Particle ID in the AHCAL
+ Single Shower Substructure (bonus)

CALICE Collaboration Meeting

Vladimir Bocharnikov, DESY
Sep 30, 2020

HELMHOLTZ RESEARCH FOR GRAND CHALLENGES

P.N.Lebedev Physical Institute of the Russian Academy of Science

NATIONAL RESEARCH NUCLEAR UNIVERSITY NRNU MEPhI

DESY.

Outline

AHCAL Particle ID using BDTs

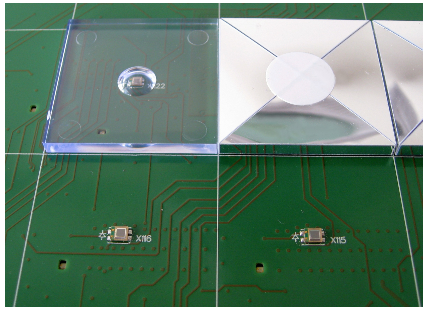
- **Particle identification**
  - Motivation and method overview
  - Data preparation
  - Boosted Decision Tree method description
    - Parameters and input
    - Resulting metrics
  - Application to test beam data
  - Summary and outlook (1st part)
- Bonus: **Detailed single hadron shower structure**
  - Motivation
  - Method and data preparation
  - Challenges
- Summary and outlook

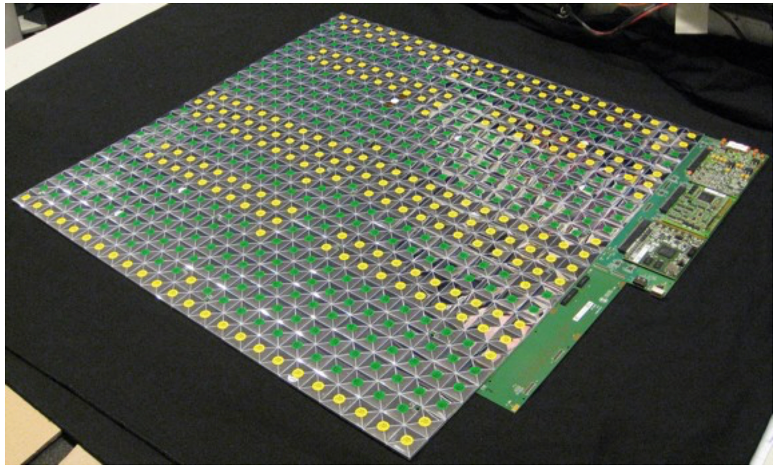
CALICE AHCAL

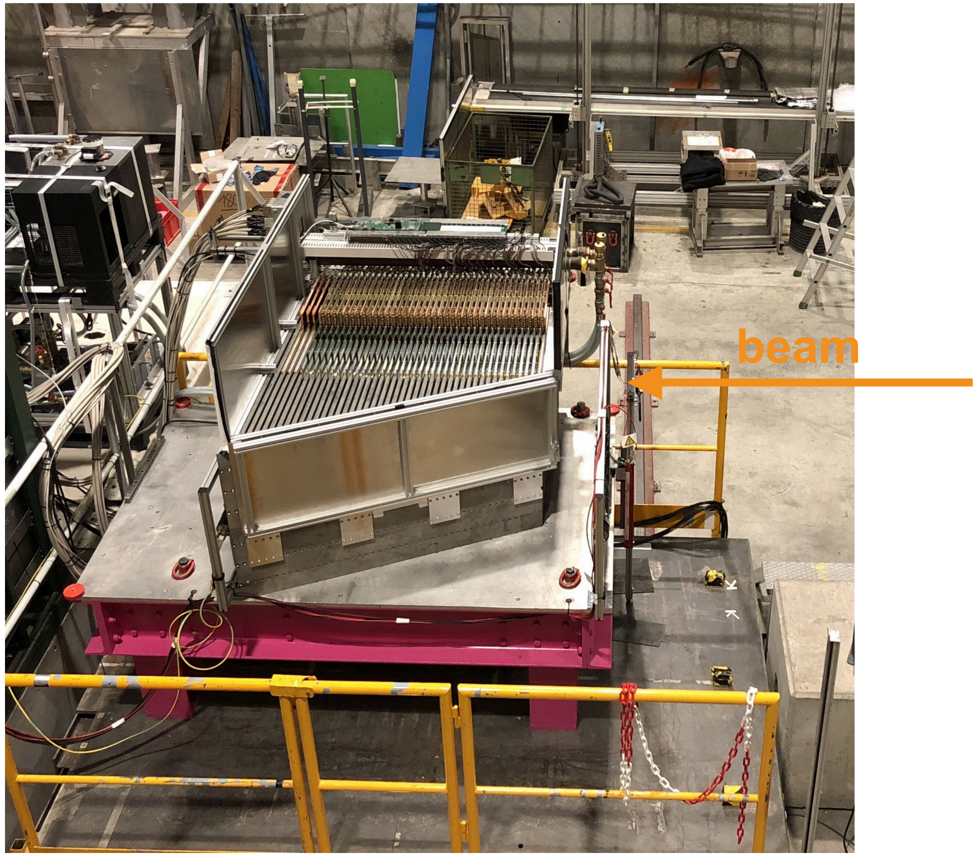
Test beam prototype.

38 active layers of 24x24 scintillator tiles (3x3 cm^2) alternating with 1.7 cm steel absorber + 1 “Tokyo” layer with 6x6 cm^2 tiles

In total: ~22000 channels, ~4 λ

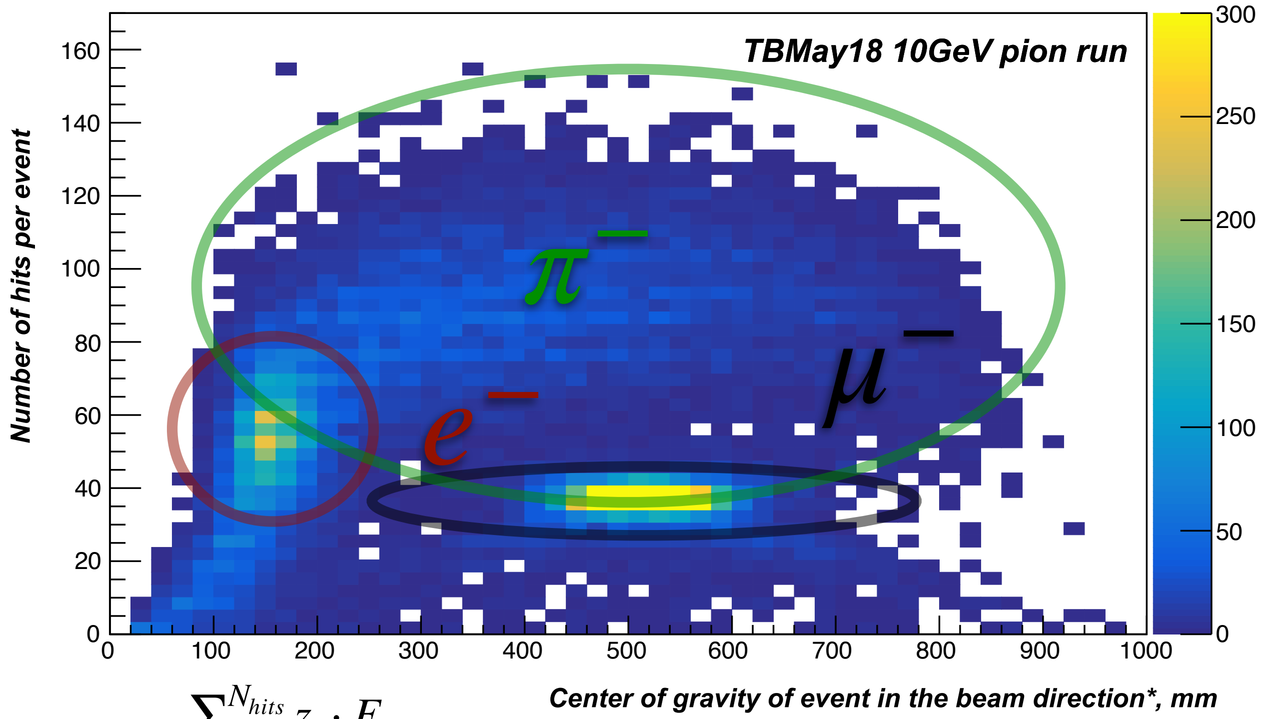






Motivation for particle ID

In test beam data



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

We always deal with admixture of other particles.

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

Particle ID workflow

Classification procedure

Event

Pre-analysis

- Calculation of common observables
- Clustering and track finding*

Event filtering

- By **number of hits**:
  nHits > nHits_min
- **multi-particle** and **upstream**

BDT multiclass model

trained on simulations (10-200GeV).

3 classifiers:

Hadron classifier
- Trained on showering pions

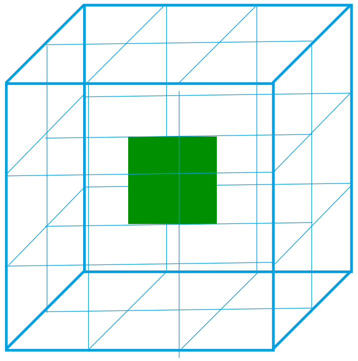
Electron classifier
- Trained on electrons

Muon (muon-like) classifier
- Trained on muons

* Described during CALICE Collaboration Meeting at CERN: https://agenda.linearcollider.org/event/8213/contributions/44343/attachments/34812/53758/VBocharnikov_CALICE_meeting_CERN.pdf

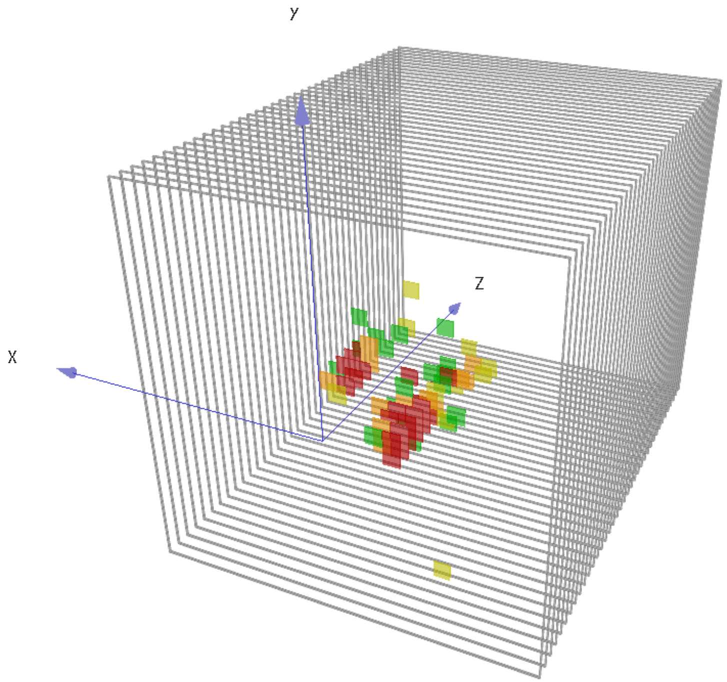
Event filtering

Simplified algorithms.



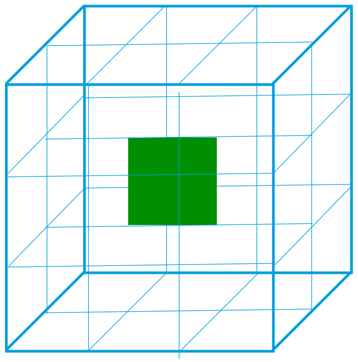
Clustering: Hits are grouped in clusters if if they are neighbours in volume. First 5 layers are taken into account

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



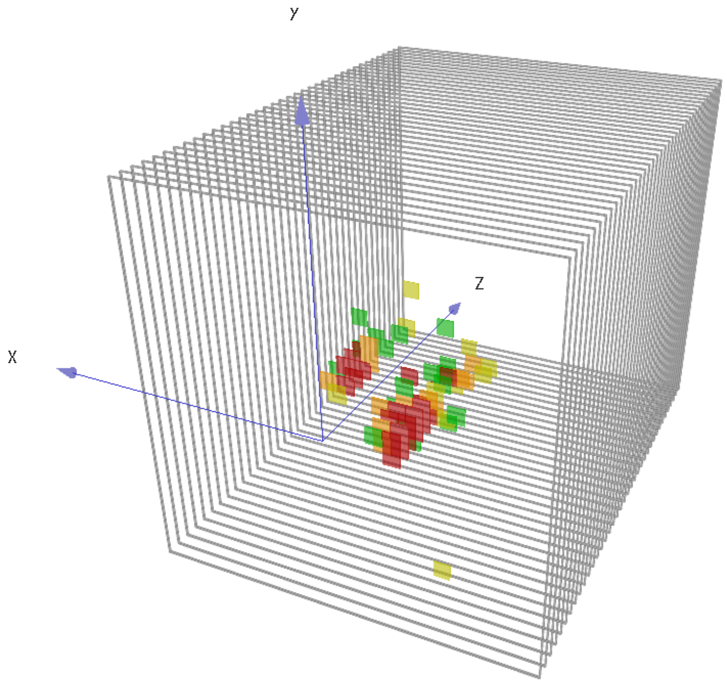
Event filtering

Simplified algorithms.



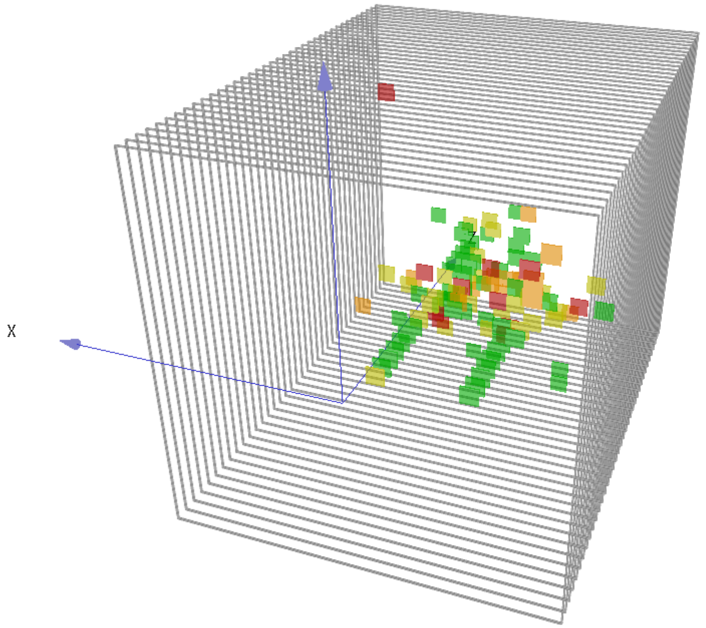
Clustering: Hits are grouped in clusters if if they are neighbours in volume. First 5 layers are taken into account

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



BDT classification

Model and input. TBJune18.

Software and model:

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

BDT classification

Model and input. TBJune18.

Software and model:

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

Gradient Boosting:
Method combines many sequential decision trees with weights. Weights are optimised during training by calculating the gradience of loss function

BDT classification

Model and input. TBJune18.

Software and model:

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

Gradient Boosting:

Method combines many sequential decision trees with weights. Weights are optimised during training by calculating the gradience of loss function

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 test sample, *3* - number of classes, Y_{ij} is binary variable with the expected labels and p_{ij} is he classification probability output by the classifier for the i-instance and the j-label.

BDT classification

Model and input. TBJune18.

Software and model:

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

Training and test set:

- **MC** particles **10-200GeV** QGSP_BERT_HP physics list simulated and reconstructed using June 2018 setup:
  - **pions (st ≤ 40)**
  - **electrons**
  - **muons**
- Simulated data is split **50/50 - test/train**

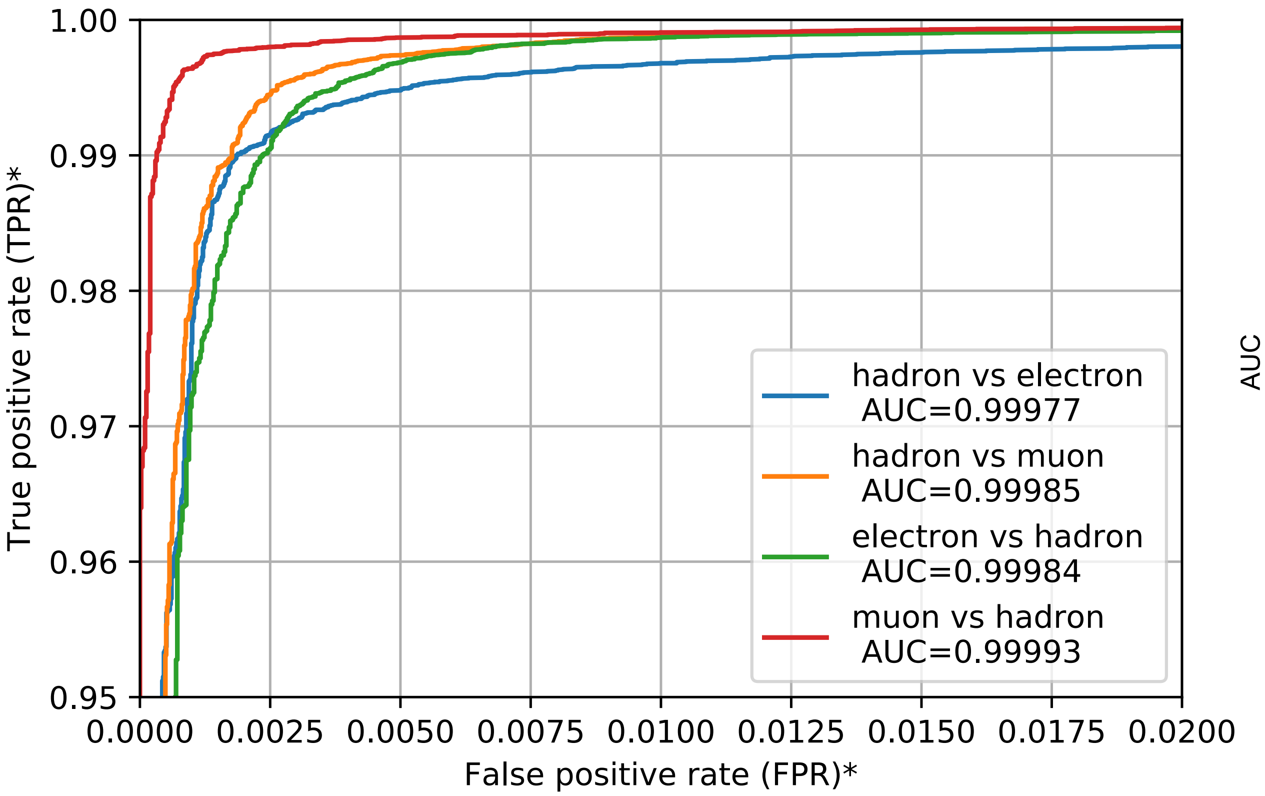
Observables:

- Number of hits
- Shower start
- Event radius
- Center of gravity in z
- Energy fraction in first 22 layers
- Energy fraction in shower center
- Energy fraction in shower core
- Fraction of track hits
- Number of track hits
- Number of layers with hits from last 5
- Mean hit energy after shower start

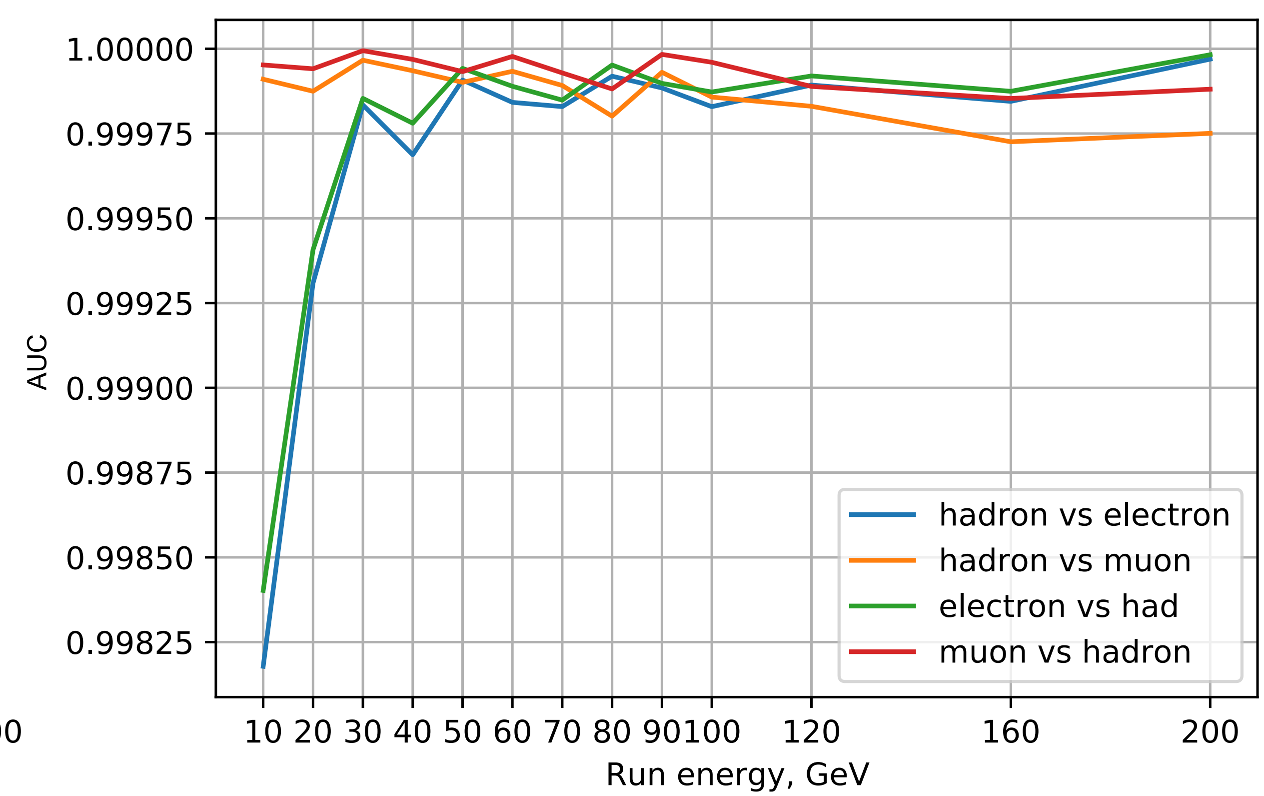
Resulting metrics

After training

ROC curves for the test data



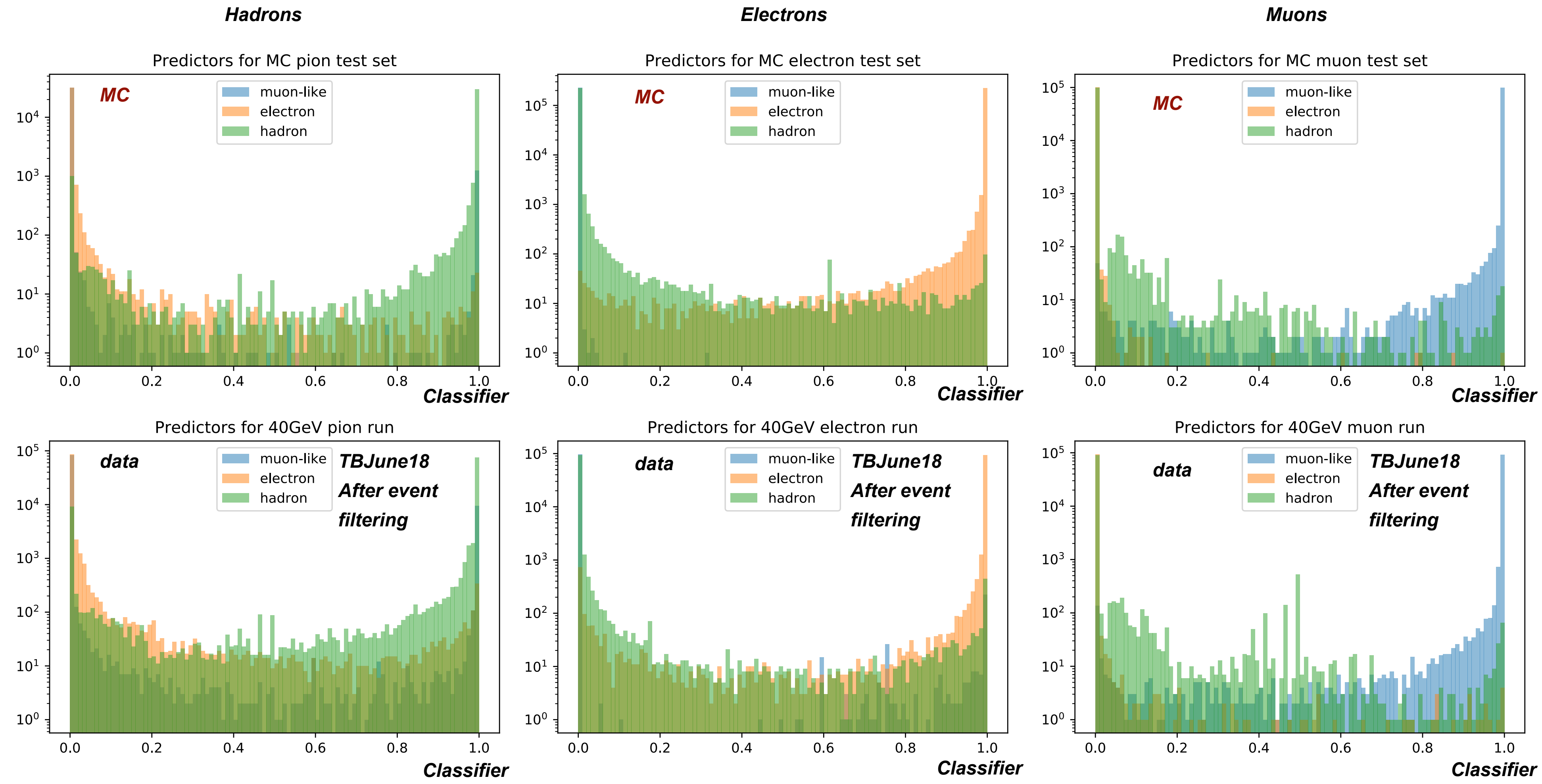
Model AUC for different energies



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

BDT classification

Output. Comparison with data.



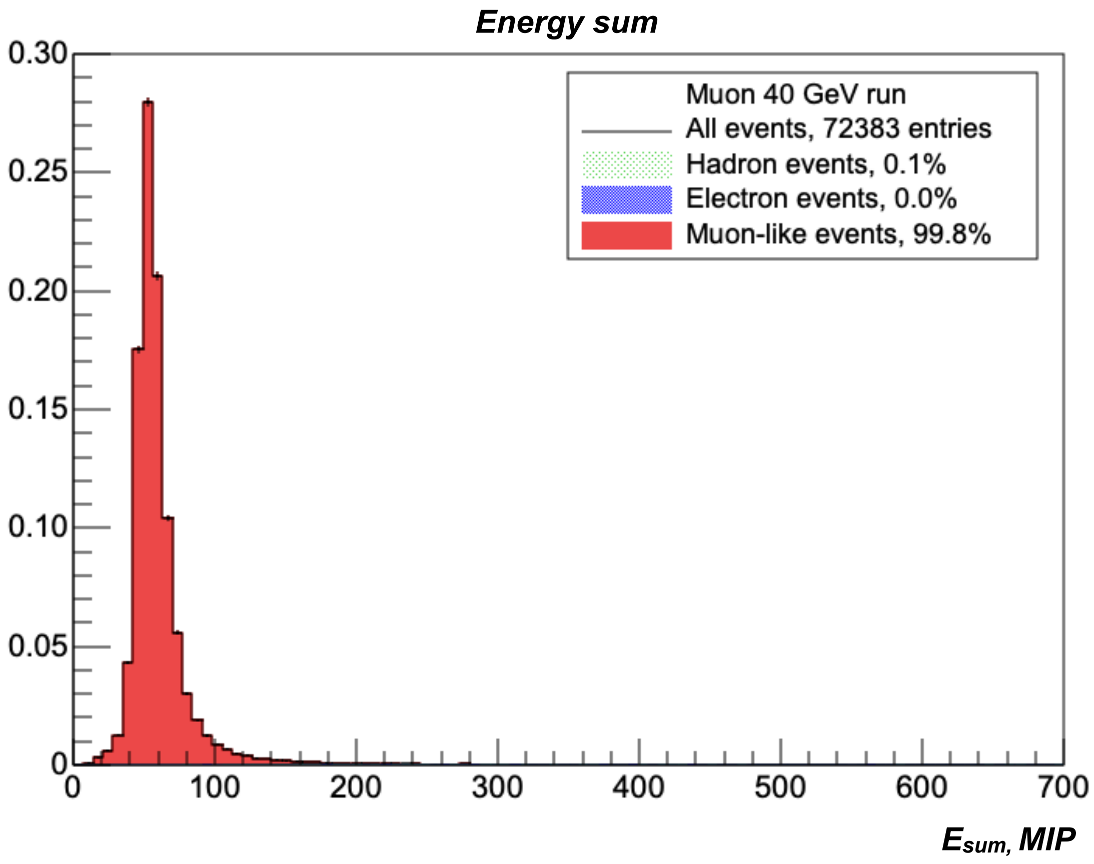
- Similar response on data and simulations

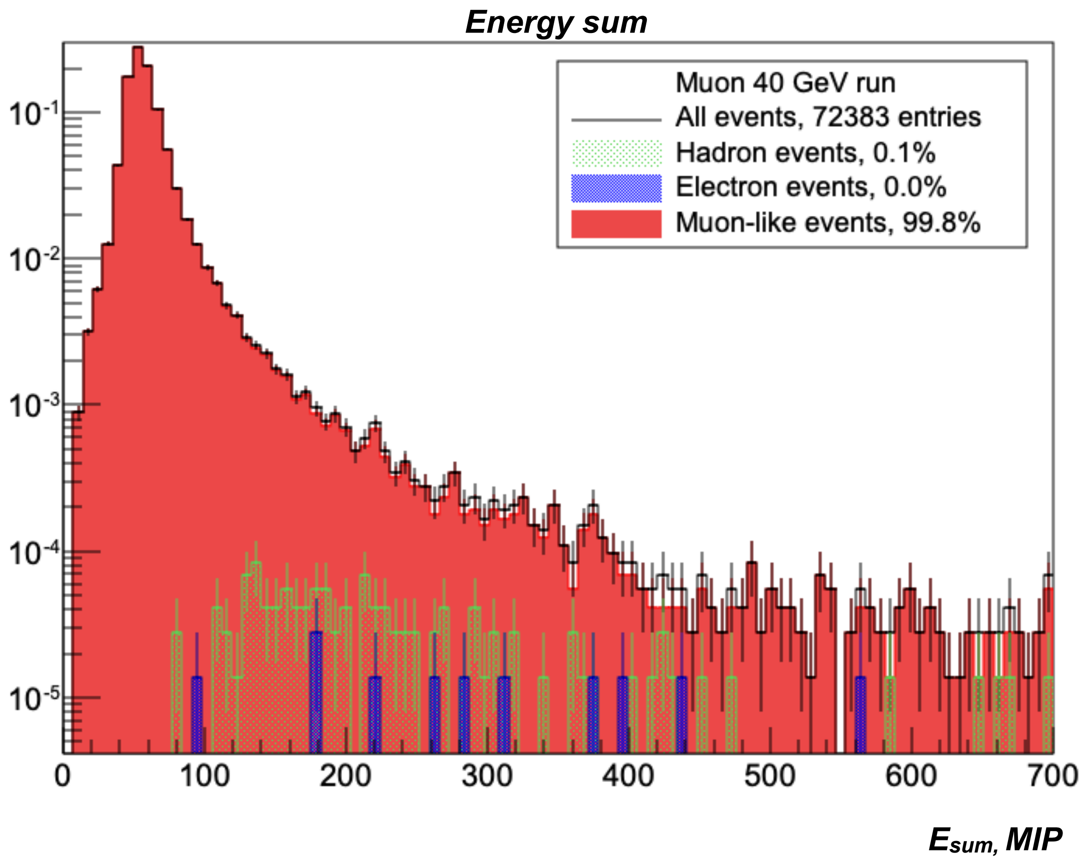
Classifiers:

muon-like
electron
hadron

Results on test beam data taken in June 2018

Energy sum distribution for 40GeV muon run

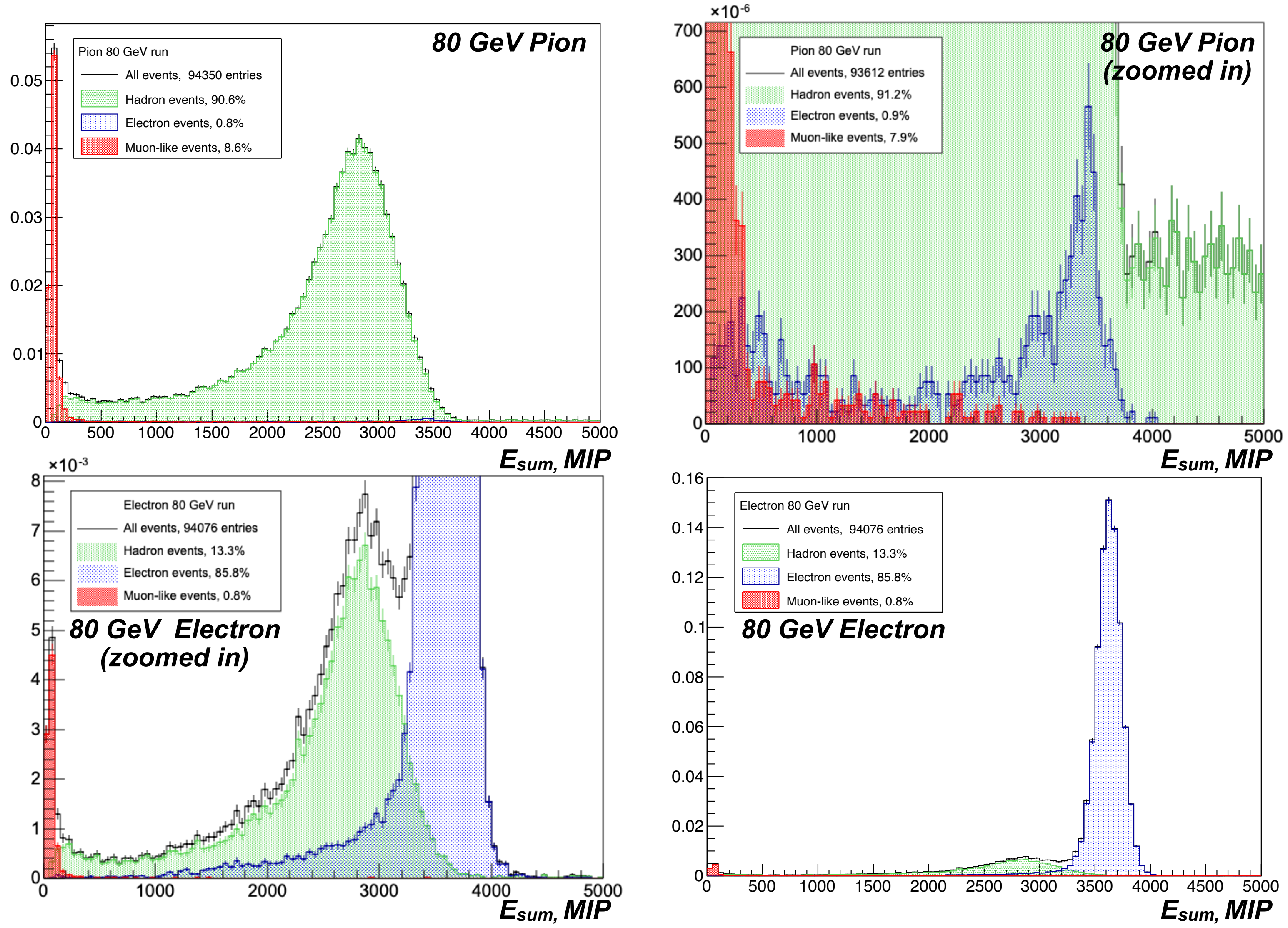




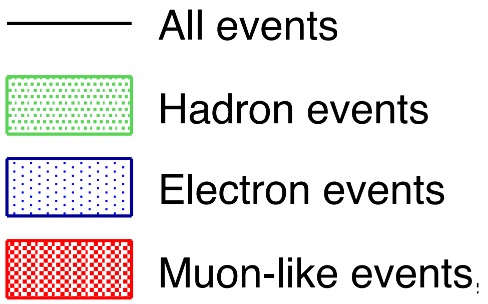
- Very low admixture of other particles
- Little fraction of delta electrons can be classified as hadron event

Results on test beam data taken in June 2018

Energy sum distributions for 80GeV runs

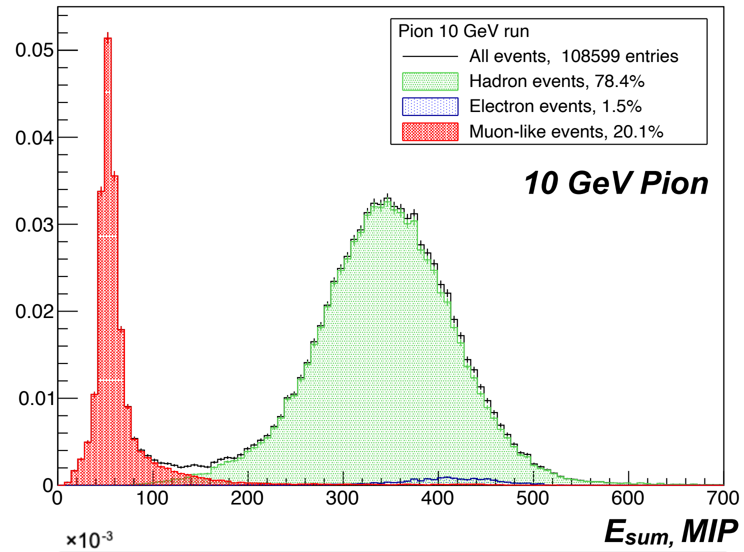


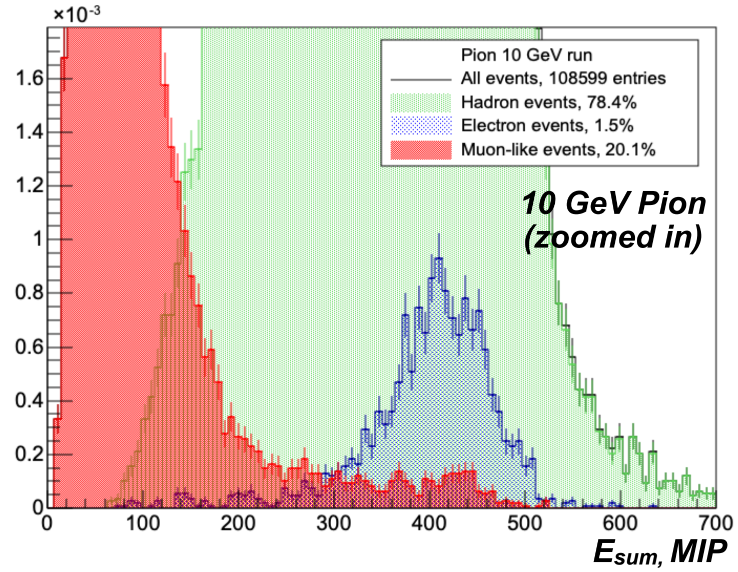
- Energy expectation for electron events in pion run is close to real electron run
- Energy distribution of hadron events in 80GeV electron run looks very similar to actual 80GeV pion

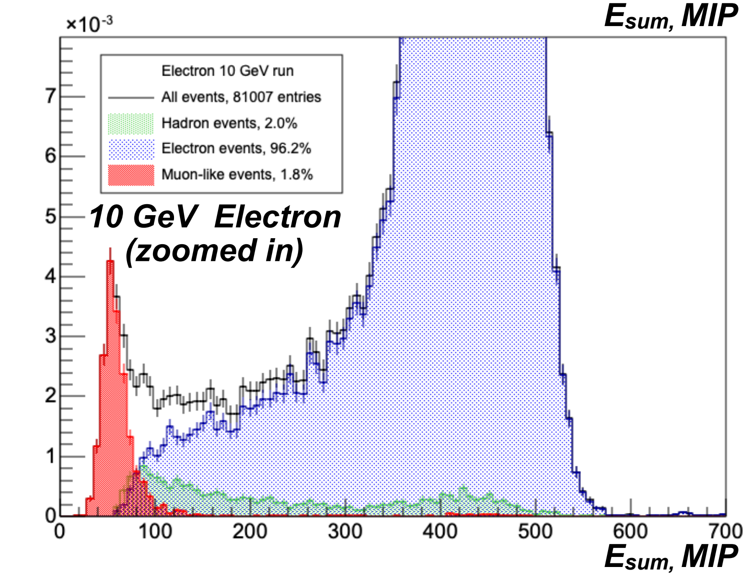


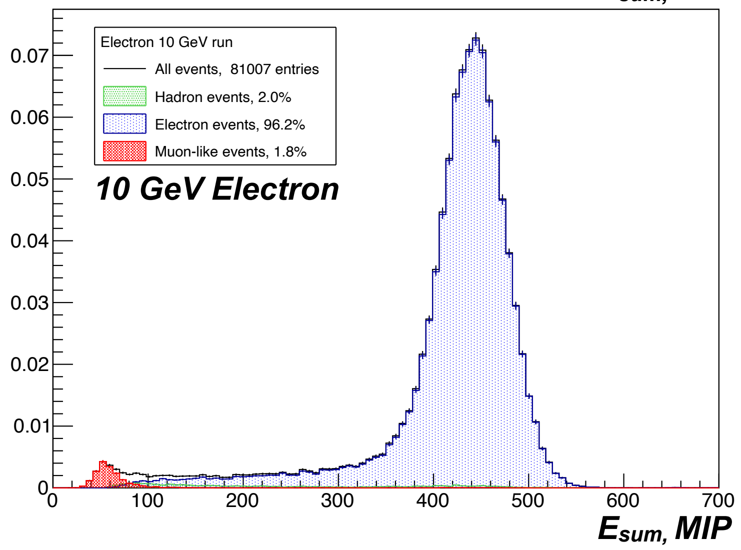
Results on test beam data taken in June 2018

Energy sum distributions for 10GeV runs

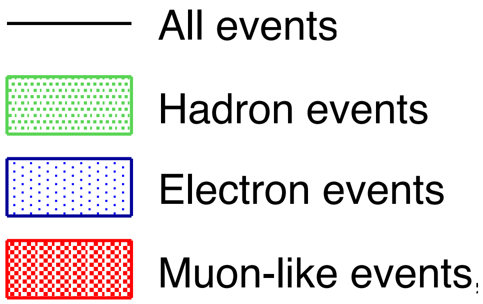








- Energy expectation for electron events in pion run is close to real electron run
- Long high energy tail of muon-like events
- Low energy tail for electrons
- Most of hadron events in electron run are at low energy



Summary and outlook (1st part)

AHCAL Particle ID using BDTs

- ☑ BDT particle ID method in the AHCAL was discussed
  - ☑ Method shows good performance
  - ☑ Similar response on data and MC
  - ☐ Sort input observables by importance to drop less useful ones
  - ☐ Final documentation

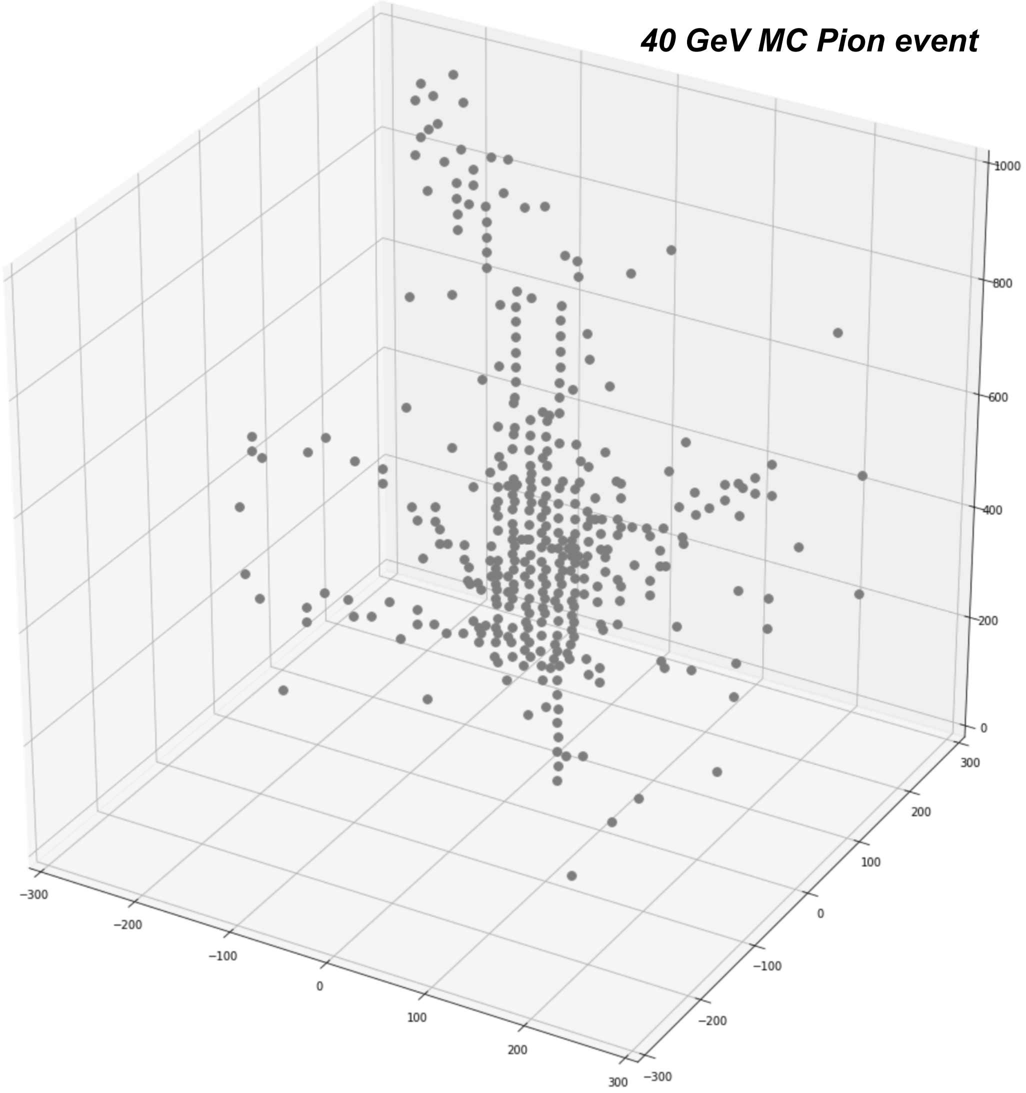
Bonus part

Detailed single hadron shower structure

Motivation

- Standard pattern recognition techniques for highly granular calorimeters provide connection between hits and particle entering a calorimeter

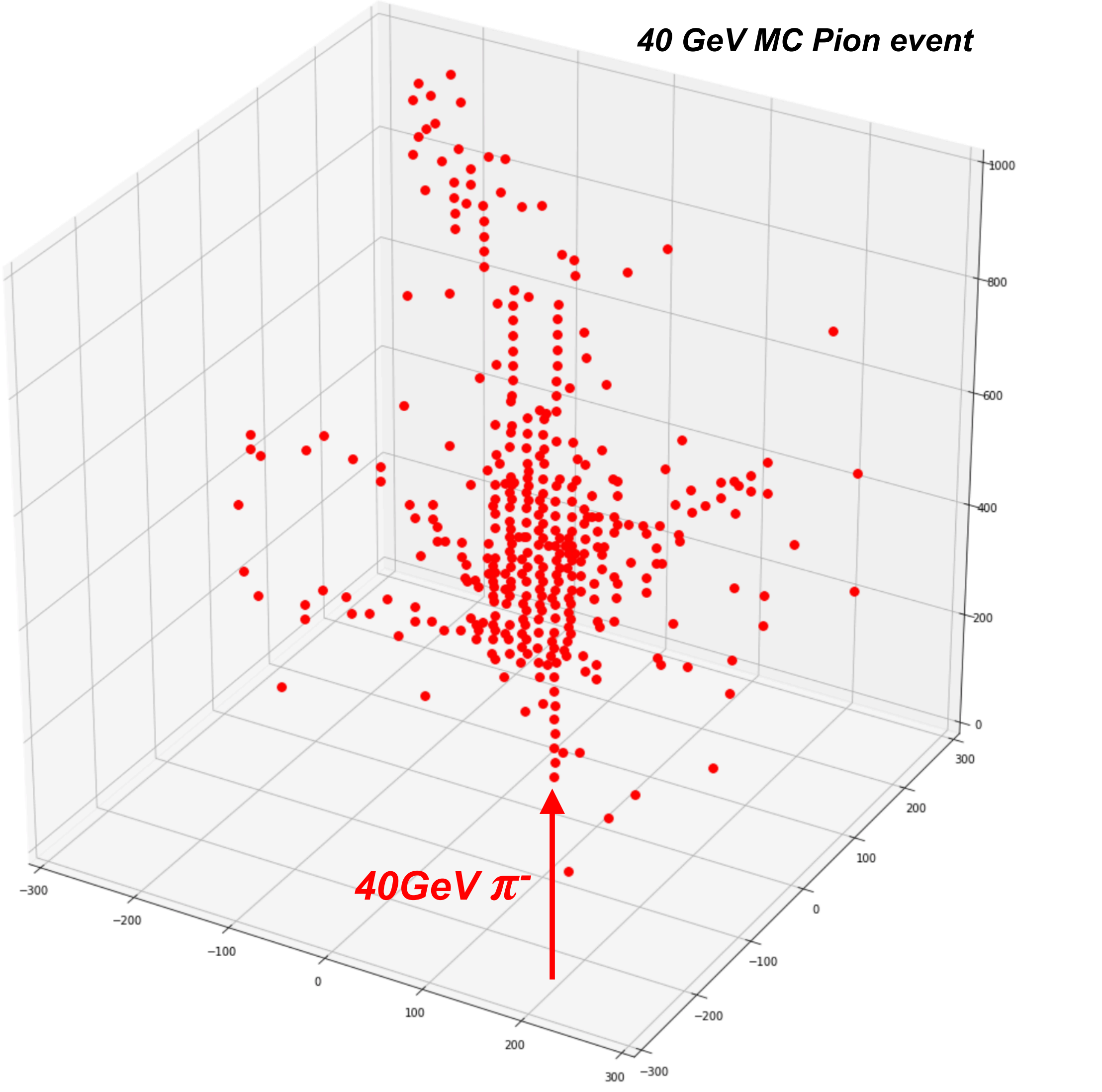
40 GeV MC Pion event



Detailed single hadron shower structure

Motivation

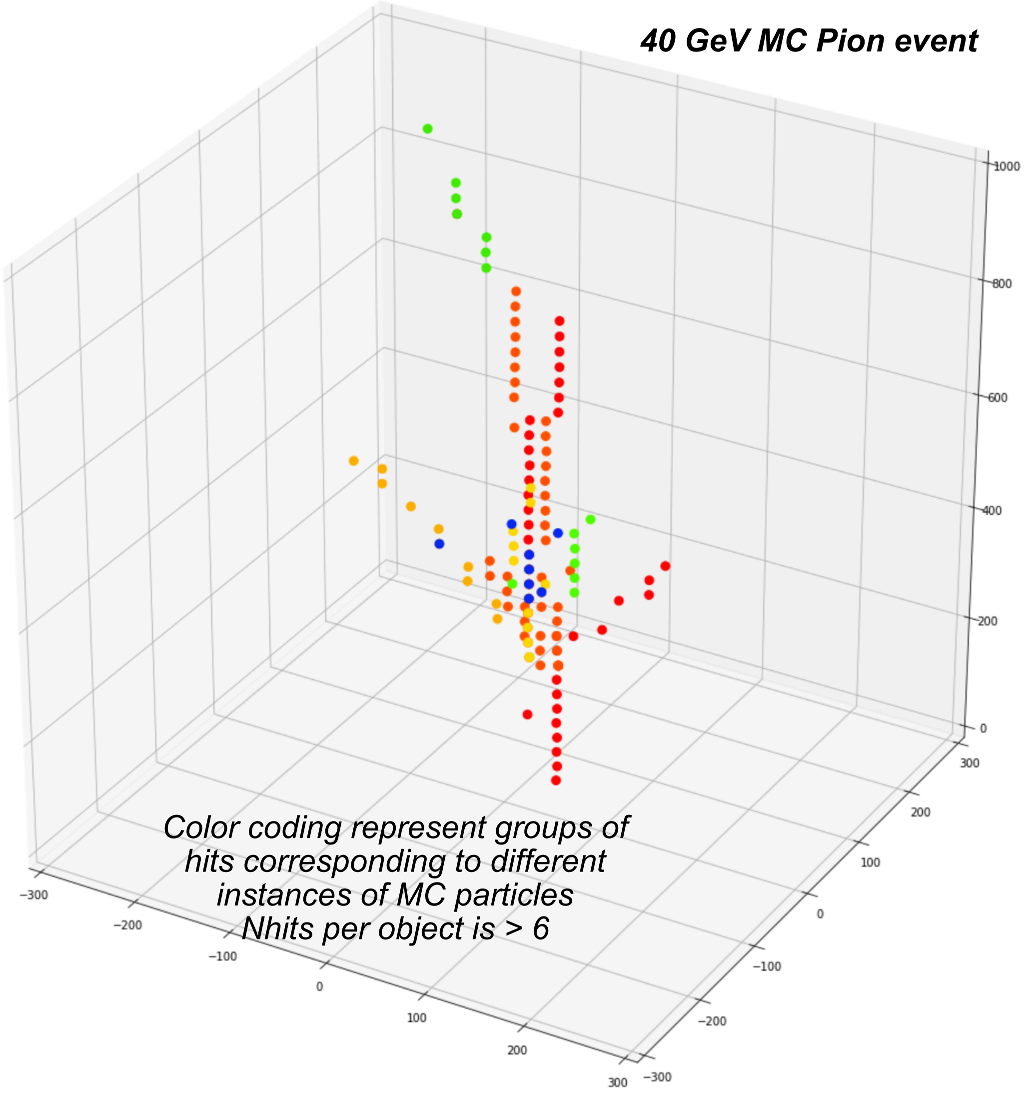
- Standard pattern recognition techniques for highly granular calorimeters provide connection between hits and particle entering a calorimeter



Detailed single hadron shower structure

Motivation

- Standard pattern recognition techniques for highly granular calorimeters provide connection between hits and particle entering a calorimeter
- High granularity combined with the latest advances in Computer Vision (CV) algorithms can allow us to **detect single particles** within a shower and fully exploit imaging capability of highly granular calorimeters



40 GeV MC Pion event

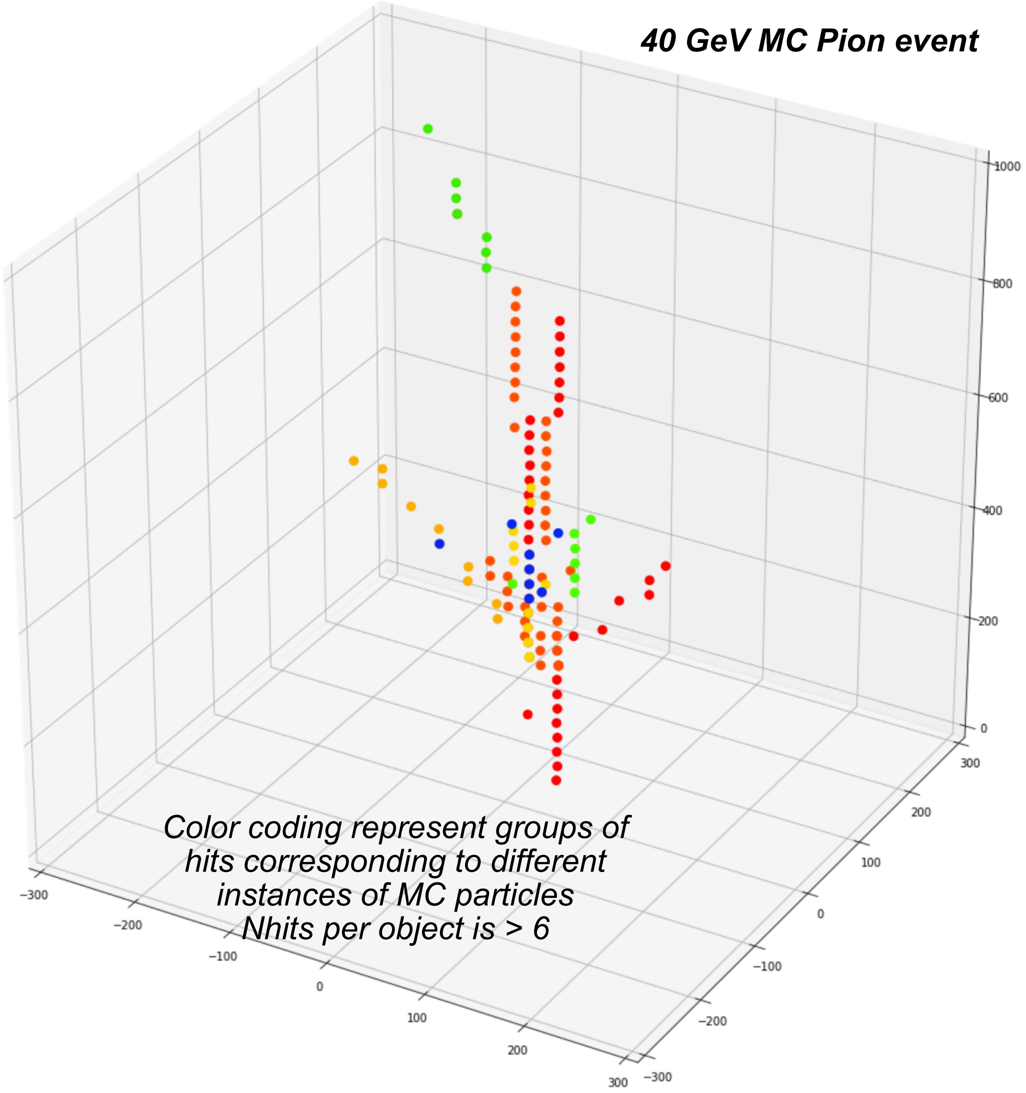
Color coding represent groups of hits corresponding to different instances of MC particles
Nhits per object is > 6

Detailed single hadron shower structure

Motivation

- Standard pattern recognition techniques for highly granular calorimeters provide connection between hits and particle entering a calorimeter
- High granularity combined with the latest advances in Computer Vision (CV) algorithms can allow us to **detect single particles** within a shower and fully exploit imaging capability of highly granular calorimeters
  - Can help to:
    - Benchmark simulations and other object detection studies (track finding)
    - Study shower shapes from different perspective
    - Improve software compensation
    - Separate showers

40 GeV MC Pion event

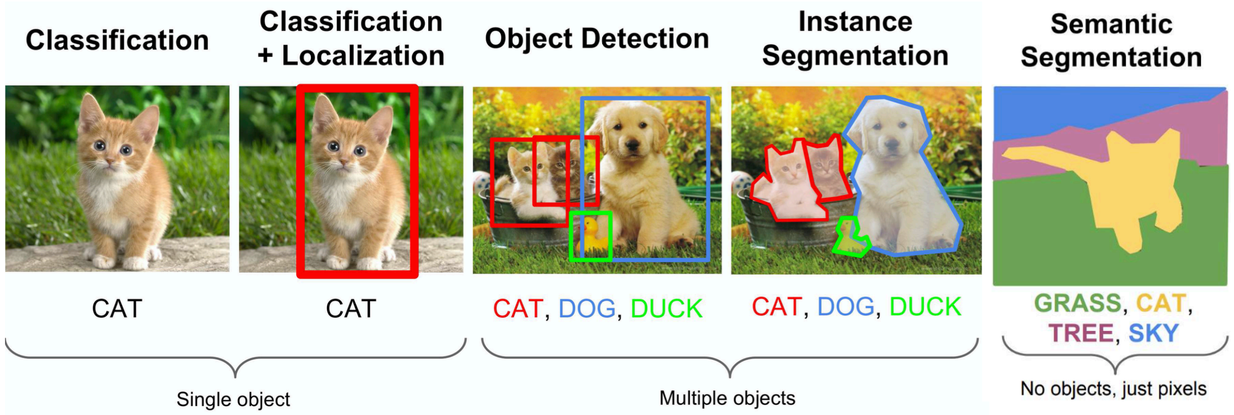


Color coding represent groups of hits corresponding to different instances of MC particles Nhits per object is > 6

Detailed single hadron shower structure

Method and data preparation

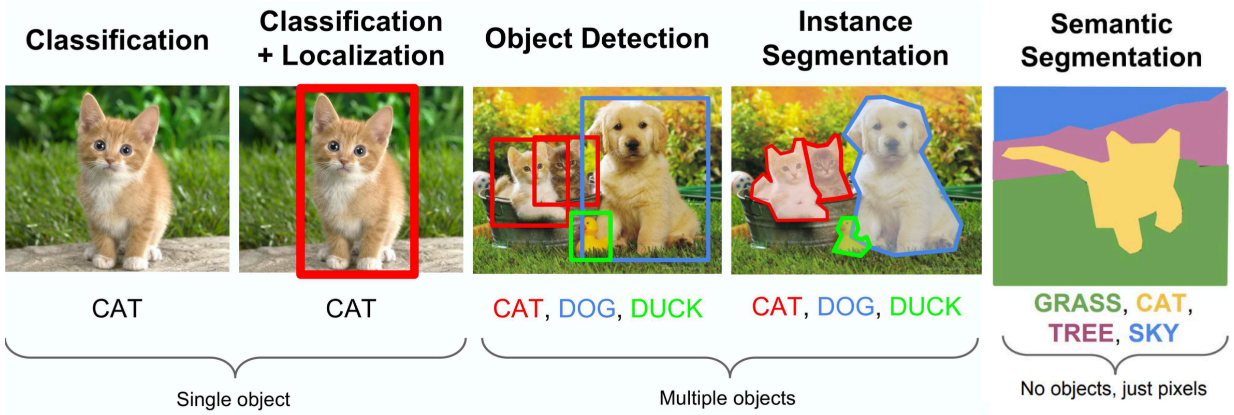
- In CV this problem is classified **instance segmentation**



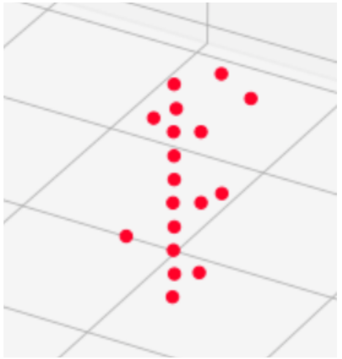
Detailed single hadron shower structure

Method and data preparation

- In CV this problem is classified **instance segmentation**
- Data set with links between simulated hits and **MC objects** has been prepared
    - **MC object** is MC particle + daughters w/o parent's endpoint (hard ionization, Bremsstrahlung, elastic interactions, etc.)



MC object



Detailed single hadron shower structure

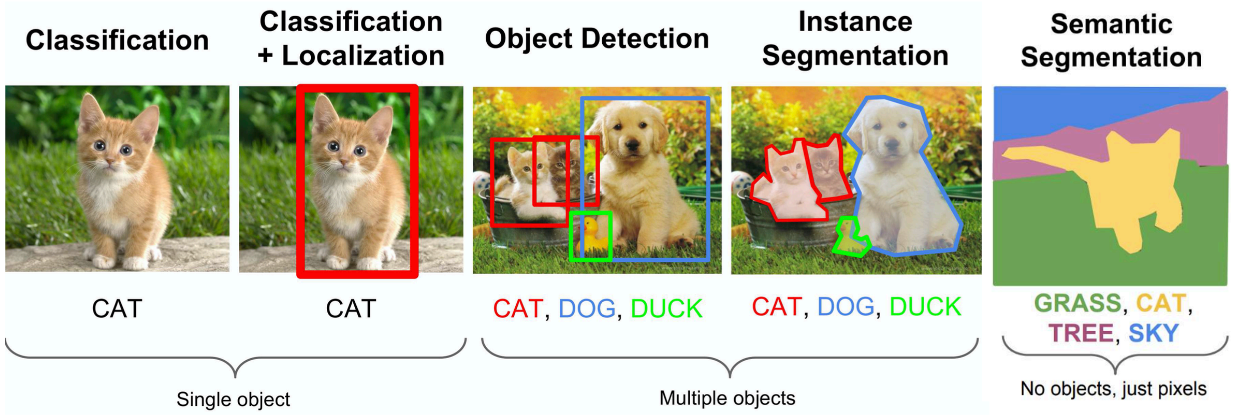
Method and data preparation

- In CV this problem is classified **instance segmentation**
- Data set with links between simulated hits and **MC objects** has been prepared
  - **MC object** is MC particle + daughters w/o parent's endpoint (hard ionization, Bremsstrahlung, elastic interactions, etc.)
- Roadmap:
  - Point Cloud input data - (X,Y,Z,E,(T))

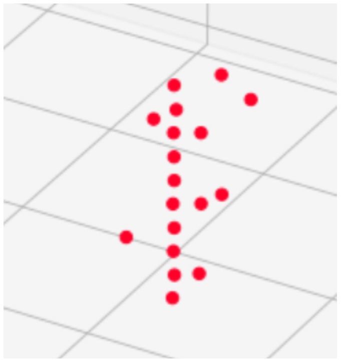
    Recent implementation in HEP:

    Jet tagging: https://arxiv.org/abs/1902.08570

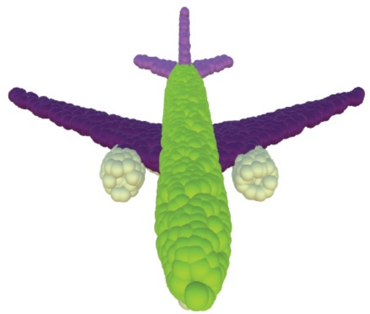
    Calorimetry: https://arxiv.org/abs/1902.07987
  - Machine learning model is being developed, existing implementations of similar problem can be adapted to the task (work in progress)



MC object



Point cloud segmentation

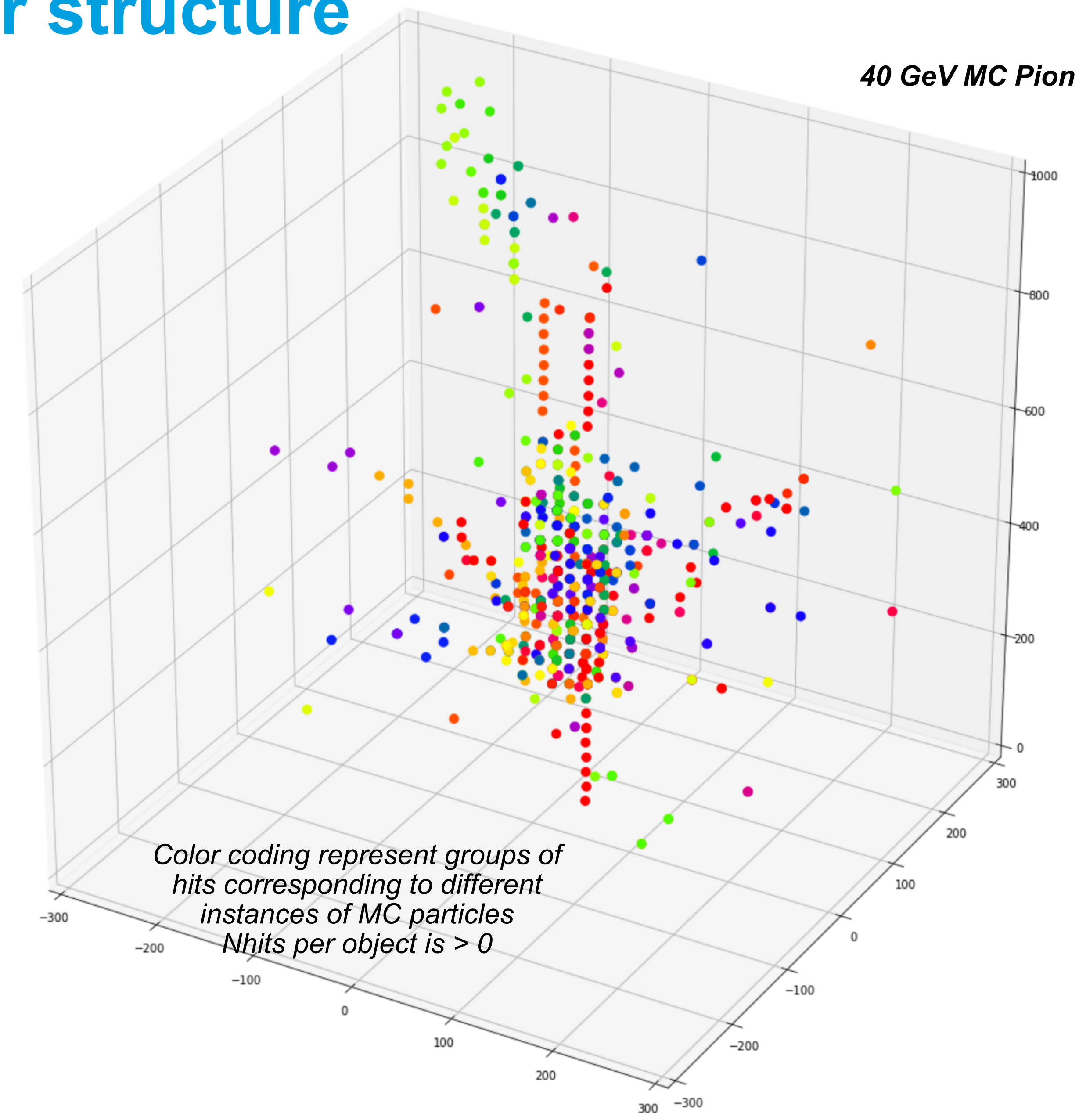


Detailed single hadron shower structure

Challenges of proposed method

- Large disk space usage
  - O(1000) MC instances per 40 GeV pion event including single hit objects
- Interpretability of segmentation output & training data
  - Many complicated objects to plot distributions
  - Hierarchical structure
- One hit can be connected to several MC objects
  - Main difference from the standard CV problems

40 GeV MC Pion event



Color coding represent groups of hits corresponding to different instances of MC particles
Nhits per object is > 0

Summary and outlook

AHCAL Particle ID using BDTs and Single Shower Substructure

- ☑ BDT particle ID method in the AHCAL was discussed
  - ☑ Method shows good performance
  - ☑ Similar response on data and MC
  - ☐ Sort input observables by importance to drop less useful ones
  - ☐ Final documentation
- ☑ Single shower substructure analysis is ongoing
  - ☑ Training data preparation is done
  - ☐ Machine learning model is being developed
  - ☐ Model performance and limitations study will help to review true segmentation on MC level
  - ☐ Physical interpretation of results

Thank you

Backup slides

Disadvantages of cut-based method

Towards BDT ID

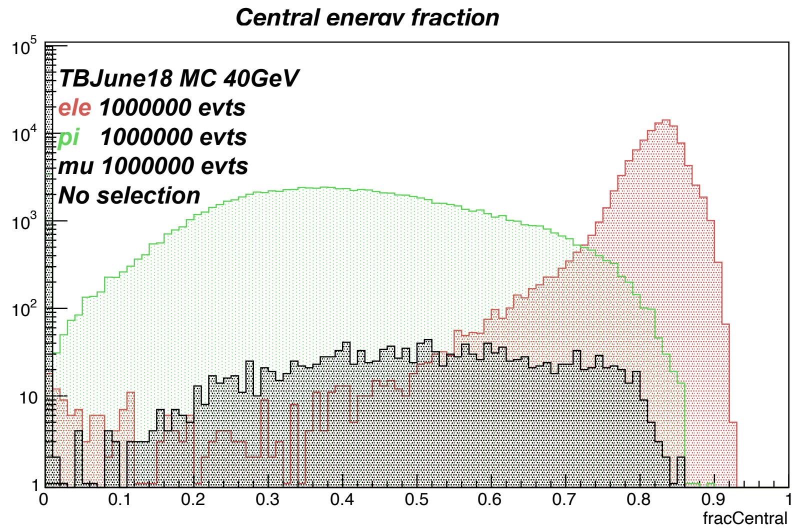
Cut-based method:

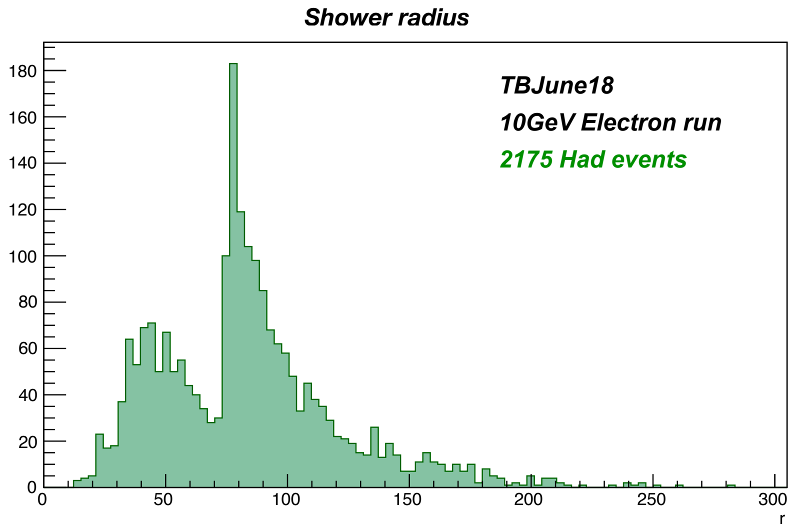
- > 10 steering parameters for each energy
- Asymmetric distributions/ long tails can be problematic

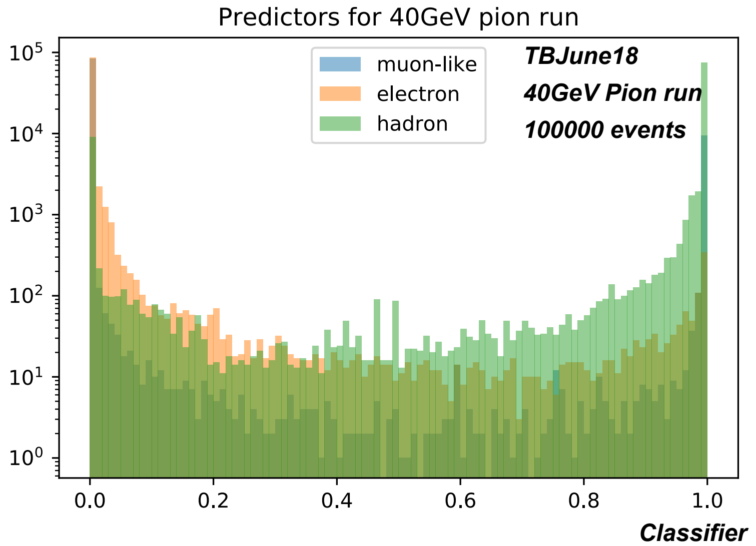
- Cut artefacts

Multivariate methods:

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



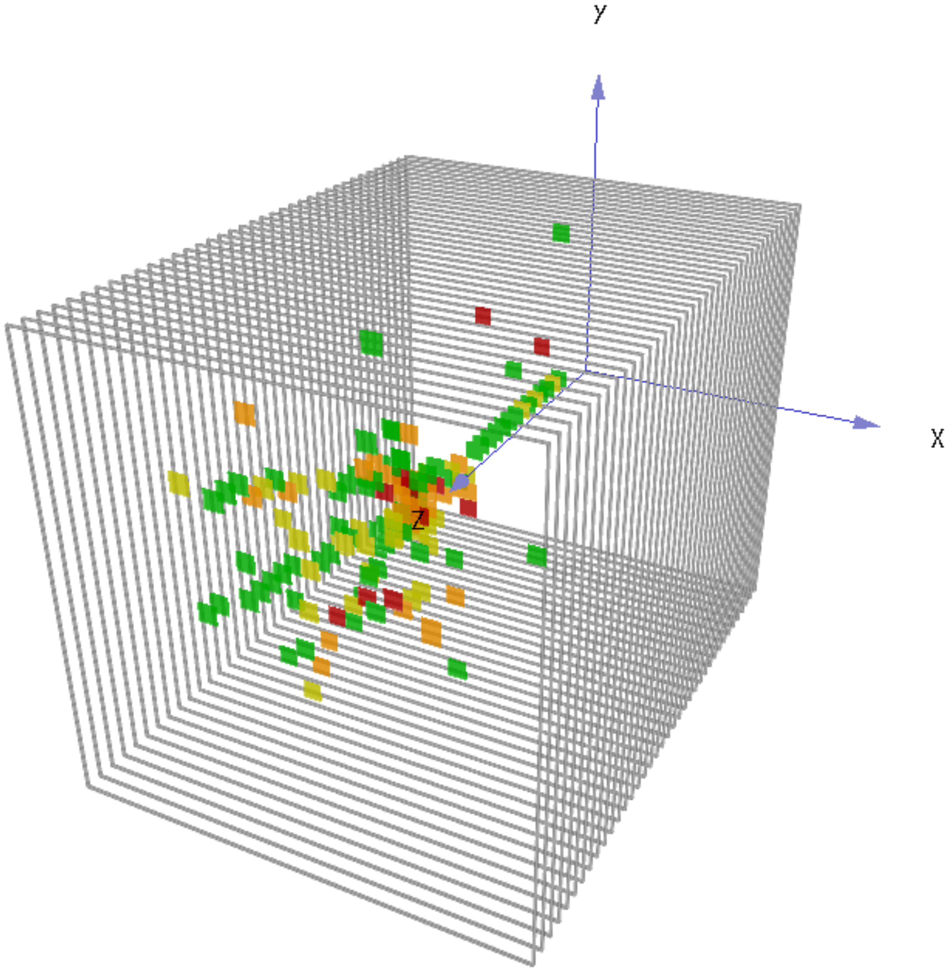




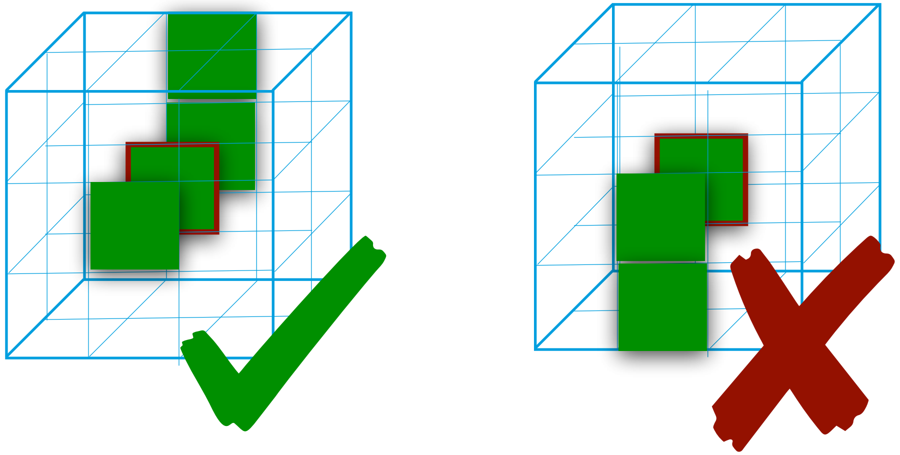
Will be discussed during one of the upcoming HGCAL meetings

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

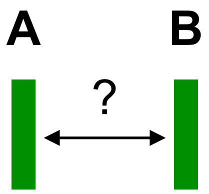
Nearest neighbour of hit#1

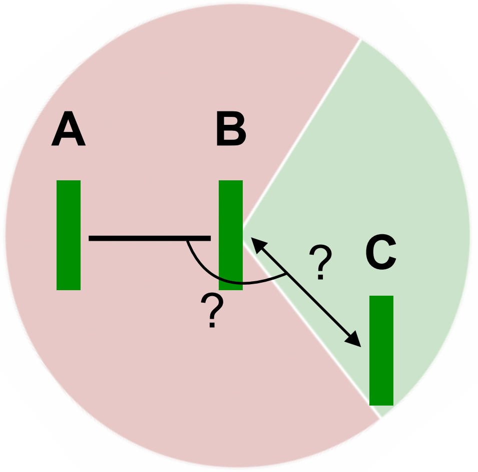
Distance check

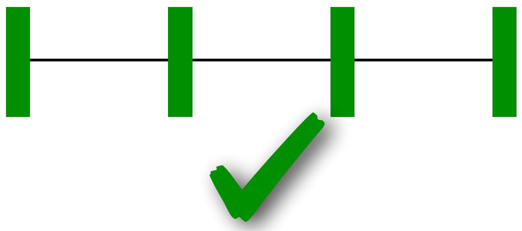
Angle check with **AC**

Angle check with **AB**

(…)





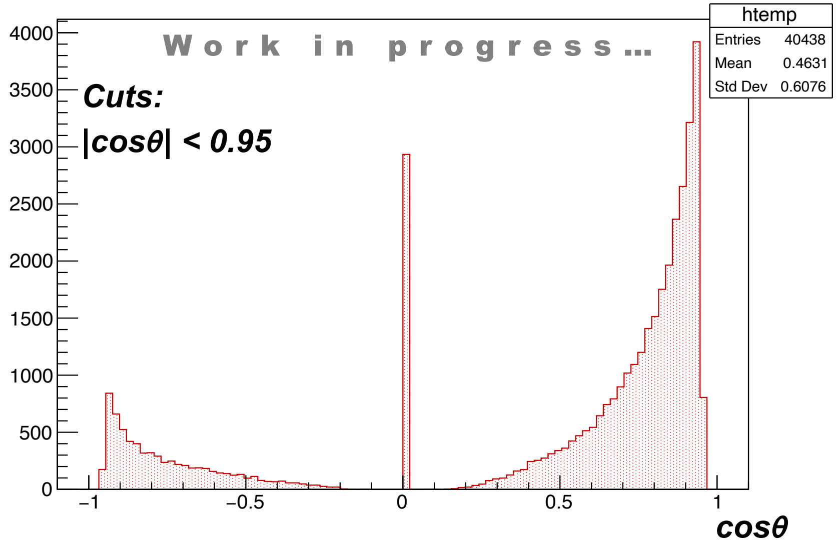


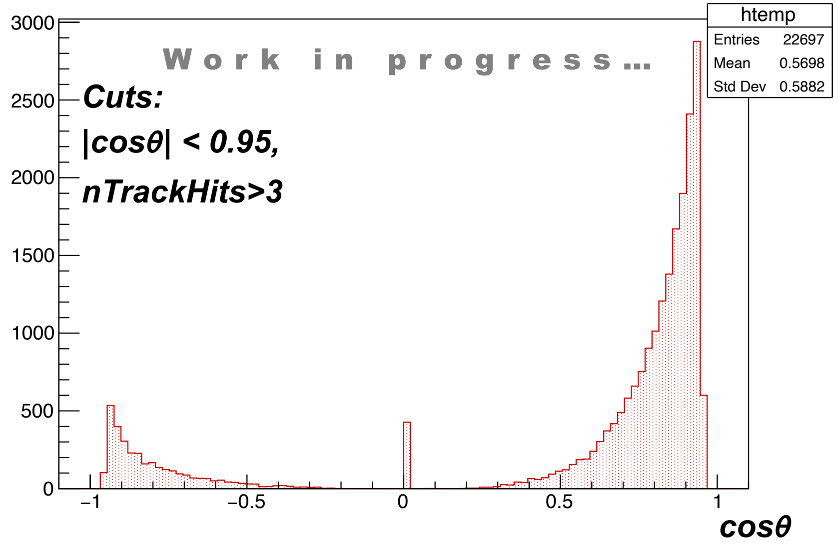
After grouping, track angle is obtained using MSE linear regression

** Procedure repeated iteratively **

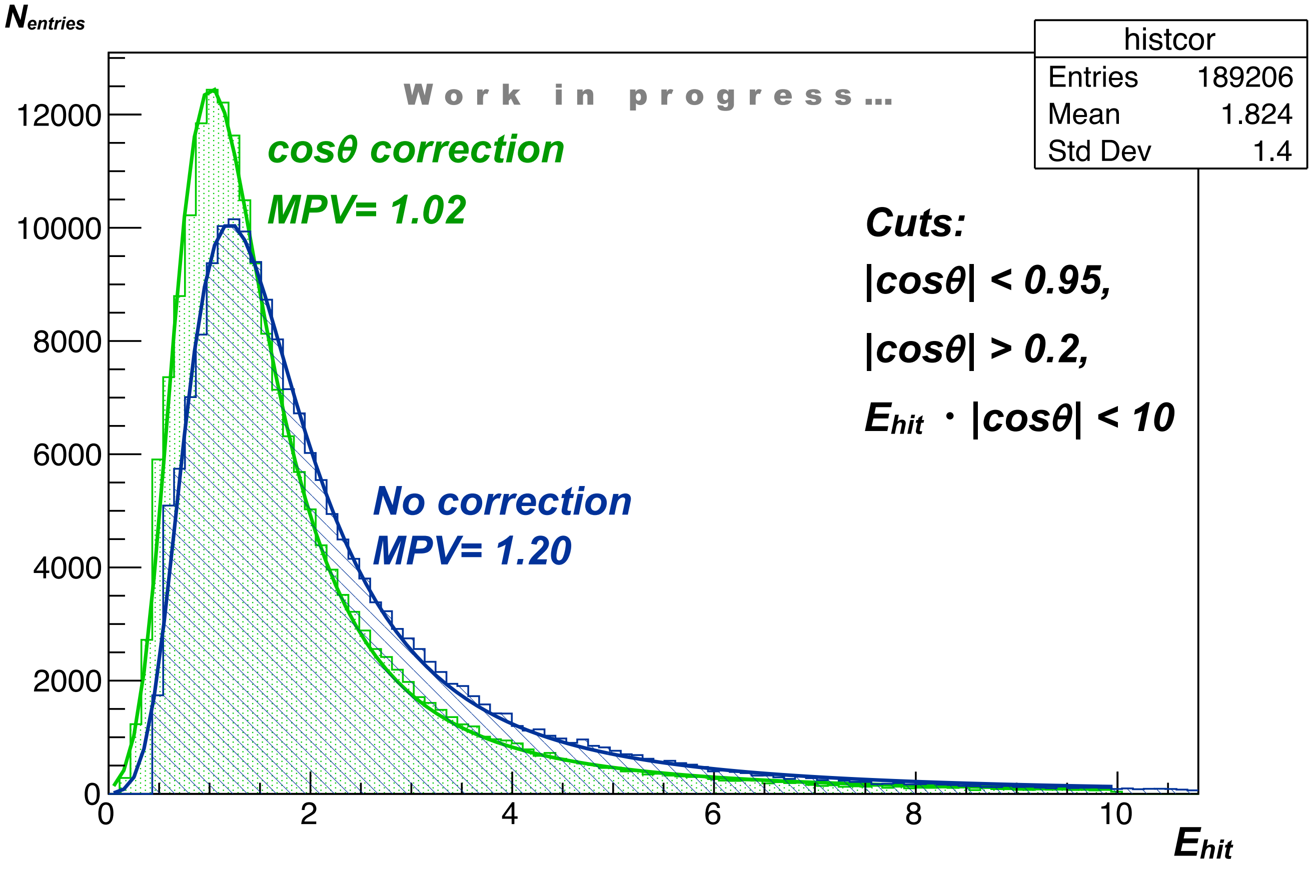
Tracking quality check

TBMay18 10GeV pion run. 50039 events.



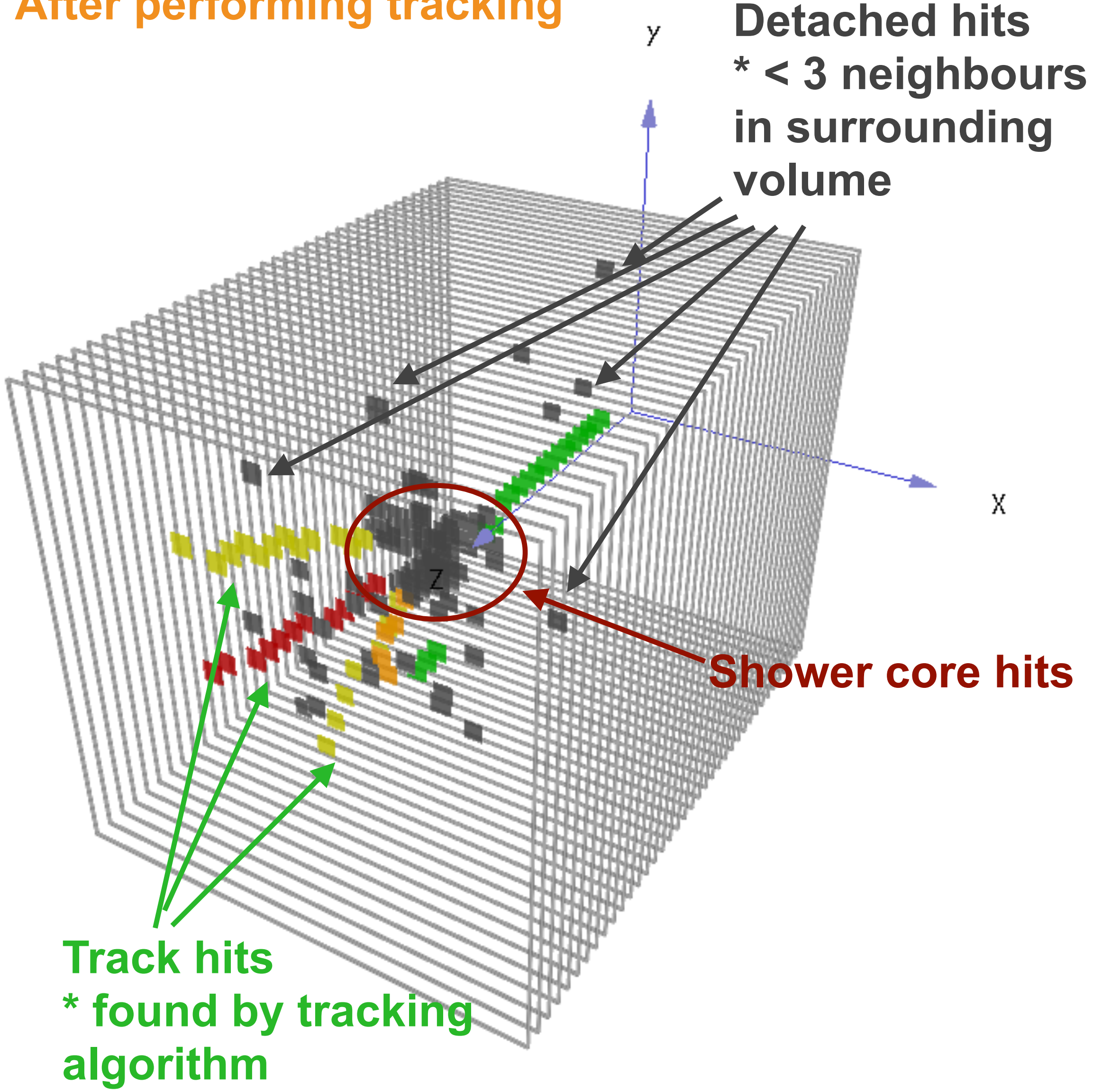


Scintillator path length correction for track hits



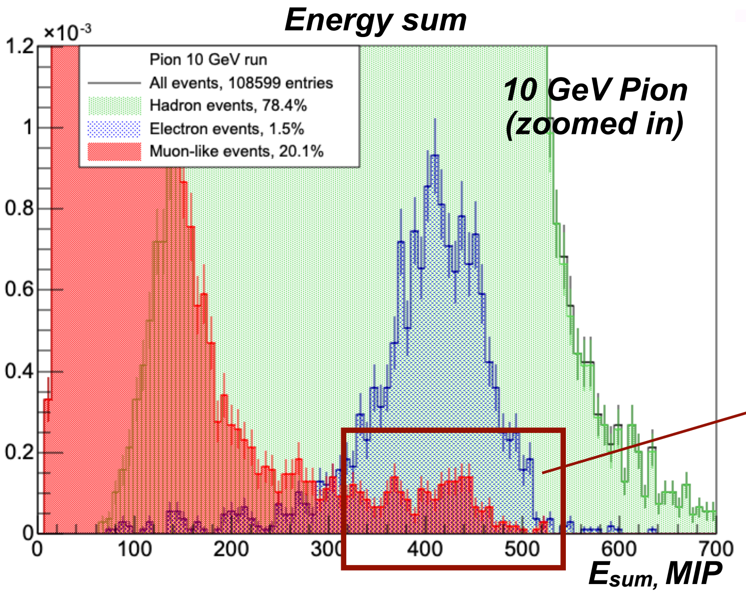
Resulting ID variables

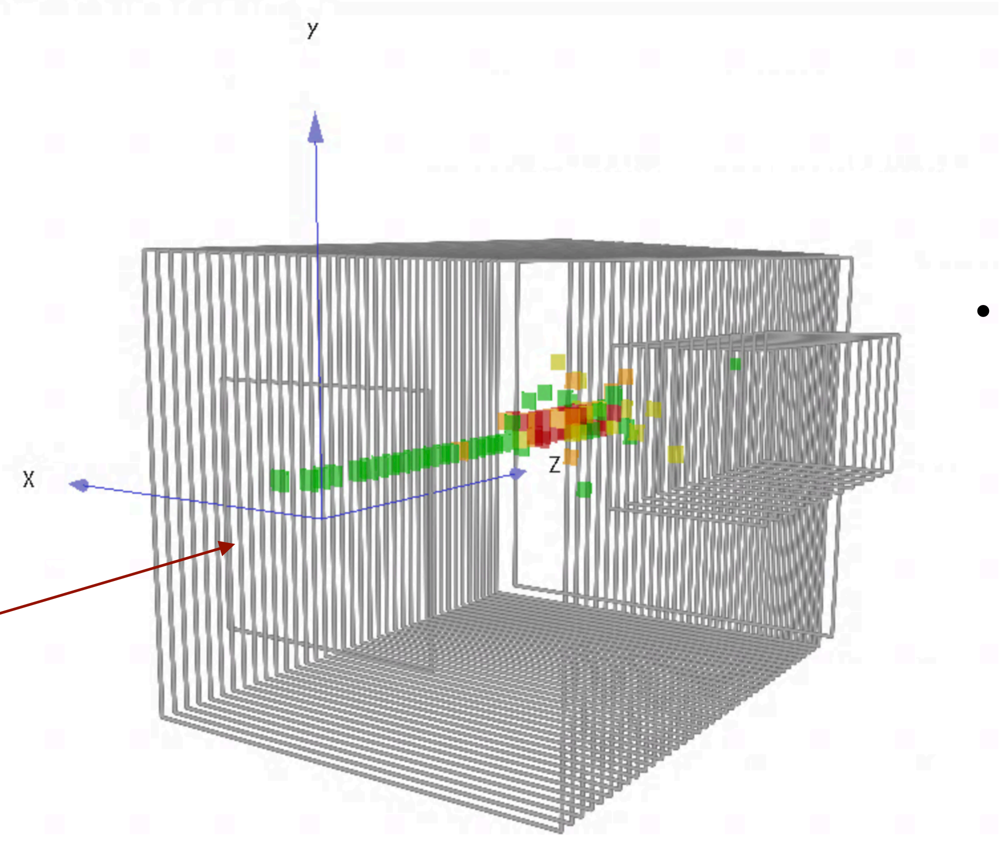
After performing tracking



Sources of confusion

From 10GeV pion run





Muon-like event :

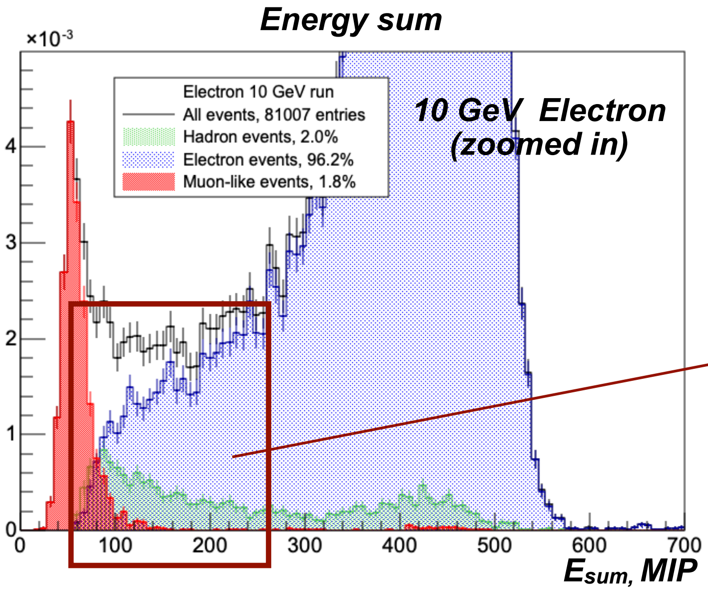
Mu-like score is 0.51

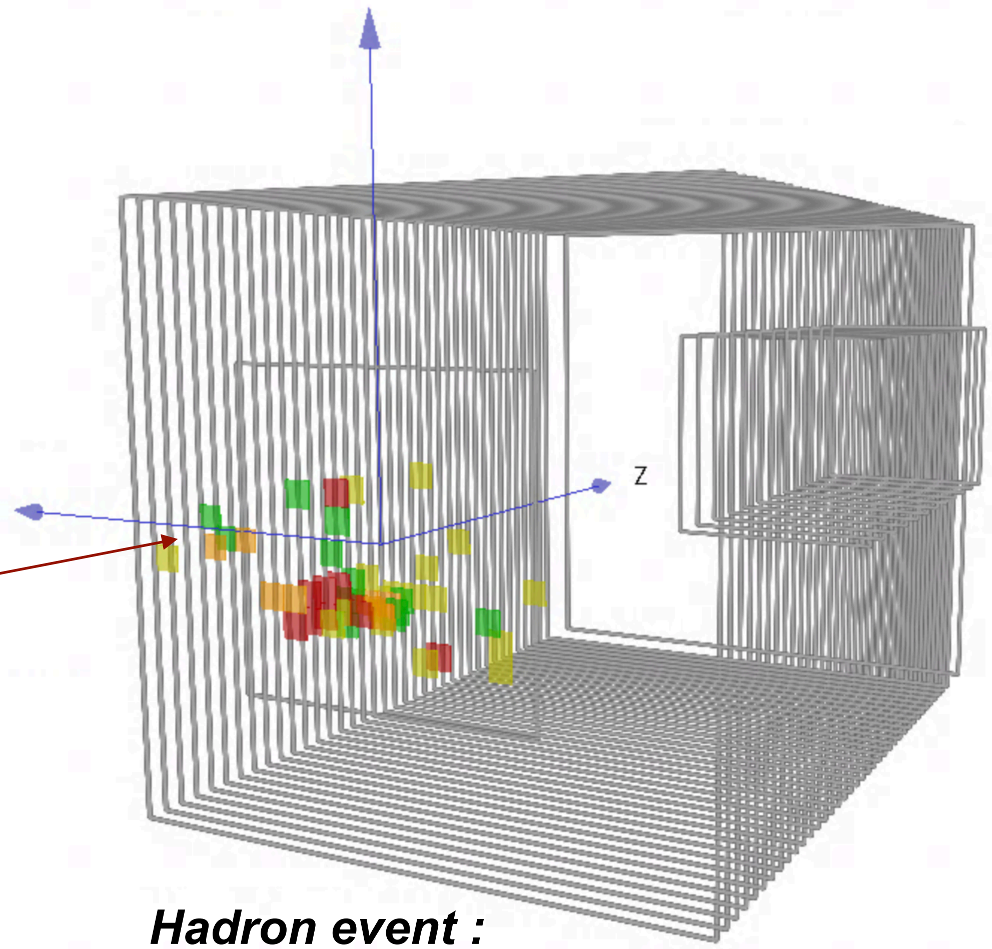
Had score is 0.48

- 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



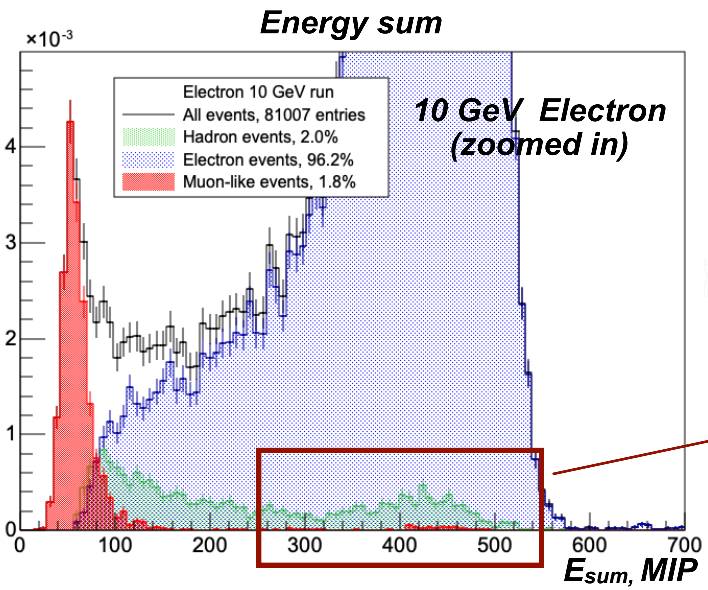


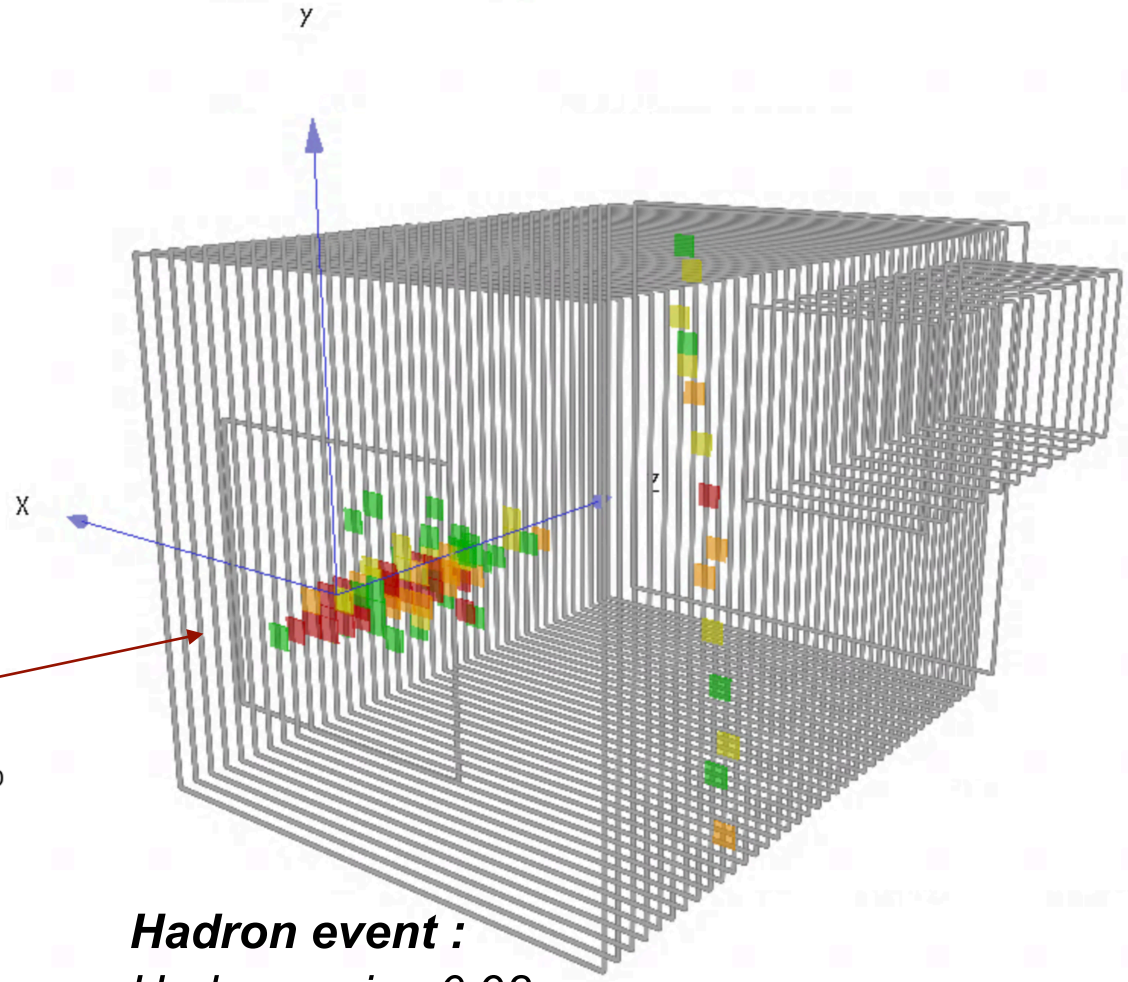
Hadron event :
Had score is ~0.9

- 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





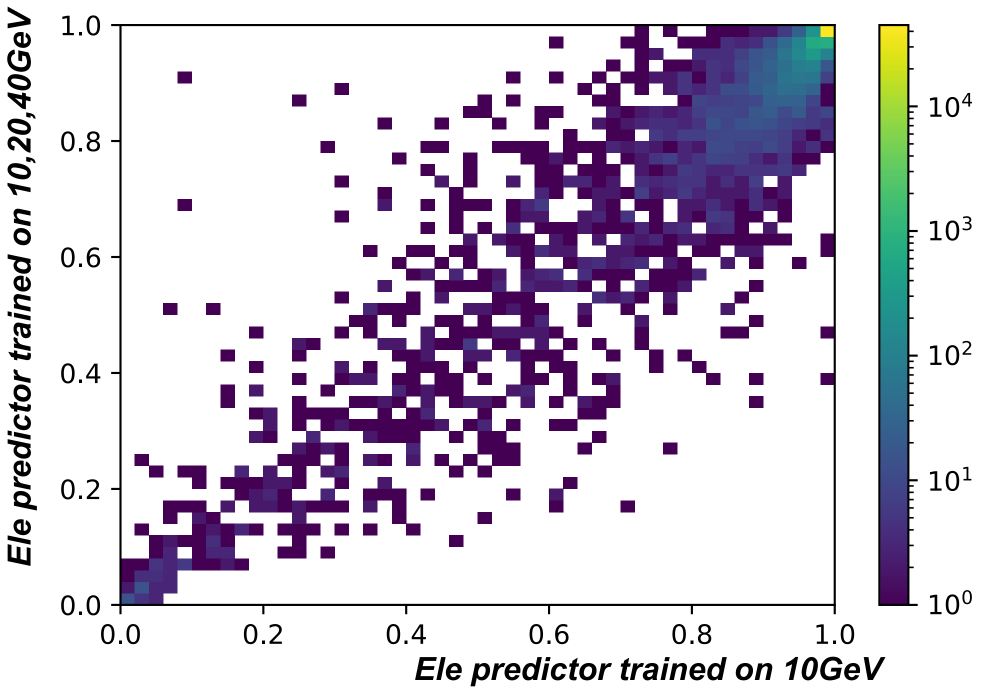
Hadron event :
Had score is ~0.98

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

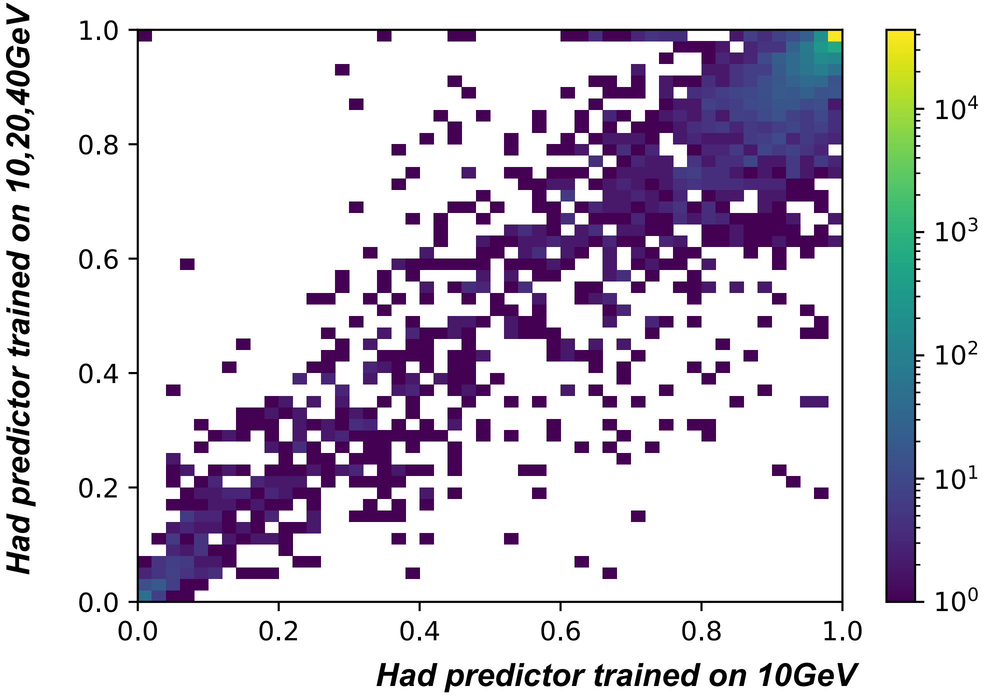
BDT output

Comparison with separate model trained only on 10GeV particles.

10GeV MC electron test sample
50000 events



10GeV MC pion test sample
50000 events



Application on electron data

Of trained BDT model