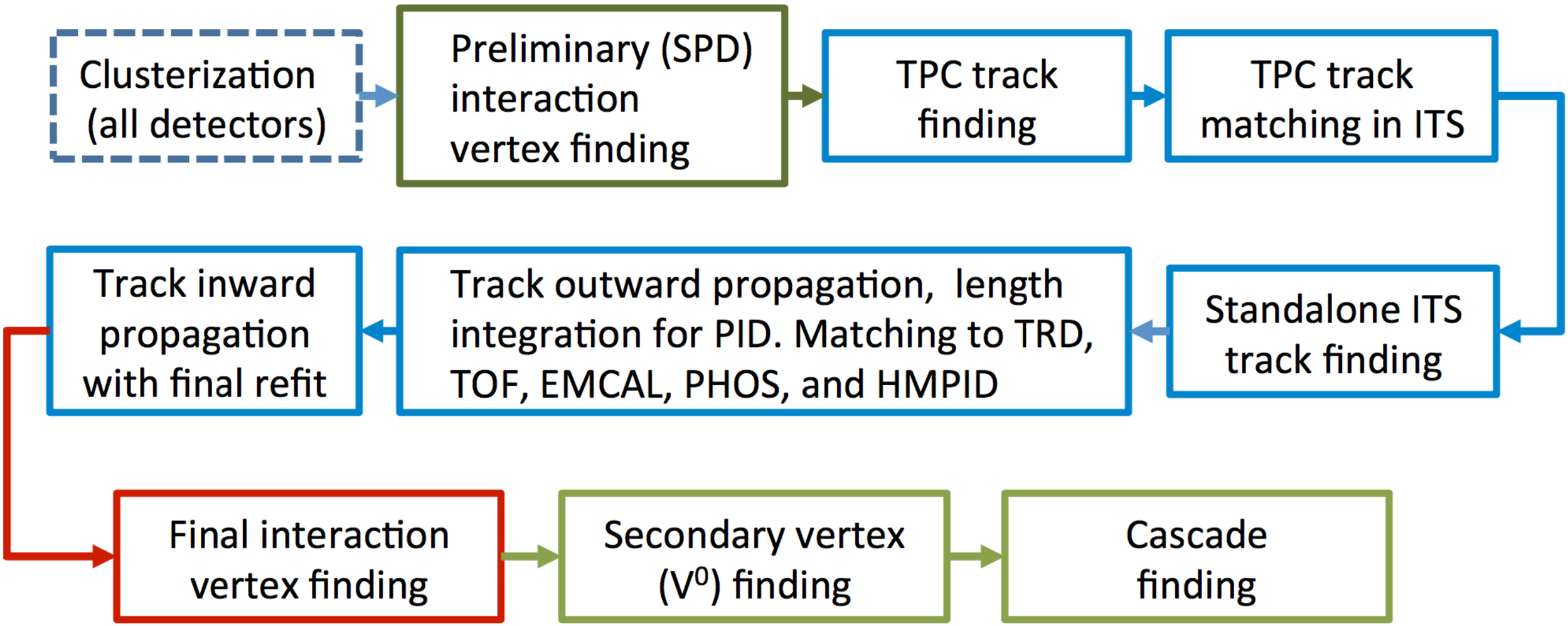
Combine **PID** in broad momentum region (0.1–20 GeV/*c*)

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

Areas of applicability of data science:
“Detector Reconstruction and Signals” &
“Signal optimization in physics analysis”

Detector: Track Reconstruction

Track reconstruction (offline)



Standard Kalman Filter

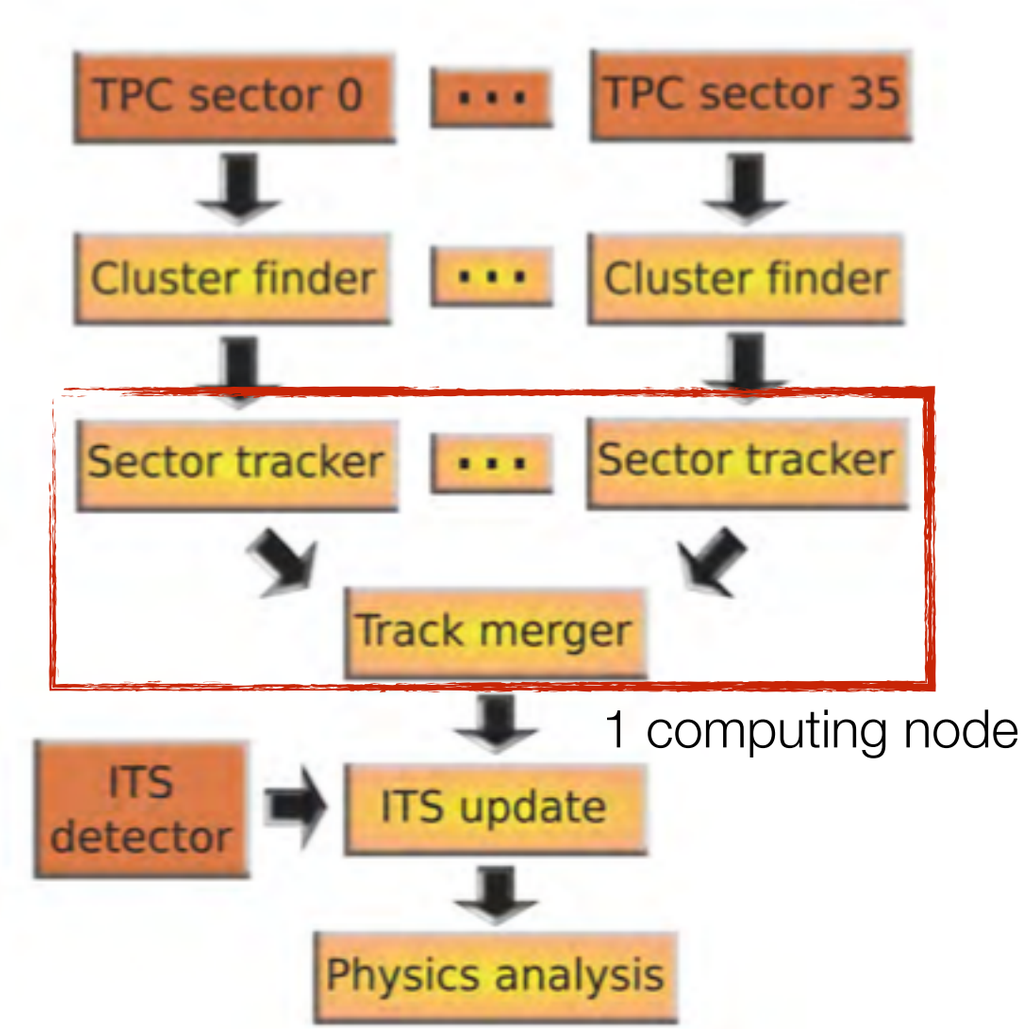
Inward-outward-inward procedure to reduce combinatorics

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

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

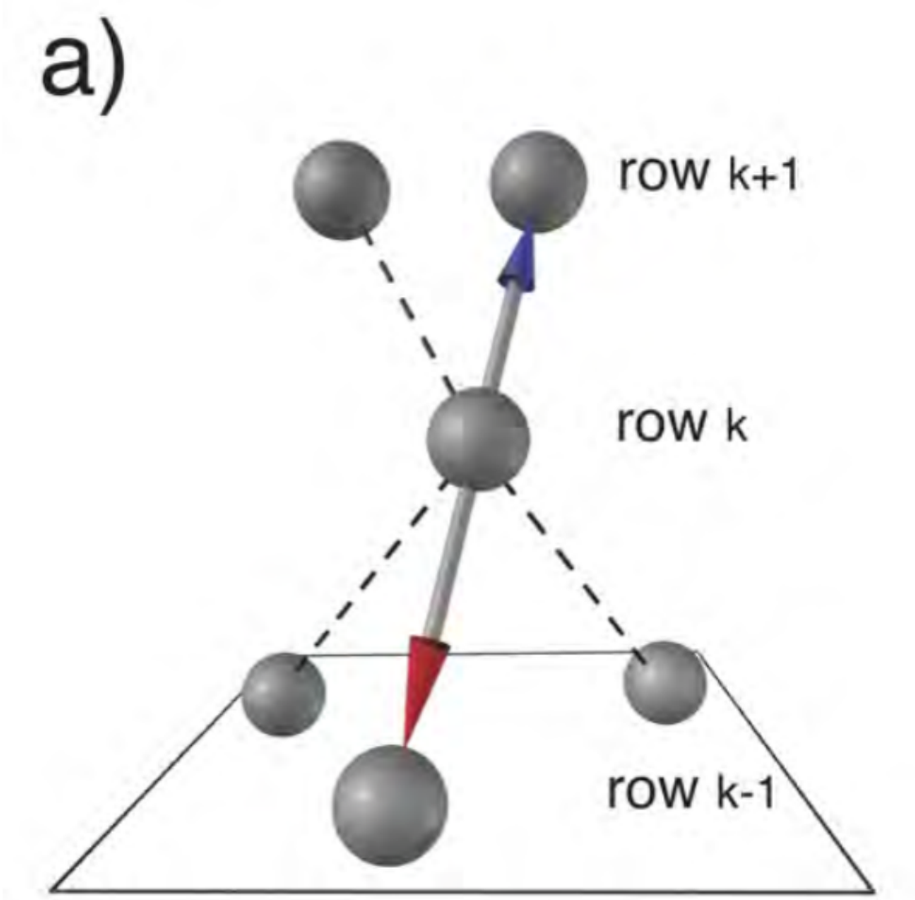
Track reconstruction in the HLT

- 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



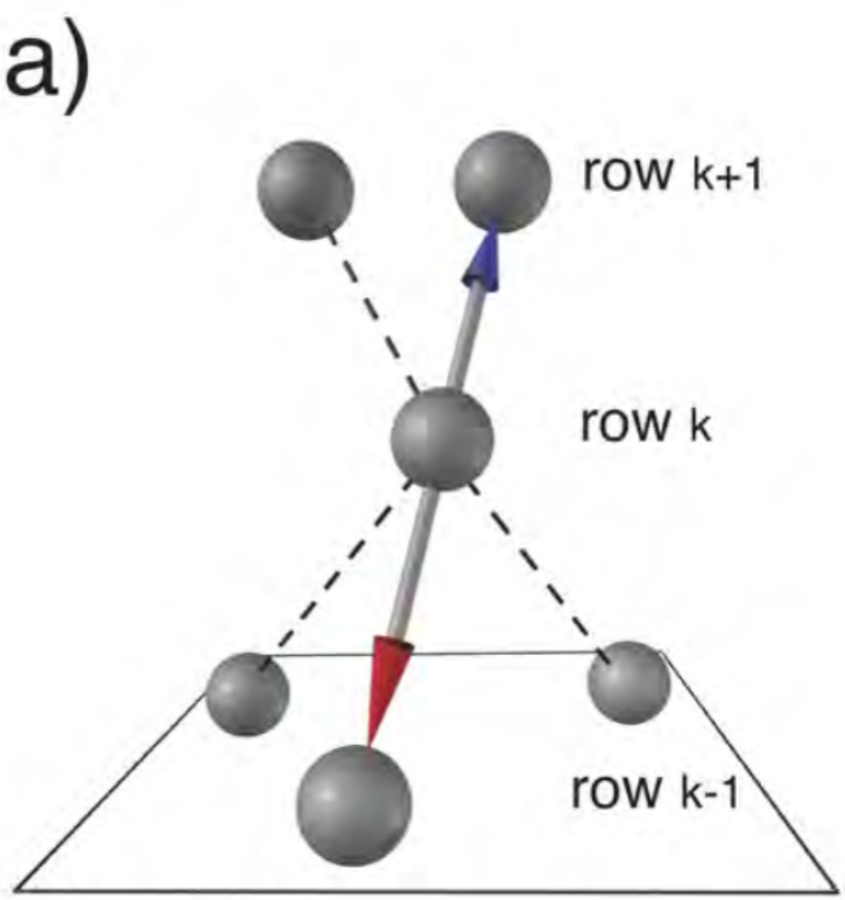
IEEE TNS, 58(4), 1845–1851, 10.1109/TNS.2011.2157702
CNNA 2012 proceedings, 10.1109/CNNA.2012.6331460

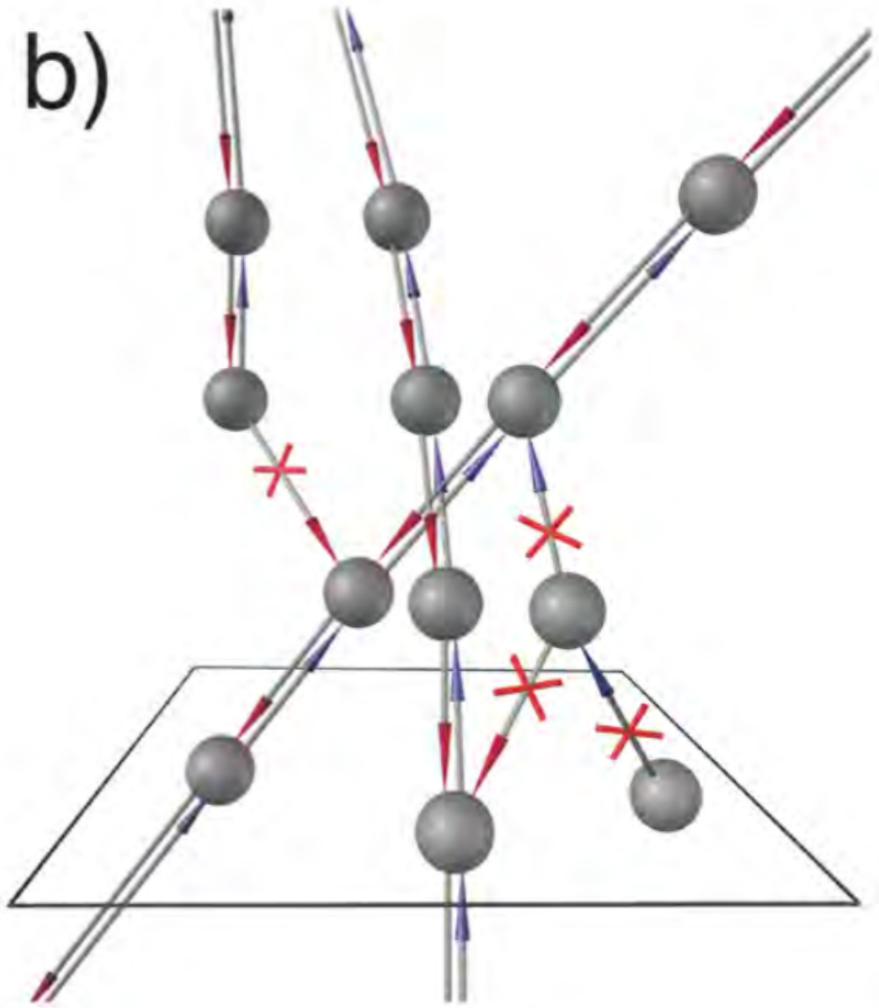
Cellular Automaton



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

Cellular Automaton

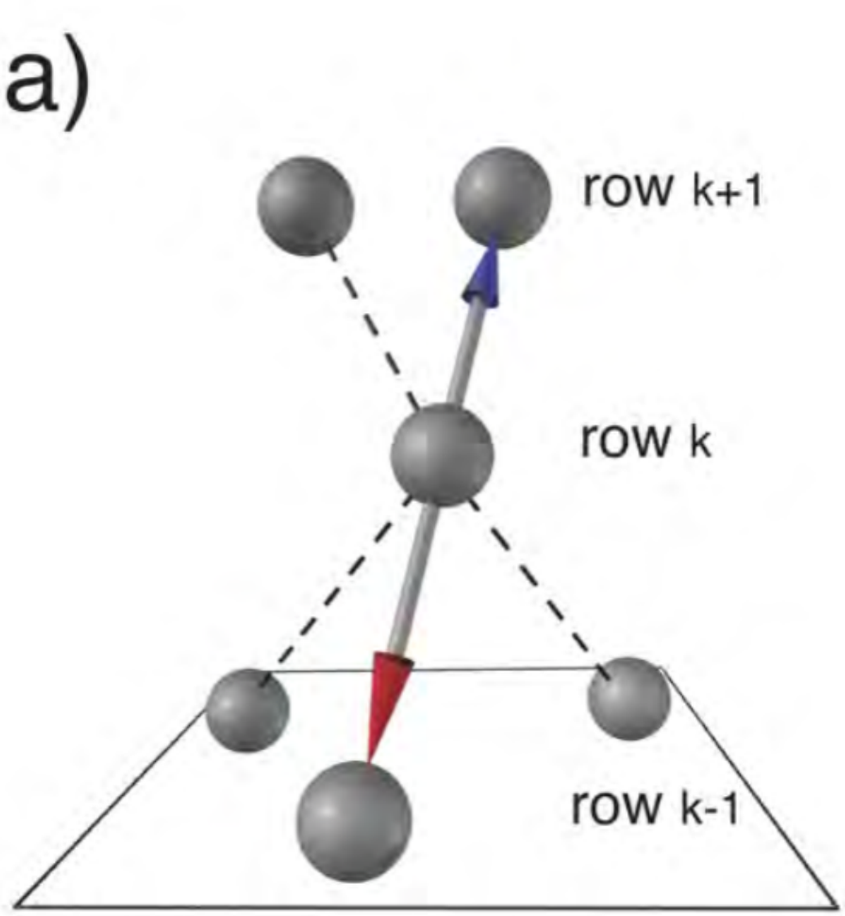


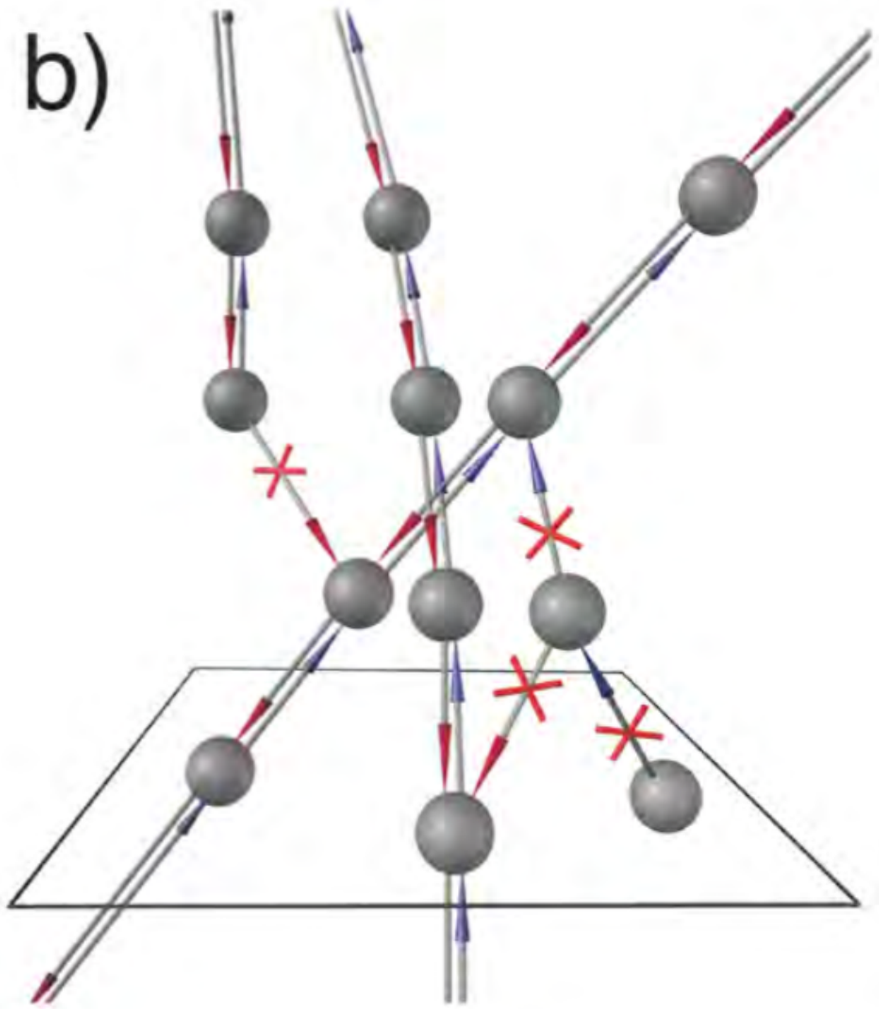


Neighbors finder:
segments of 3 clusters forming a straight line (“link”)

Evolution step:
Only reciprocal links are kept

Cellular Automaton



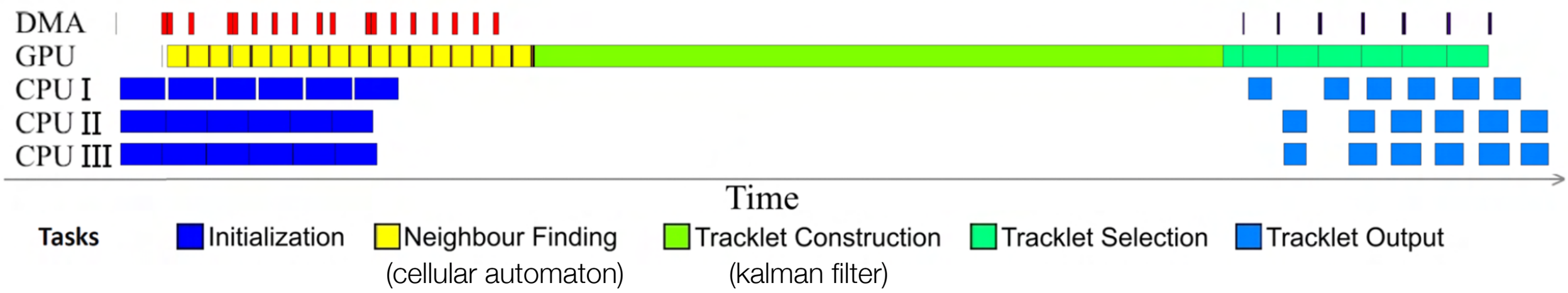


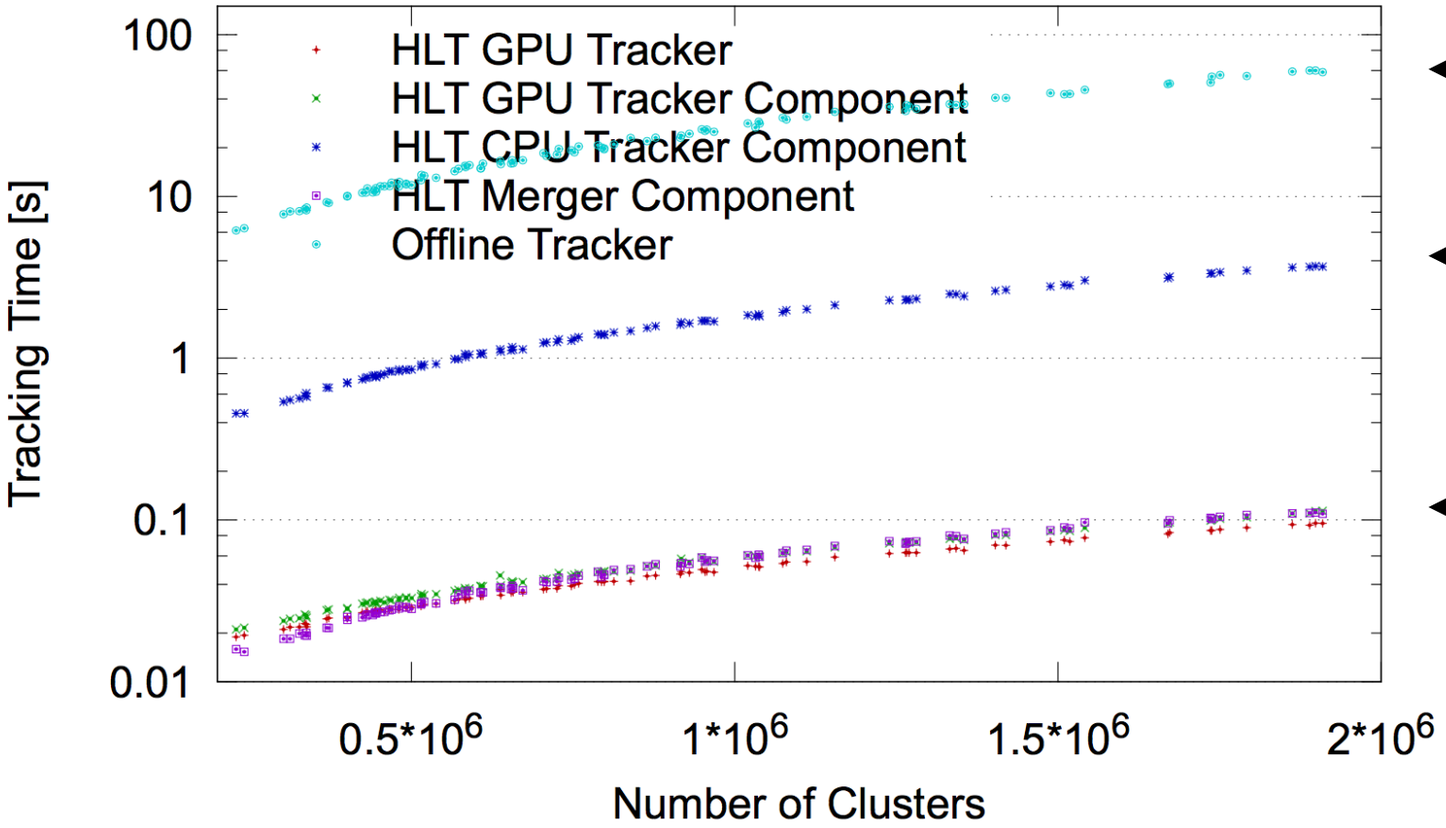
Neighbors finder:
segments of 3 clusters forming a straight line (“link”)

Evolution step:
Only reciprocal links are kept

Chain of links for the track candidates → Kalman Filter

Performance





← Offline tracker

← HLT tracker (CPU)

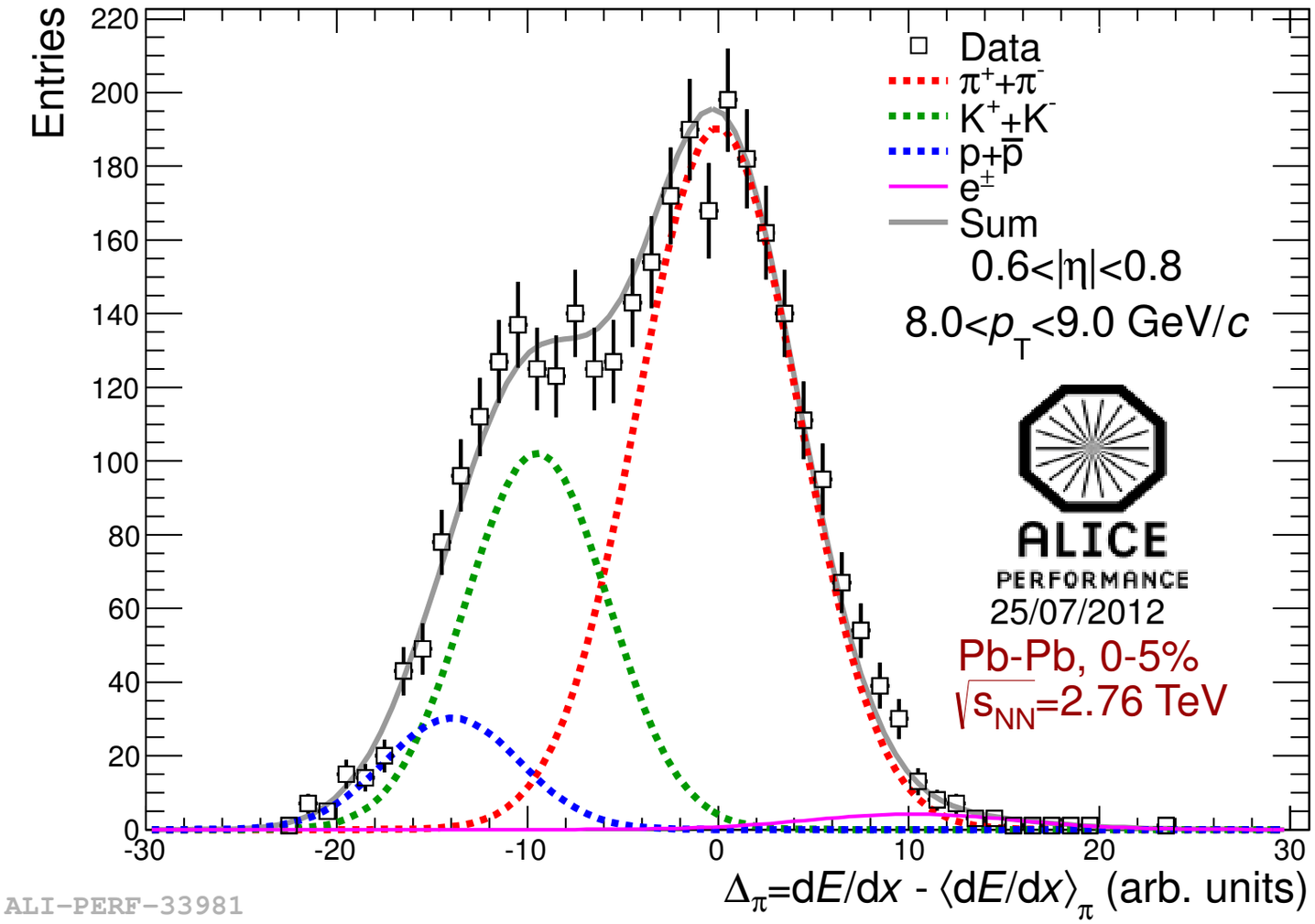
← HLT tracker (GPU)

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

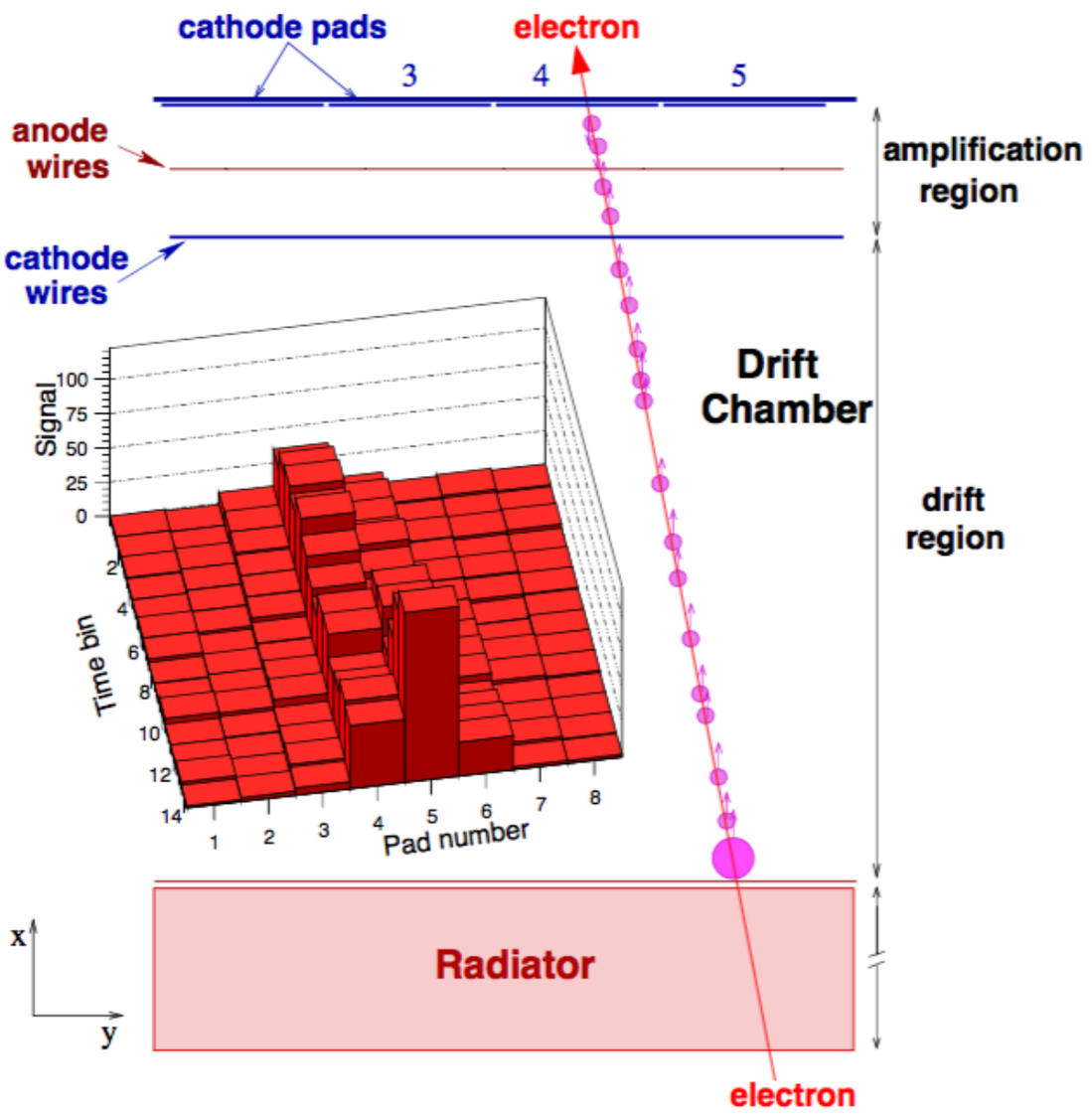
Detector: processing of (PID) signals

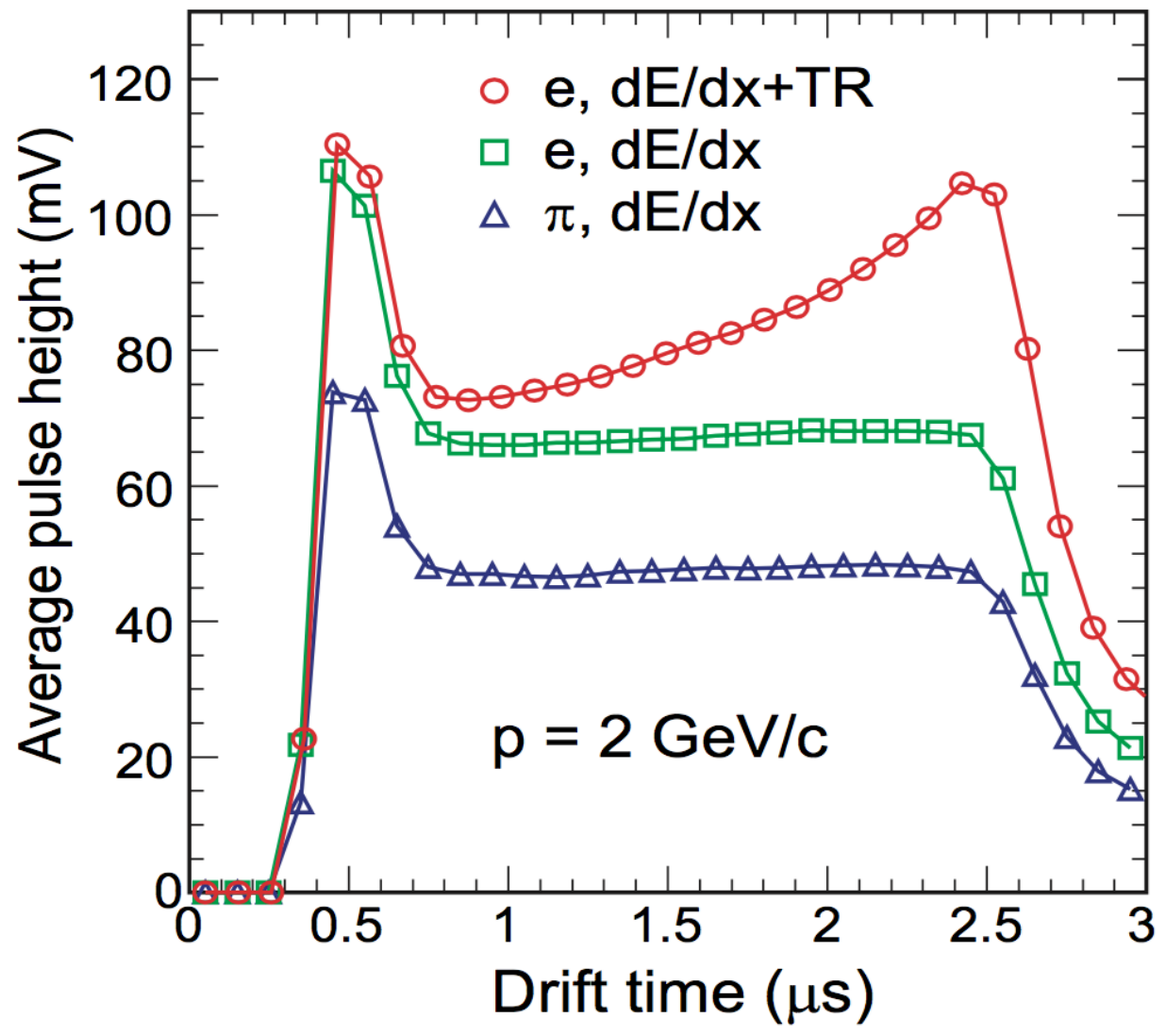
PID and heavy ions, analysis

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



Electron identification in the TRD



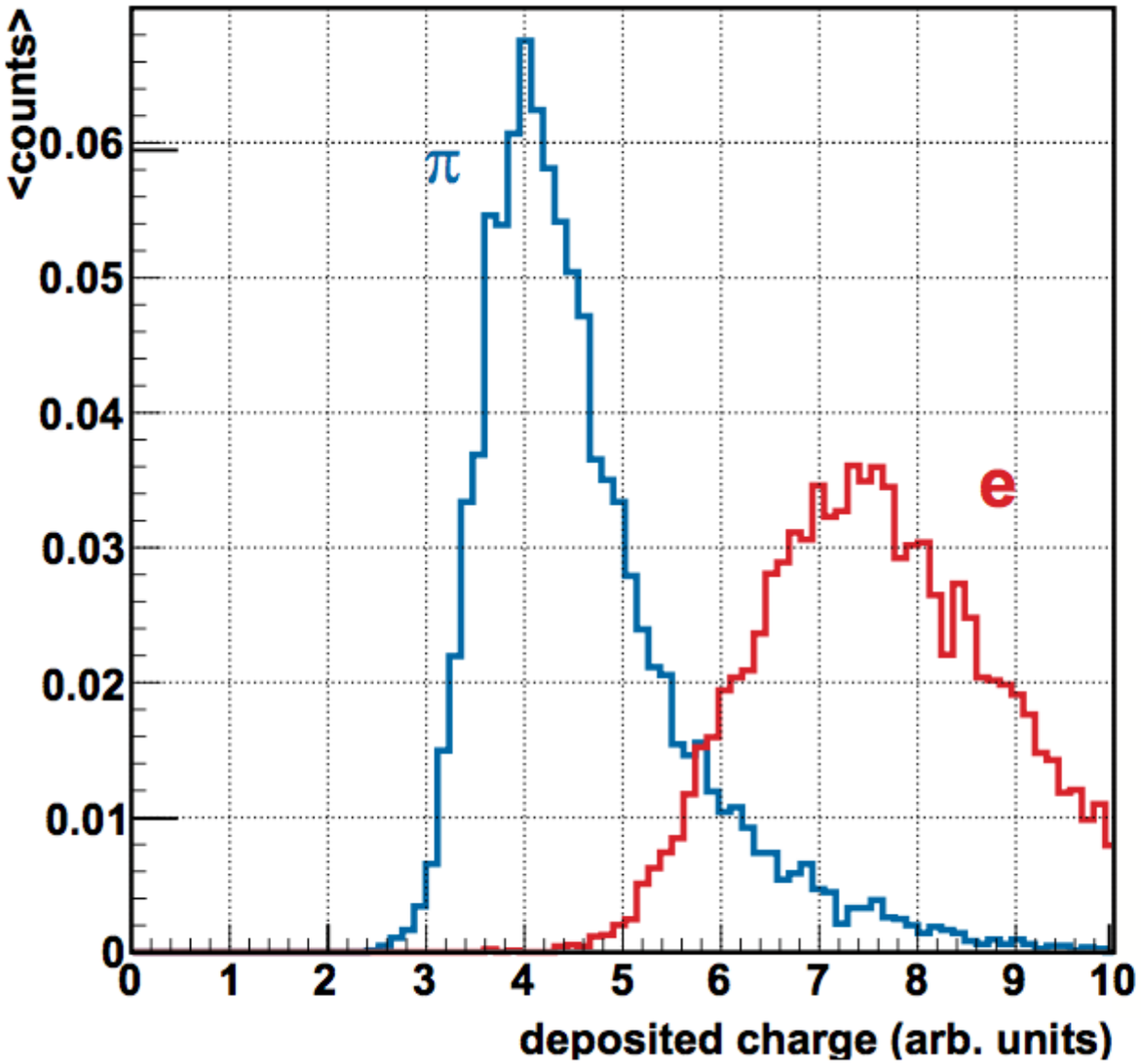


ALICE TRD: stack of 6 identical layers
Electrons: larger signal and different time dependence

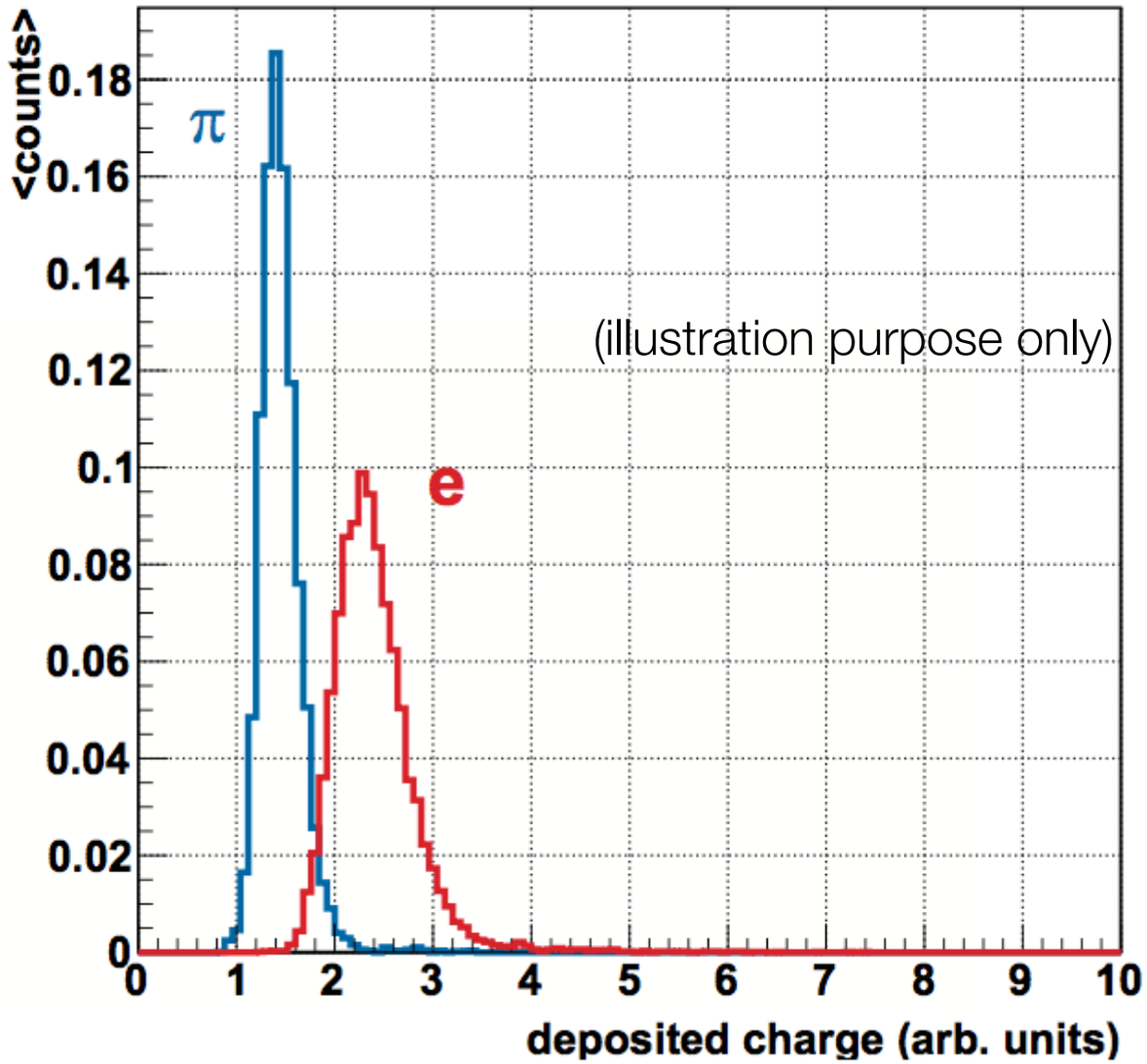
2008 JINST 3 S08002

Truncated mean

All clusters



Truncated mean (60%)



Tail at large deposited charge ⇒ contamination

Truncated mean (60% lowest charge clusters) significantly reduces tail and contamination

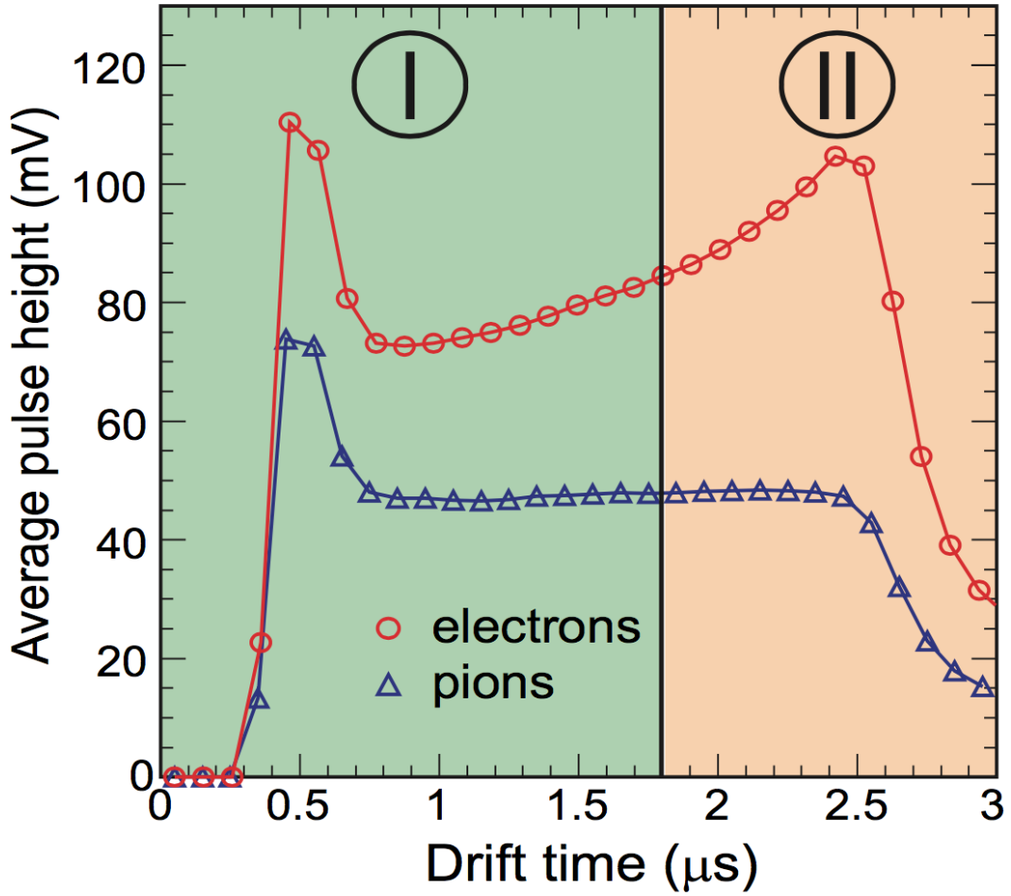
http://nbn-resolving.de/urn:nbn:de:hbz:6-97469411383

Likelihood and Neural Networks

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, …

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

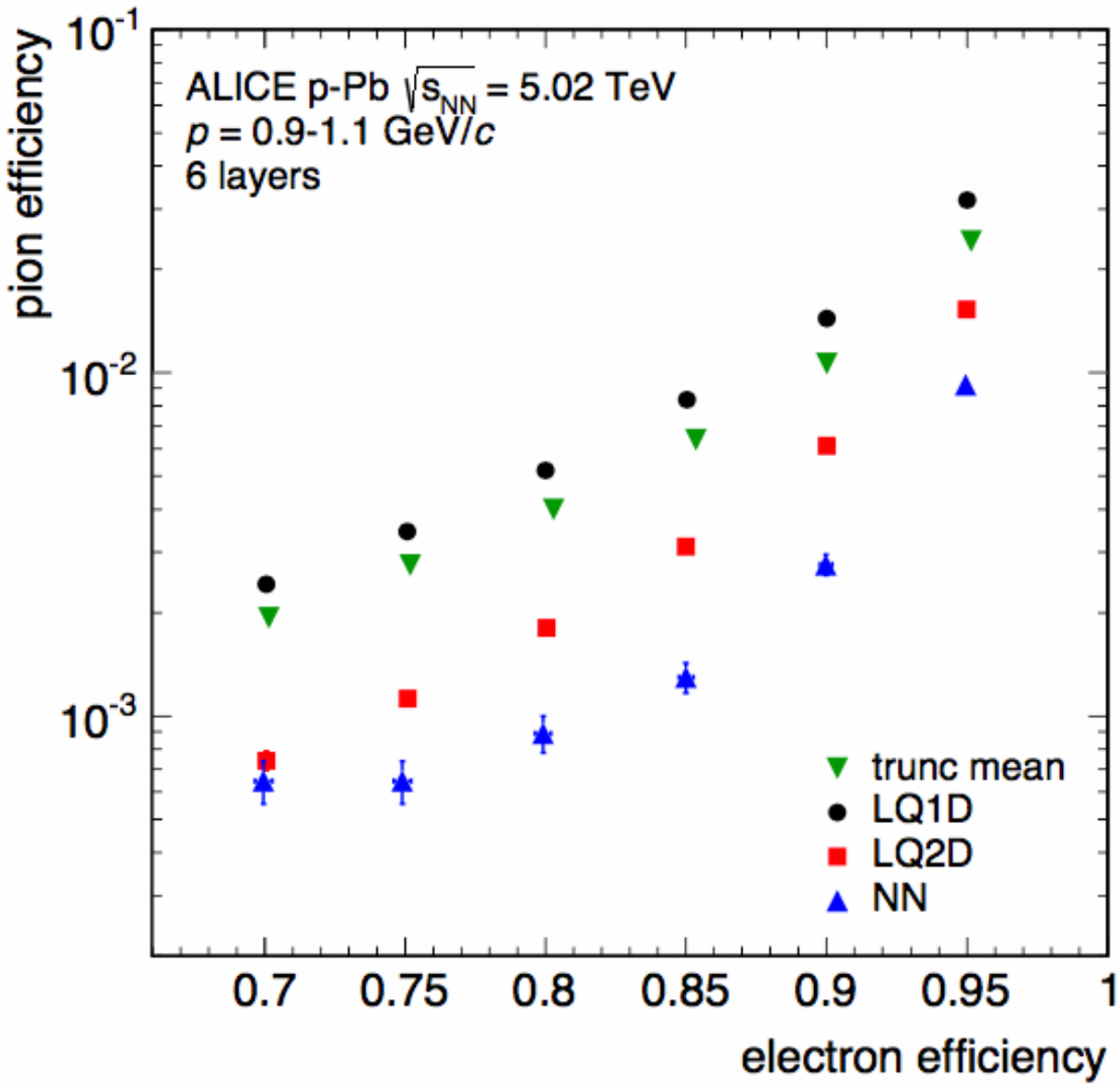


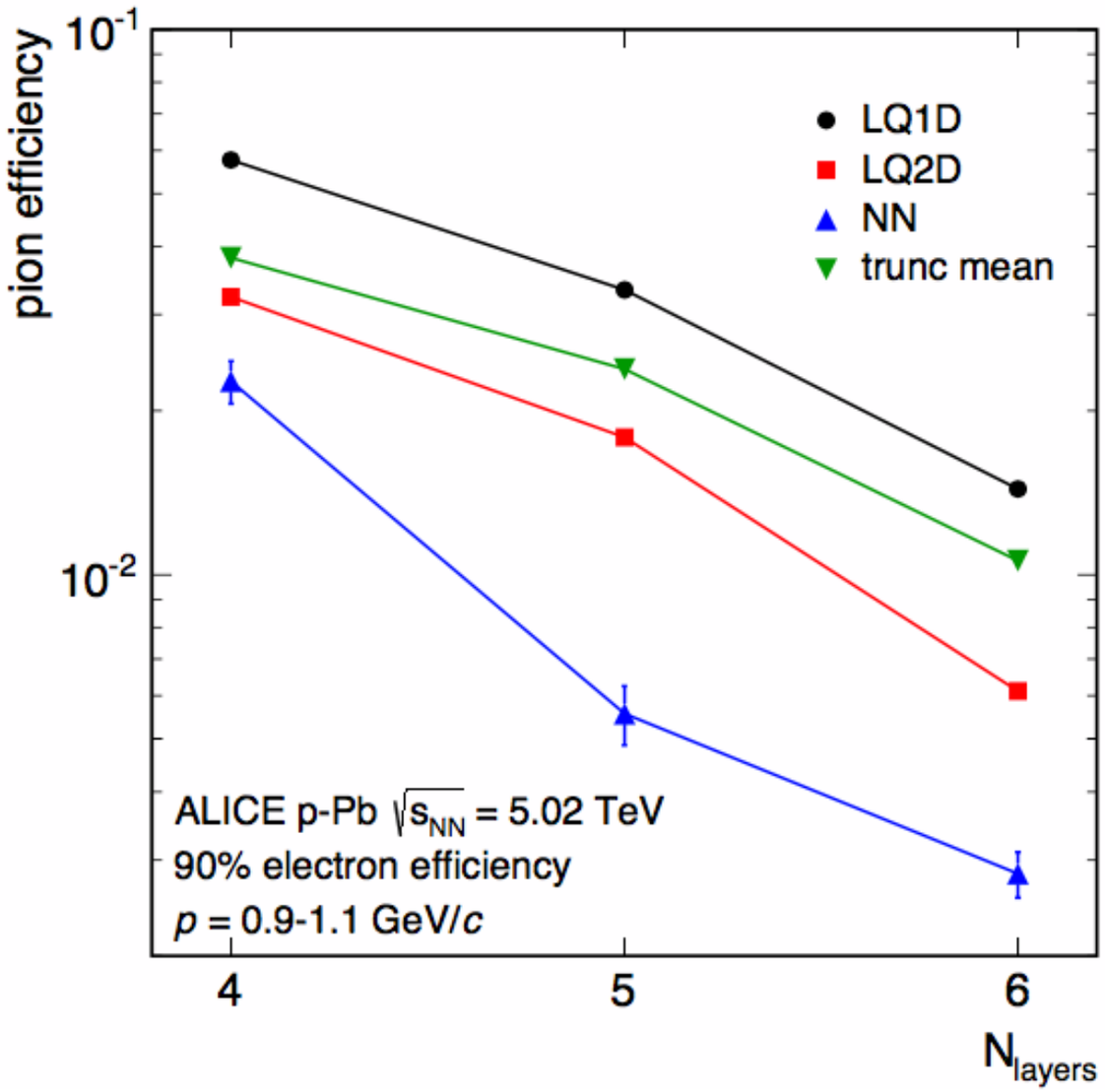
2D Likelihood: charged deposition in 2 time bins

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

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

Comparison of e/π distrimination methods



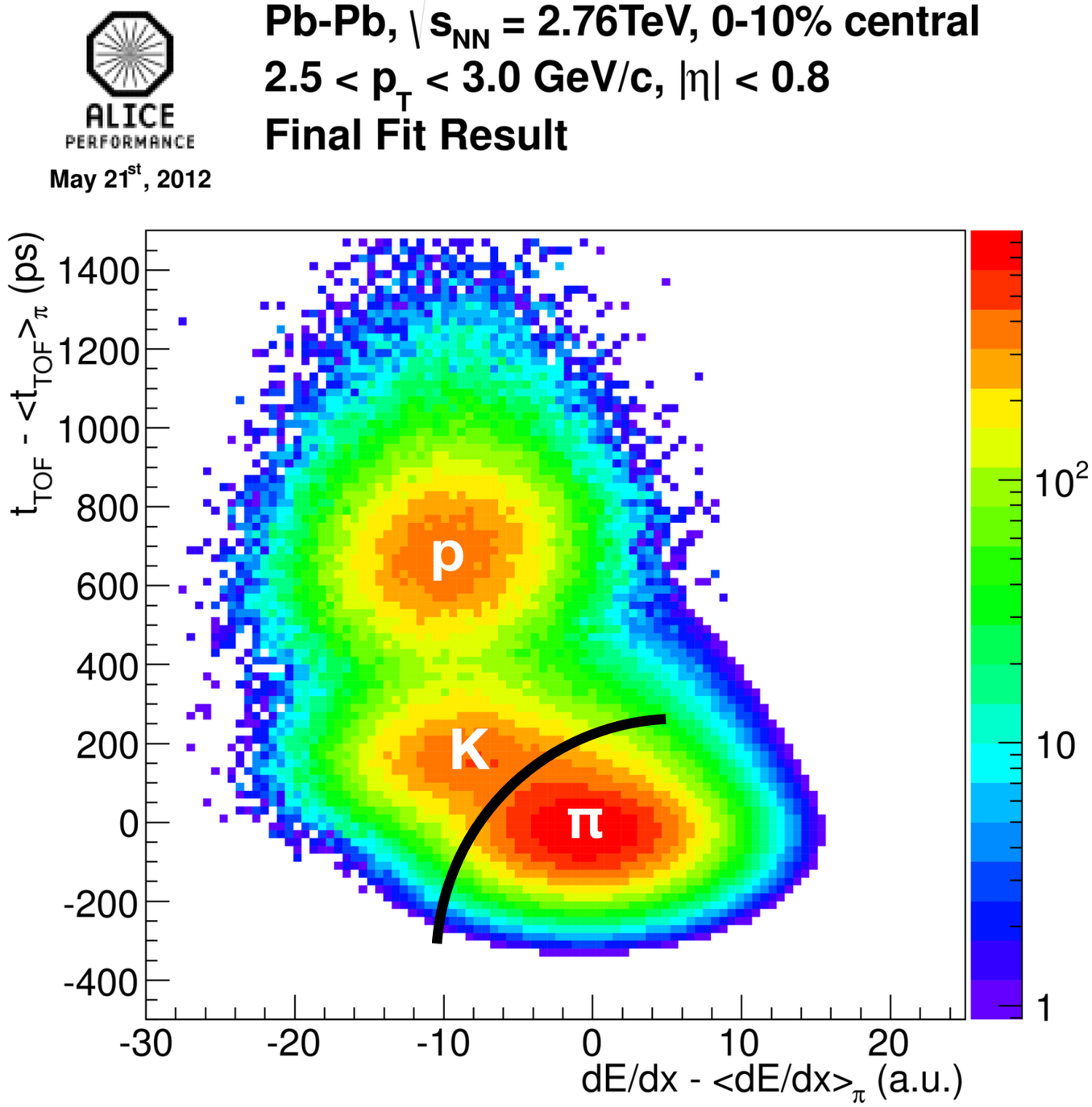


MLP works better, but uses more information.
7-dim likelihood performs as NN?
Next: include track properties

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



TPC+TOF

Bayesian PID

$$P(S|H_i) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\mathrm{n}_\sigma{}^2} = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(S-\hat{S}(H_i))^2}{2\sigma^2}}$$

Probability that a particle *i* produces a signal *S* in a given detector

$$P(\boldsymbol{S}|H_i) = \prod_{\alpha=\mathrm{ITS,TPC,...}} P_\alpha(S_\alpha|H_i)$$

Combined *P* of many detectors

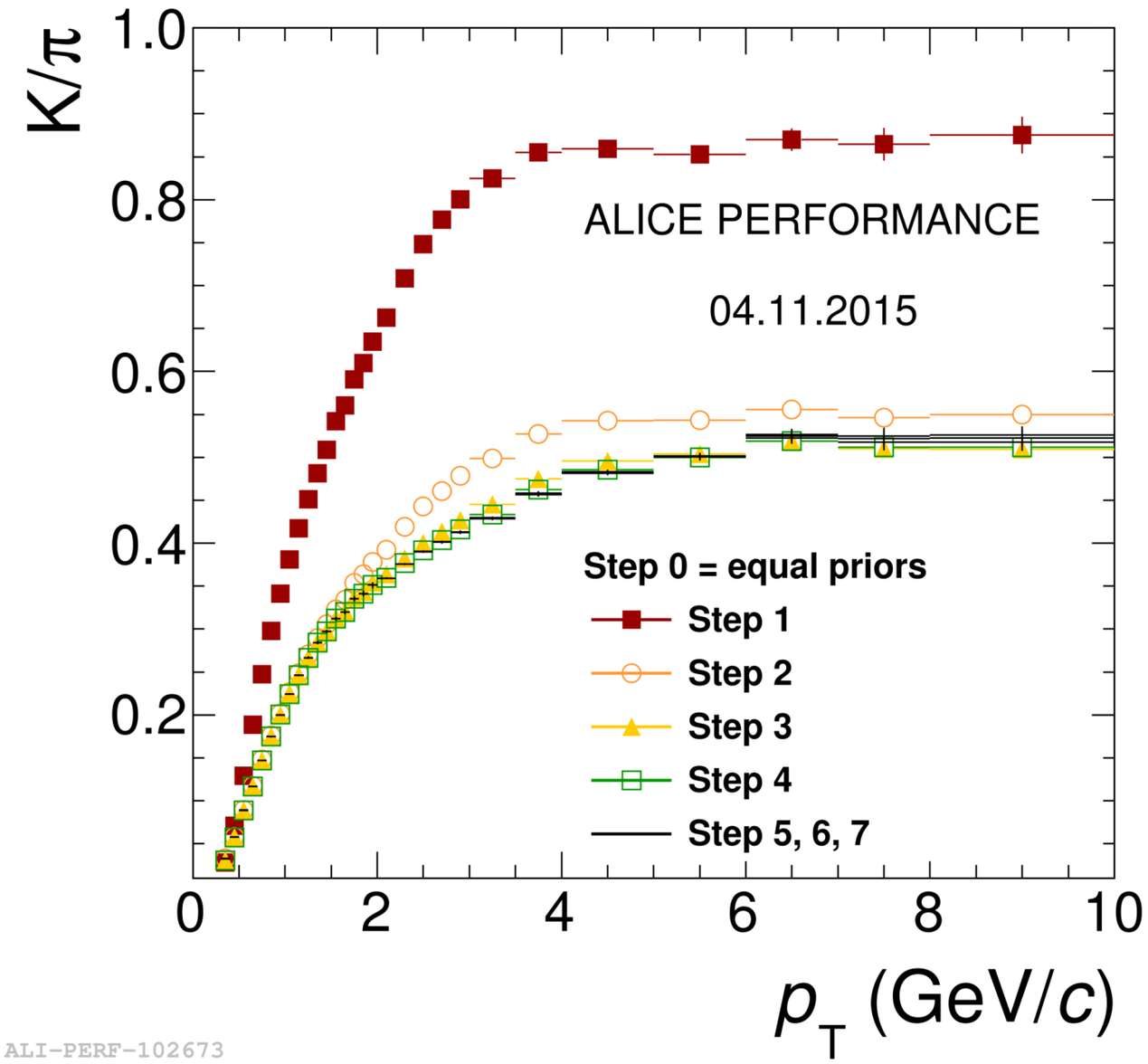
$$P(H_i|\boldsymbol{S}) = \frac{P(\boldsymbol{S}|H_i)C(H_i)}{\sum_{k=e,\mu,\pi,\ldots} P(\boldsymbol{S}|H_k)C(H_k)}.$$

Invert with Bayes theorem, needs prior

Subtle effects (e.g. mismatch in the TOF) can be easily incorporated

Probabilities can be used in **physics analysis** in various way:
Fixed threshold, maximum probability, weights

Iterative priors



ALI-PERF-102673

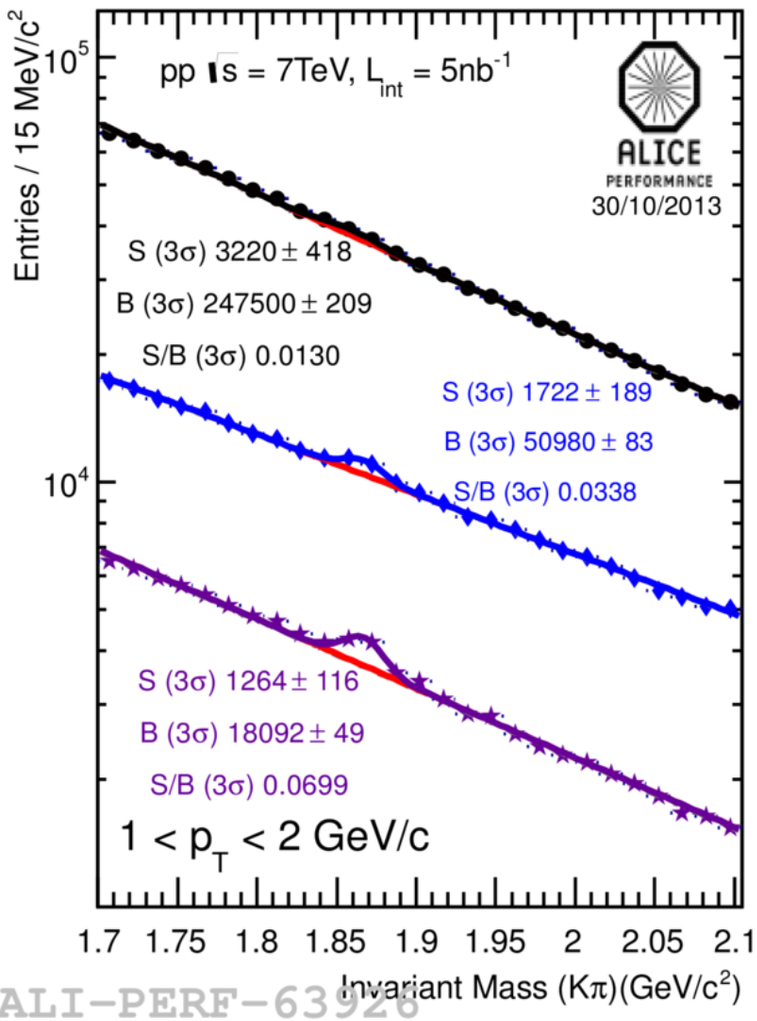
Priors can be determined **iteratively**

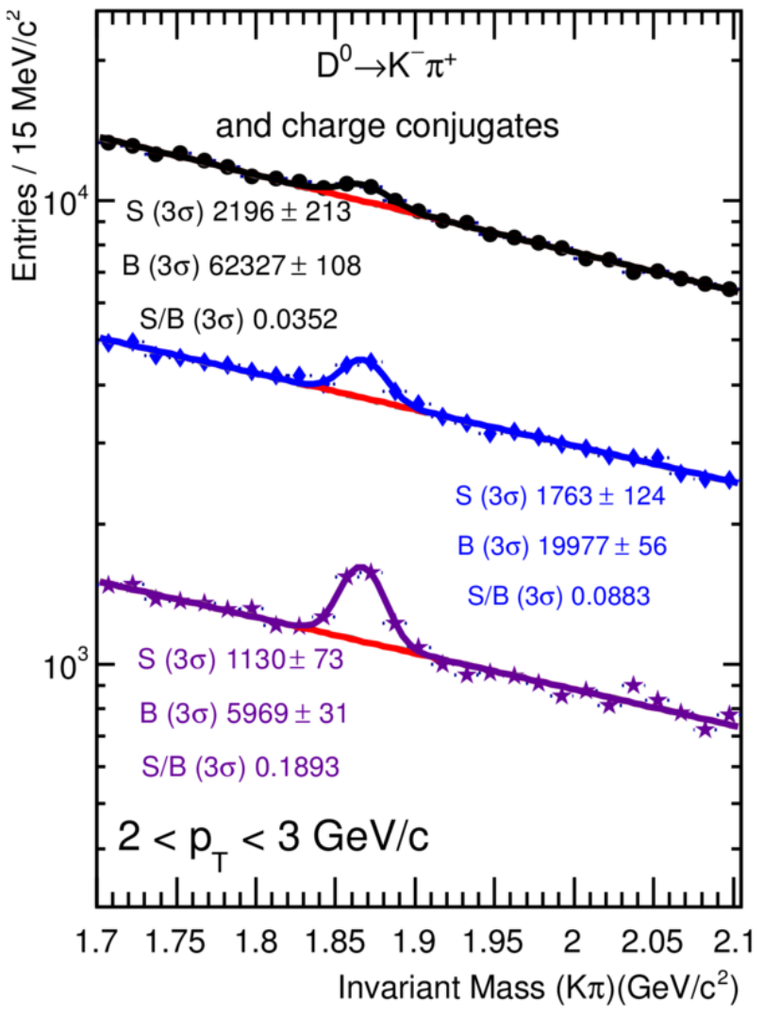
$$Y_{n+1}(H_i, p_T) = \sum_{all} P_n(H_i|S),$$

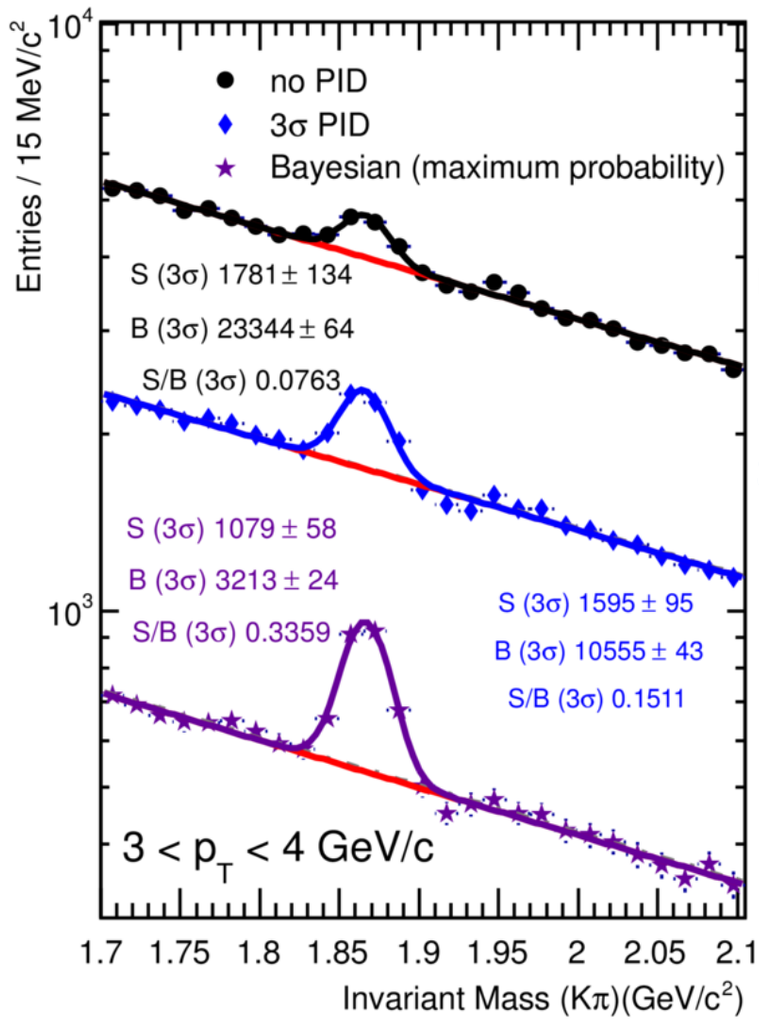
$$C_{n+1}(H_i, p_T) = \frac{Y_{n+1}(H_i, p_T)}{Y_{n+1}(H_\pi, p_T)}.$$

Rapid **convergence**, consistent with “unfolding” measurement
Exact value of **priors** not critical for **efficiency**, but important if not negligible **contamination**

Bayesian PID and the $D^0 \rightarrow K\pi$







no PID

3σ PID

Bayesian

Bayesian PID (maximum probability) improves significantly S/B in the study of the $D^0 \rightarrow K\pi$ decay

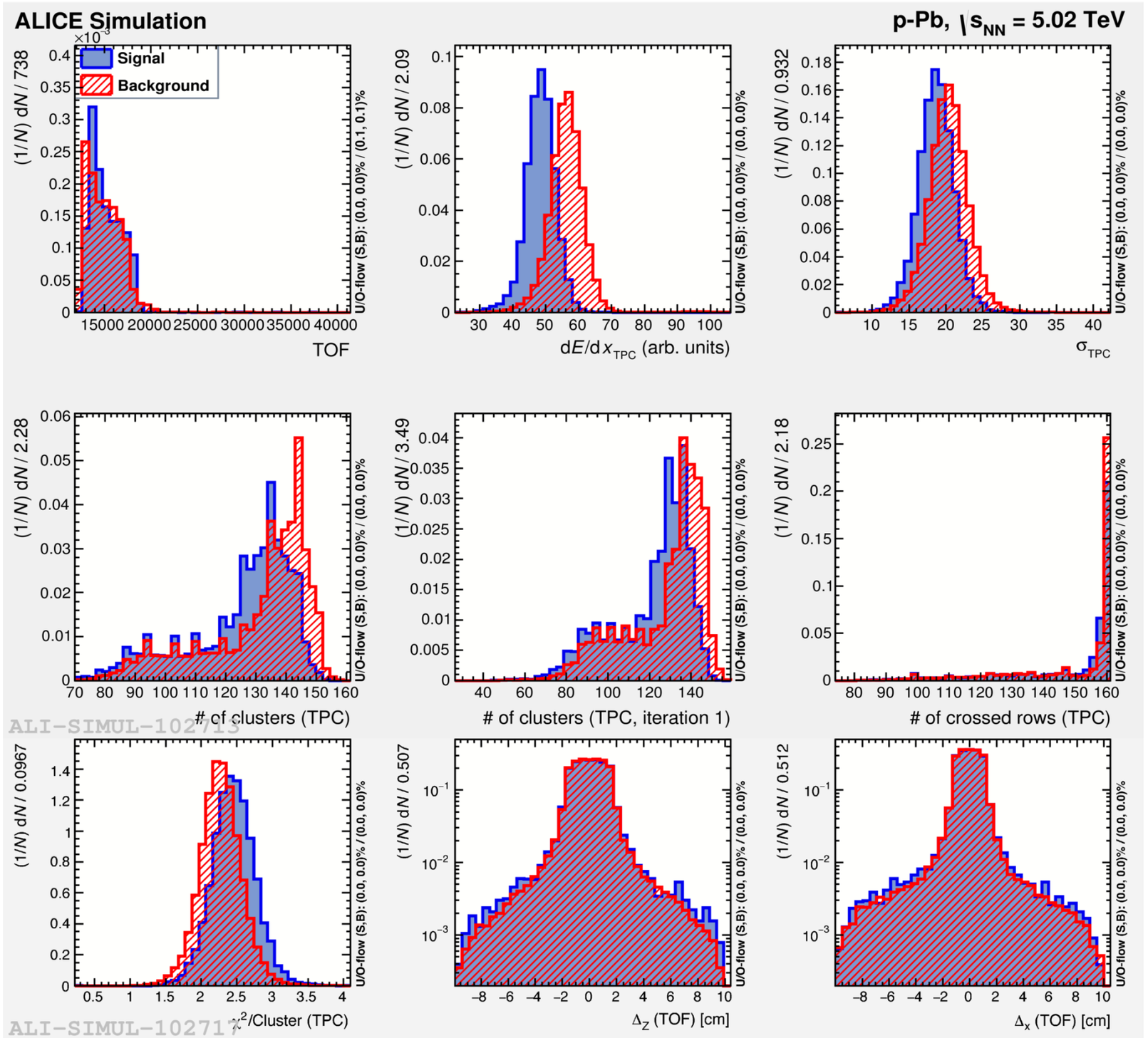
First attempts to TMVA PID

Monte Carlo, Signal = $K^{\pm}$, $1.6 < p < 1.8$ GeV/c

PID

Track Length

Track Quality

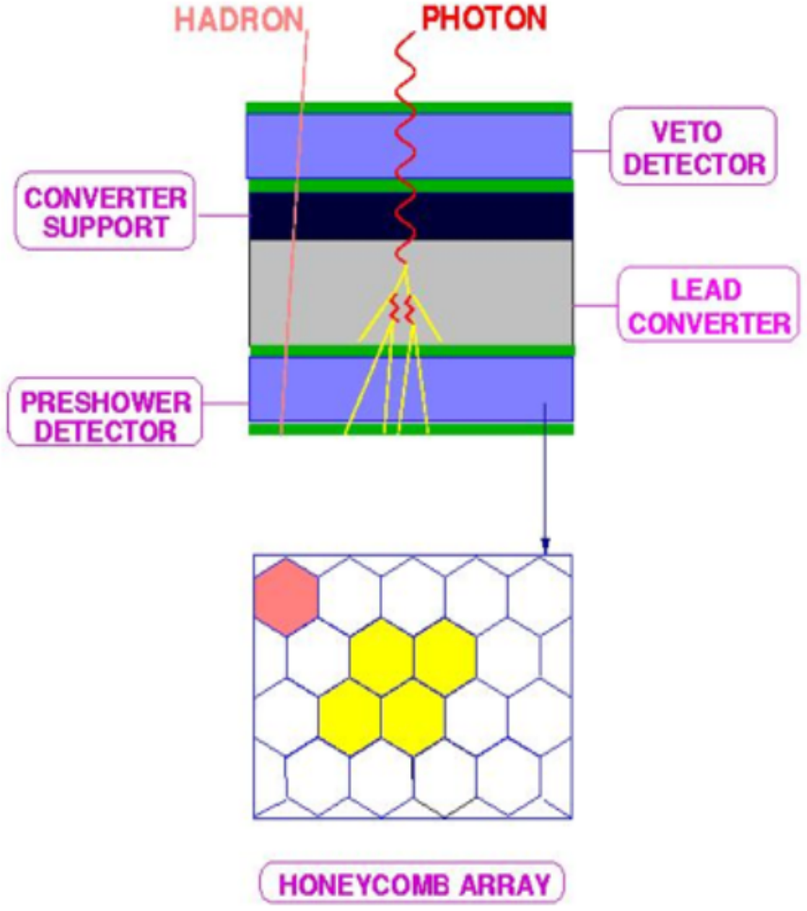


Idea: combine PID signals and info related to PID signals with MVA

KNN, MLP, BDT tested, early results not conclusive

γ/h and e/h Discrimination

- **Photon/Hadron** and **Electron/Hadron** cluster discrimination: promising area for the application of multivariate methods
- **Current analysis** based on:
  - E/p, matching to tracks and shape variables (EMCAL/PHOS)
  - Energy deposition and cluster size (PMD)
- **MVA can improve purity**, early attempts ongoing



Photon Multiplicity Detector (PMD)
$2.3 < \eta < 3.7$

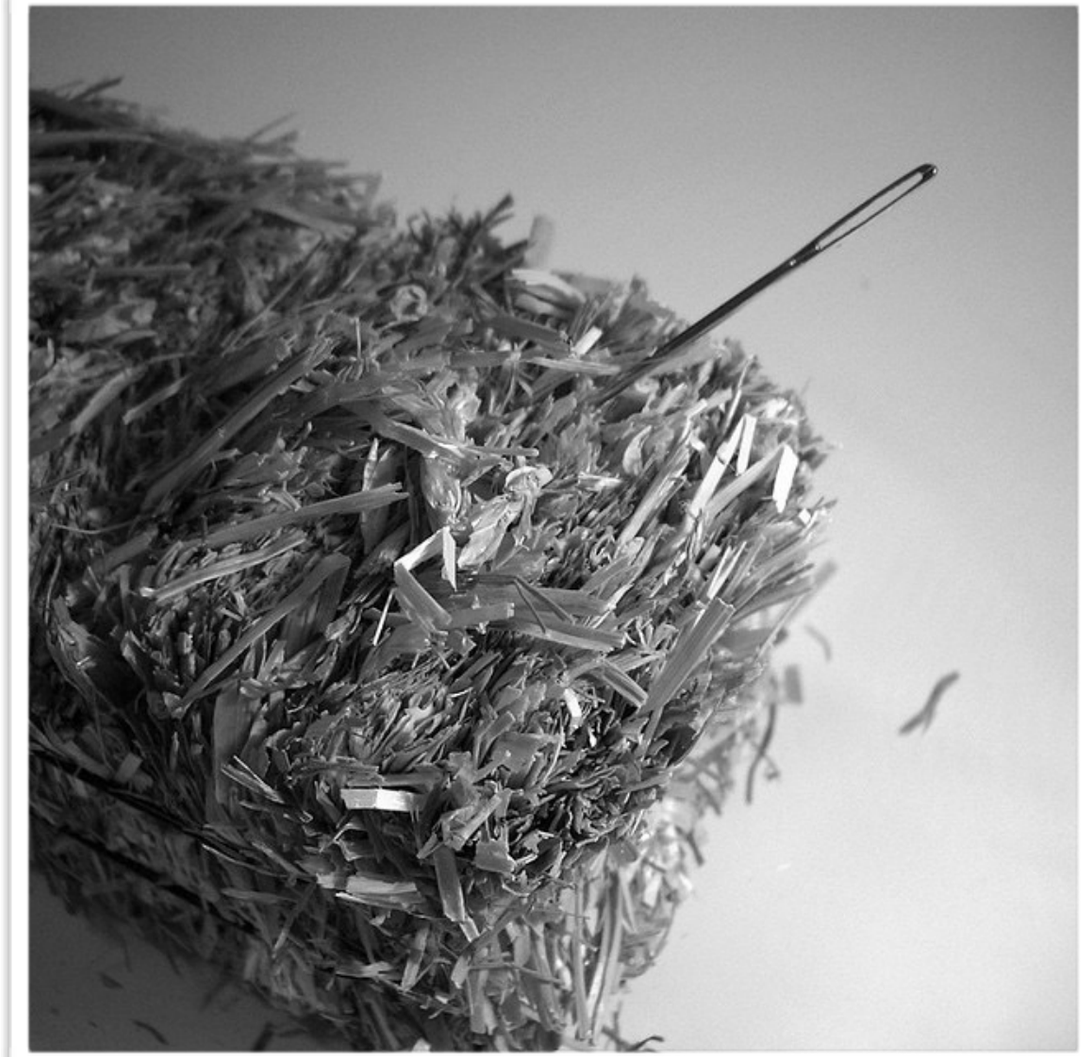
Improve photon identification using SVD or MLP from TMVA fed with 4 variables (cluster size and energy in veto detector and preshower)

Improvement over traditional analysis:
efficiency/purity 50-70% → 95%
Caveat: more info used

Signal extraction

Signal extraction in low S/B environment

- Reconstruction of **2- and 3-prong decays** in heavy ion collisions is challenging: **large combinatorics**
- 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

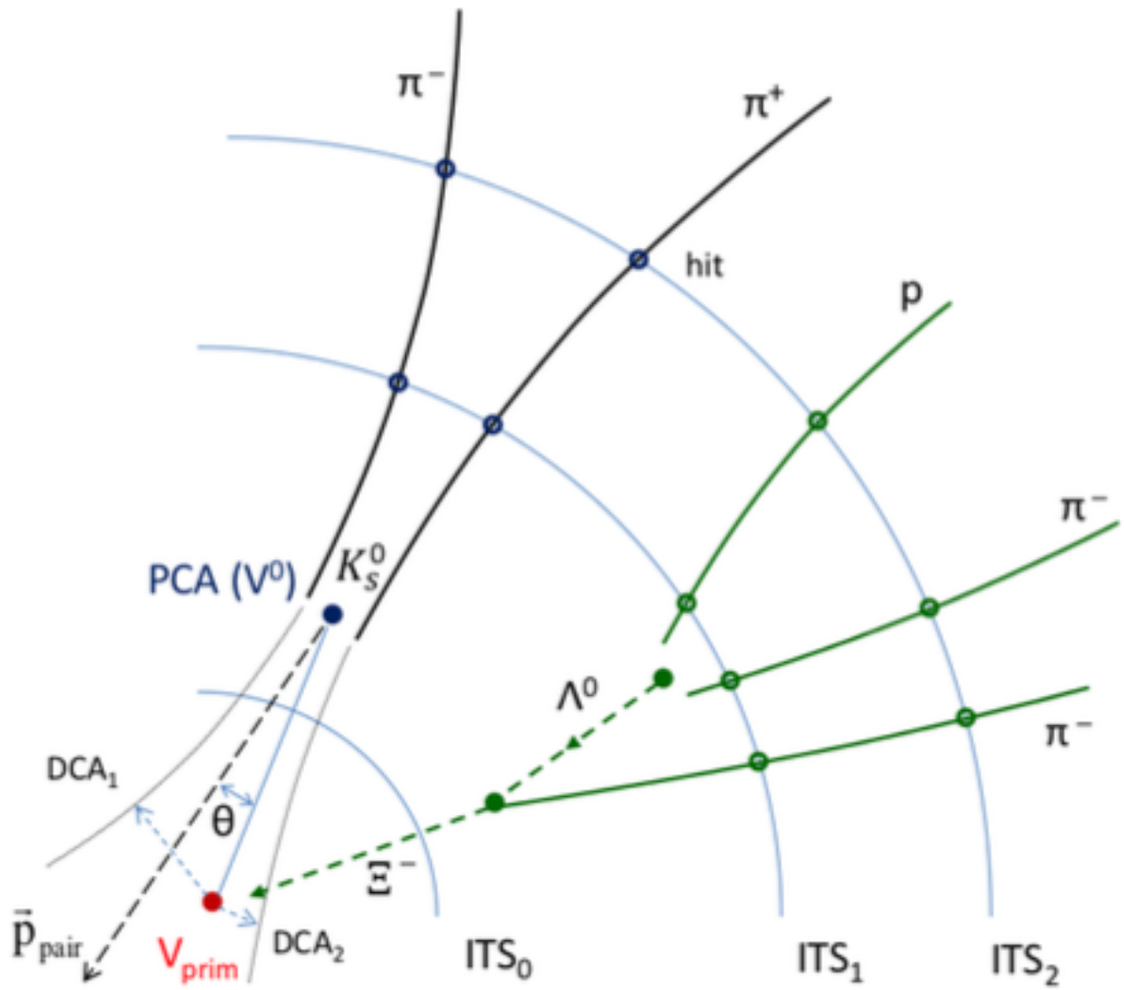
Invariant mass reconstruction

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

Topological reconstruction of weakly decaying particles:

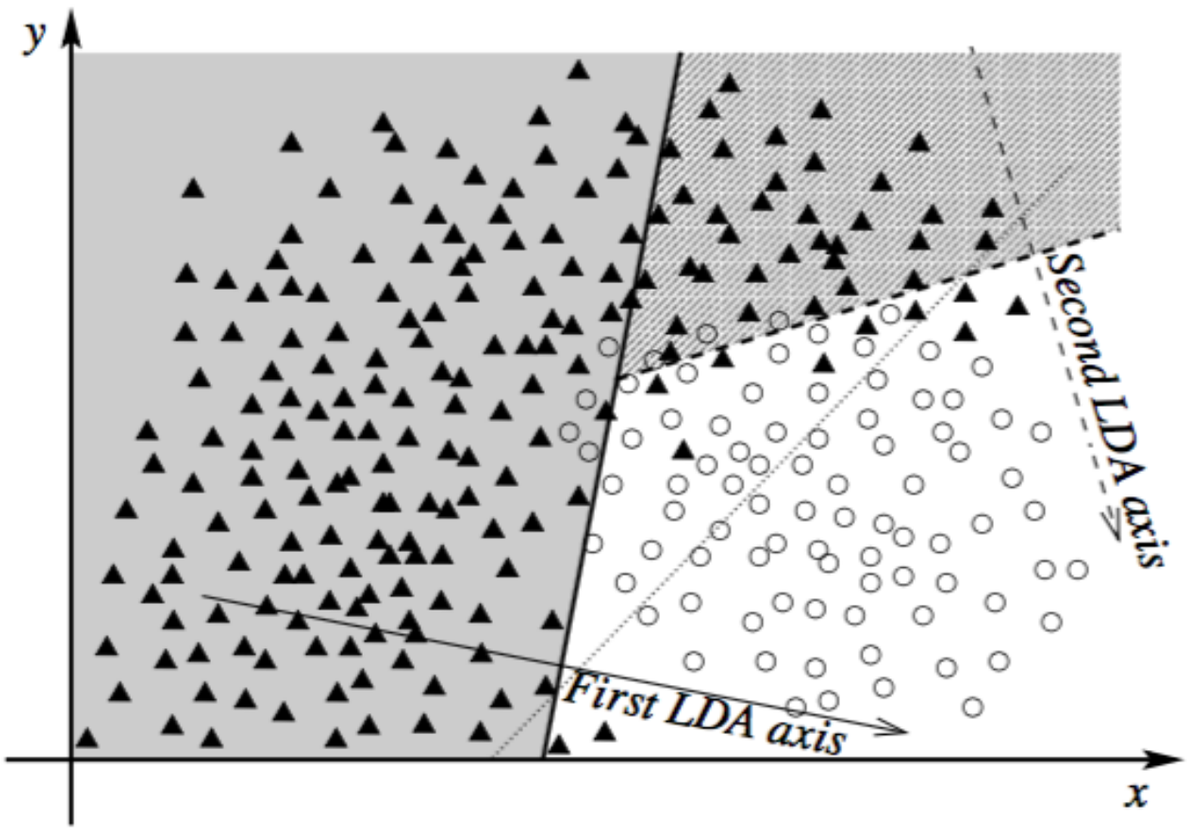
- Decay radius
- cos(θ) – pointing angle
- Distance of their closest approach (DCA1 and DCA2) to V_{prim}
- Distance of daughters at the point of closest approach (PCA)
- Armenteros-Podolansky variables

Correlations among the cut variables



Early attempts (~2006): Multicut-LDA

1. Determine **first LDA** direction on full sample
2. Determine **second LDA** direction on remaining directions
3. Repeat

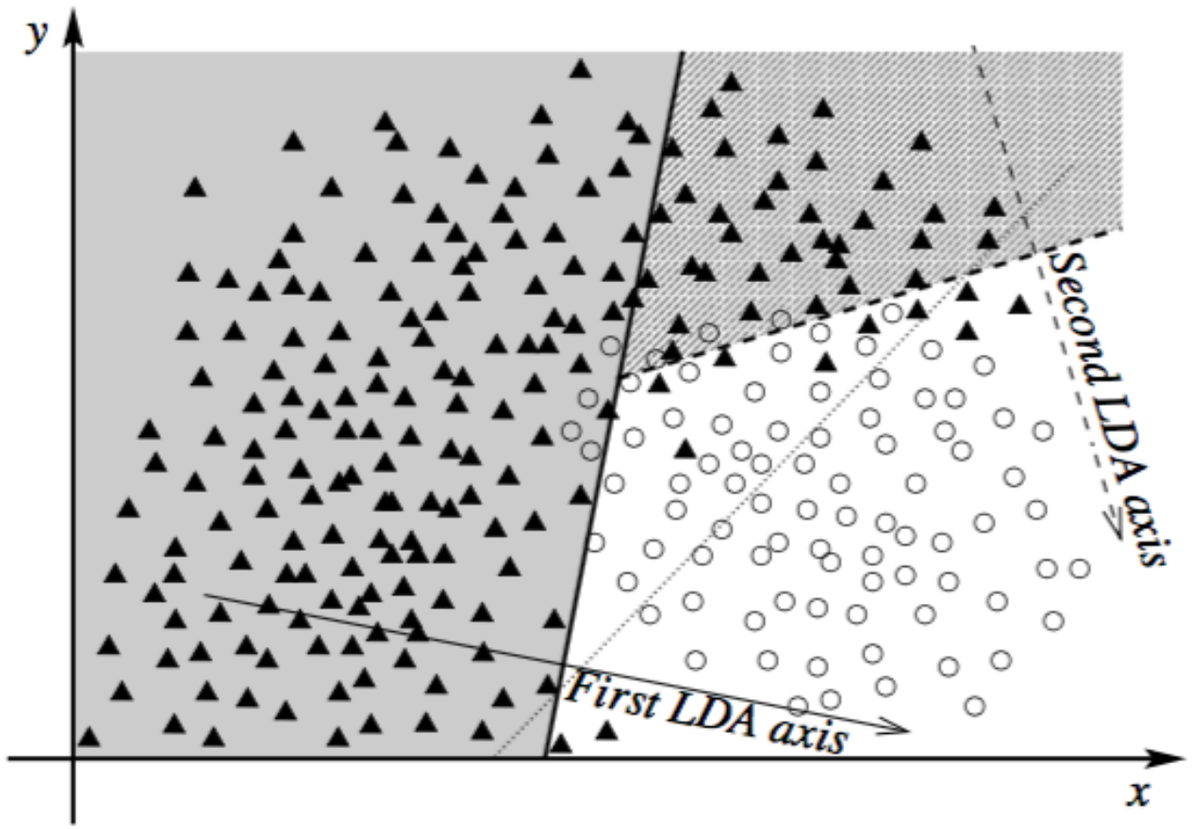


N.B.: before data taking and before TMVA

ALICE-INT-2007-002, https://cds.cern.ch/record/1027337

Early attempts (~2006): Multicut-LDA

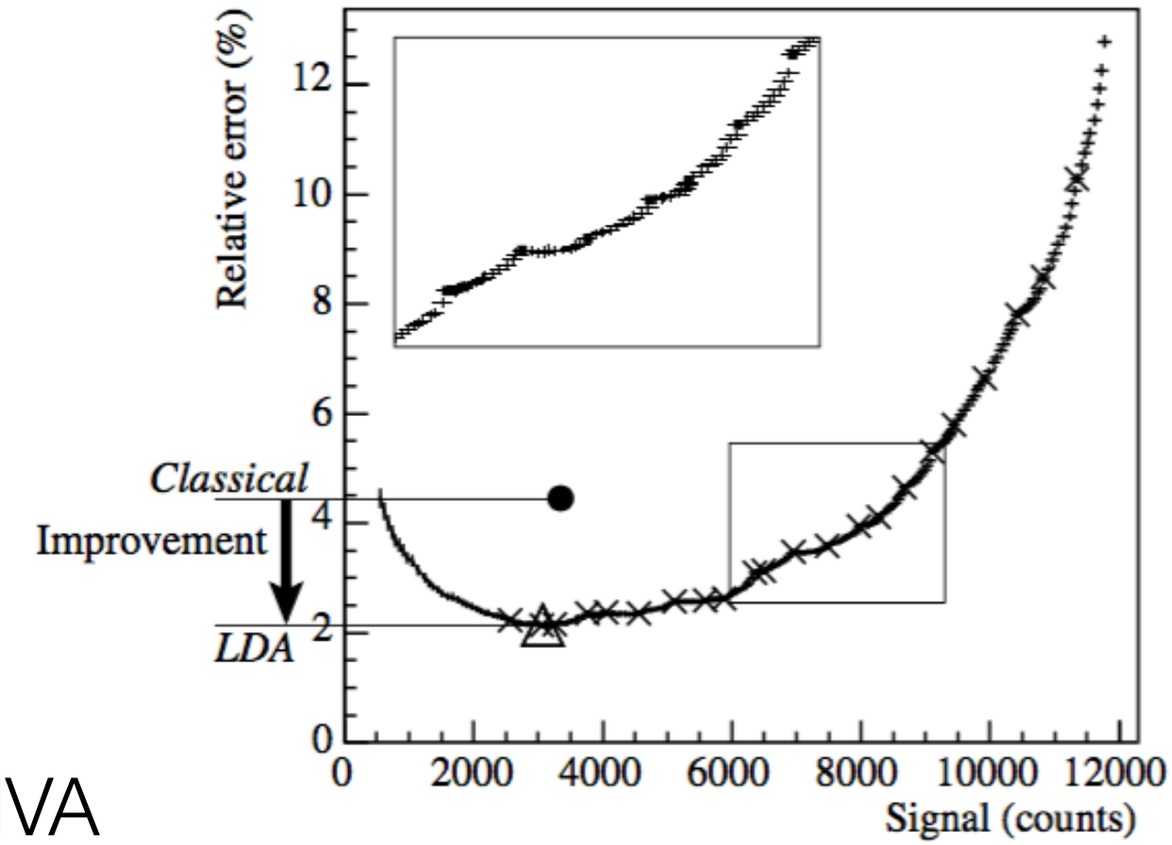
1. Determine **first LDA** direction on full sample
2. Determine **second LDA** direction on remaining directions
3. Repeat



Fisher criterion replaced by **optimization criterion**: given the desired efficiency, maximize BG removed

Number of cuts tuned based e.g. on relative error

MC, $D^0 \rightarrow K\pi$

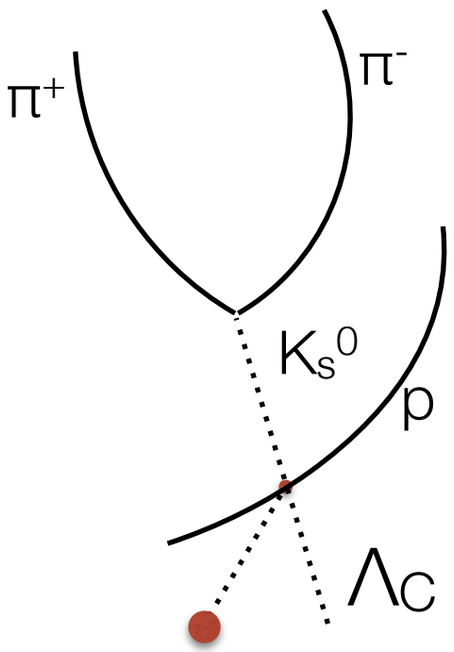


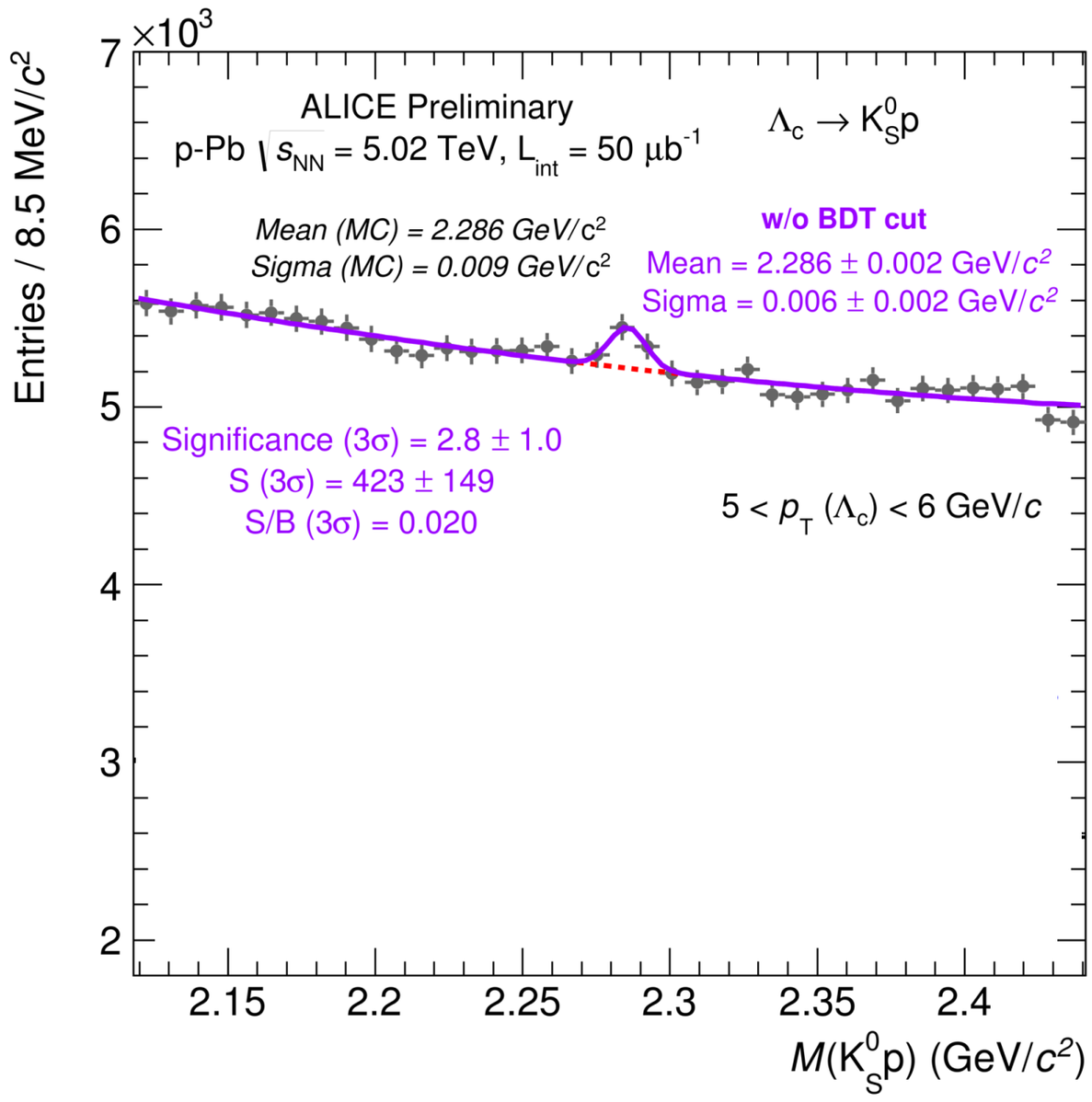
N.B.: before data taking and before TMVA

ALICE-INT-2007-002, https://cds.cern.ch/record/1027337

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

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

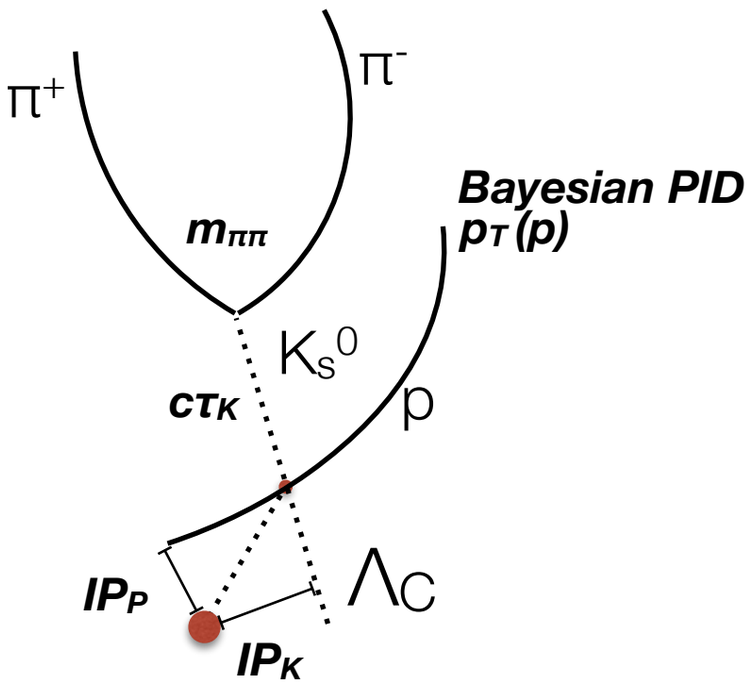


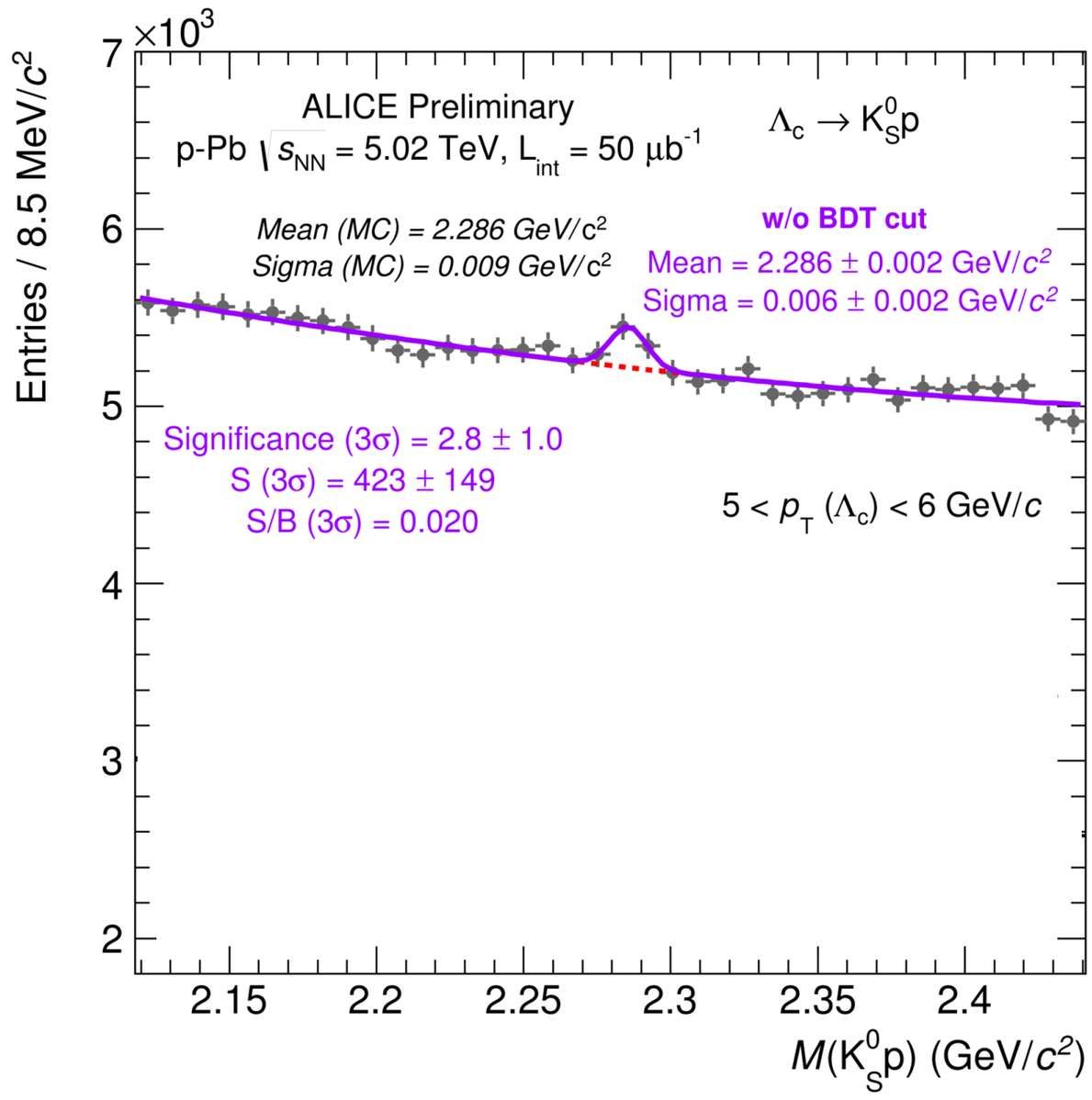


ALI-PREL-76134

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

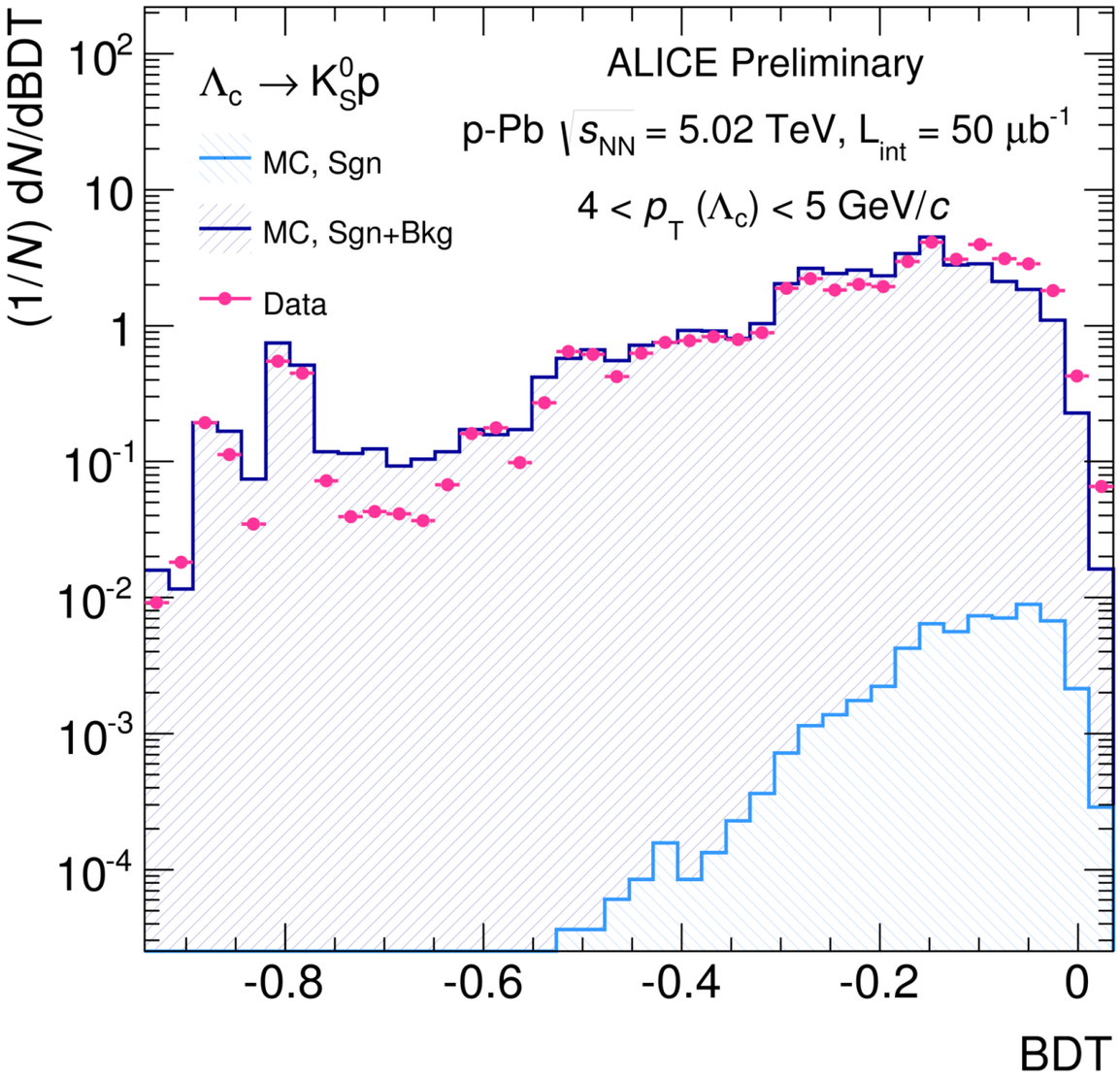
- **Recent attempts based on TMVA,** mostly **BDTs**
- Several channels studied:
  - $\Lambda \rightarrow p\pi$, $K_s^0 \rightarrow \pi\pi$, $\Lambda_C \rightarrow \pi Kp$, …
- Example discussed here: **$\Lambda_C \rightarrow K_s^0 p$**
- 3-prong decay: large combinatorial BG





ALI-PREL-76134

BDT output



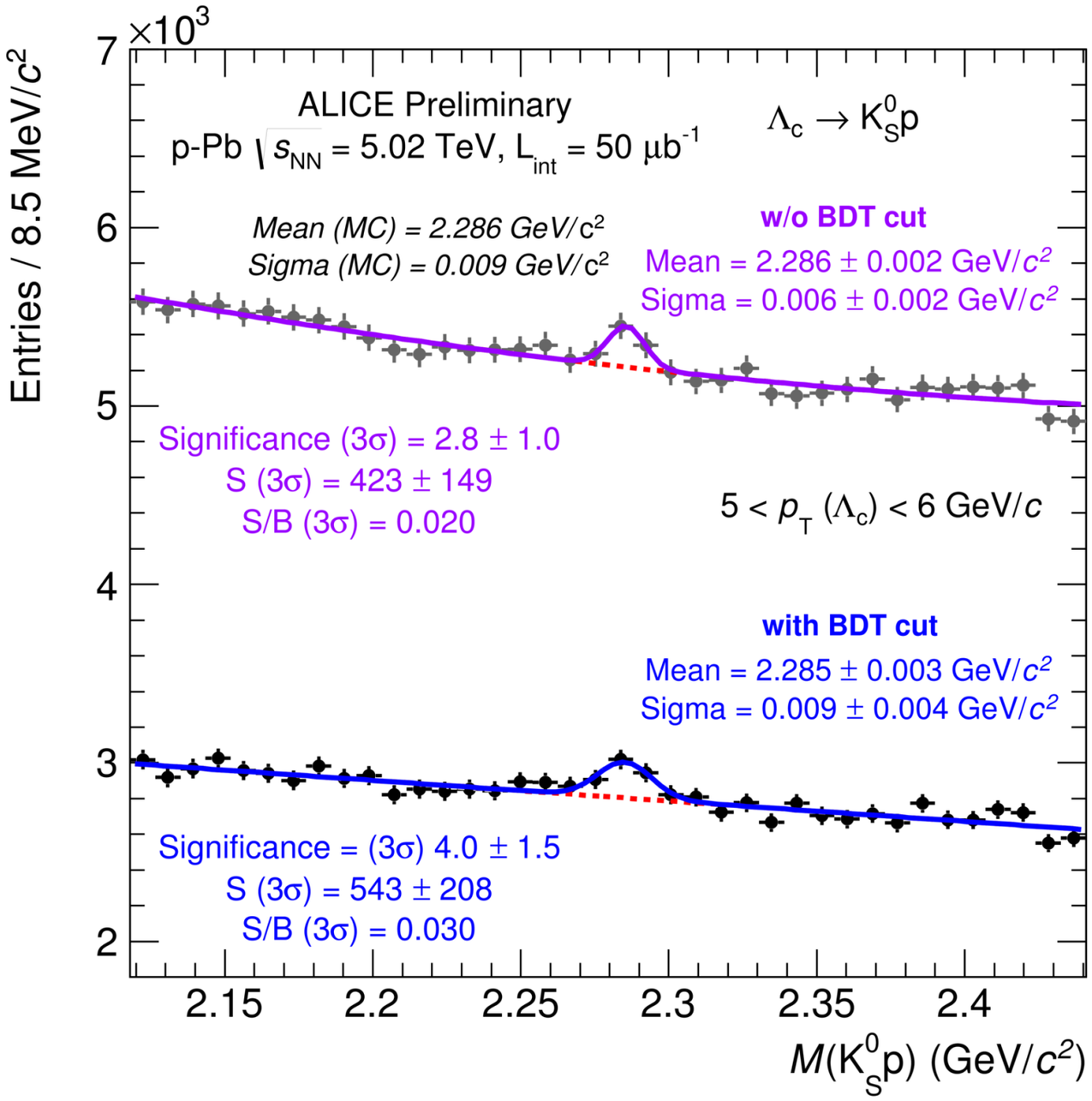
BDT output distribution in data and MC reasonably similar

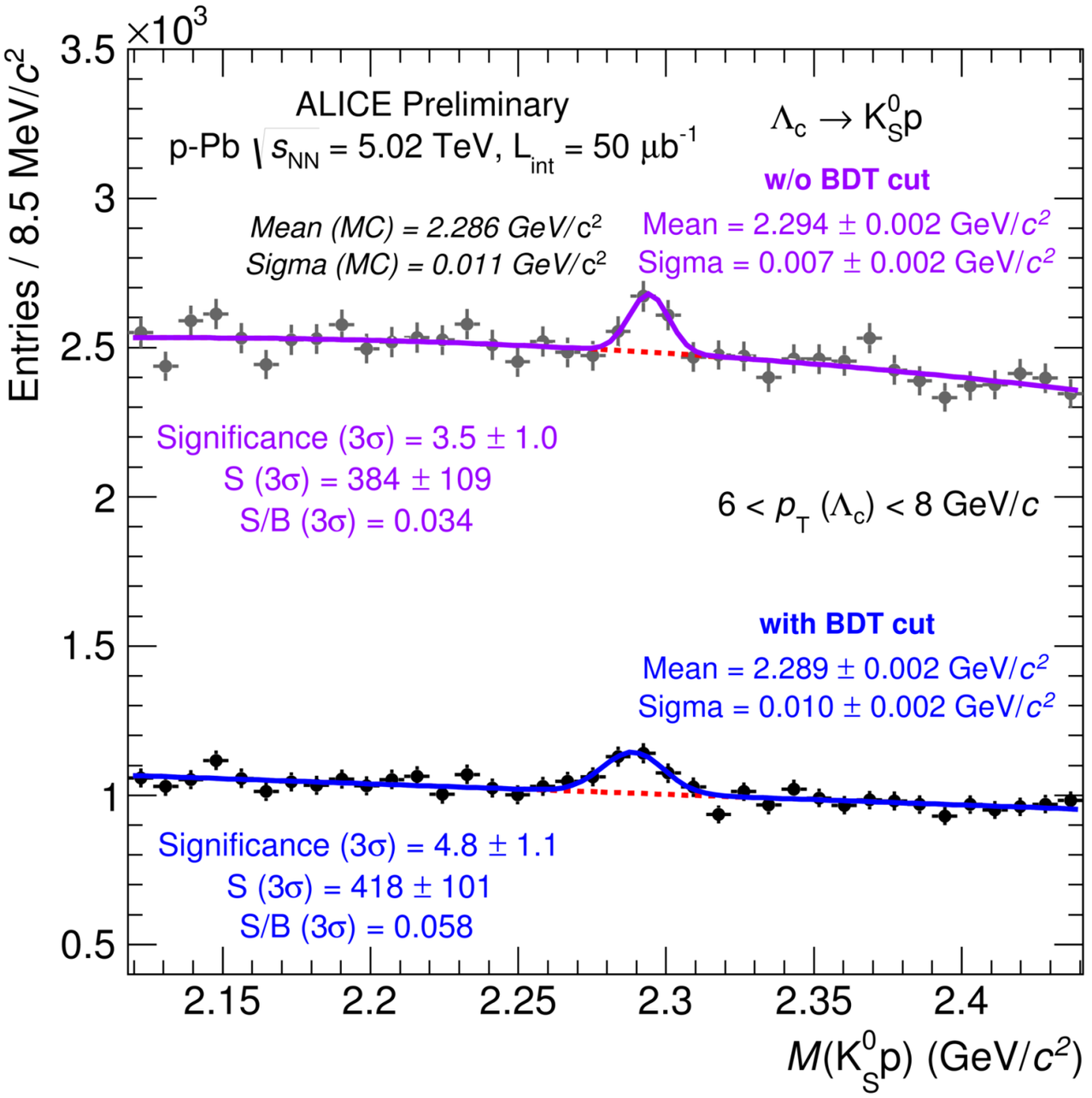
Tuning repeated with BG from data (side bands)

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

Optimization of BDT parameters in progress

$\Lambda_c \rightarrow Ks^0$: Results





-PREL-76134 -PREL-76142

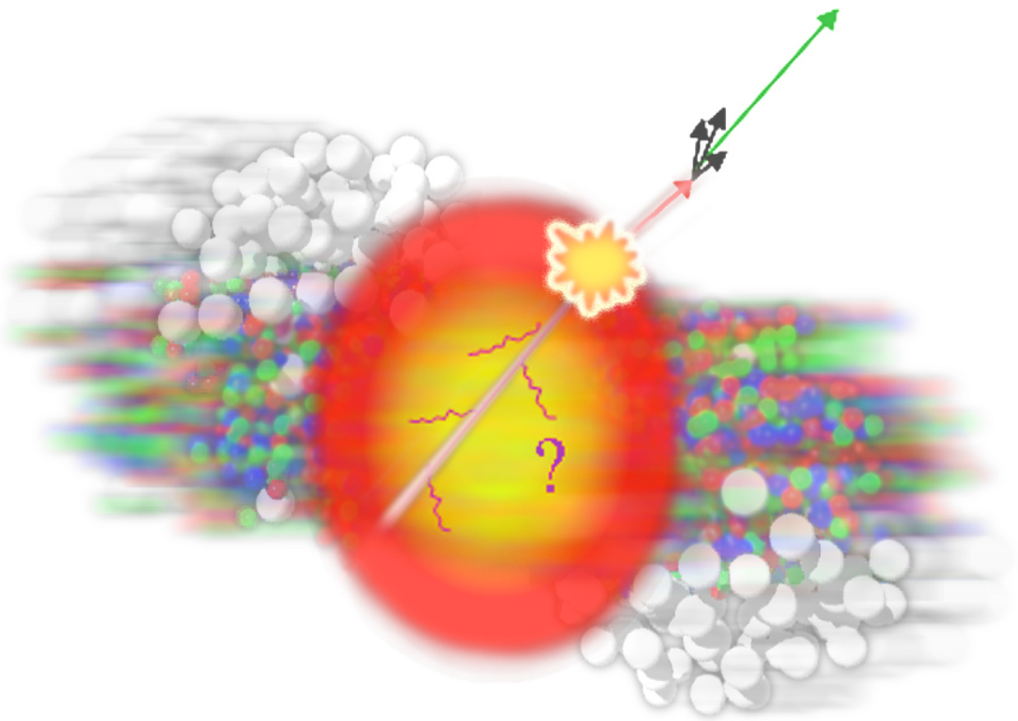
Significance improved by BDT

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

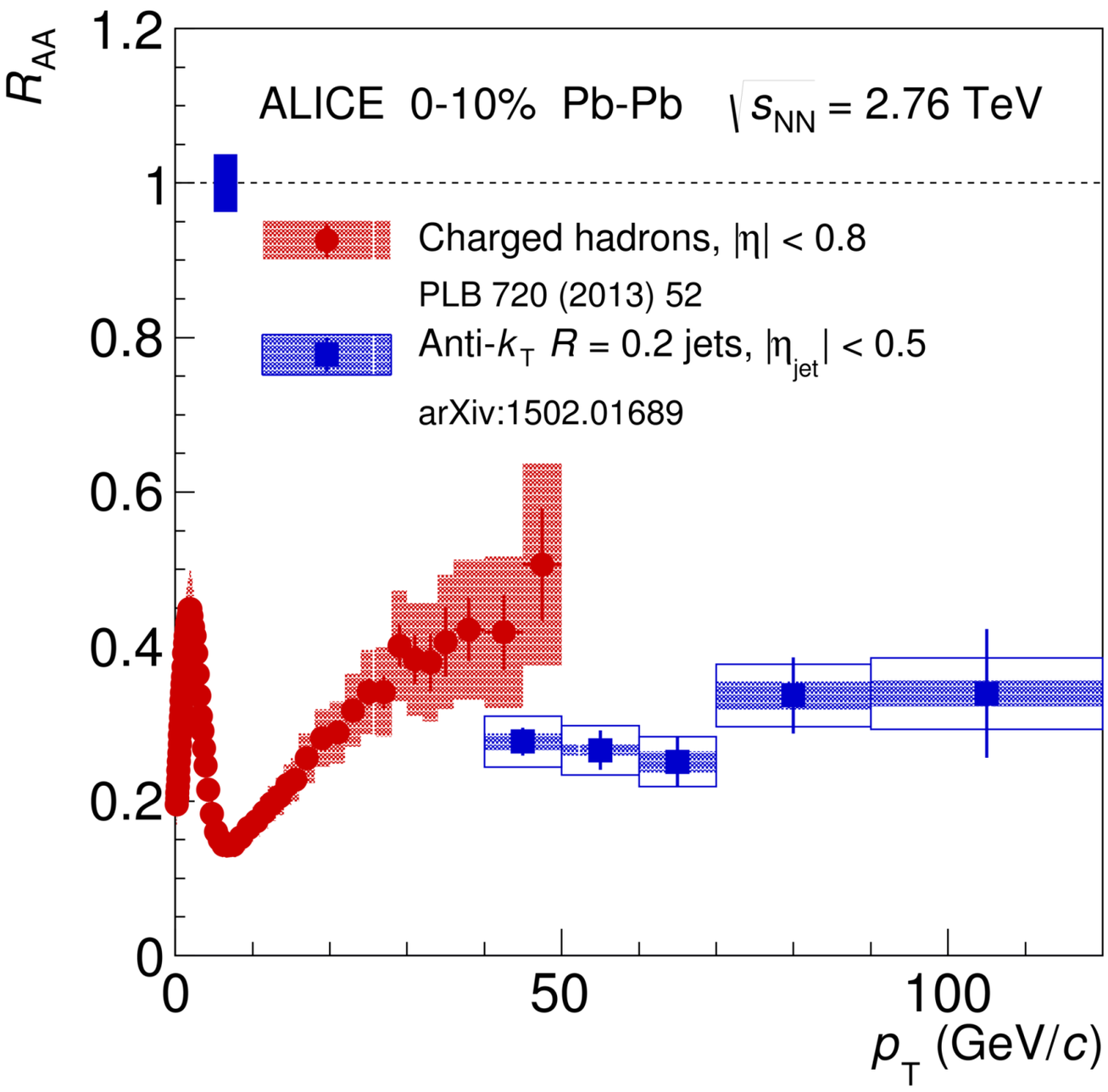
$\Delta E_g > \Delta E_{u,d,s}$ (Color factors)

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

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

A primer on jet quenching

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



LI-DER-92185

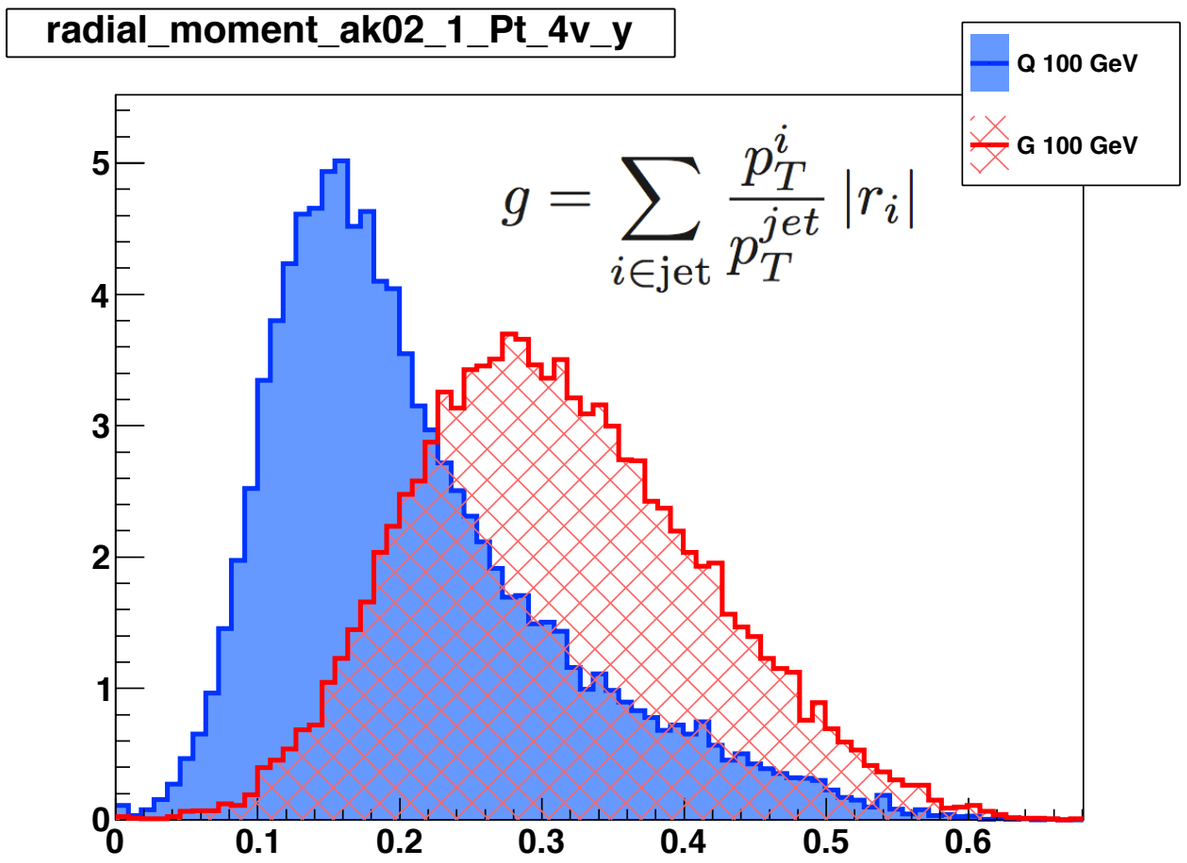
Recoil **jet loses energy** when traversing the medium "Radiative" and "Collisional" energy loss

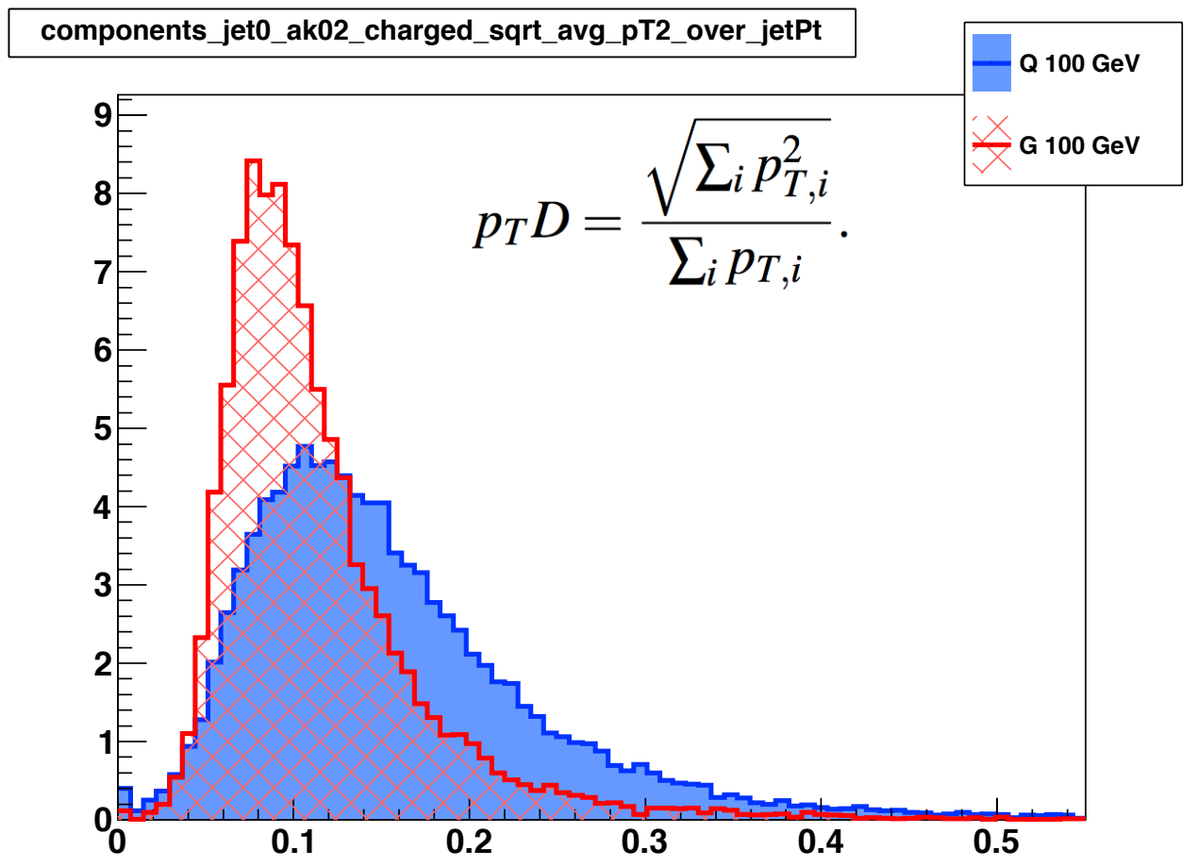
$\Delta E_g > \Delta E_{u,d,s}$ (Color factors)

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

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

Quark-Gluon Jet Discrimination

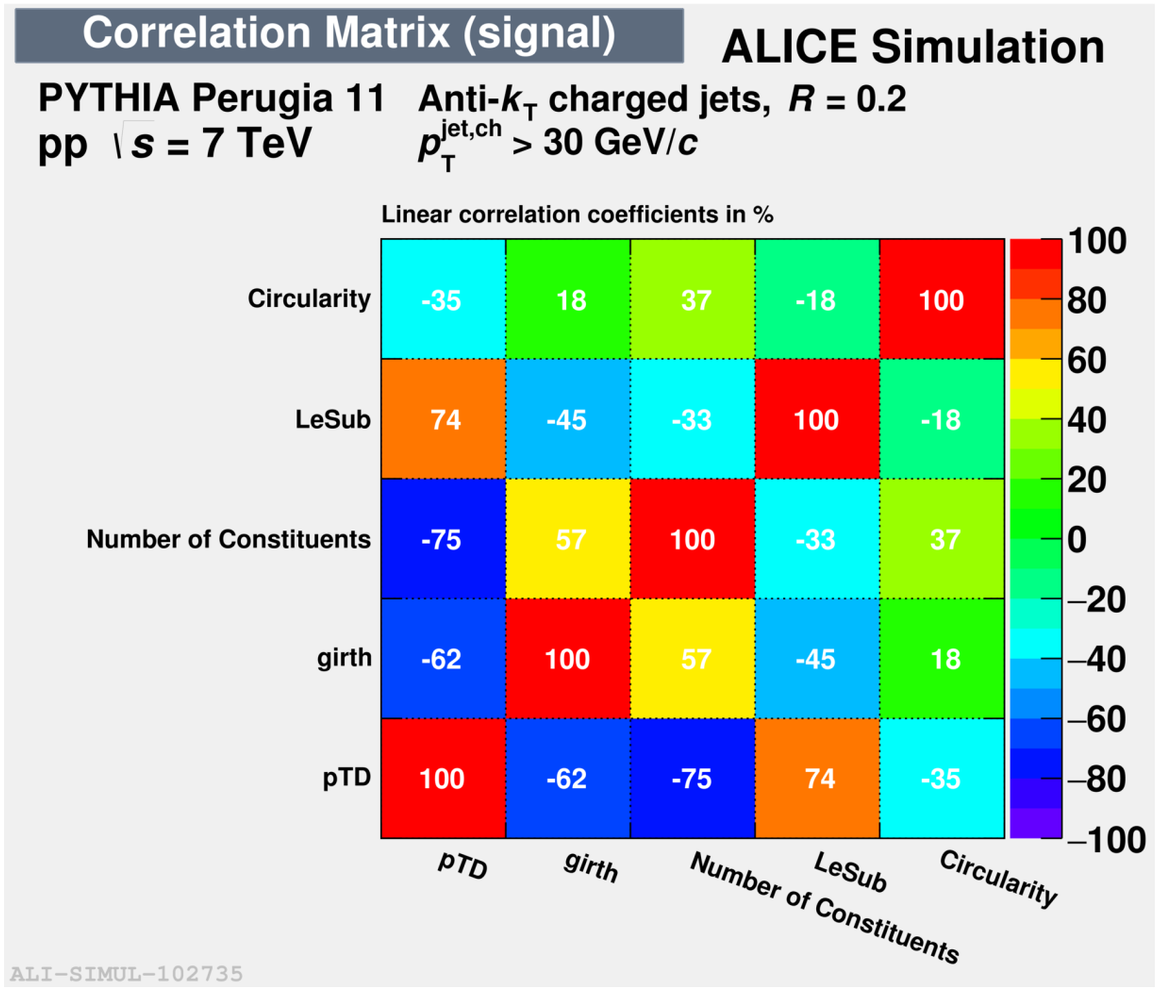


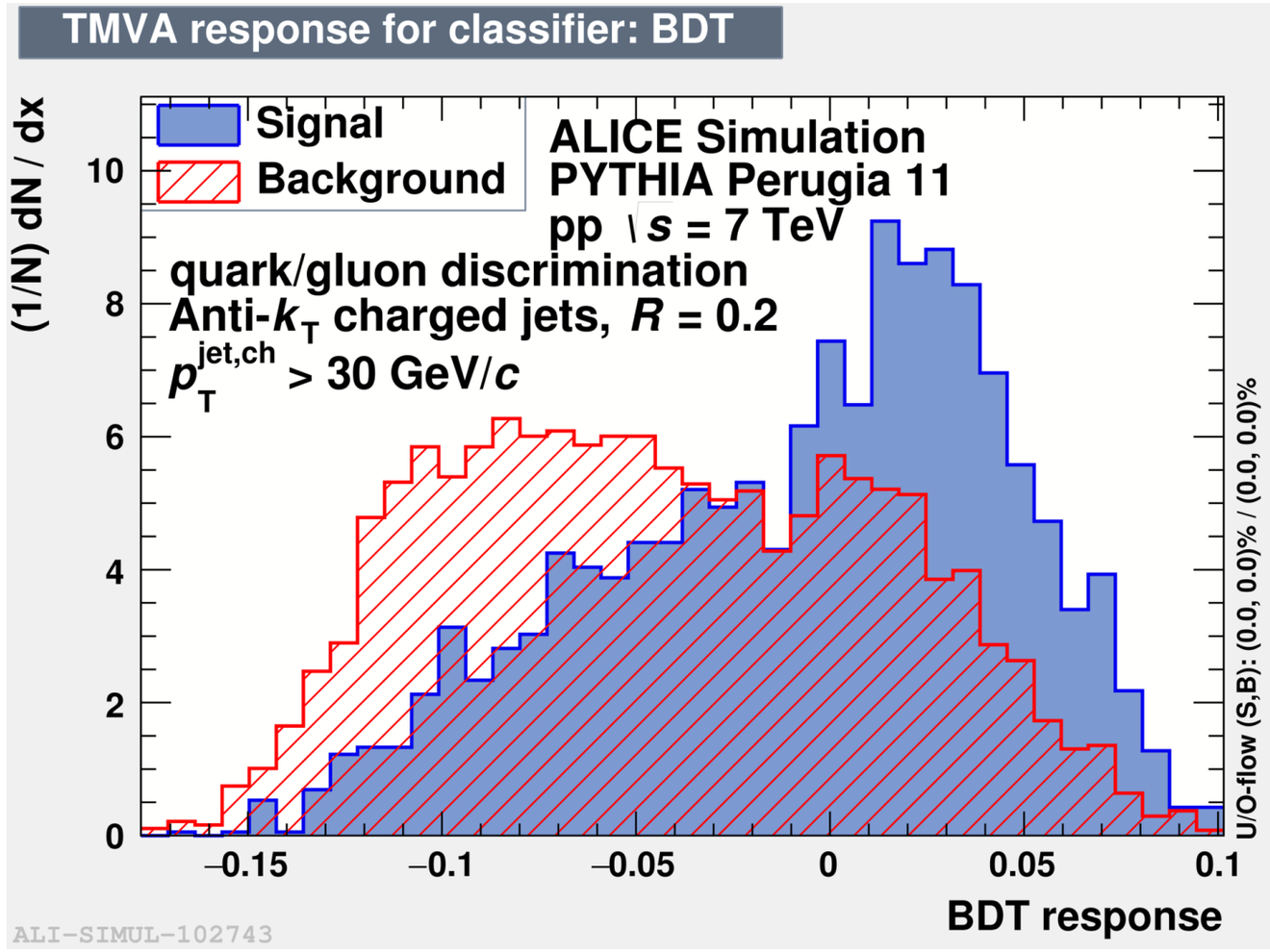


Jet shapes like angularities, radial moment or p_TD show sensitivity to differences between quark and gluon fragmentation

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

Tagging Jets with BDT





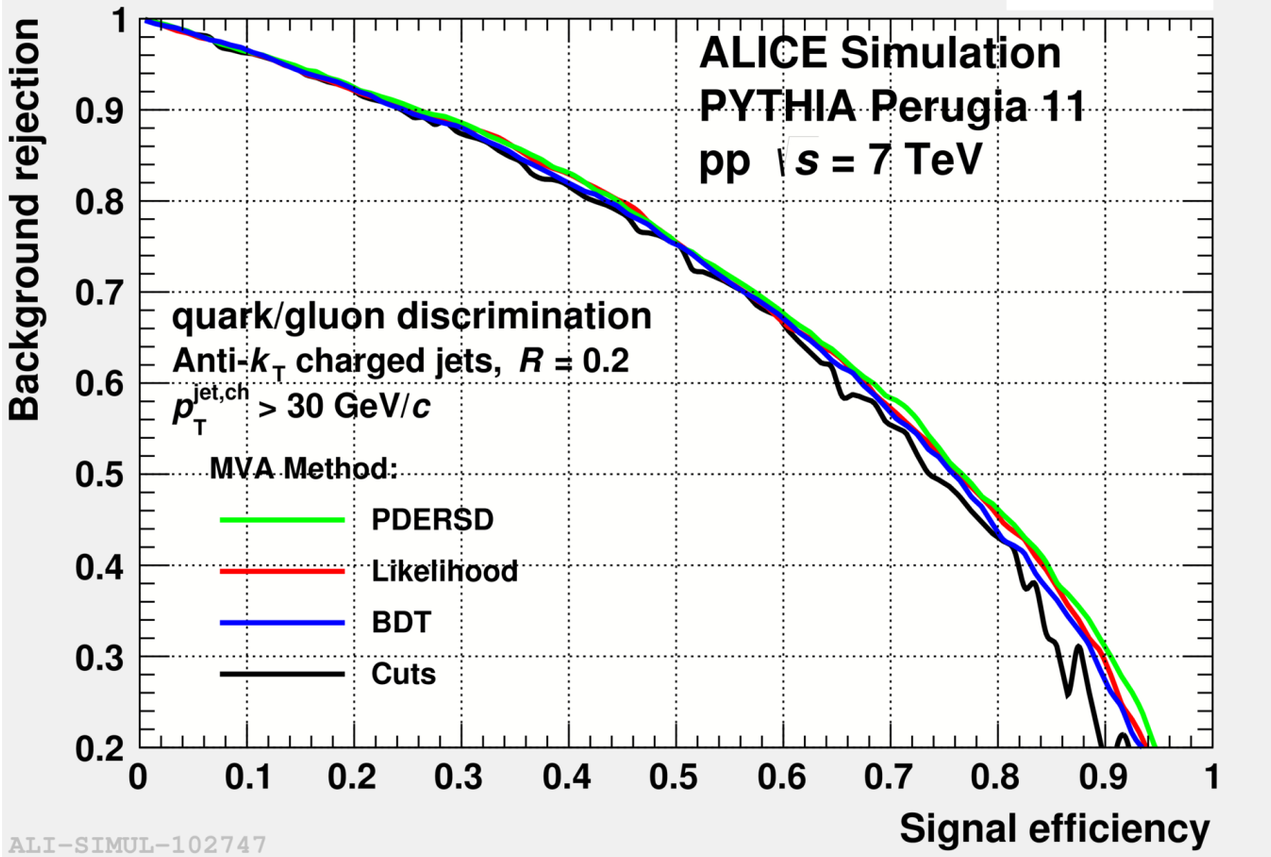
Pythia Perugia 2011, particle level

Anti-kT, R=0.2

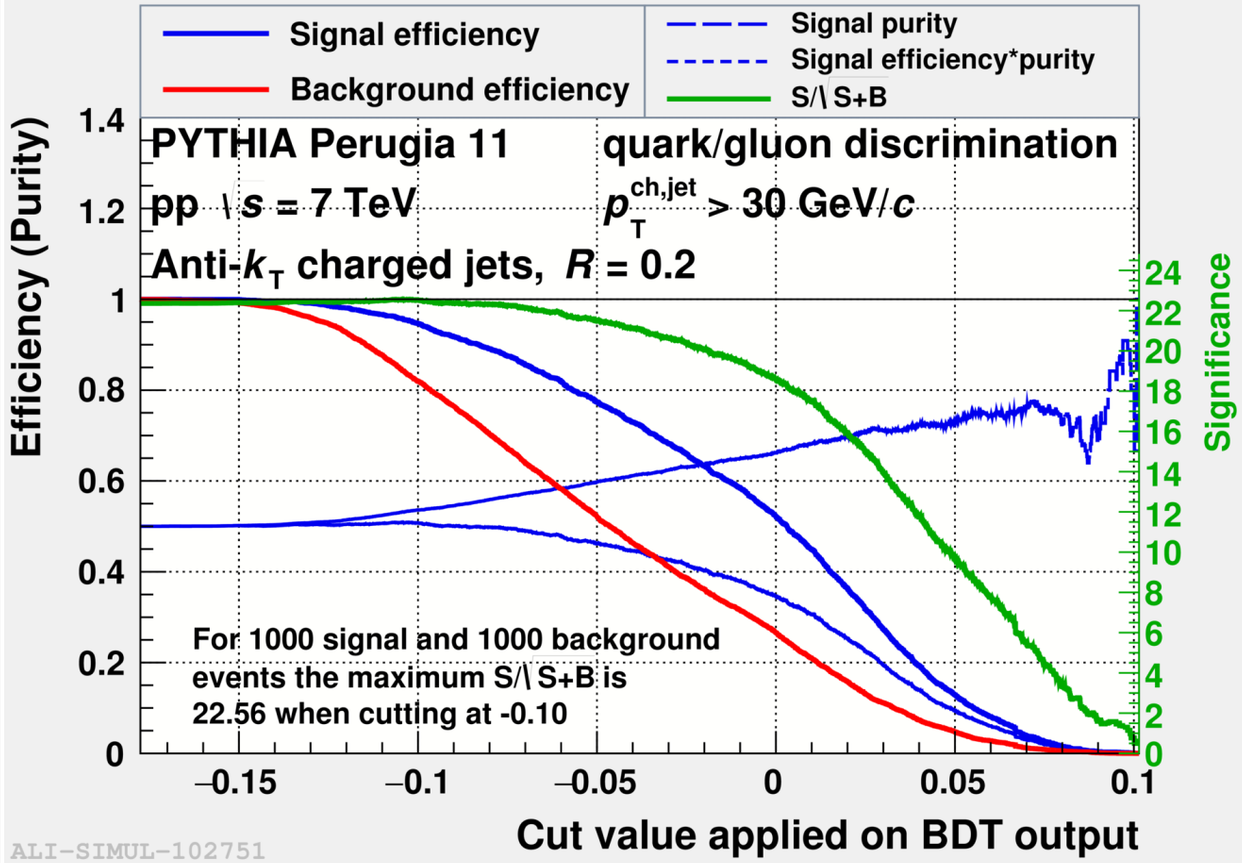
Variables input to BDT: p_TD, girth, constituents, LeSub, Circularity

MVA Jet Results

Background rejection versus Signal efficiency



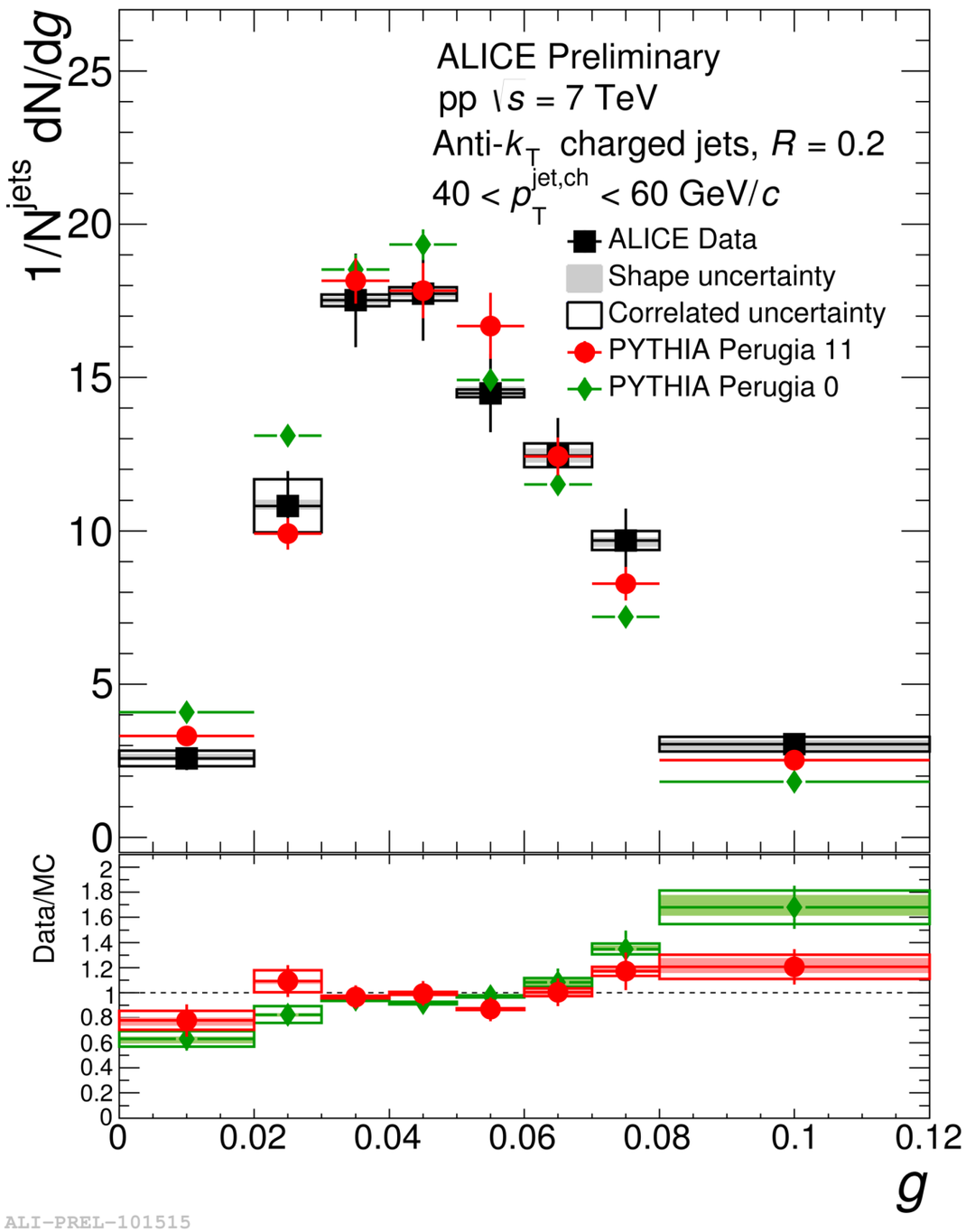
Cut efficiencies and optimal cut value



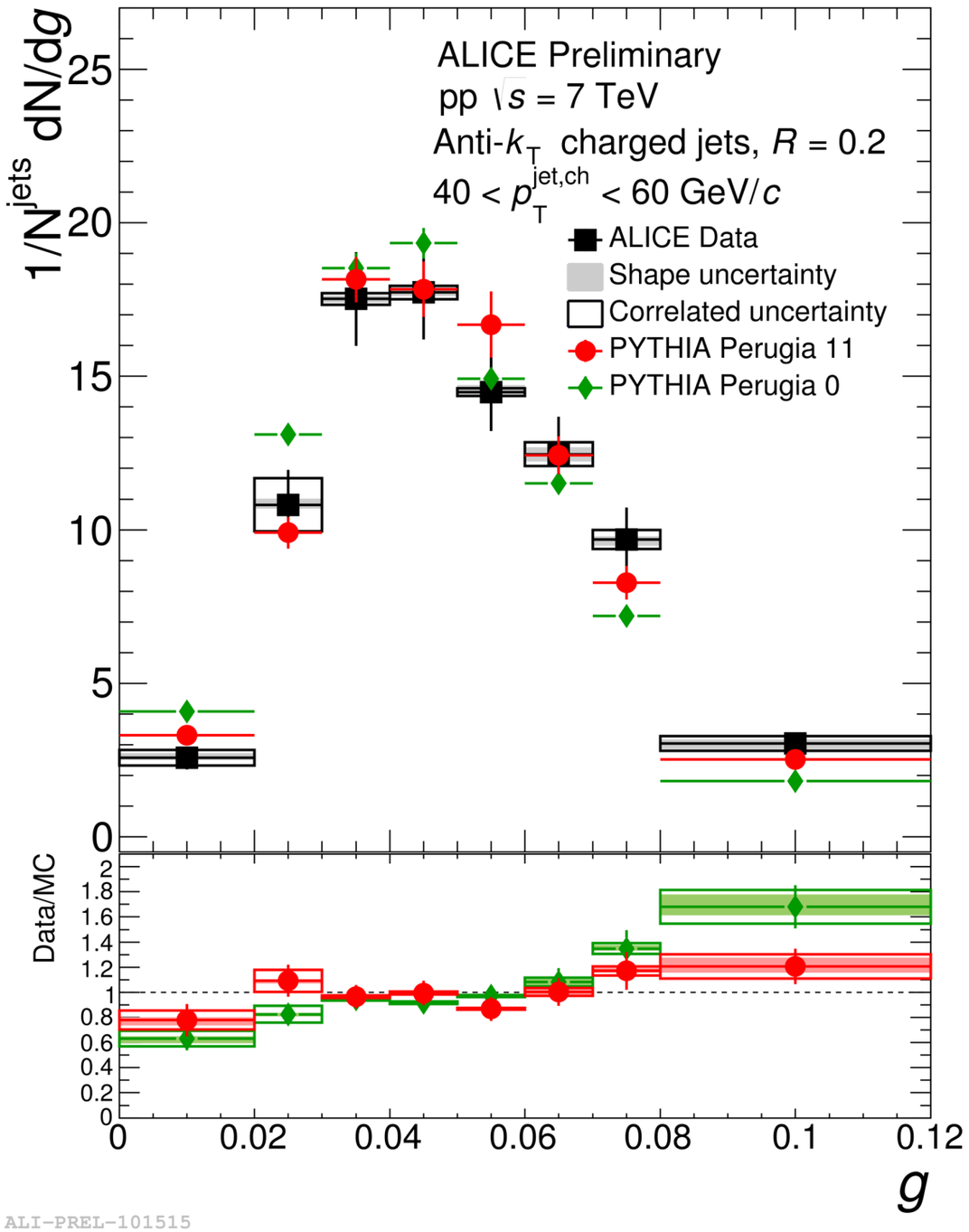
Work in progress! Methods tested: BTD, PDERSD, Likelihood

Similar performance of the various methods

Purity and efficiency ~ 60%

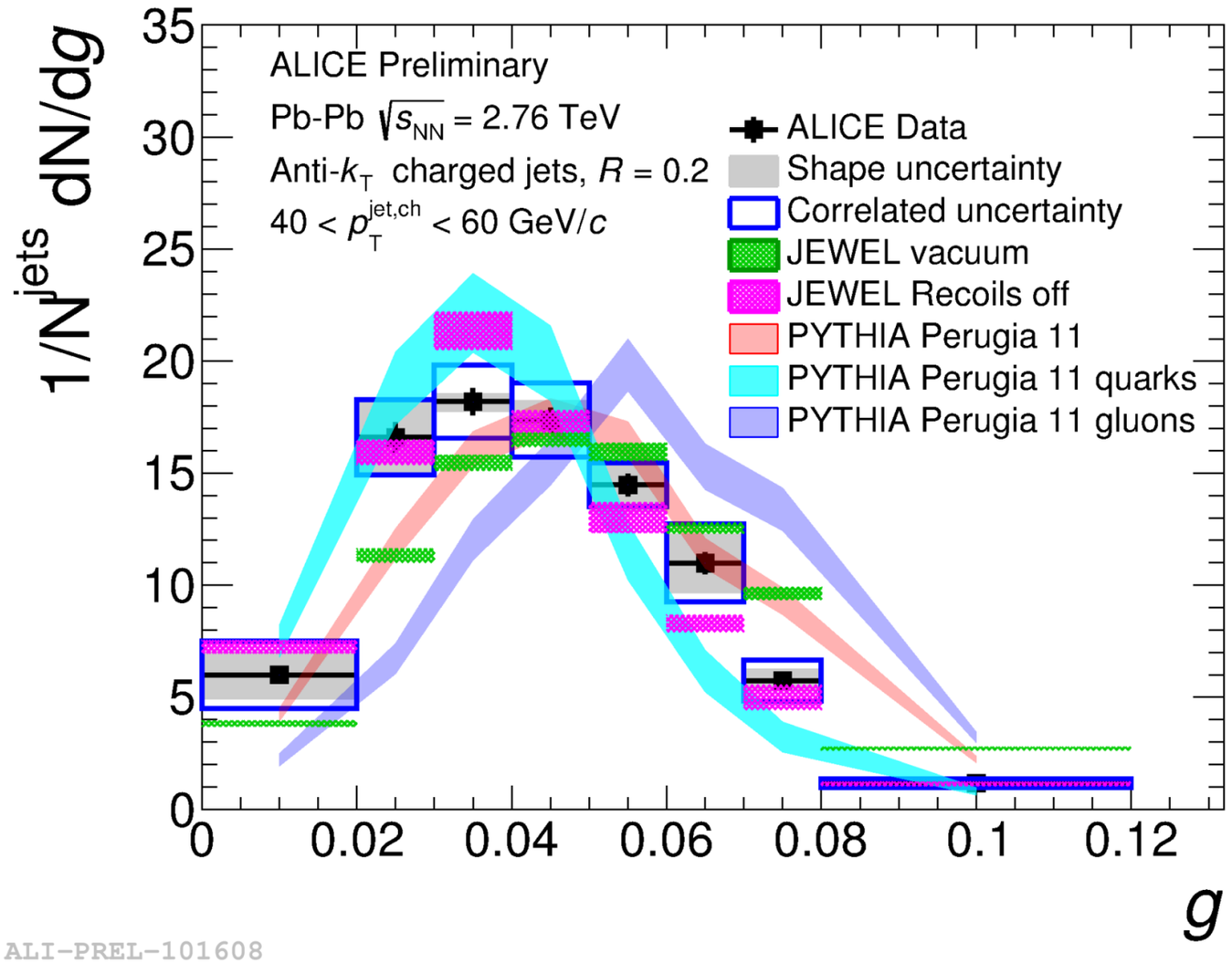


Pythia reproduces jet **shapes**
(e.g. girth) in pp collisions



ALI-PREL-101515

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

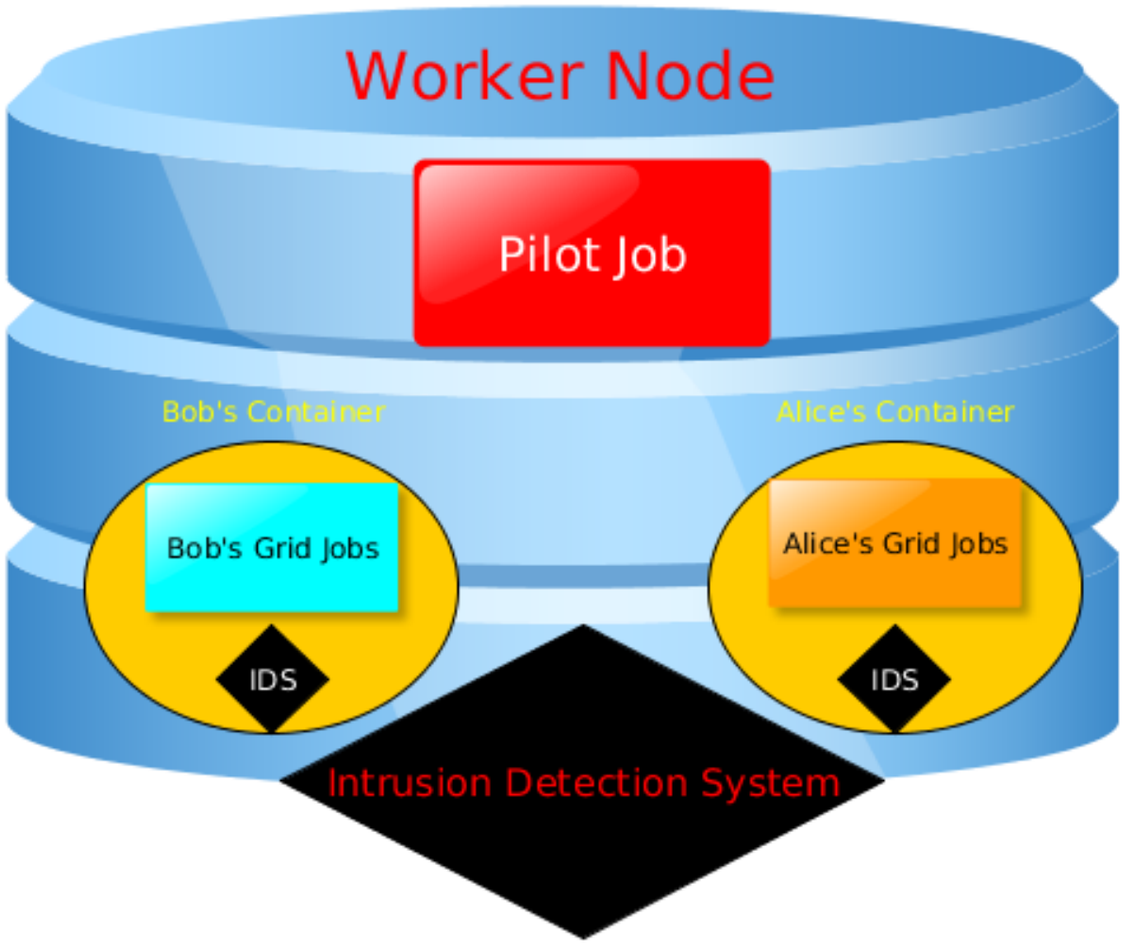


ALI-PREL-101608

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
- Signal extraction in analysis:

👍 Easier inclusion of additional information, seemingly better S/B performance

👎 “Black Box”: hidden systematics? Is it really better than traditional approach?

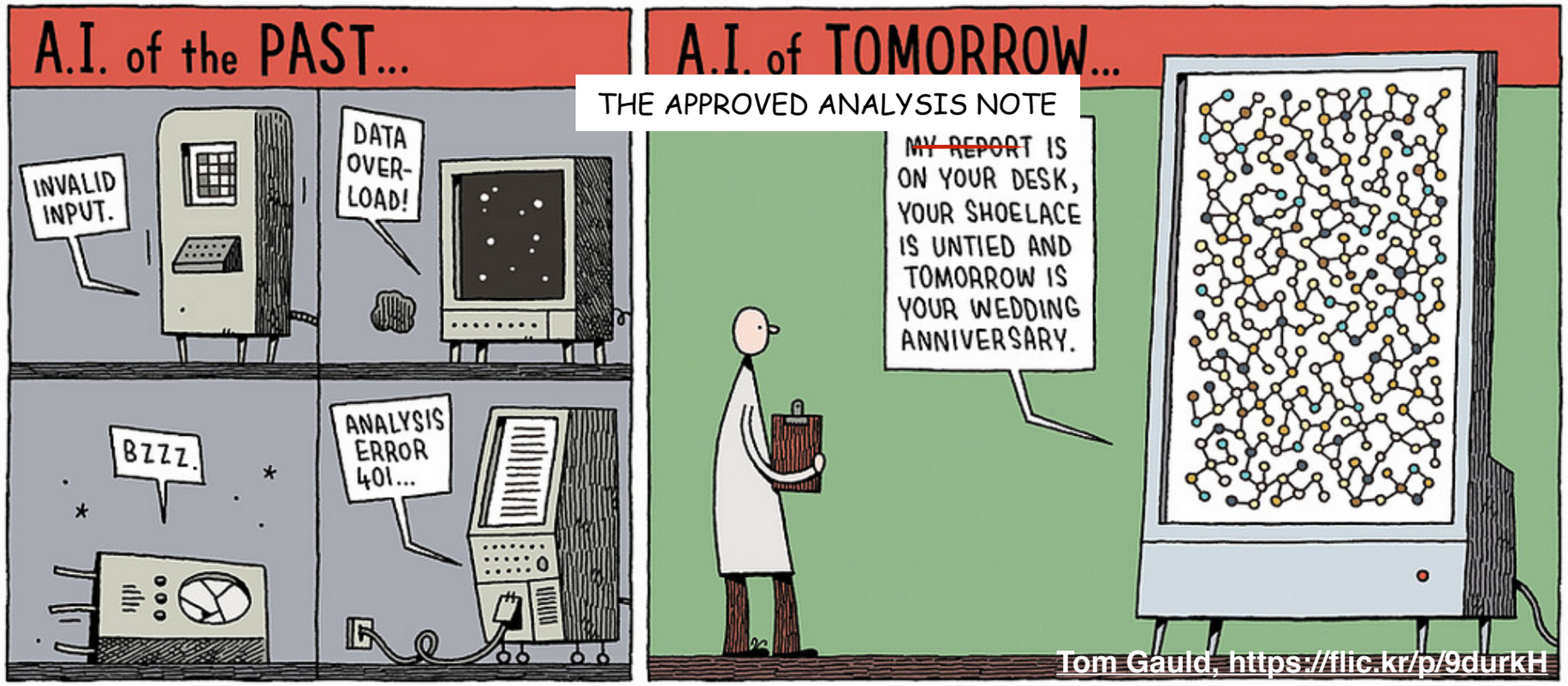
Thanks! Andrea Alici, Andres Gomez, Andrew Lowe, Chiara Zampolli, David Rohr, Davide Caffarri, Georgios Krintiras, Jaime Norman, Julien Faivre, Leticia Cunqueiro, Mike Sas, Michael Weber, Yvonne Pachmayer, Zaida Conesa Del Valle

Summary

- Several potential applications for machine learning techniques in ALICE
  - Detector, reconstruction, physics analysis, computing
- Early attempts, no widespread use yet
- Signal extraction in analysis:

👍 Easier inclusion of additional information, seemingly better S/B performance

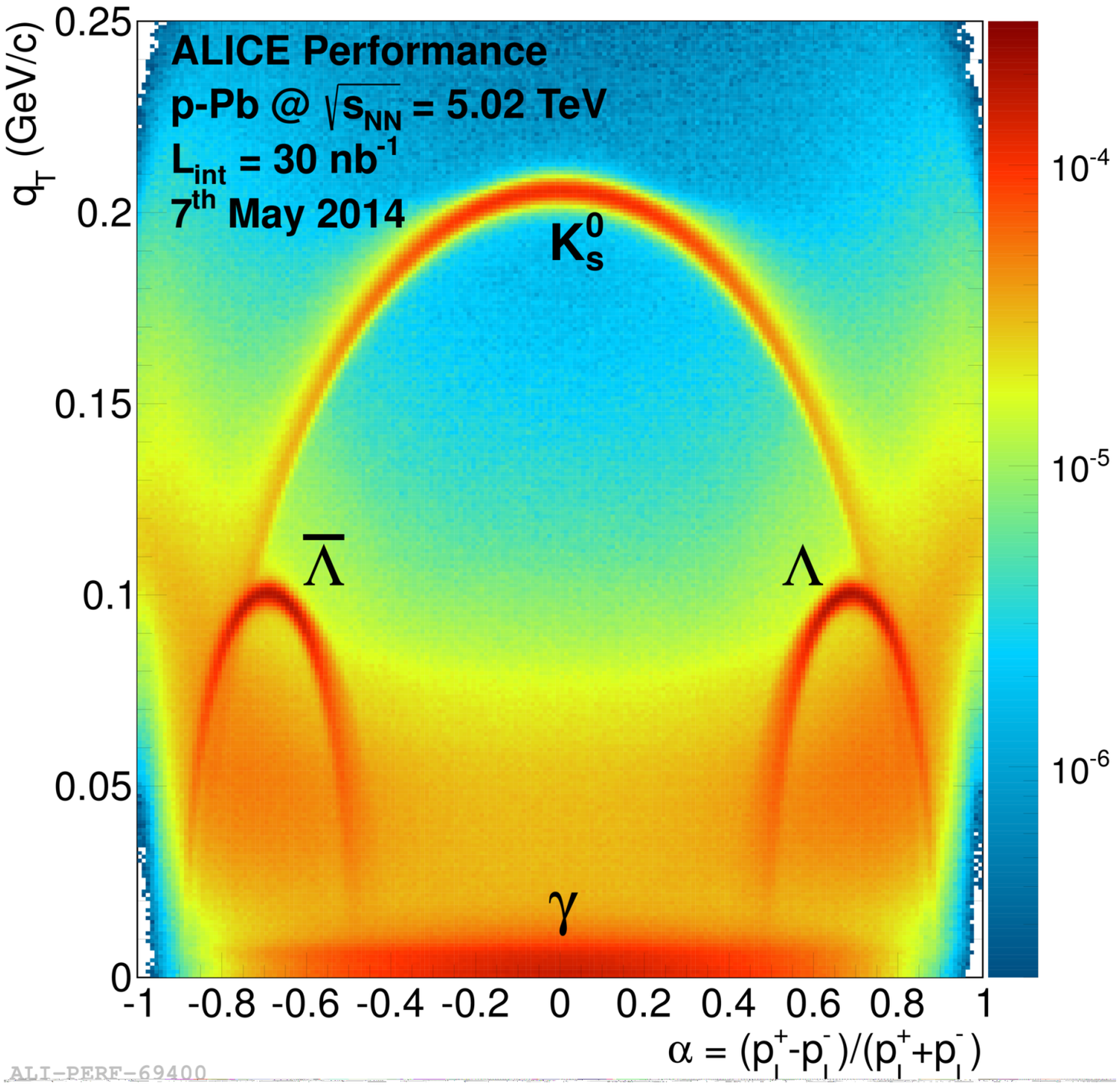
👎 “Black Box”: hidden systematics? Is it really better than traditional approach?

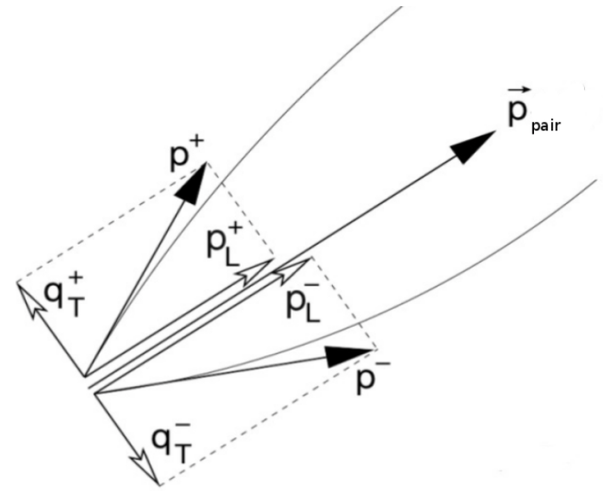


Thanks! Andrea Alici, Andres Gomez, Andrew Lowe, Chiara Zampolli, David Rohr, Davide Caffarri, Georgios Krintiras, Jaime Norman, Julien Faivre, Leticia Cunqueiro, Mike Sas, Michael Weber, Yvonne Pachmayer, Zaida Conesa Del Valle

Backup

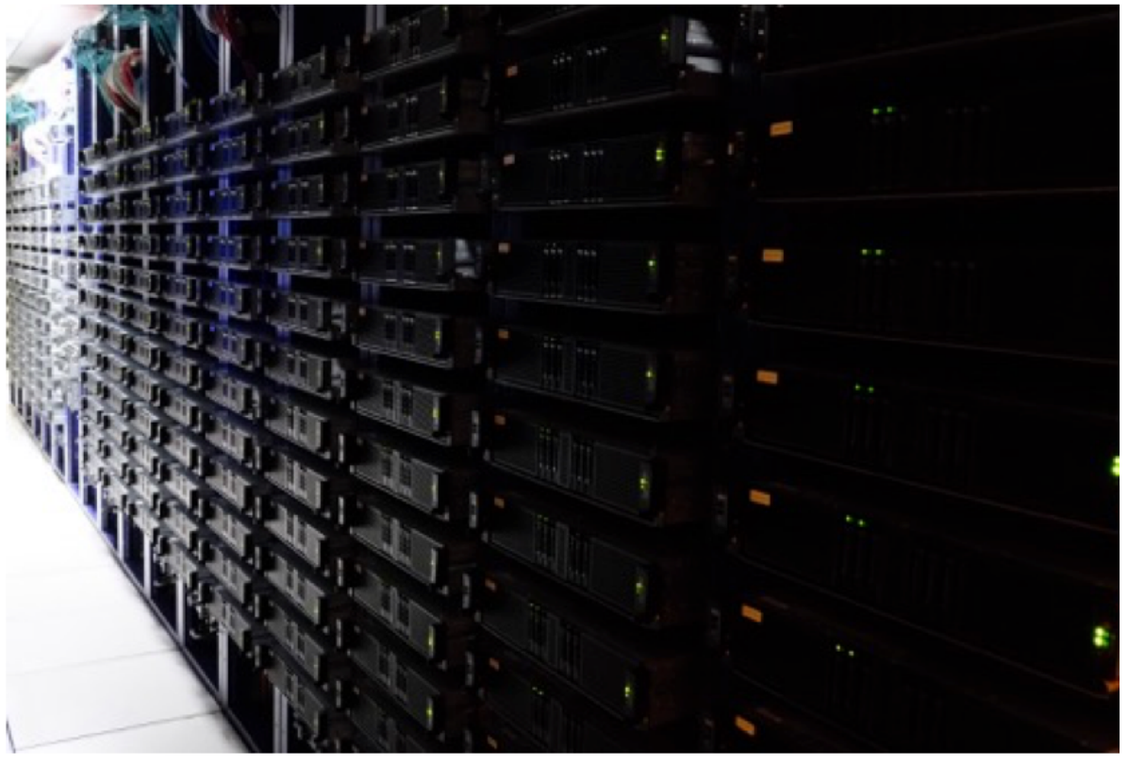
Armenteros-Podolanski





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:
  - Data: Infiniband in IPoIB Mode (FDR with 56Gb/s, full bisection bandwidth).
  - Management: gigabit ethernet with sideband IPMI - one physical ethernet port per node.
    - 10Gbit backbone.