Particle identification using boosted decision trees for the CALICE highly granular SiPM-on tile calorimeter

DPG Spring Meeting, Dortmund 2021

Vladimir Bocharnikov on behalf of CALICE-D Collaboration

Mar 17, 2021

HELMHOLTZ RESEARCH FOR GRAND CHALLENGES

MEPhI National Research Nuclear University

DESY.

CALICE AHCAL

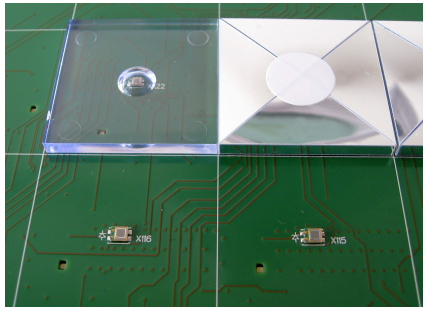
Test beam prototype.

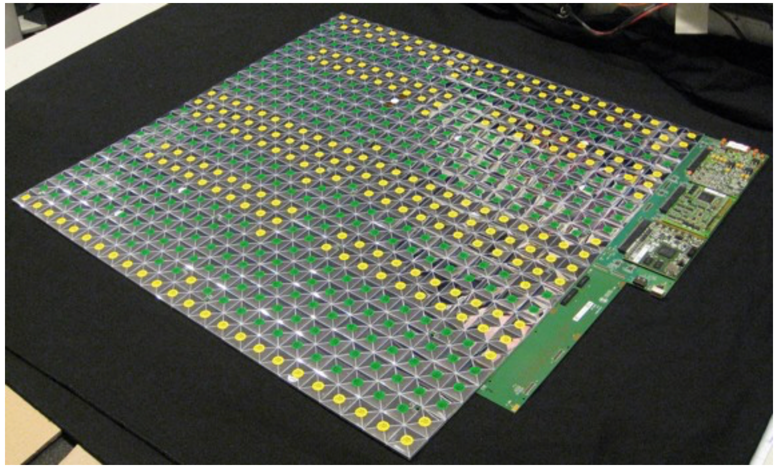
38 active layers of *24x24* scintillator tiles (*3x3* cm^2) alternating with *1.7 cm* steel absorber

In total: *~22000* channels, *~4* λ

Beam particles: **muons, electrons, pions**

Energy range: ***10-200 GeV***



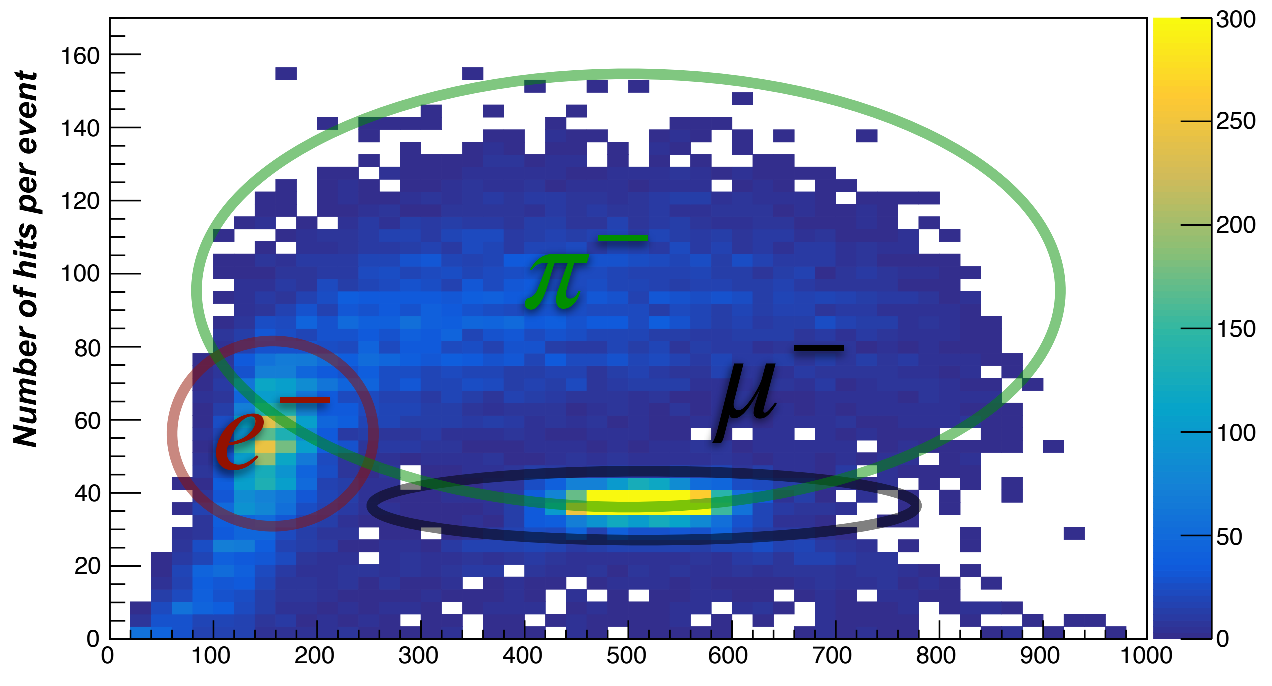


beam

Particle ID for beam tests

Motivation and goal

Example of standard data quality monitoring plot for 10GeV pion run



$$* \, z_{CoG} = \frac{\sum_{i=1}^{N_{hits}} z_i \cdot E_i}{E_{sum}}$$

We always deal with admixture of other particles in data runs.

⇒To investigate detector response to particles of given type we need to perform particle identification

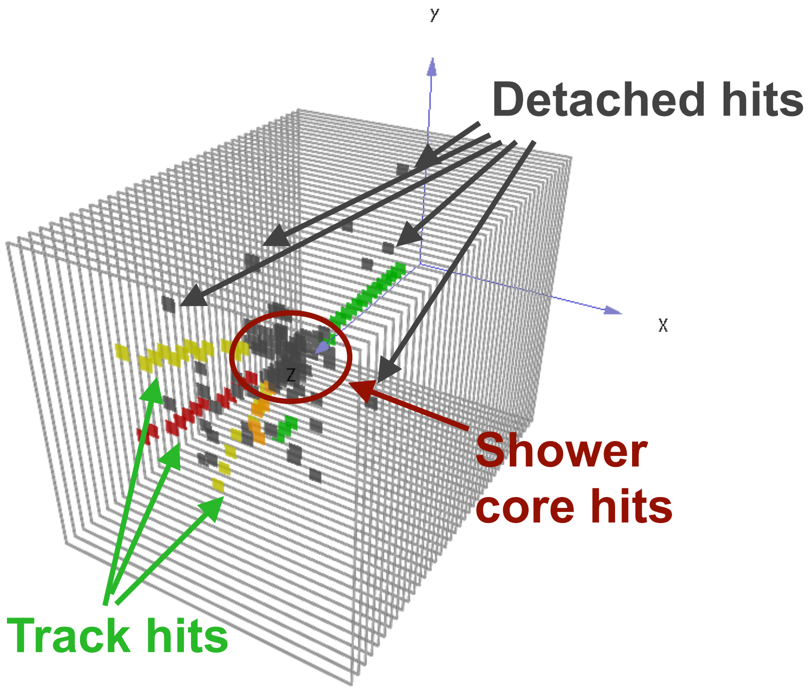
3 main categories:

- Hadron events (showering hadrons)
- Electron events
- Muon-like events (including punch-through hadrons)

Data pre-processing

Pre-analysis

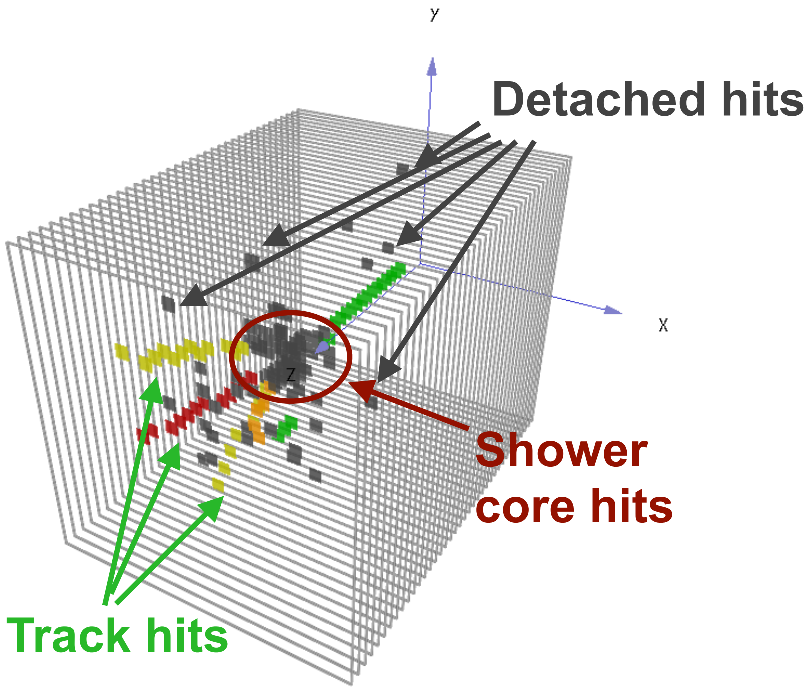
- Simple clustering and track finding algorithms to estimate event structure
- Calculation of observables used for training



Data pre-processing

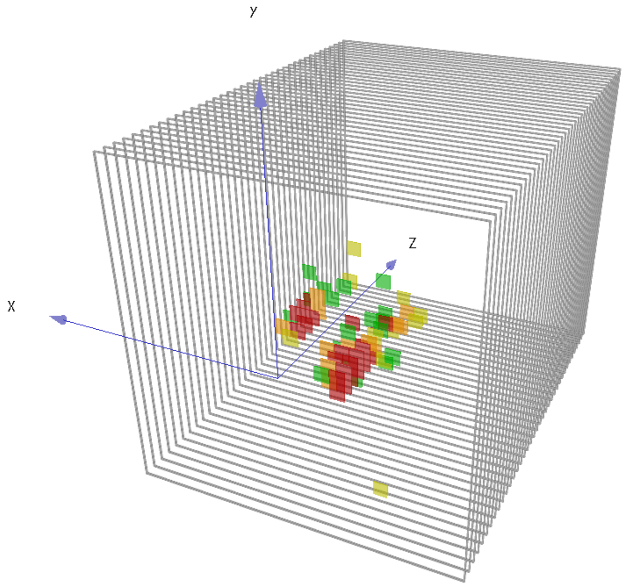
Pre-analysis

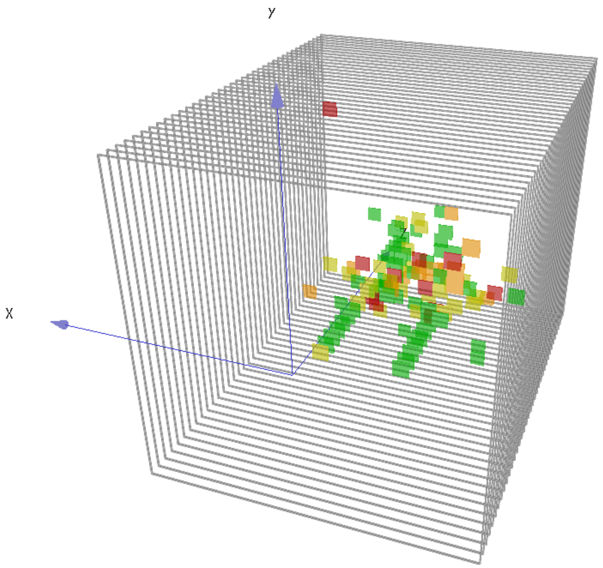
- Simple clustering and track finding algorithms to estimate event structure
- Calculation of observables used for training



Event filtering

- By **number of hits**:
  nHits > nHits_min
- **multi-particle** event rejection (analysing activity in first layers)





BDT classification

Model and input.

Software and model:

- **LightGBM** package
- Multi-class **Gradient Boosted Decision Tree**
- **Multi log** loss function

BDT classification

Model and input.

Software and model:

- **LightGBM** package
- Multi-class **Gradient Boosted Decision Tree**
- **Multi log** loss function

Decision Tree

Simplest machine learning predictive model that in case of classification splits labeled dataset by observable values (or features) in to separated leafs corresponding to given class labels.

Gradient Boosting:

Method combines many sequential decision trees. Each tree is trained to predict loss of previous one thus improving it's accuracy.

BDT classification

Model and input.

Software and model:

- **LightGBM** package
- Multi-class **Gradient Boosted Decision Tree**
- **Multi log** loss function

Training and test set:

- **MC** particles **10-200GeV** sumulated using *Geant4 (v10.03.p02)* QGSP_BERT_HP physics list:
  - **pions (st ≤ 40)**
  - **electrons**
  - **muons**
- Simulated data is split **50/50 - test/train**

Observables (sorted by importance):

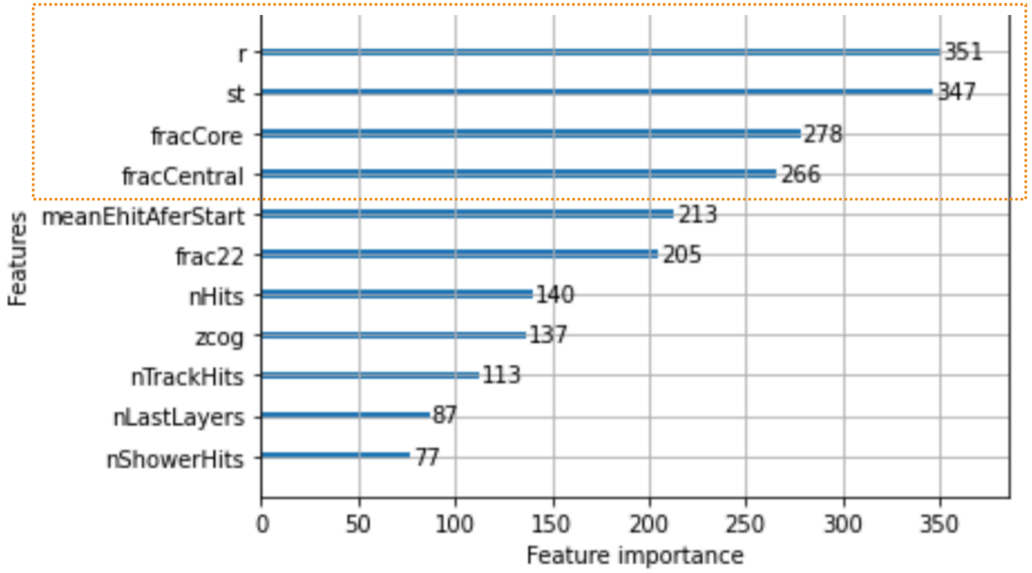
- Event radius
- Shower start layer number
- Energy fraction in shower core
- Energy fraction in shower central region (in XY plane)
- Mean hit energy after shower start
- Energy fraction in first 22 layers
- Number of hits
- Center of gravity in z
- Number of track hits
- Number of layers with hits from last 5
- Number of hits after shower start

BDT classification

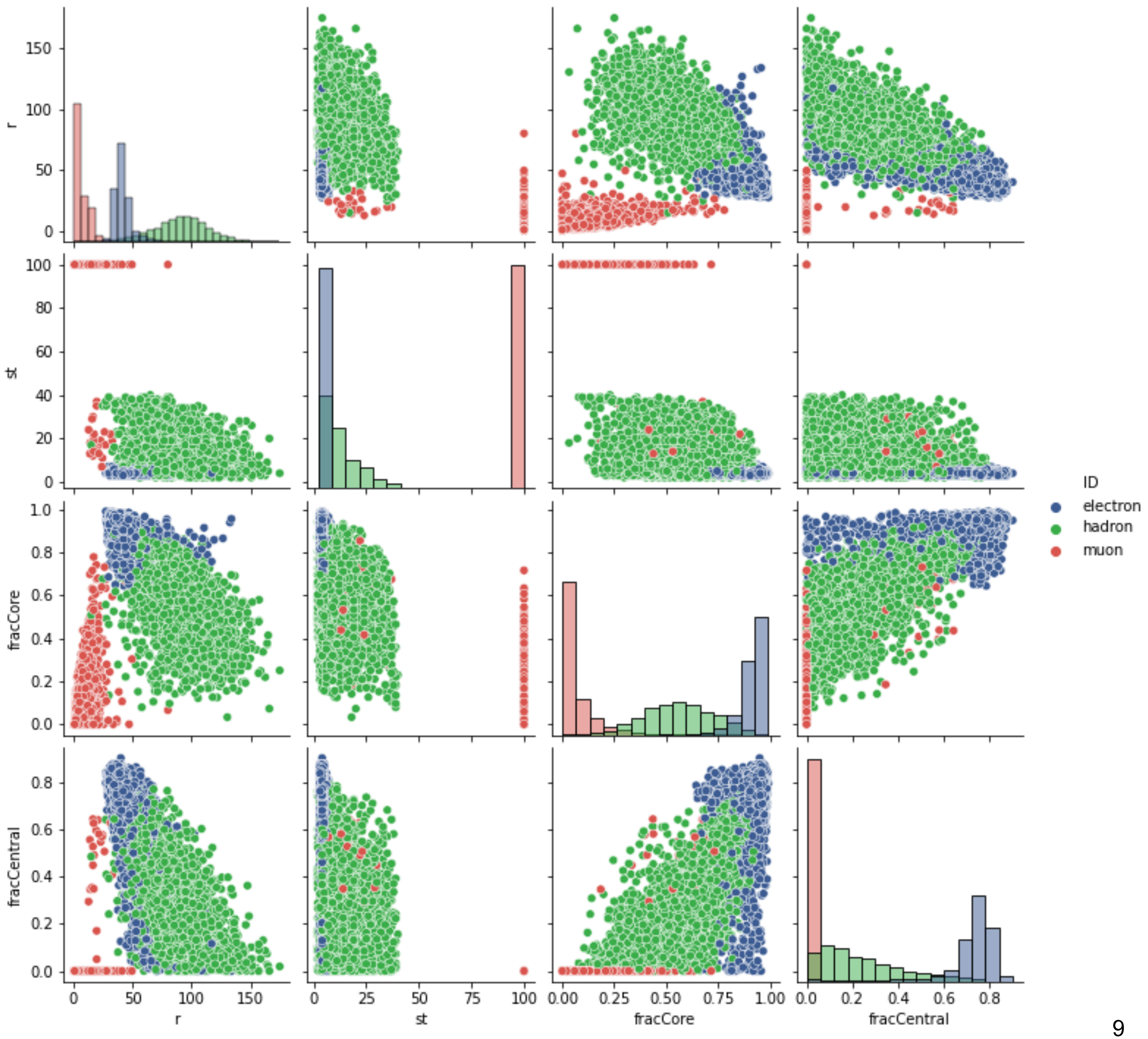
Input variables.

Observables (4 most important):
- Event radius - **r**
- Shower start layer number - **st** (if shower start was not found *st=100*)
- Energy fraction in shower core - **fracCore**
- Energy fraction in shower central region after shower start in XY plane - **fracCentral**

Observable importance calculated by number of splits



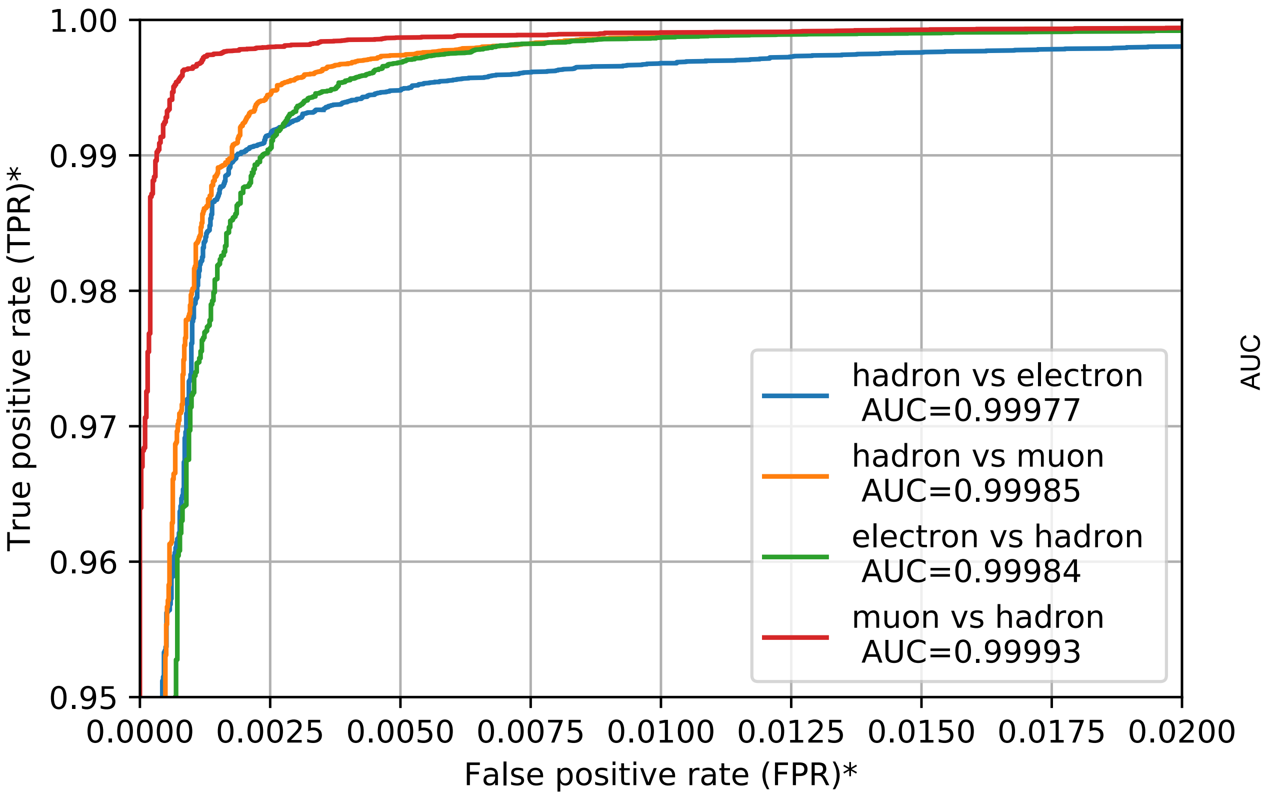
Pairplot for most important observables



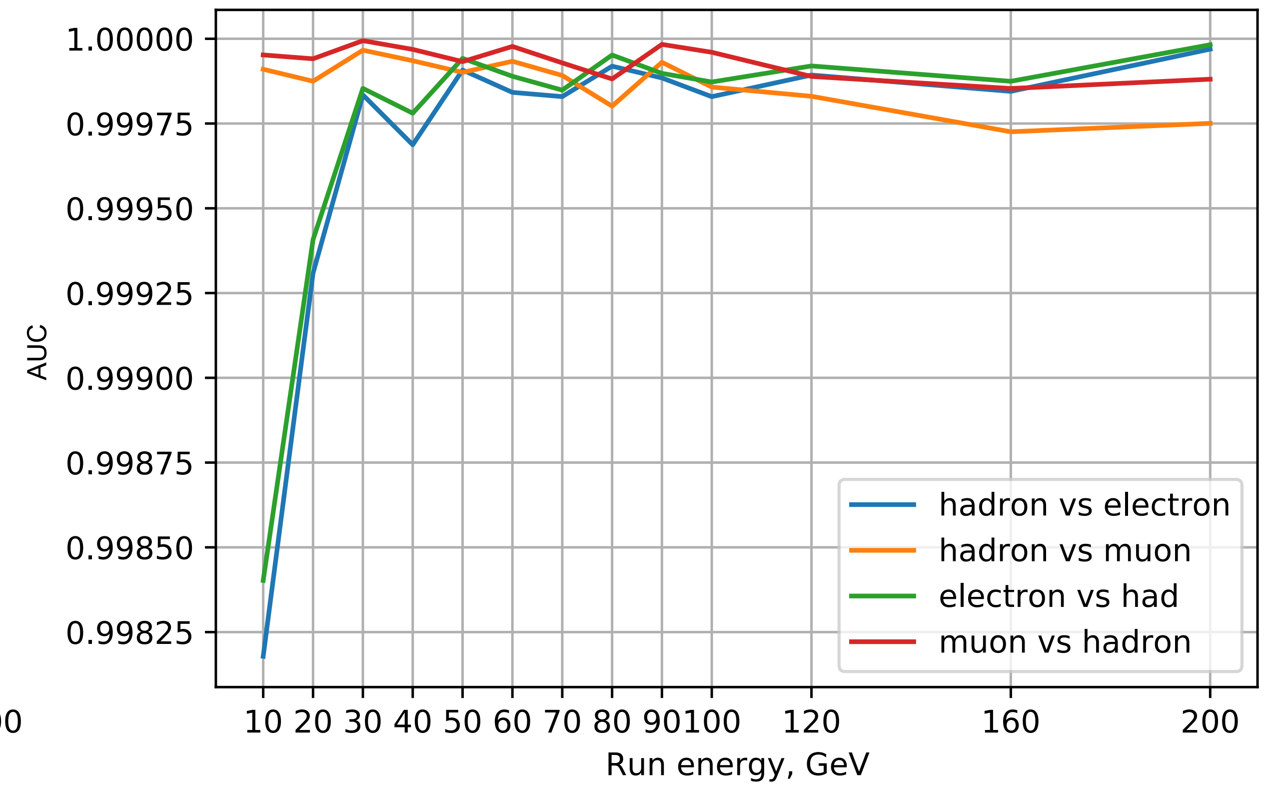
Resulting metrics

On Monte-Carlo test sample

ROC curves for the test data



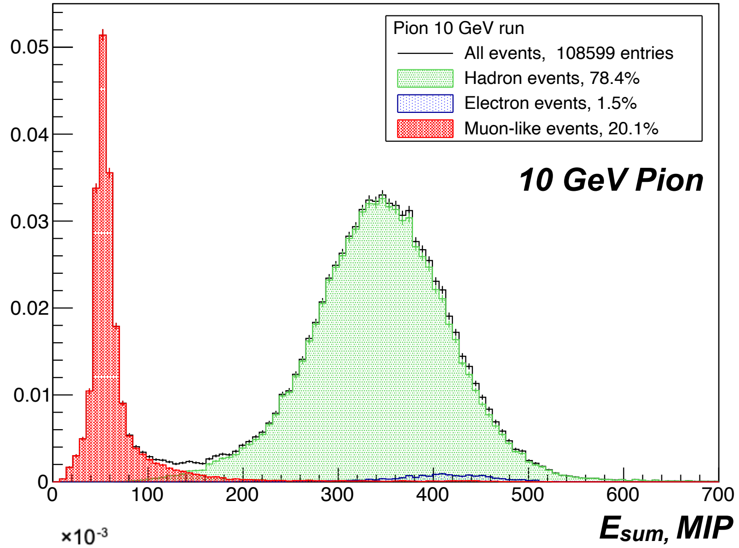
Model AUC for different energies

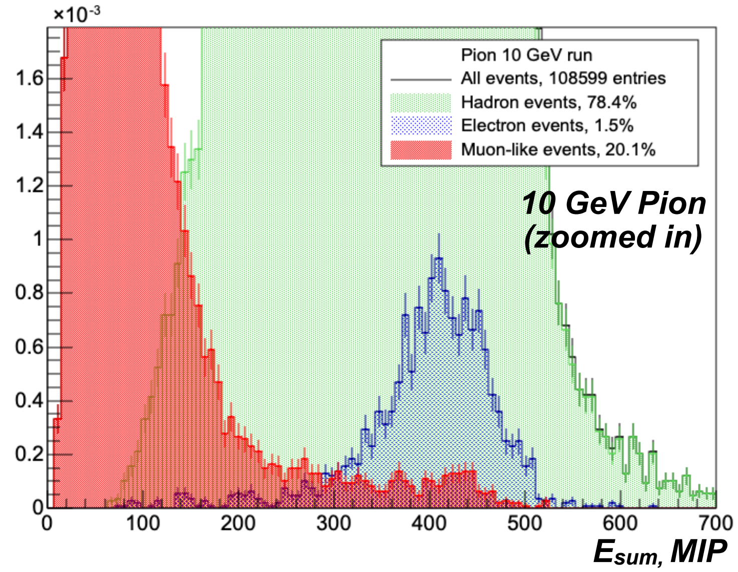


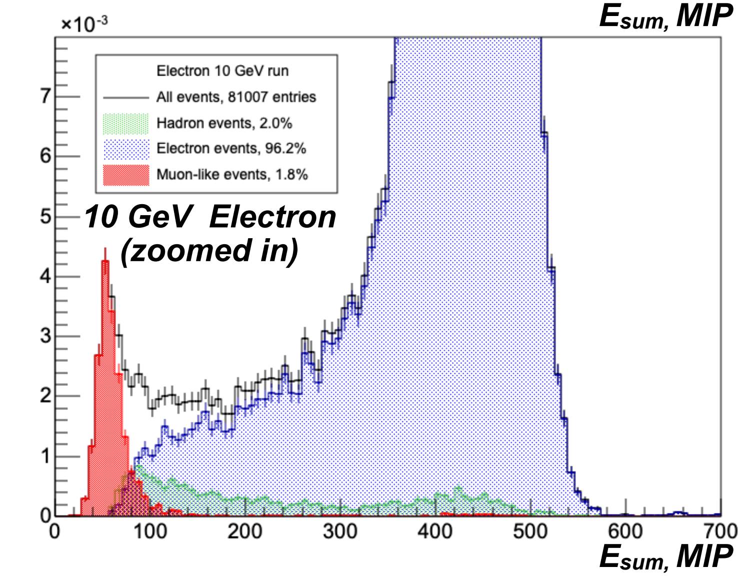
$$*TPR = \frac{TP}{TP + FN}, \quad FPR = \frac{FP}{FP + TN}$$

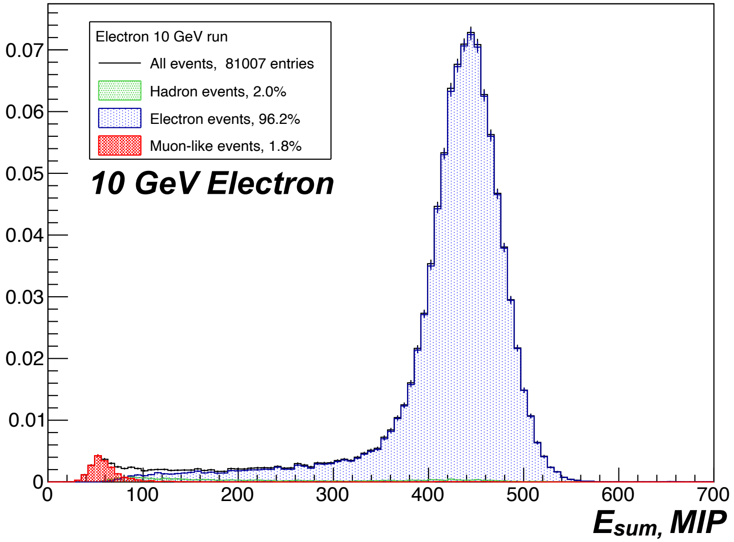
Results on test beam data taken in June 2018

Energy sum distributions for 10GeV runs







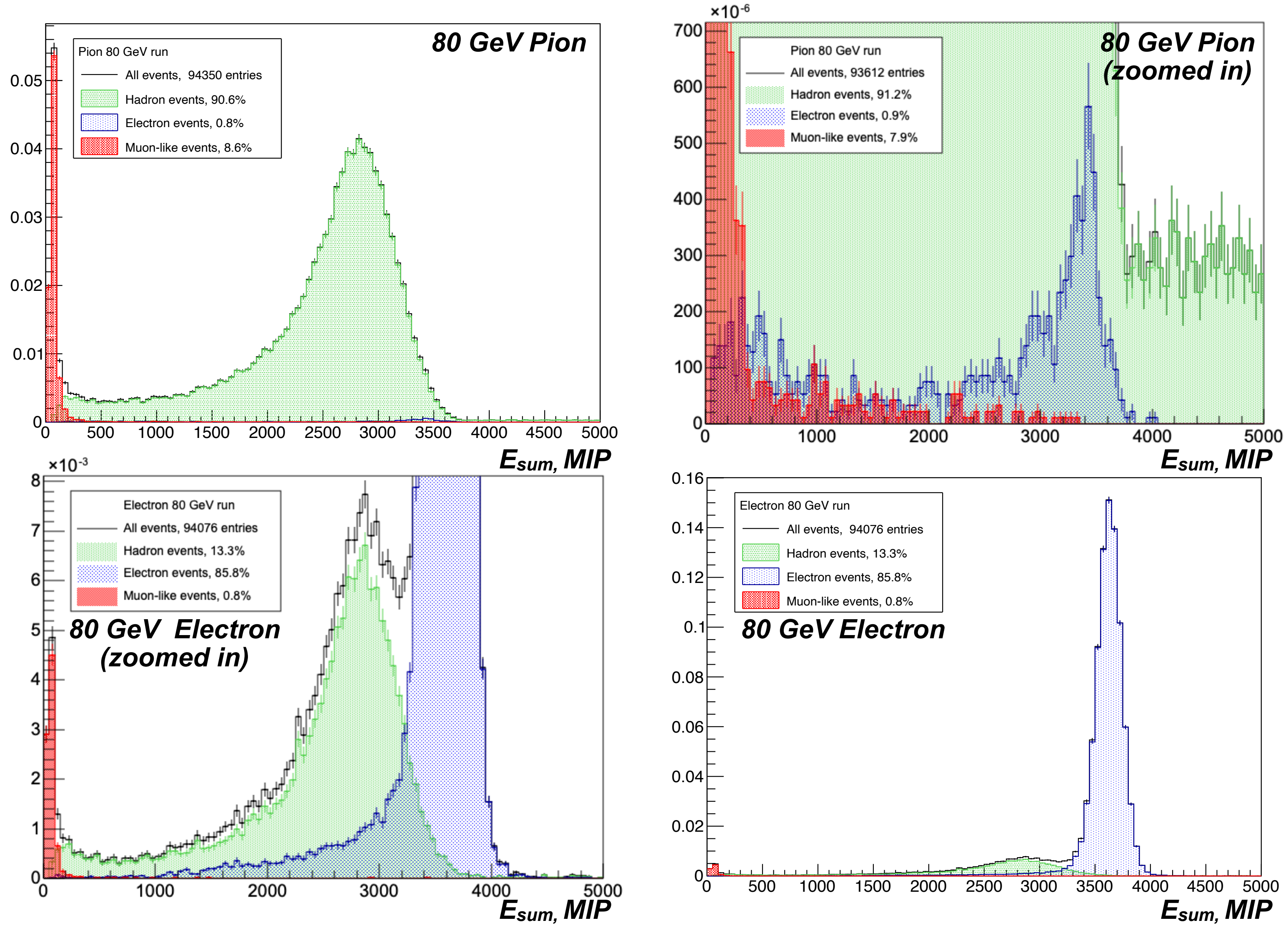


- Energy expectation for electron events in pion run is close to real electron run
- Long high energy tail of muon-like events
- Low energy tail for electrons
- Most of hadron events in electron run are at low energy

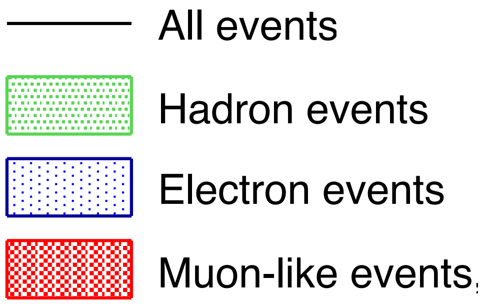
All events
Hadron events
Electron events
Muon-like events

Results on test beam data taken in June 2018

Energy sum distributions for 80GeV runs

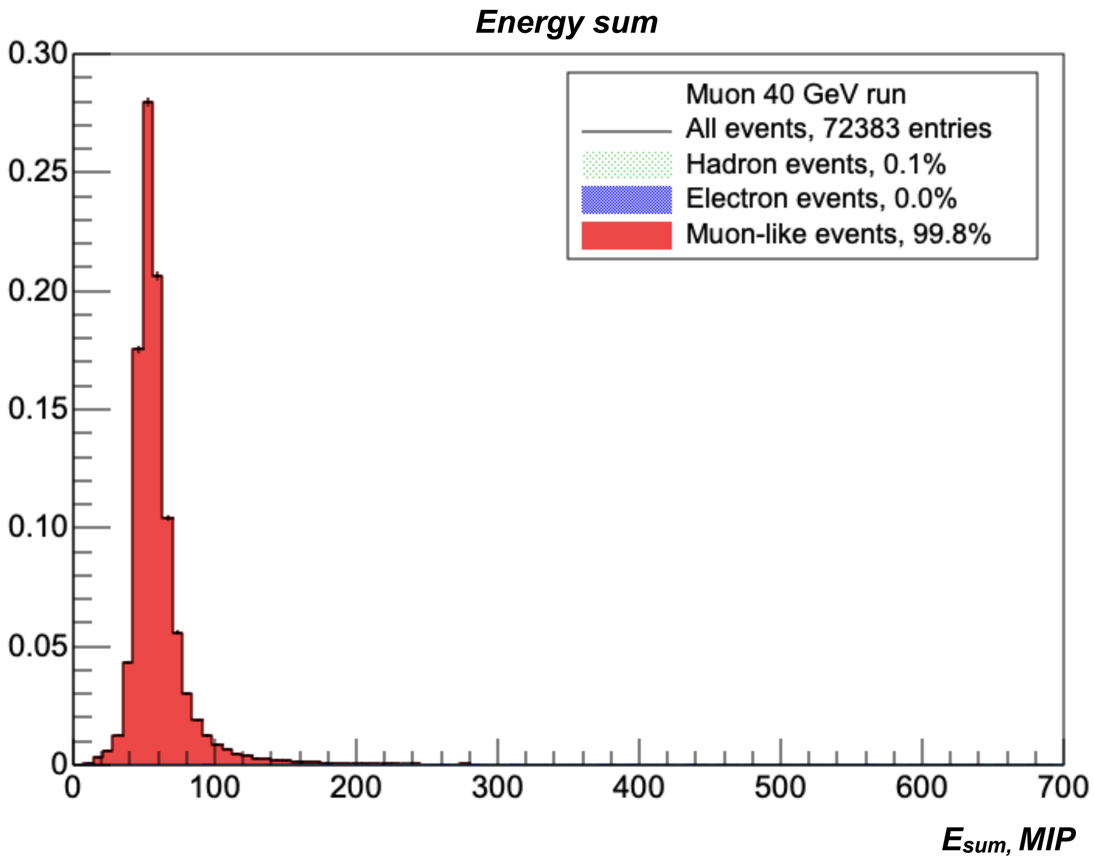


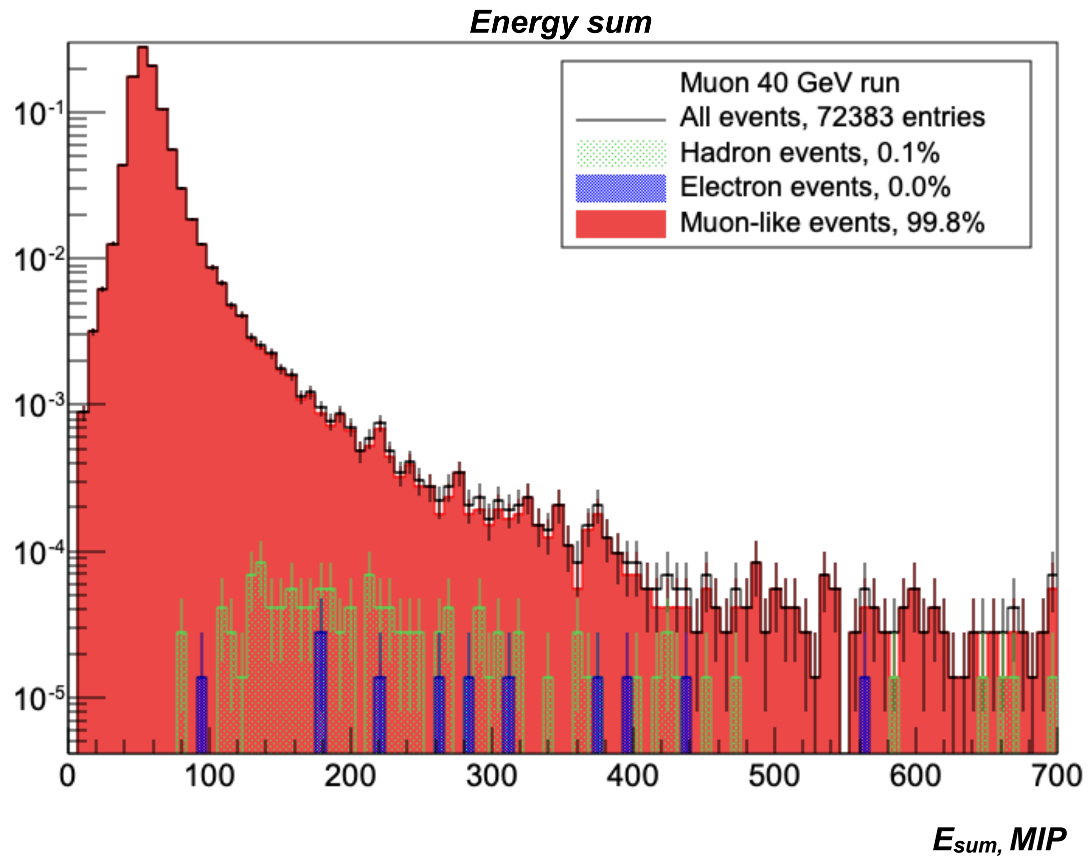
- Energy expectation for electron events in pion run is close to real electron run
- Energy distribution of hadron events in 80GeV electron run looks very similar to actual 80GeV pion



Results on test beam data taken in June 2018

Energy sum distribution for 40GeV muon run

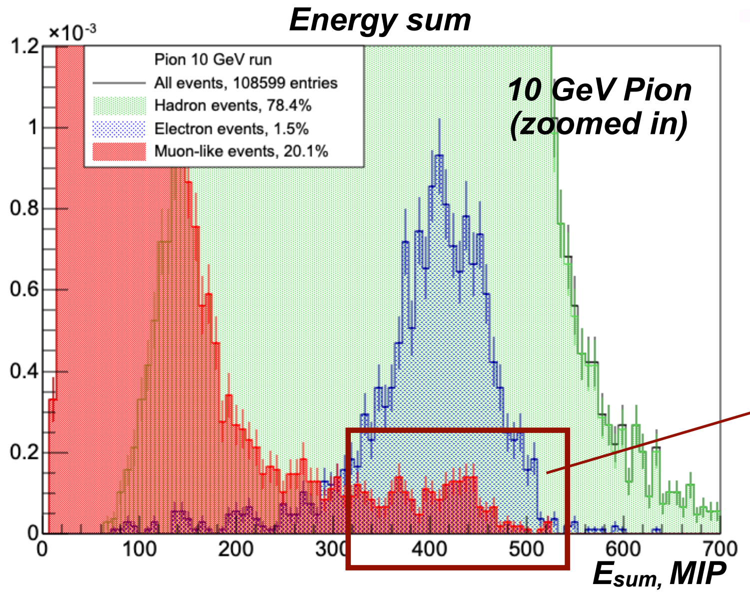


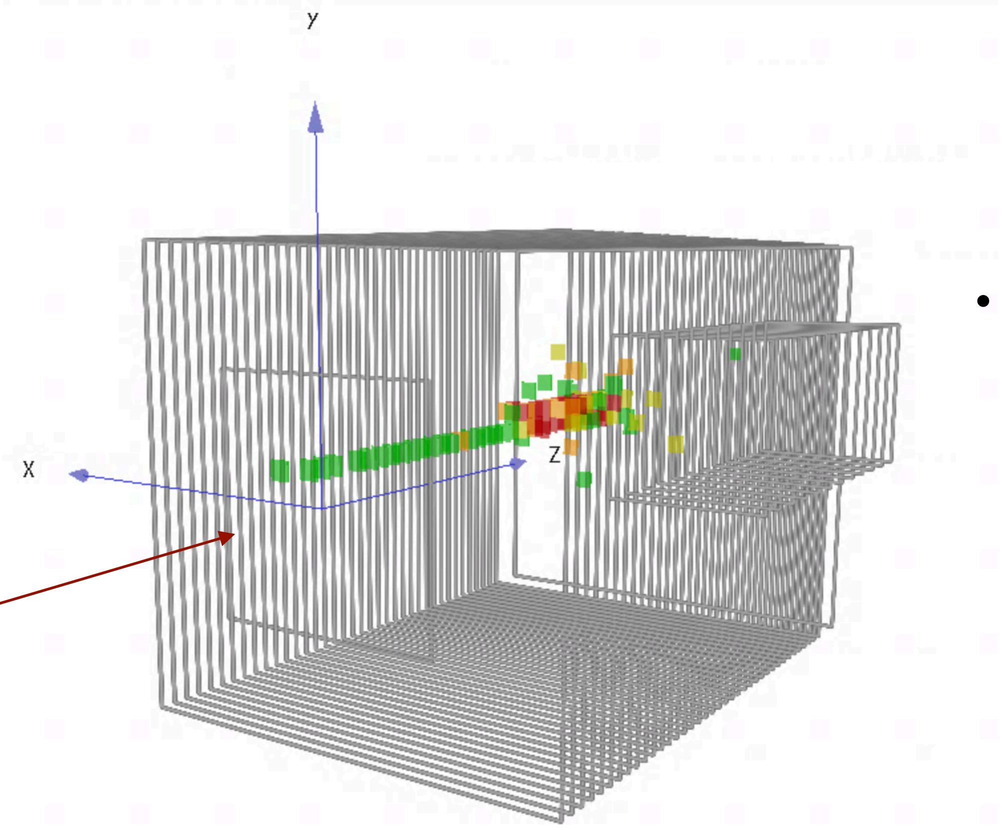


- Very low admixture of other particles
- Little fraction of delta electrons can be classified as hadron event

Sources of confusion

From 10GeV pion run





Muon-like event :

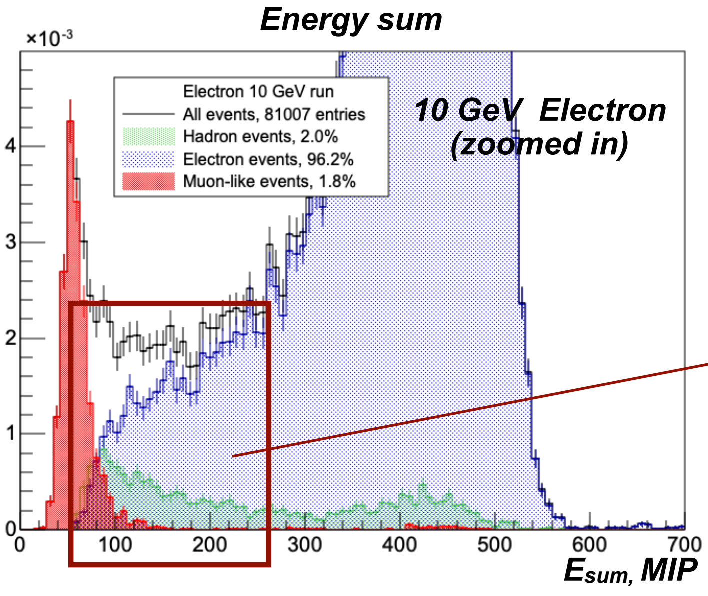
Mu-like score is 0.51

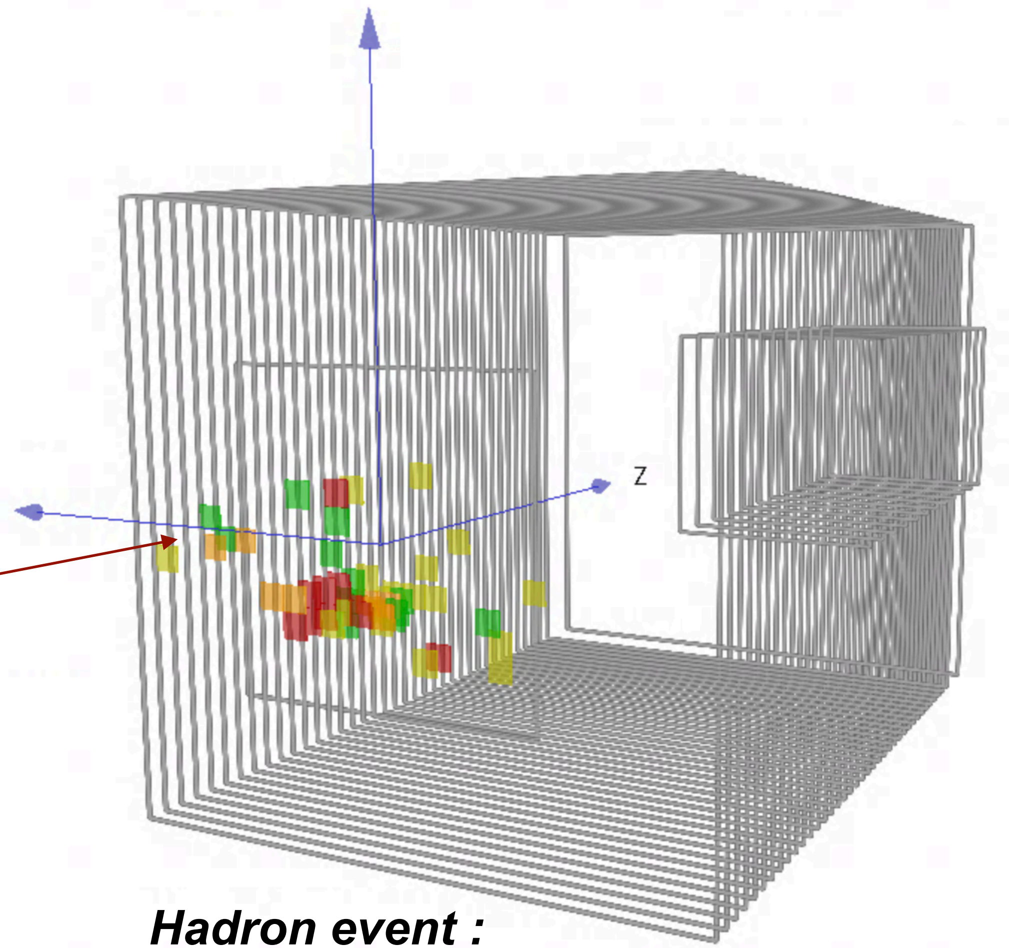
Had score is 0.48

- Compact pion showers with late shower start can be classified as muons
  - Additional variables can improve identification
  - Fraction << 1%

Sources of confusion

From 10GeV electron run





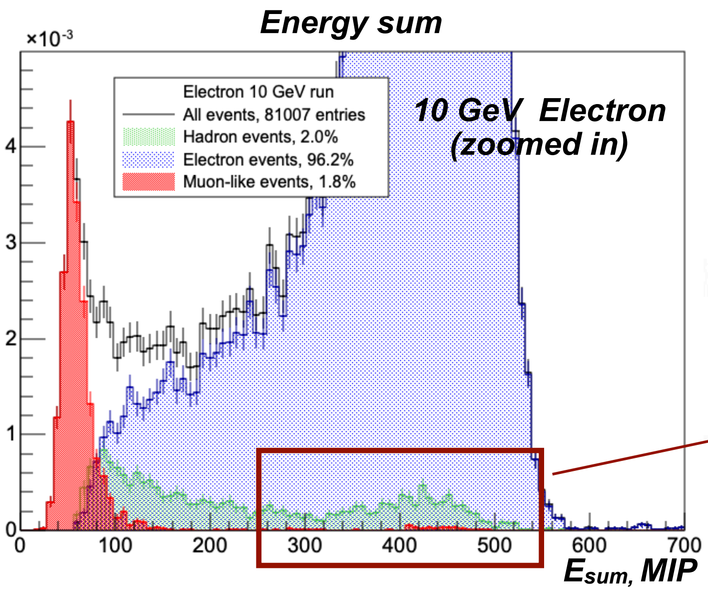
Hadron event :

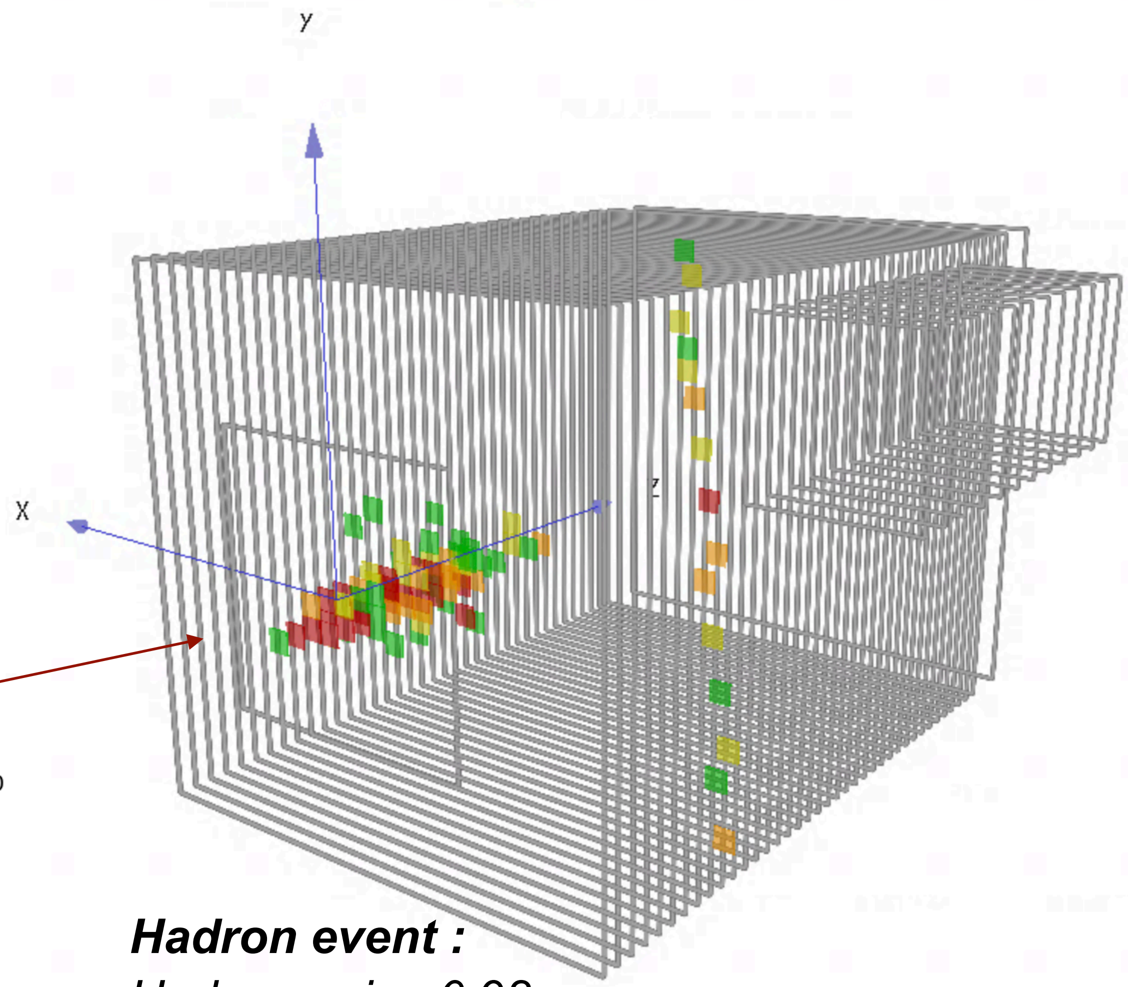
Had score is ~0.9

- Multi-particle/upstream shower events with small fragments can be classified as hadron events
    - Multi-particle events can be partly filtered out using timing information

Sources of confusion

From 10GeV electron run





Hadron event :

Had score is ~0.98

- Some events are contaminated with cosmic muons
    - Multi-particle events can be partly filtered out using timing information

Summary

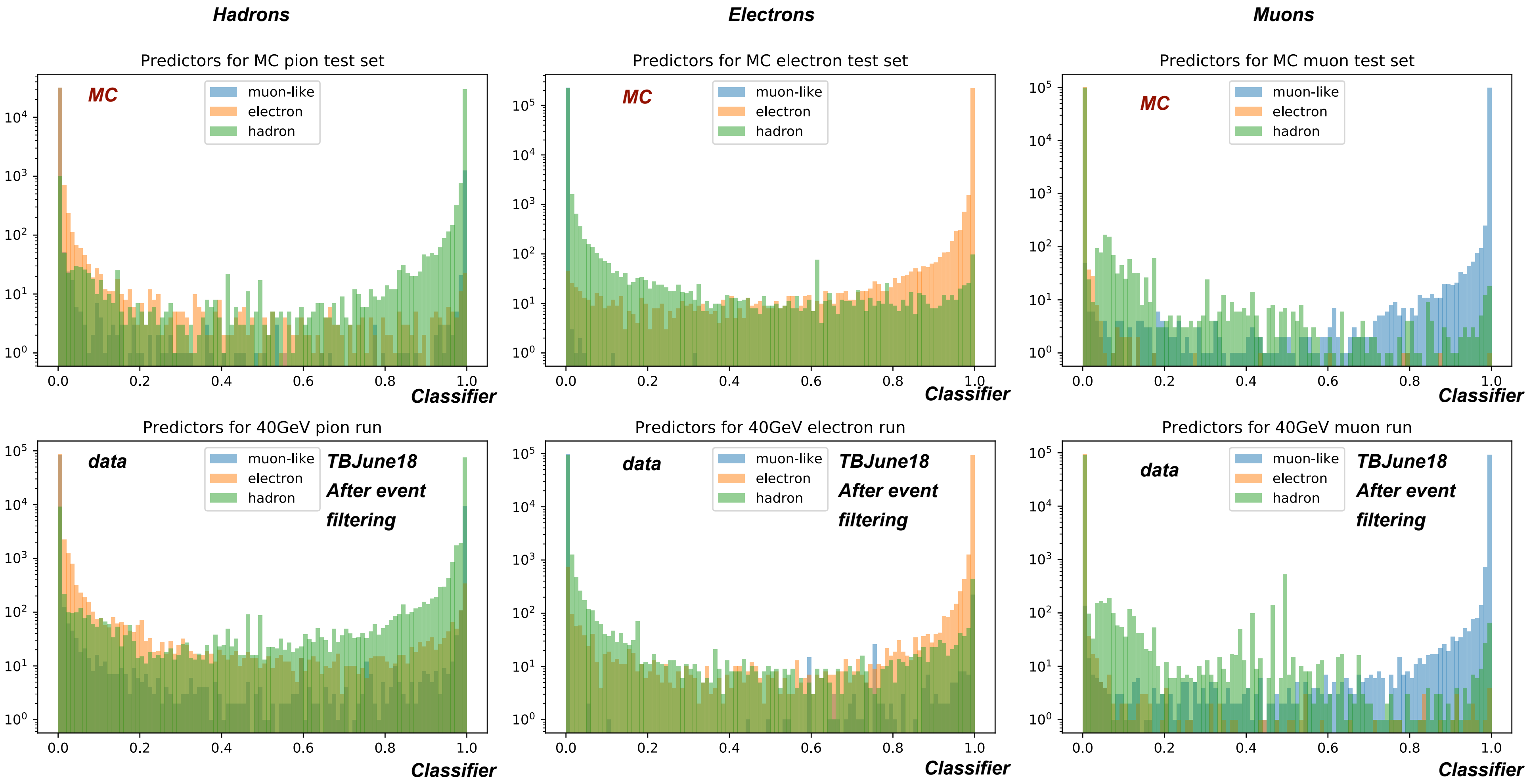
AHCAL Particle ID using BDTs

- High granularity provides detailed information of event structure to separate different particle type
- BDT particle ID method shows excellent performance on simulations and reasonable results on data
  - Main sources of confusion are understood and can be improved with more advanced event filtering

Backup

BDT classification

Output. Comparison with data.



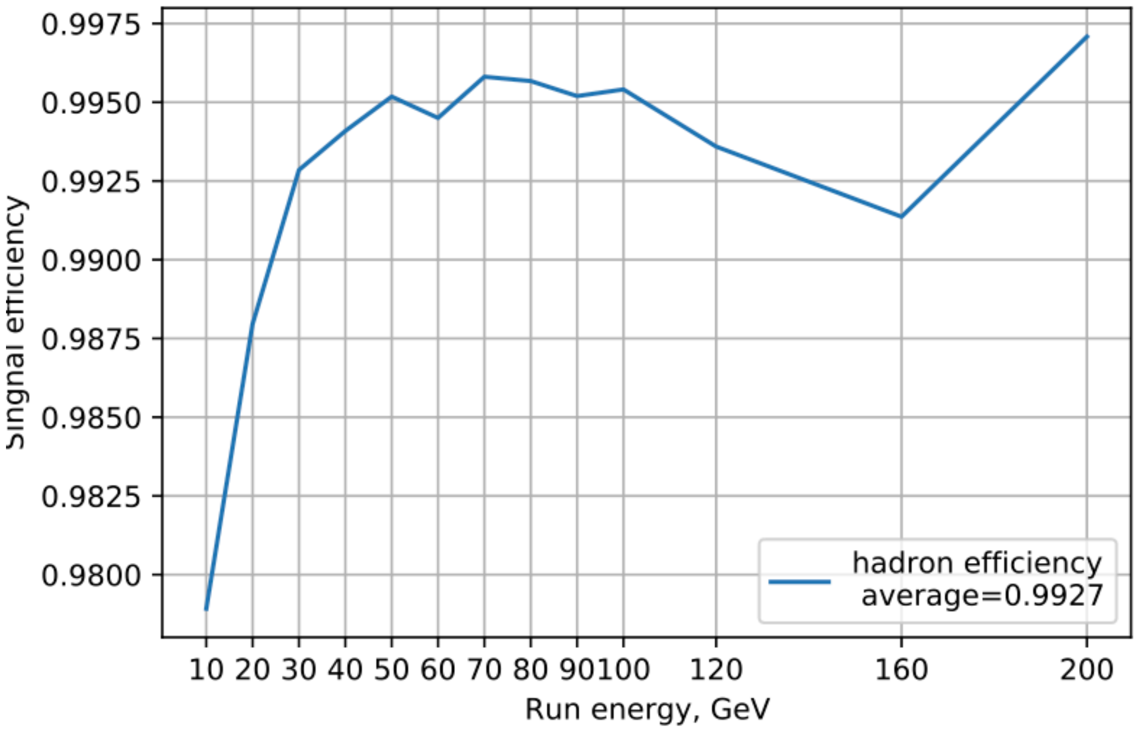
- Similar response on data and simulations

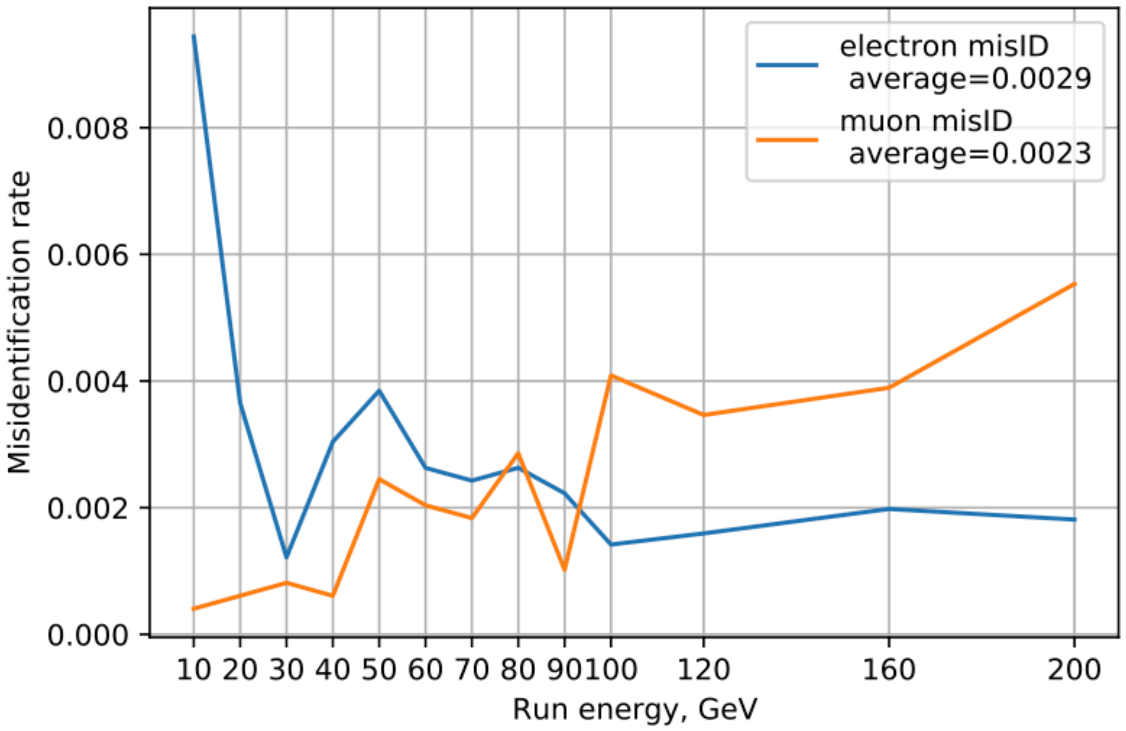
Classifiers:

- muon-like
- electron
- hadron

Resulting metrics

On Monte-Carlo test sample





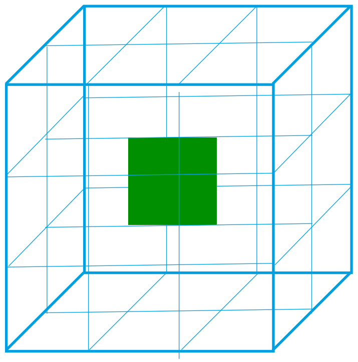
Multi log loss:

$$L = -\frac{1}{N}\sum_{i}^{N}\sum_{j}^{3} Y_{ij} ln(p_{ij})$$

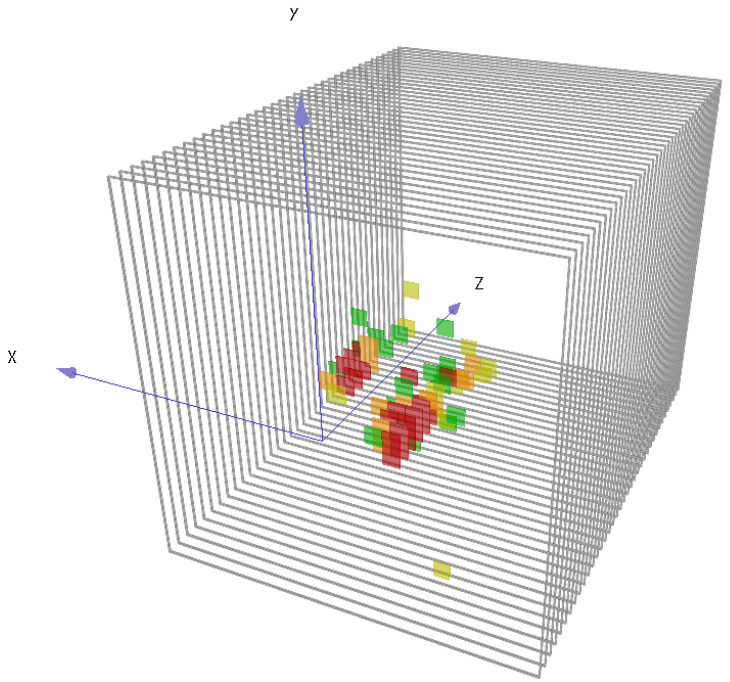
Where N - number of events in the data sample, 3 - number of classes, Y_{ij} is binary variable with the expected labels and p_{ij} is the classification probability output by the classifier for the i-instance and the j-label.

Event filtering

Simplified algorithms.



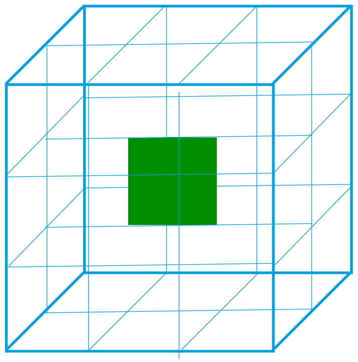
Clustering: Hits are grouped in clusters if if they are neighbours in volume. First 5 layers are taken into account



If $N_{Clusters}$ > *1* => multi-particle event (or upstream shower)

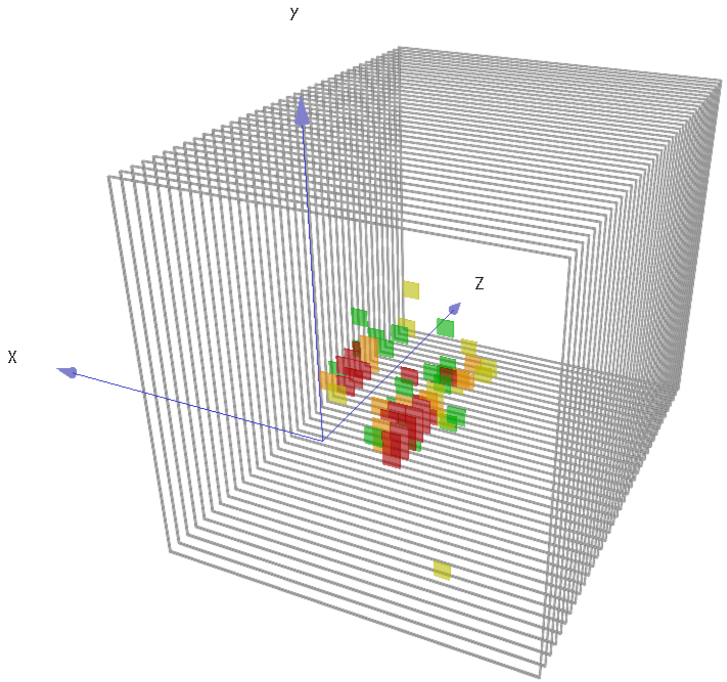
Event filtering

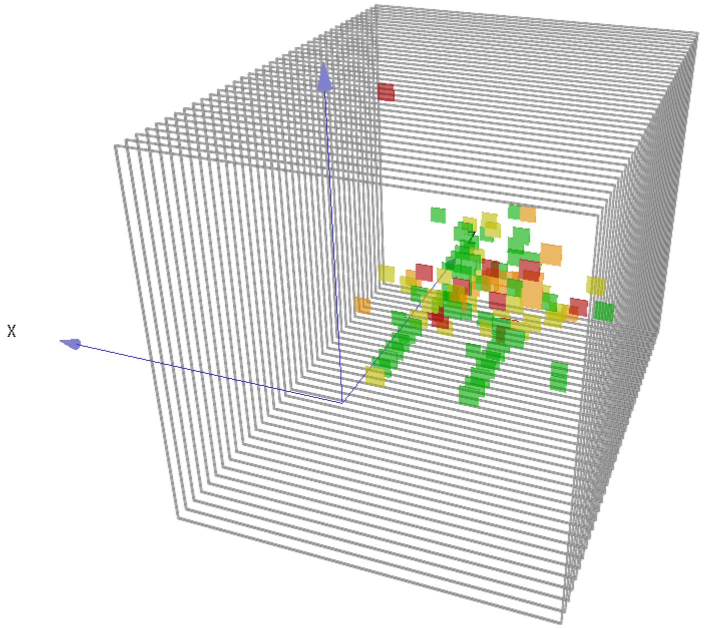
Simplified algorithms.



Clustering: Hits are grouped in clusters if if they are neighbours in volume. First 5 layers are taken into account

If $N_{Clusters} > 1$ => multi-particle event (or upstream shower)





MIP tracking: Construct towers with same x and y coordinates. First 5 layers are taken into account.

If $N_{MIPTracks} > 1$ => multi-particle event

Disadvantages of cut-based method

Towards BDT ID

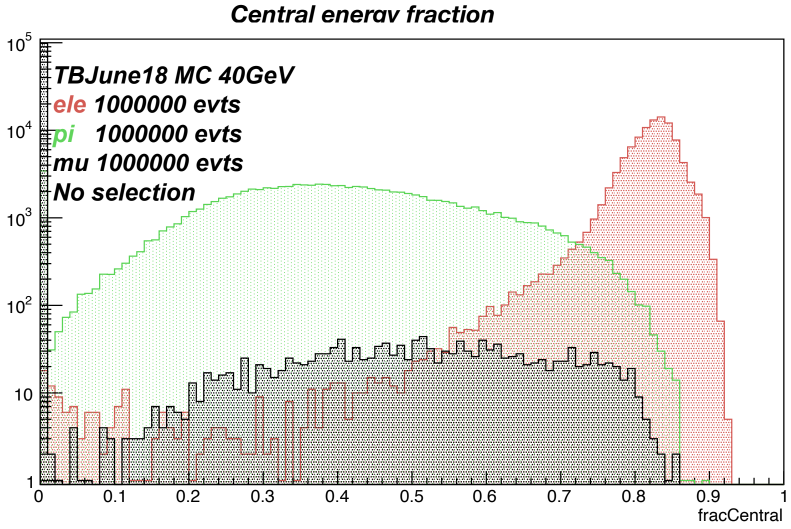
Cut-based method:

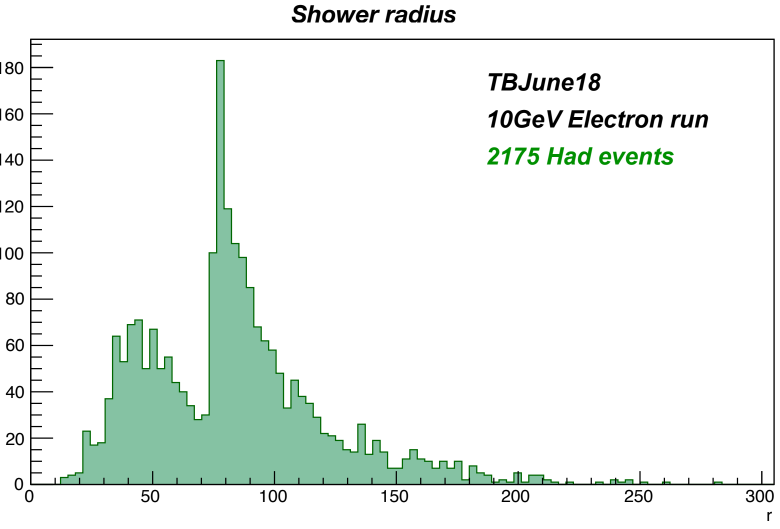
- > 10 steering parameters for each energy
- Asymmetric distributions/ long tails with overlay can be problematic

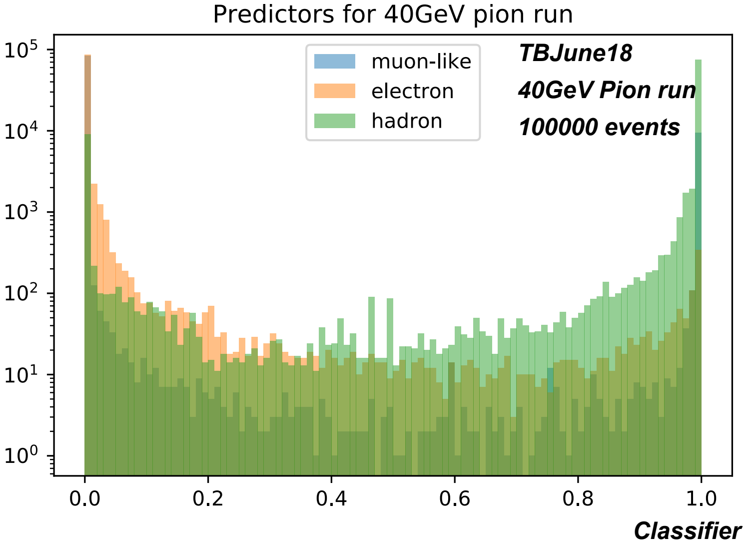
- Cut artefacts

Multivariate methods:

- Can provide probabilistic classifier trained on given distributions of observables
- One model can be used for whole dataset



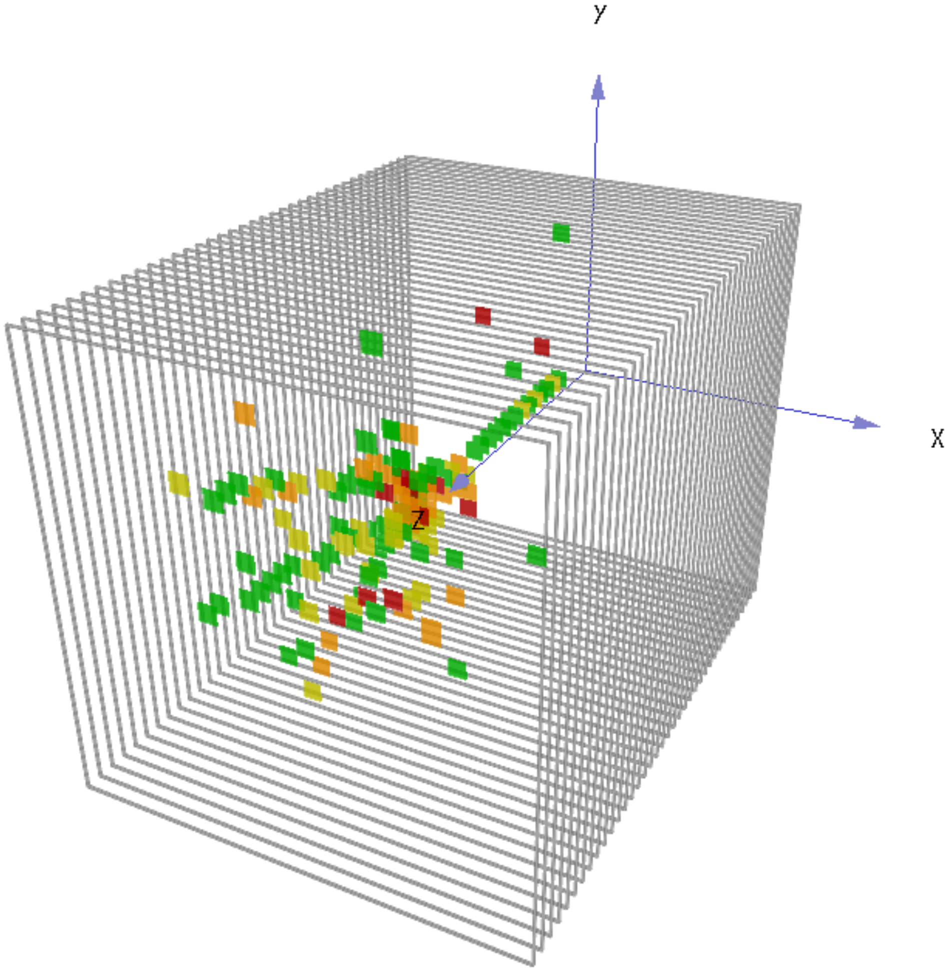




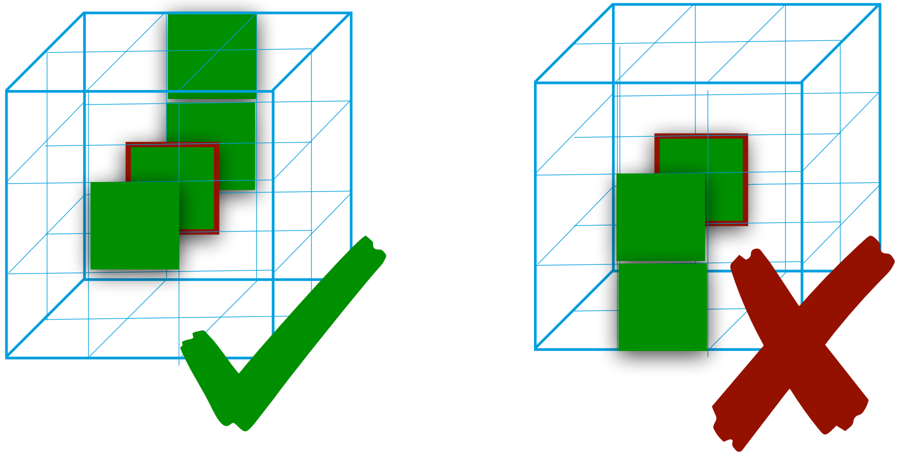
Will be discussed during one of the upcoming HGCAL meetings

Track finding

Important tool for shower characterisation,
Can be used for particle ID



Track candidates:
2/3 neighbours in surrounding volume. 2 of them on different sides



Candidates ordered:
- z-coordinate
- Distance to (0,0,z) in same layer

Track finding

Grouping candidates into tracks

Hit#1
A
Seed
(first from candidates)

Nearest neighbour of hit#1

Distance check

Hit#2
B

Nearest neighbour of hit#1

Distance check

Angle check with **AB**

Hit#3
C

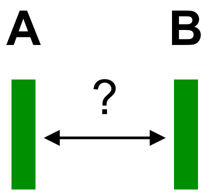
Nearest neighbour of hit#1

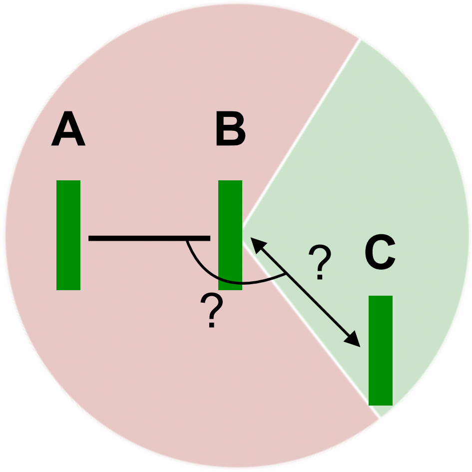
Distance check

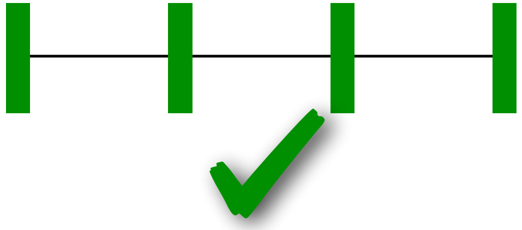
Angle check with **AC**

Angle check with **AB**

(…)





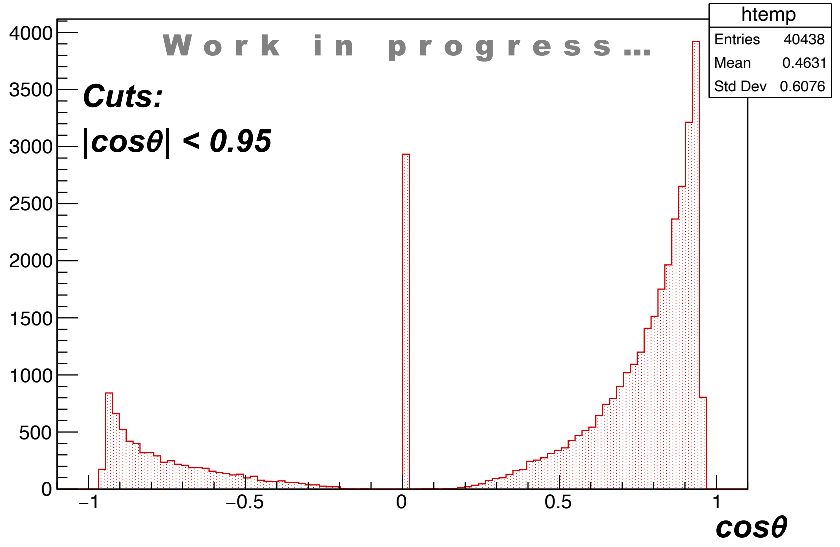


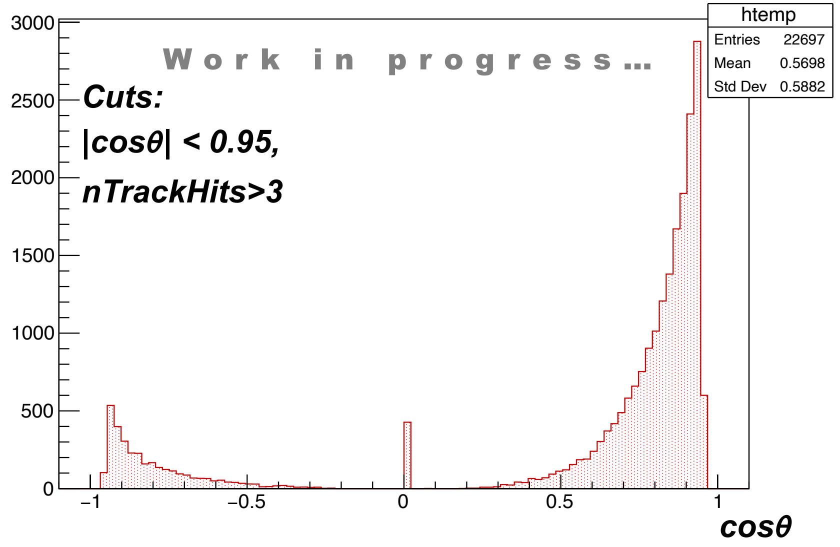
After grouping, track angle is obtained using MSE linear regression

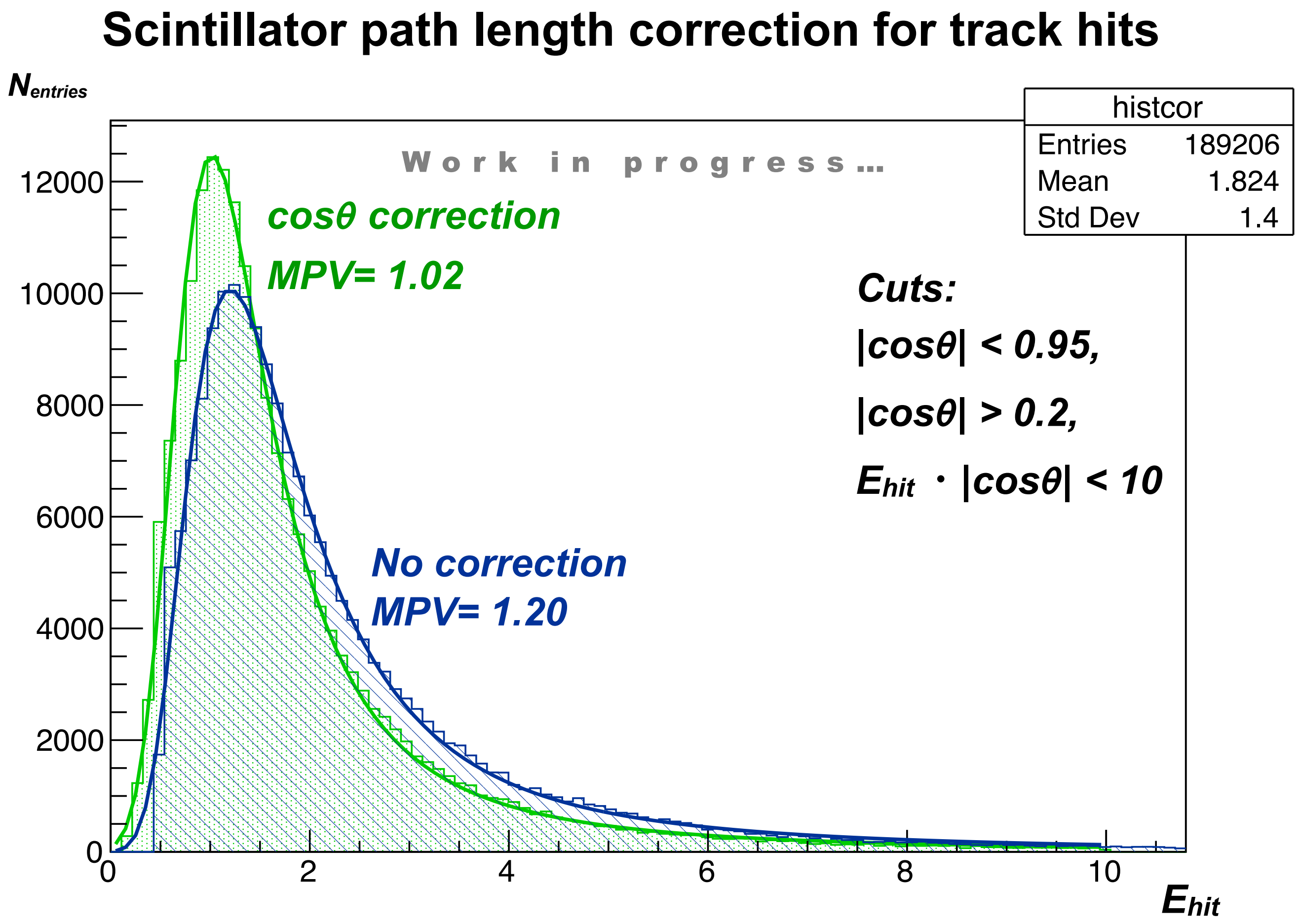
** Procedure repeated iteratively **

Tracking quality check

TBMay18 10GeV pion run. 50039 events.

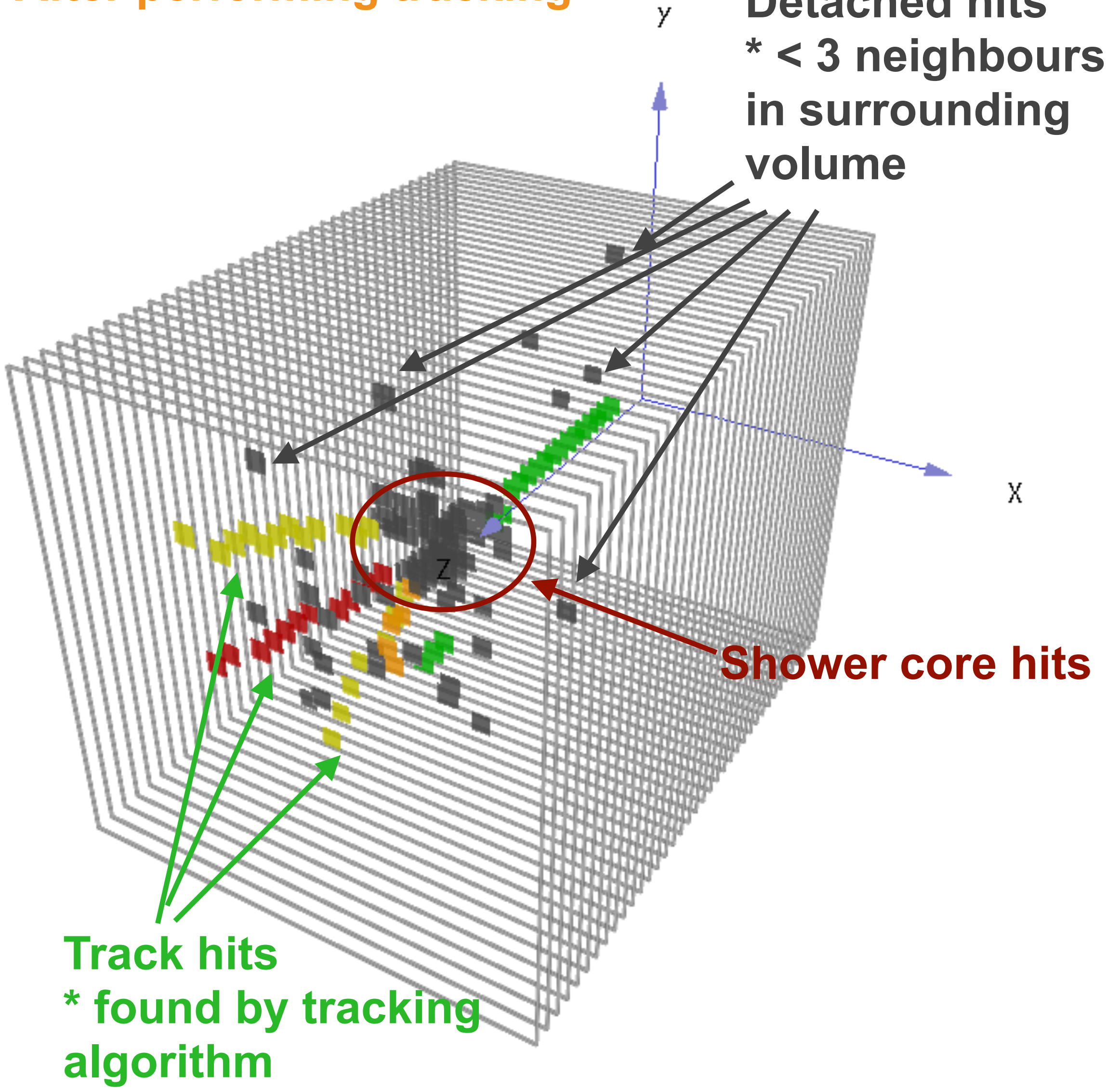






Resulting ID variables

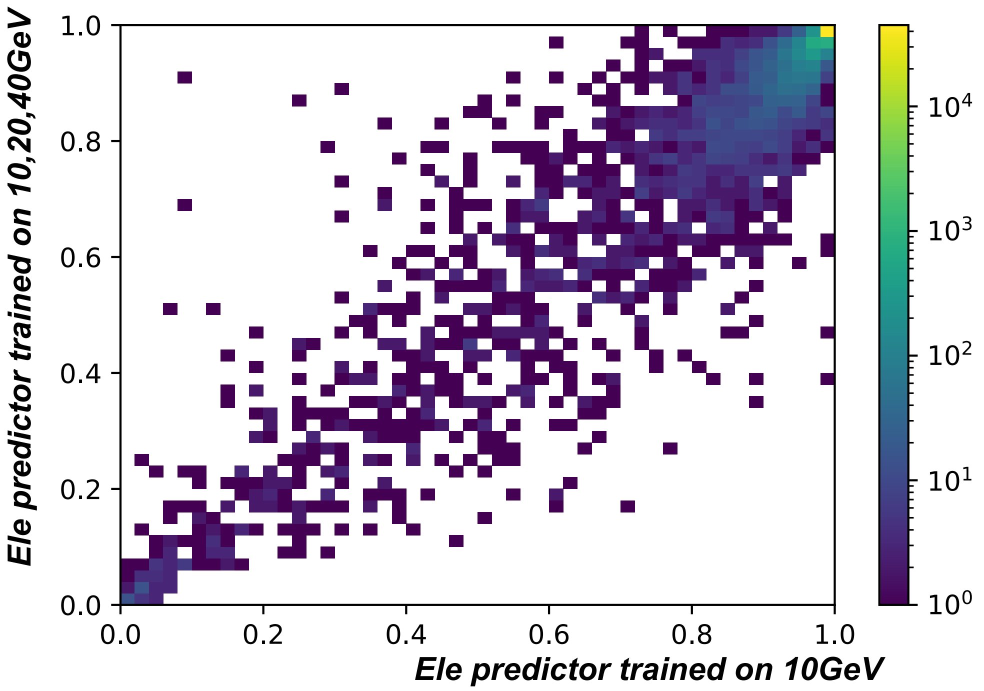
After performing tracking



BDT output

Comparison with separate model trained only on 10GeV particles.

10GeV MC electron test sample
50000 events



10GeV MC pion test sample
50000 events