Photon/Pi0 ID at ALLEGRO

Giovanni Marchiori, Tong Li (APC-Paris CNRS)

with Gregorio Bernardi, Nicolas Morange

FCC Detector Full Sim Meeting, 26 June 2024

FUTURE CIRCULAR COLLIDER

cnrs NUCLÉAIRE & PARTICULES

Introduction

- Shape parameters calculated for photon/pi0 identification with ALLEGRO: Pull Request

- Generated samples of sliding-window clusters from photon/pi0
  - `ddsim` for simulation
  - Detector: ALLEGRO v03 (11 layers with projective cell corners)
  - Energy of incident particle: 1 – 100 GeV, theta: 40 – 140 degrees, phi: 0 – 2pi
  - Implement default and custom versions, for strip in different layers:
    - L1 (default), L2, L3, L4, L5

- Trained BDT using these shape parameters, to see the separation of photon and pi0

- Implemented the algorithm that runs the photon ID in Gaudi: Pull Request

Shape parameters (1/2)

- Cluster level:
  - Energy
  - Mass
  - Number of cells

```
*............................................................................*
*Br   22 :_AugmentedCaloClusters_clusters.index :                            *
*         | Int_t index[_AugmentedCaloClusters_clusters_]                    *
*Entries :   100000 : Total  Size=     404585 bytes  File Size  =       8037 *
*Baskets :       30 : Basket Size=     120320 bytes  Compression=  50.20     *
*............................................................................*
*Br   23 :_AugmentedCaloClusters_clusters.collectionID :                     *
*         | UInt_t collectionID[_AugmentedCaloClusters_clusters_]            *
*Entries :   100000 : Total  Size=     404823 bytes  File Size  =       8246 *
*Baskets :       30 : Basket Size=     120320 bytes  Compression=  48.95     *
*............................................................................*
*Br   24 :_AugmentedCaloClusters_hits : Int_t _AugmentedCaloClusters_hits_   *
*Entries :   100000 : Total  Size=    1065461 bytes  File Size  =     678975 *
*Baskets :       50 : Basket Size=      32000 bytes  Compression=   1.19     *
*............................................................................*
*Br   25 :_AugmentedCaloClusters_hits.index :                                *
*         | Int_t index[_AugmentedCaloClusters_hits_]                        *
*Entries :   100000 : Total  Size=  271376441 bytes  File Size  =   32610806 *
*Baskets :     6412 : Basket Size=   11182460 bytes  Compression=   8.32     *
*............................................................................*
*Br   26 :_AugmentedCaloClusters_hits.collectionID :                         *
*         | UInt_t collectionID[_AugmentedCaloClusters_hits_]                *
*Entries :   100000 : Total  Size=  271421353 bytes  File Size  =    4041675 *
*Baskets :     6412 : Basket Size=   11182460 bytes  Compression=  67.12     *
*............................................................................*
*Br   27 :_AugmentedCaloClusters_shapeParameters : vector<float>             *
*Entries :   100000 : Total  Size=   85228043 bytes  File Size  =   74838844 *
*Baskets :     1966 : Basket Size=   11182460 bytes  Compression=   1.14     *
*............................................................................*
```

- Calculated in each layer:
  - Maximum energy of cell
  - Energy fraction: E(i) / E
    - E(i) is energy in layer i, E is cluster energy
  - Width in theta: sqrt(sum(theta_i^2*E(i))/sum(E(i))-(sum(theta_i E_i)/sum(E_i))^2)
    - theta_i is theta ID of cell
  - Width in phi (module): sqrt(sum(module_i^2*E(i))/sum(E(i))-(sum(module_i E_i)/sum(E_i))^2)
    - module_i is module ID of cell

Shape parameters (2/2)

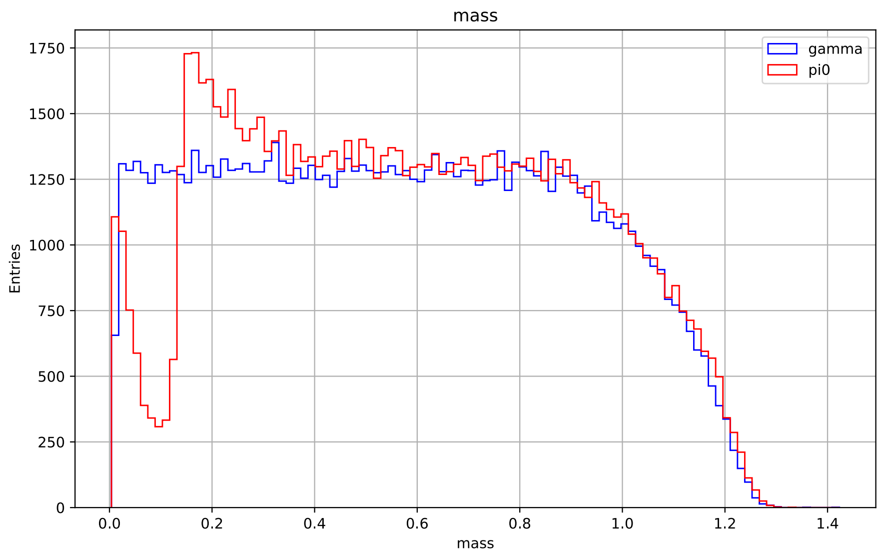
- Calculated in each layer, expected to have good separation especially in the strip:

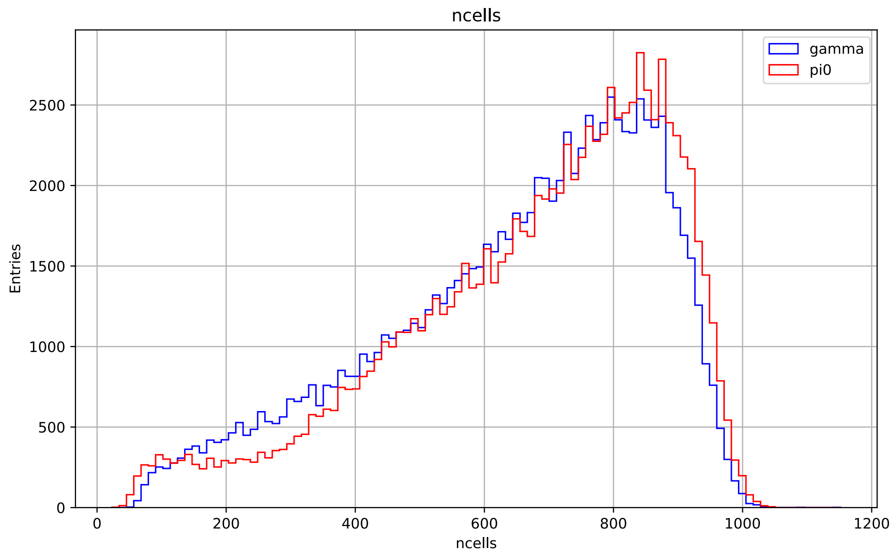
- Ratio_E vs. theta: (E_max - E_2ndmax) / (E_max + E_2ndmax)   [will be 1 if no E_2ndmax found]
- Delta_E vs. theta: E_2ndmax - E_min
  - E_max and E_2ndmax found in 1-D theta spectrum
  - E_min found in the theta range of E_max and E_2ndmax
- Ratio_E vs. phi and Delta_E vs. phi, similarly as in theta:
  - E_max and E_2ndmax found in 1-D module spectrum
- Width in theta, taking account only N bins around the cell with E_max
  - N = 3, 5, 7, 9
- E fraction side: E(within up to +-N cells around E_max) / E(within up to +-1 cells around E_max) - 1.0
  - N = 2, 3, 4
  - Performed with 1-D theta spectrum

Distributions

- In the following shape parameter distributions:
  - Blue: photon (100k events) Red: pi0 (100k events)
  - Strip in L3
- Full set of distributions: LINK

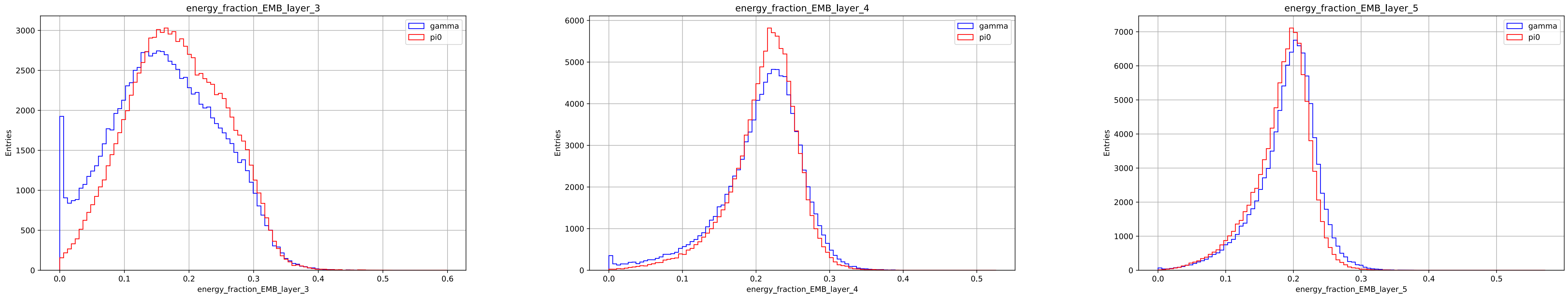
- Mass and number of cells in clusters



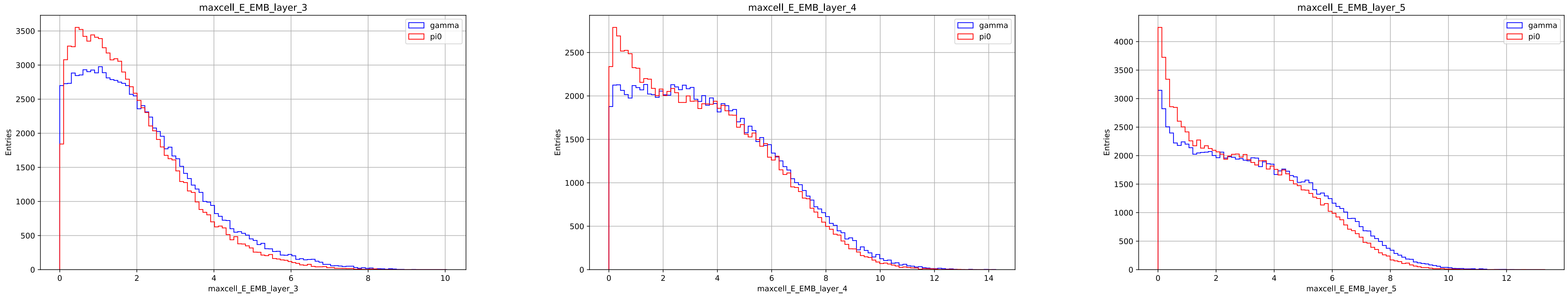


Distributions

- Energy fraction in L3, L4, L5

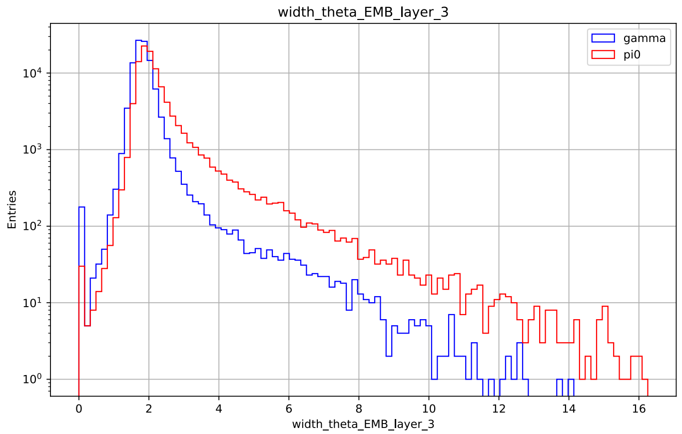


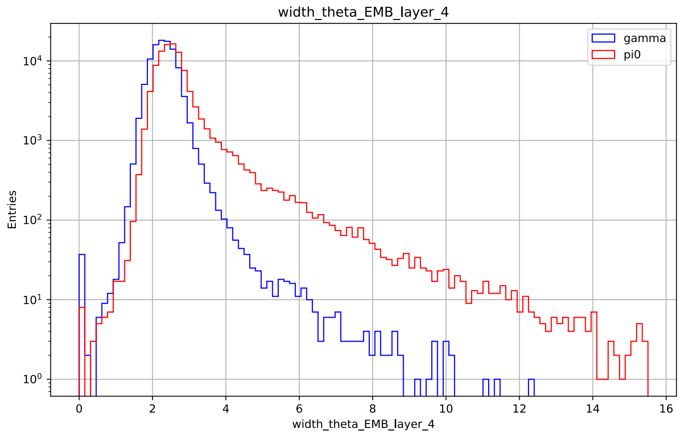
- Maximum energy of cell in L3, L4, L5

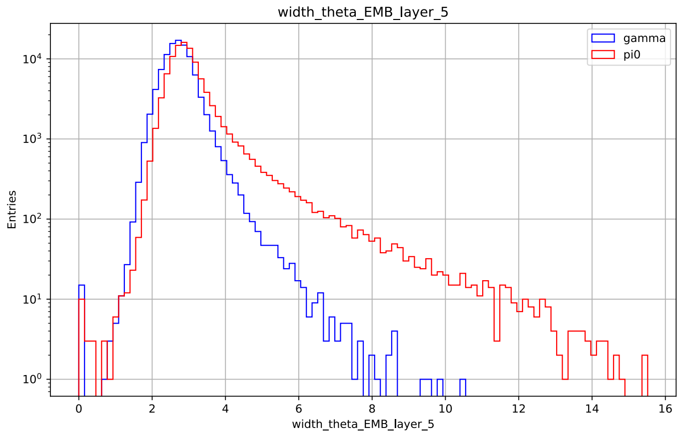


Distributions

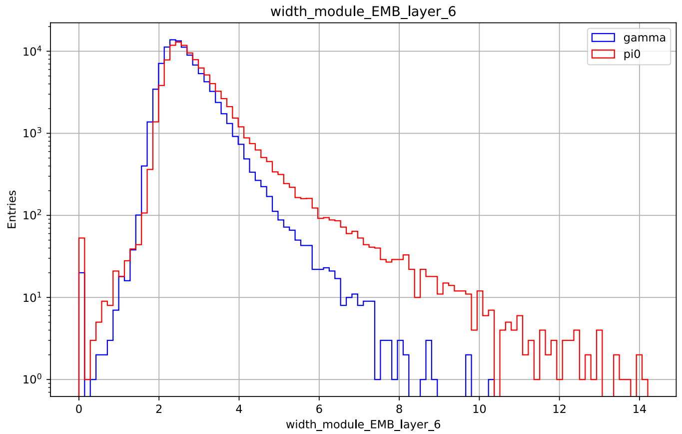
- Width in theta in L3, L4, L5

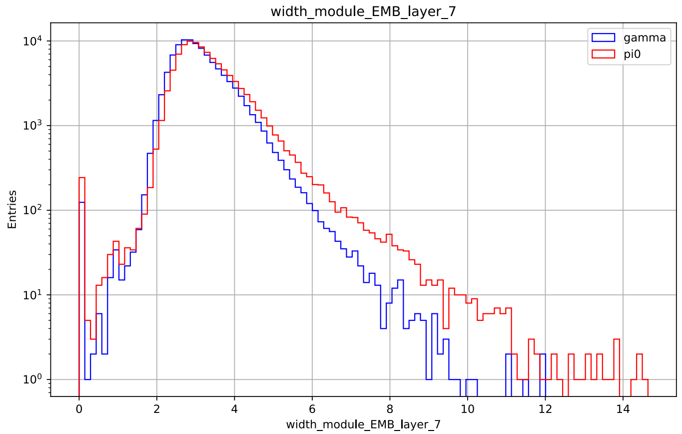


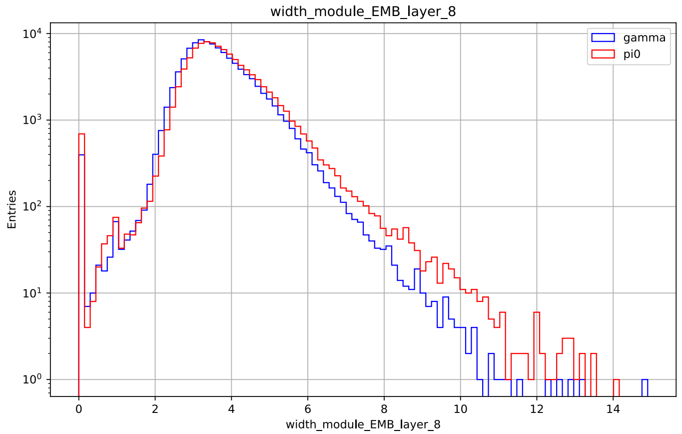




- Width in module in L6, L7, L8

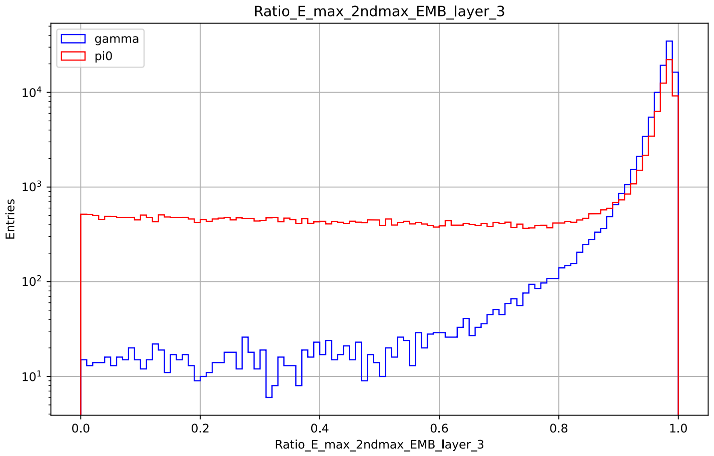


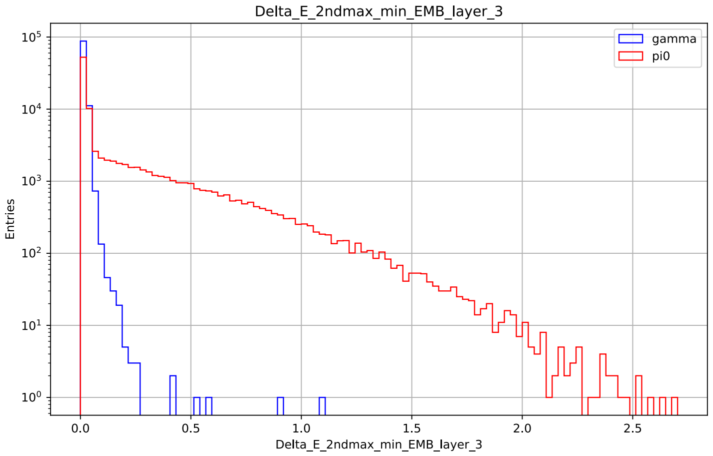




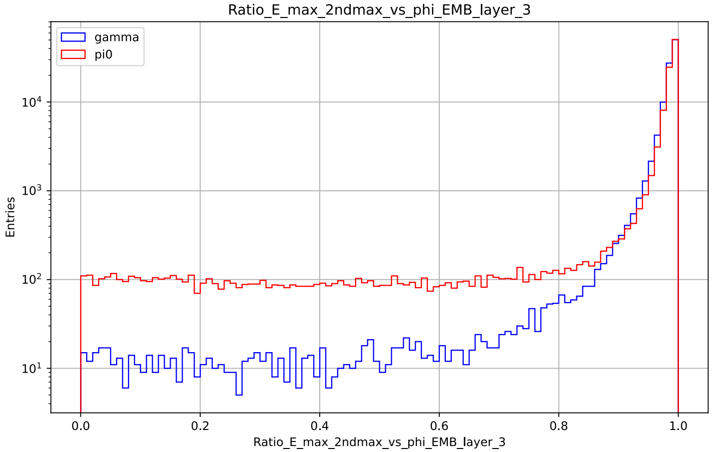
Distributions

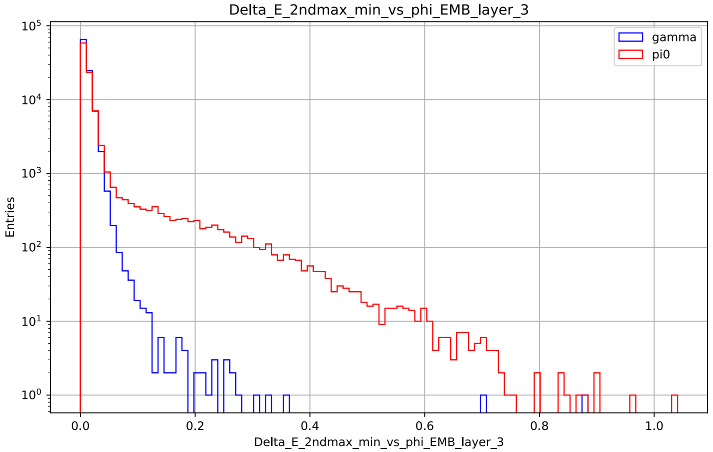
- Ratio_E and Delta_E vs. theta in L3





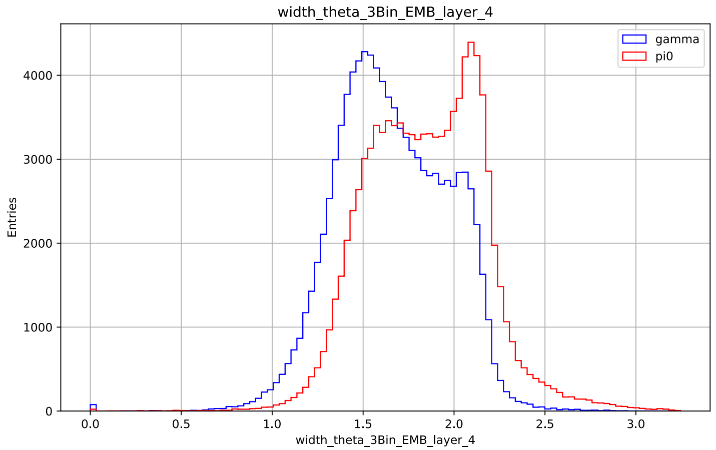
- Ratio_E and Delta_E vs. phi in L3

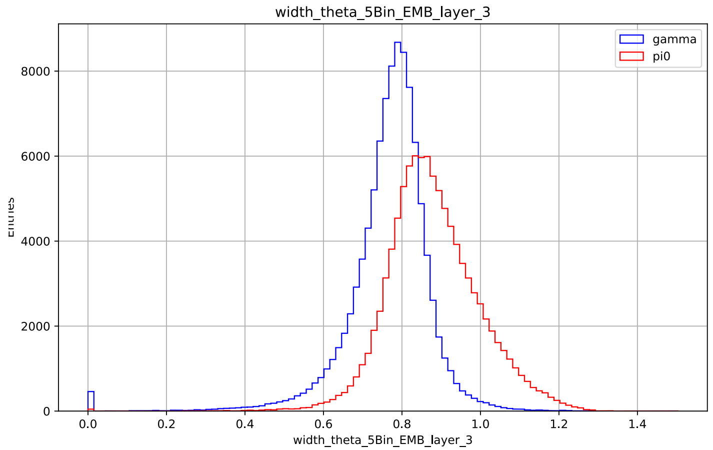


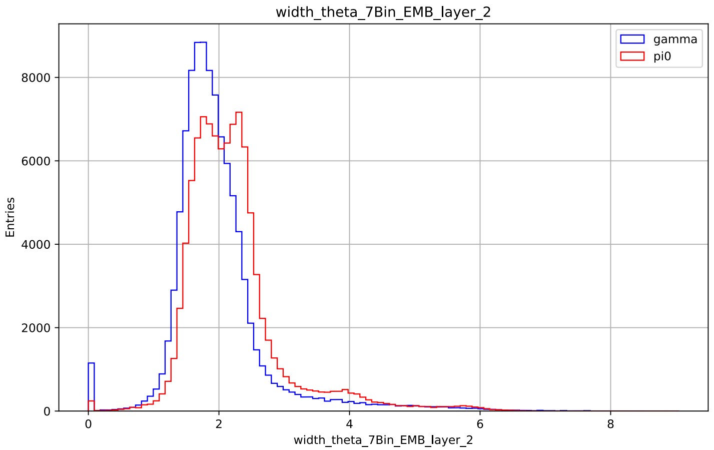


Distributions

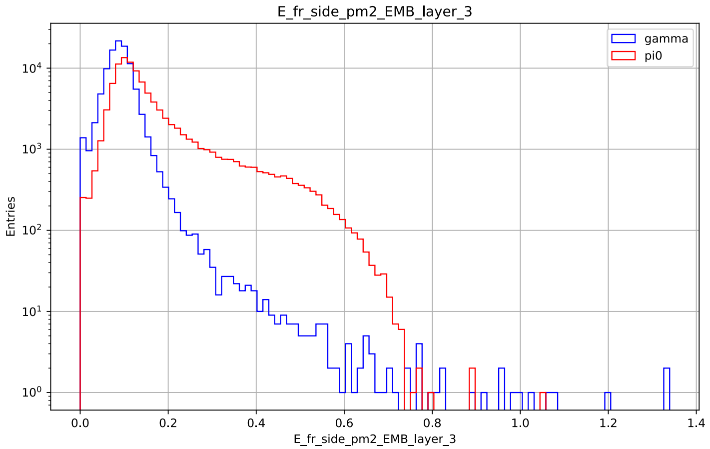
- Width in theta calculated from 3, 5, 7 cells in L4, L3, L2

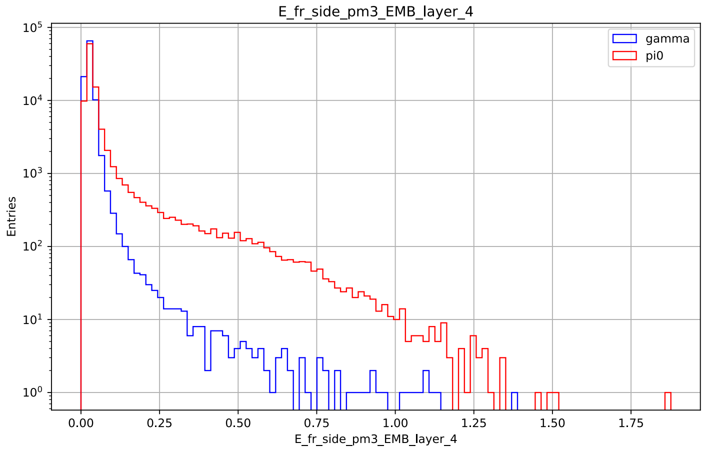


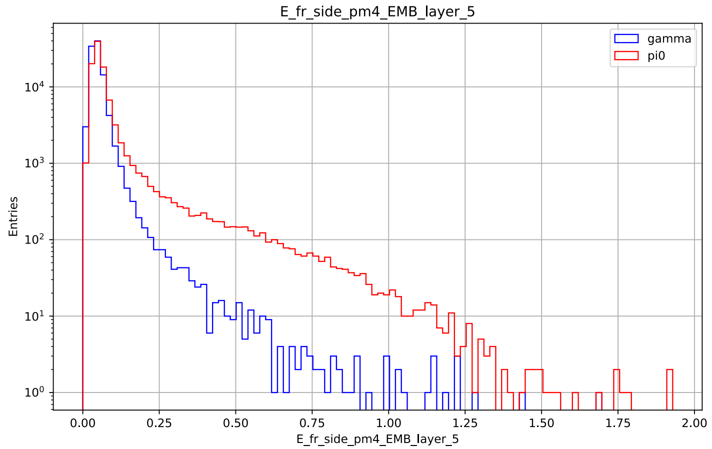




- E fraction side calculated up to +- 2, 3, 4 cells in L3, L4, L5







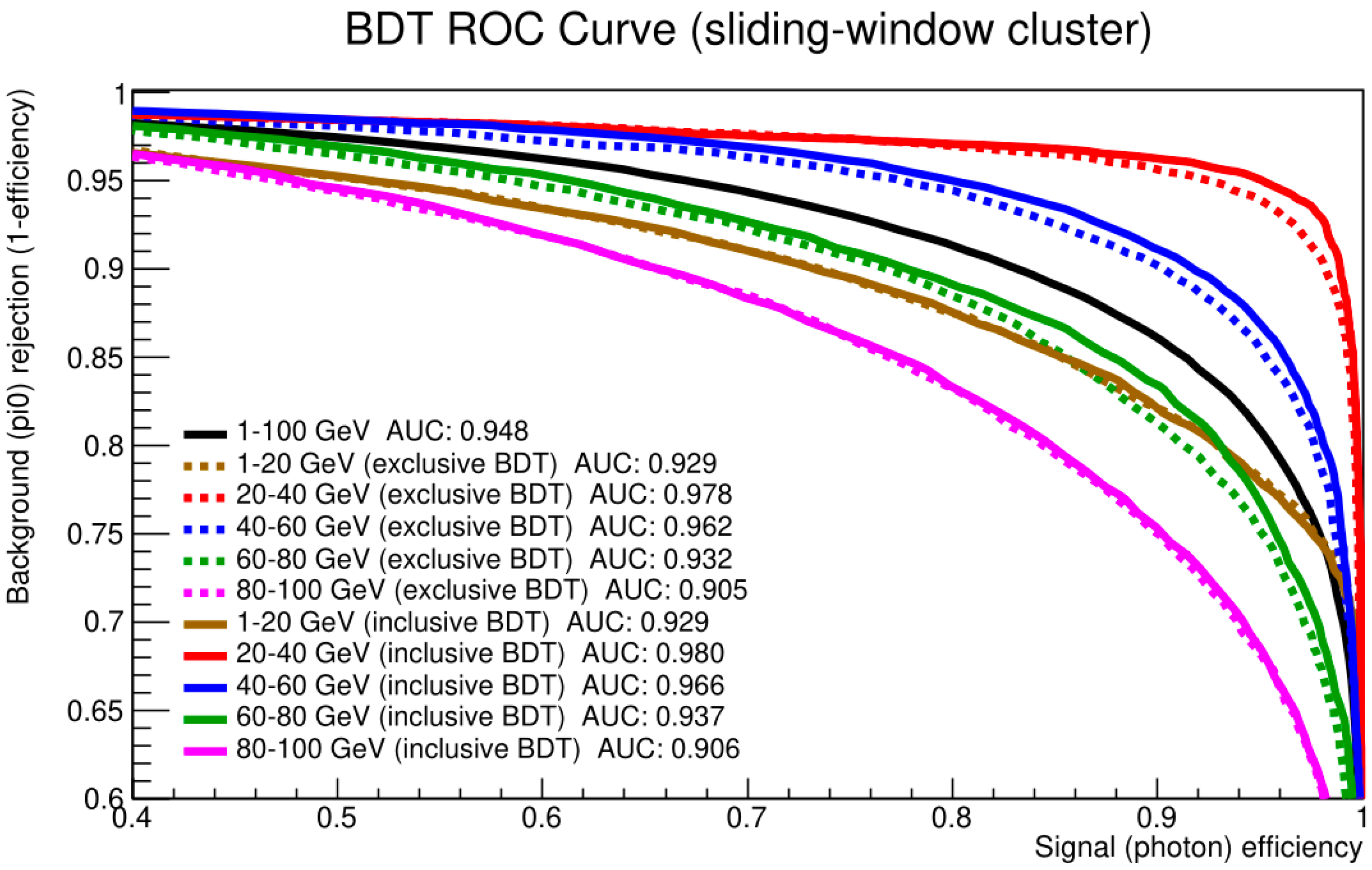
BDT setup

- Train the BDT using a selected set of shape parameters (83 in total):
  - (Sliding-window) cluster energy, mass, number of cells 3
  - Ratio_E vs. theta / phi, L1 to L5 10
  - Delta_E vs. theta / phi, L1 to L5 10
  - Maximum energy of cell, L1 to L5 5
  - Energy fraction, L1 to L8 8
  - Width in theta / phi, L1 to L6 12
  - Width in theta of 3 / 5 / 7 / 9 cells, L1 to L5 20
  - Energy fraction side of +- 2 / 3 / 4 cells, L1 to L5 15

- Photon as signal (100k), pi0 as background (100k)
- Half for training, the other half for test
- BDT hyper parameter optimised

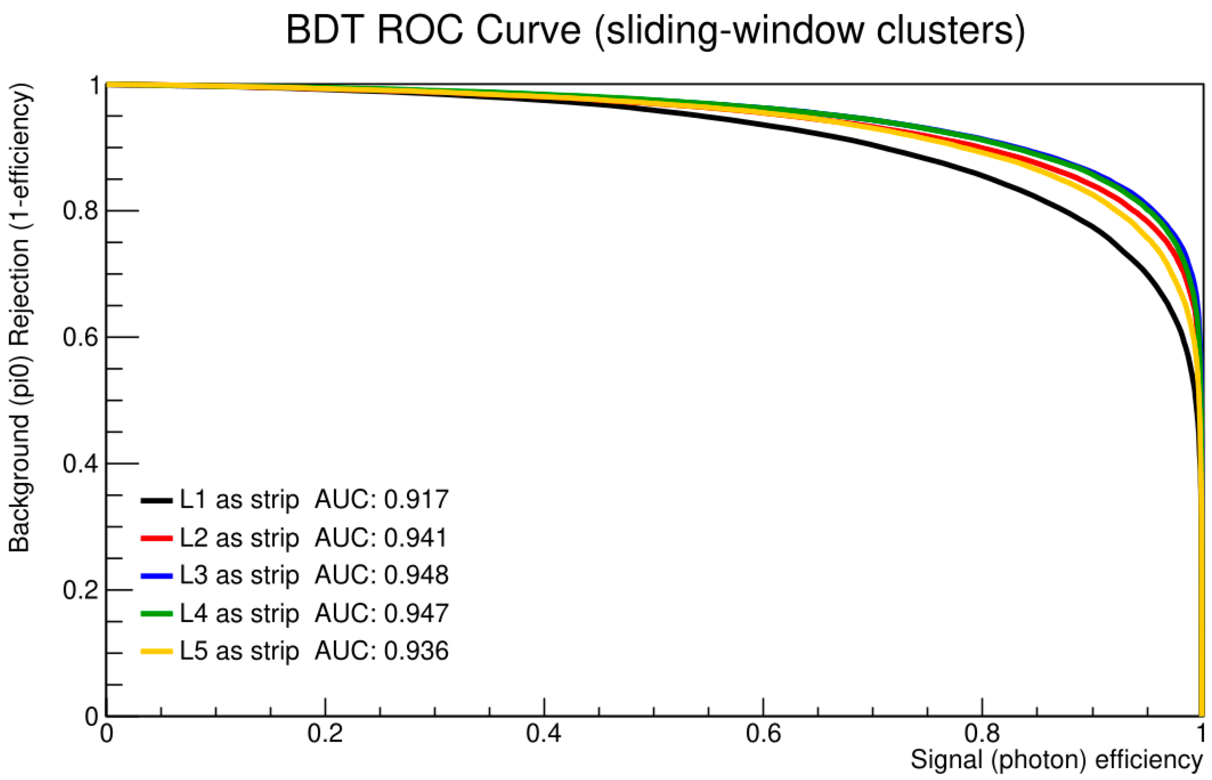
BDT: inclusive and exclusive vs. E

- Train inclusive BDT (1-100 GeV) and exclusive BDTs in 5 E_cluster intervals:
  - 1-20 GeV
  - 20-40 GeV
  - 40-60 GeV
  - 60-80 GeV
  - 80-100 GeV
- ROC curve derived from test sample
- For clusters with very low (< 20) and high energy (> 60), BDT performances get worse
- Inclusive BDT as good as exclusive BDTs



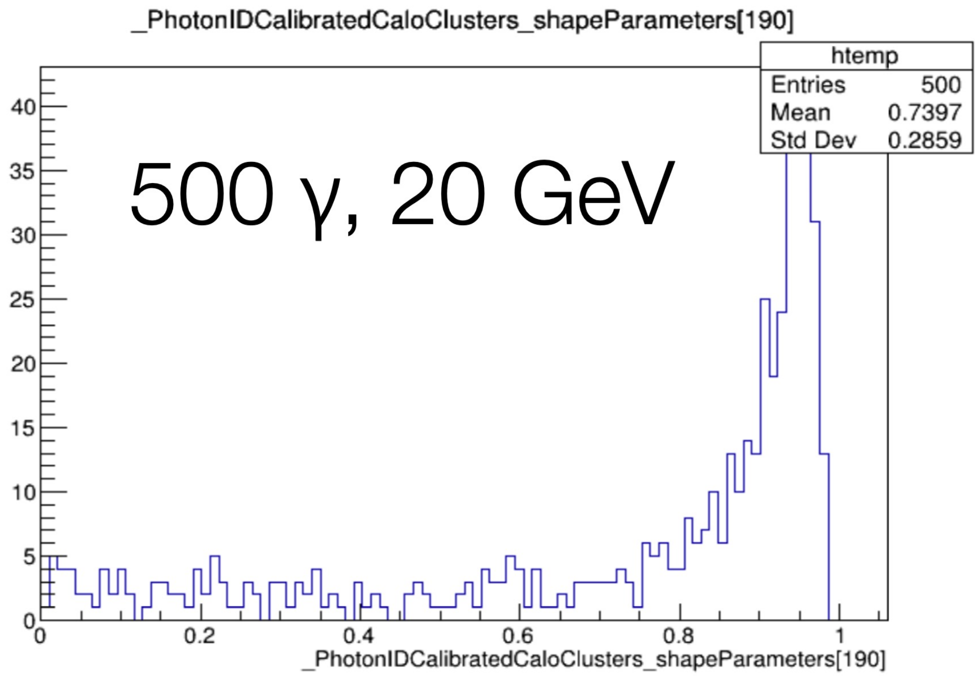
BDT: shift of strip layer

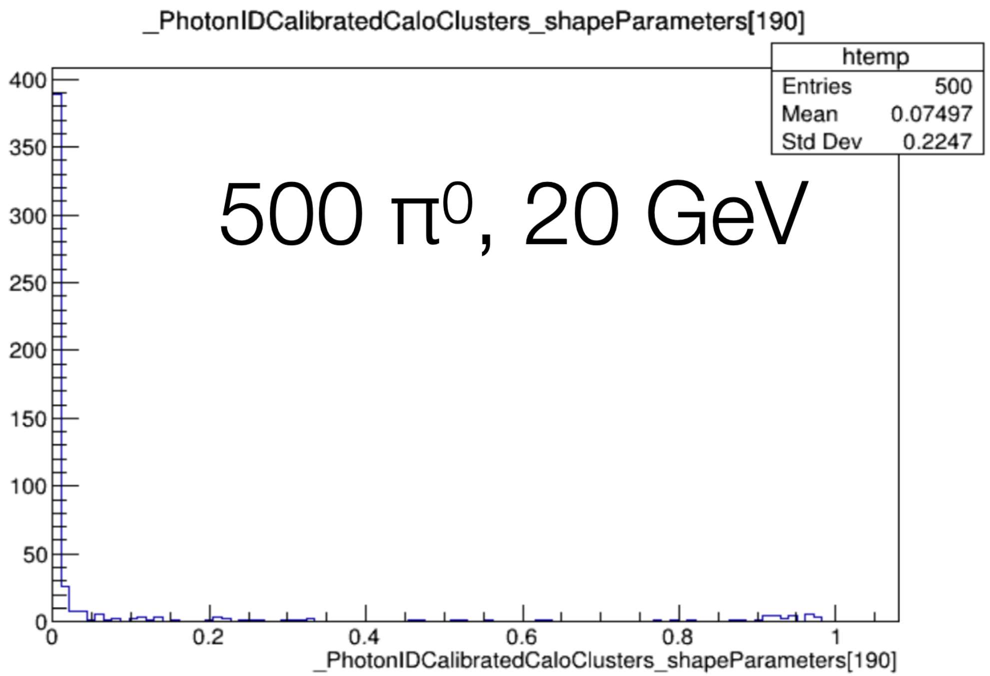
- In default geometry, strip is in L1
  - 4 times finer theta granularity than others
- From observing distributions of shape parameters, L1 might not the best choice of strip
- Generate samples using custom detector versions
  - Shift strip layer to L2, L3, L4, L5
  - 100k events for photon / pi0 each
- From the ROC curve:
  - L3 has the best performance (AUC 0.948)
    - L4 is very close (AUC 0.947)



Running the photon ID algorithm in Gaudi

- Implemented a Gaudi algorithm that Pull Request: PR
    - Reads the list of input features from a JSON file and the trained BDT model from an ONNX file
    - For each cluster, reads the input features from the shapeParameters vector (if available), runs the inference, and saves the photon probability as an additional shape parameter of the new output cluster collection
    - Works with ONNX files created with either XGBoost and LightGBM (tested with both)
    - Could probably work also with different models based on features (DNN..)





Summary and outlook

- Shape parameters calculated for photon/pi0 identification with ALLEGRO: Pull Request
  - Full set of shape parameter distributions: LINK

- BDTs trained using shape parameters
  - to test photon/pi0 separations vs. energy of cluster
  - to find a better/the best position of strip

- Implemented the algorithm that runs the photon ID in Gaudi: Pull Request

➢ Optimise further the input list for BDT

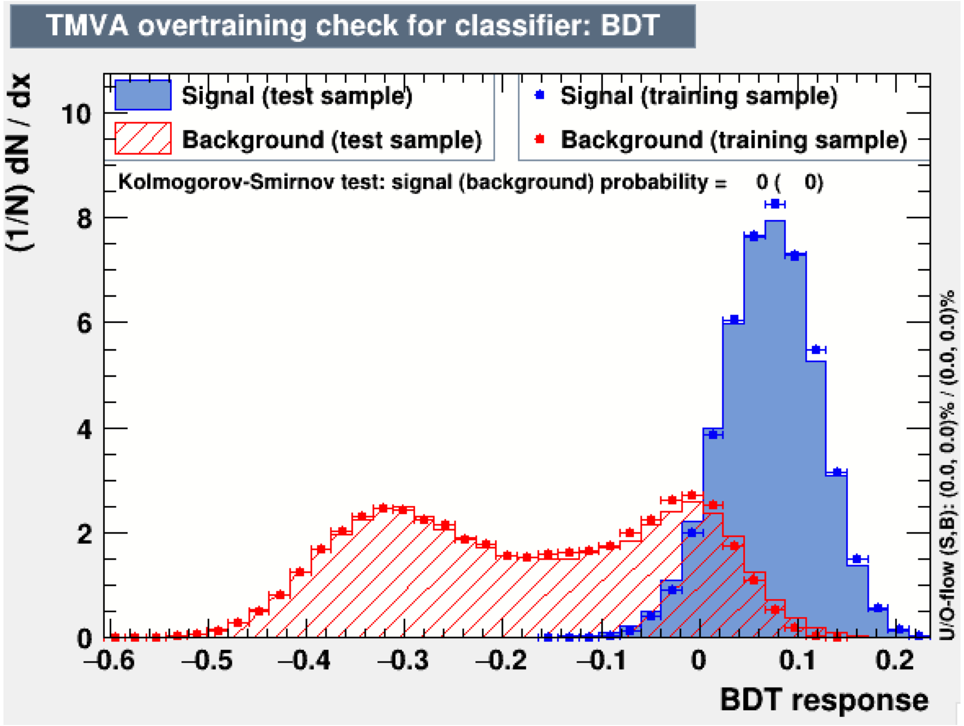
Back Up

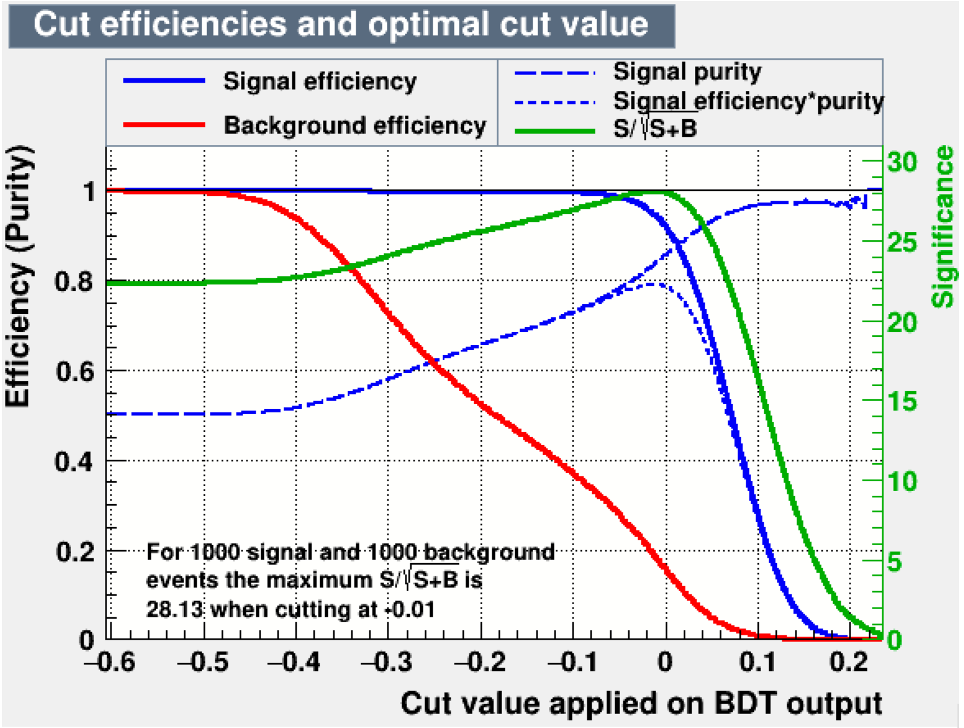
BDT

- Optimised BDT hyper parameters:

```
nTrees=1000:MaxDepth=4:BoostType=AdaBoost:AdaBoostBeta=0.6:SeparationType=GiniIndex:nCuts=20:MinNodeSiz
e=1:PruneMethod=NoPruning"
```

- Inclusive BDT (1-100 GeV), L3 as strip





Variable importance (Top 15)

- Delta_E_2ndmax_min_EMB_layer_3
- maxcell_E_EMB_layer_4
- mass
- maxcell_E_EMB_layer_3
- width_theta_5Bin_EMB_layer_3
- width_theta_3Bin_EMB_layer_3
- E_fr_side_pm2_EMB_layer_3
- width_theta_7Bin_EMB_layer_3
- ncells
- width_theta_3Bin_EMB_layer_2
- Ratio_E_max_2ndmax_EMB_layer_3
- width_module_EMB_layer_3
- Energy
- width_module_EMB_layer_2
- width_theta_3Bin_EMB_layer_4

Strip variables ranked higher

References

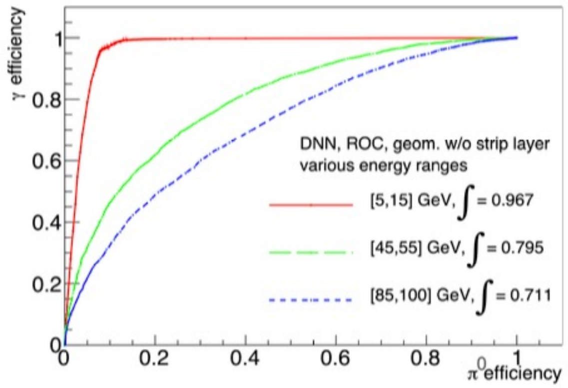
- Pavlo & Brieuc study (DNN, CNN, hybrid..): https://cds.cern.ch/record/2836383/files/LAr_particle_separation.pdf
  - A neutral pion mis-identification probability of 10% for a 95% photon efficiency working point is achieved with a regular geometry and a Hybrid Neural Network approach.

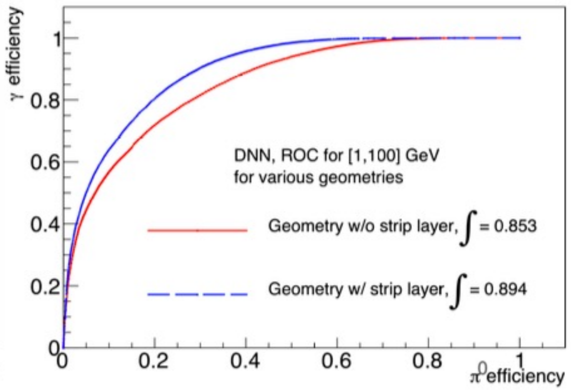
◇ 50k (out of total 100k; see further the reason) events per particle

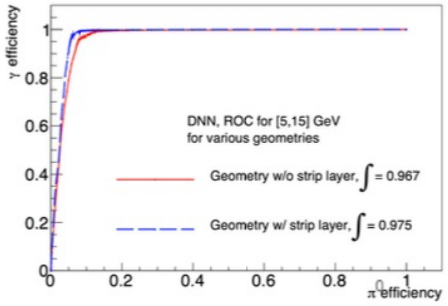
◇ No noise included

◇ Geometries with/without 2^{nd} layer geometries (explanation follows)

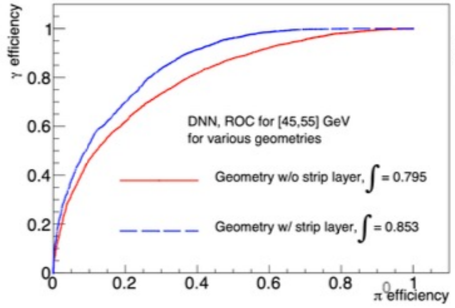
- 2 geometries: one w/o strip layer (referred as to uniform celling geometry) denotes celling with uniform size in η and ϕ (Δη = 0.01 and Δϕ ≈ 8 mrad), and one geometry w/ strip 2nd layer uses finer (4x) resolution in η along the 2nd radial layer



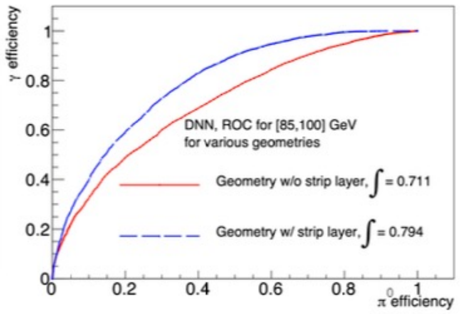




(a) {5, 15} GeV



(b) {45, 55} GeV



(c) {85, 100} GeV

References

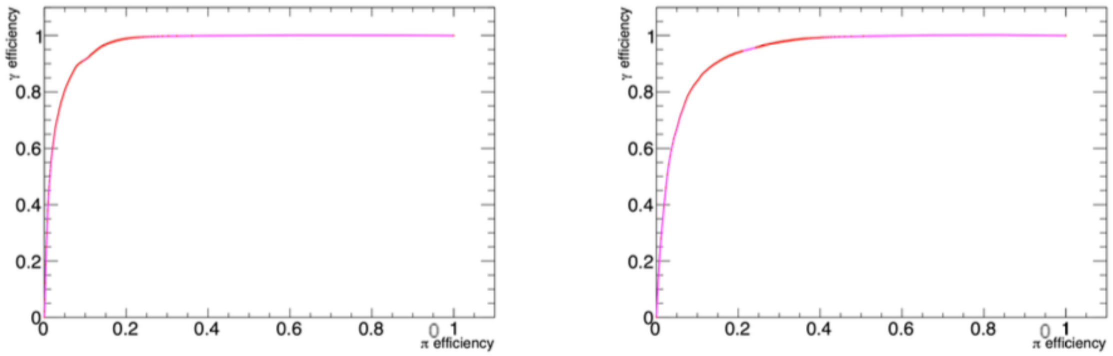
- Pavlo & Brieuc study (DNN, CNN, hybrid..): https://cds.cern.ch/record/2836383/files/LAr_particle_separation.pdf
  - A neutral pion mis-identification probability of 10% for a 95% photon efficiency working point is achieved with a regular geometry and a Hybrid Neural Network approach.

◇ 50k (out of total 100k; see further the reason) events per particle

◇ No noise included

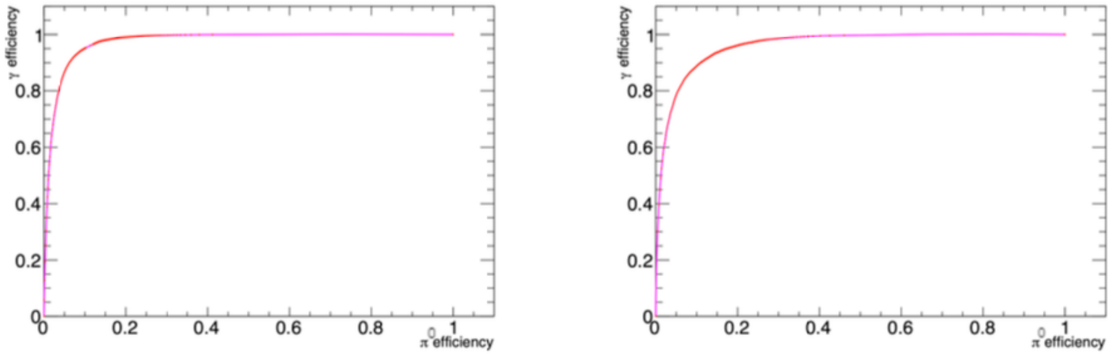
◇ Geometries with/without 2^{nd} layer geometries (explanation follows)

  - 2 geometries: one w/o strip layer (referred as to uniform celling geometry) denotes celling with uniform size in η and ϕ (Δη = 0.01 and Δϕ ≈ 8 mrad), and one geometry w/ strip 2^{nd} layer uses finer (4x) resolution in η along the 2^{nd} radial layer



(a) uniform celling , $\int = 0.968$ (b) strip 2^{nd} layer , $\int = 0.944$

Figure 23: Inclusive ROC curve for CNNs built on top of various geometries.



(a) Uniform celling , $\int = 0.974$ (b) Strip 2^{nd} layer , $\int = 0.960$

Figure 26: Inclusive ROC curve for HNNs built on top of various geometries.]