Development of particle flow algorithm with GNN for Higgs factories

Tatsuki Murata

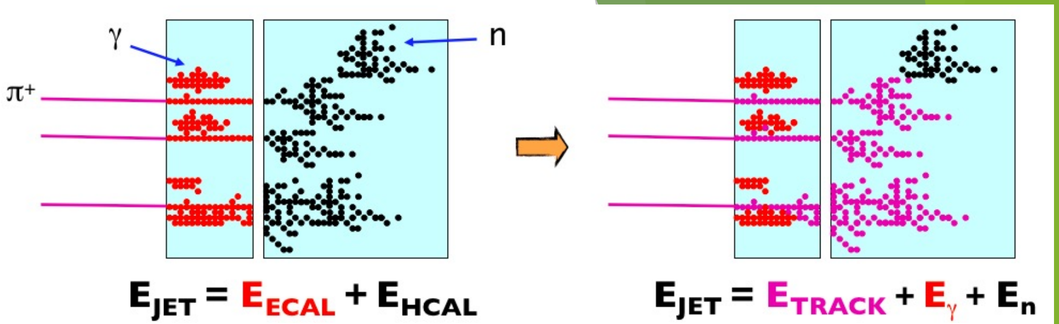
July 10th, 2024

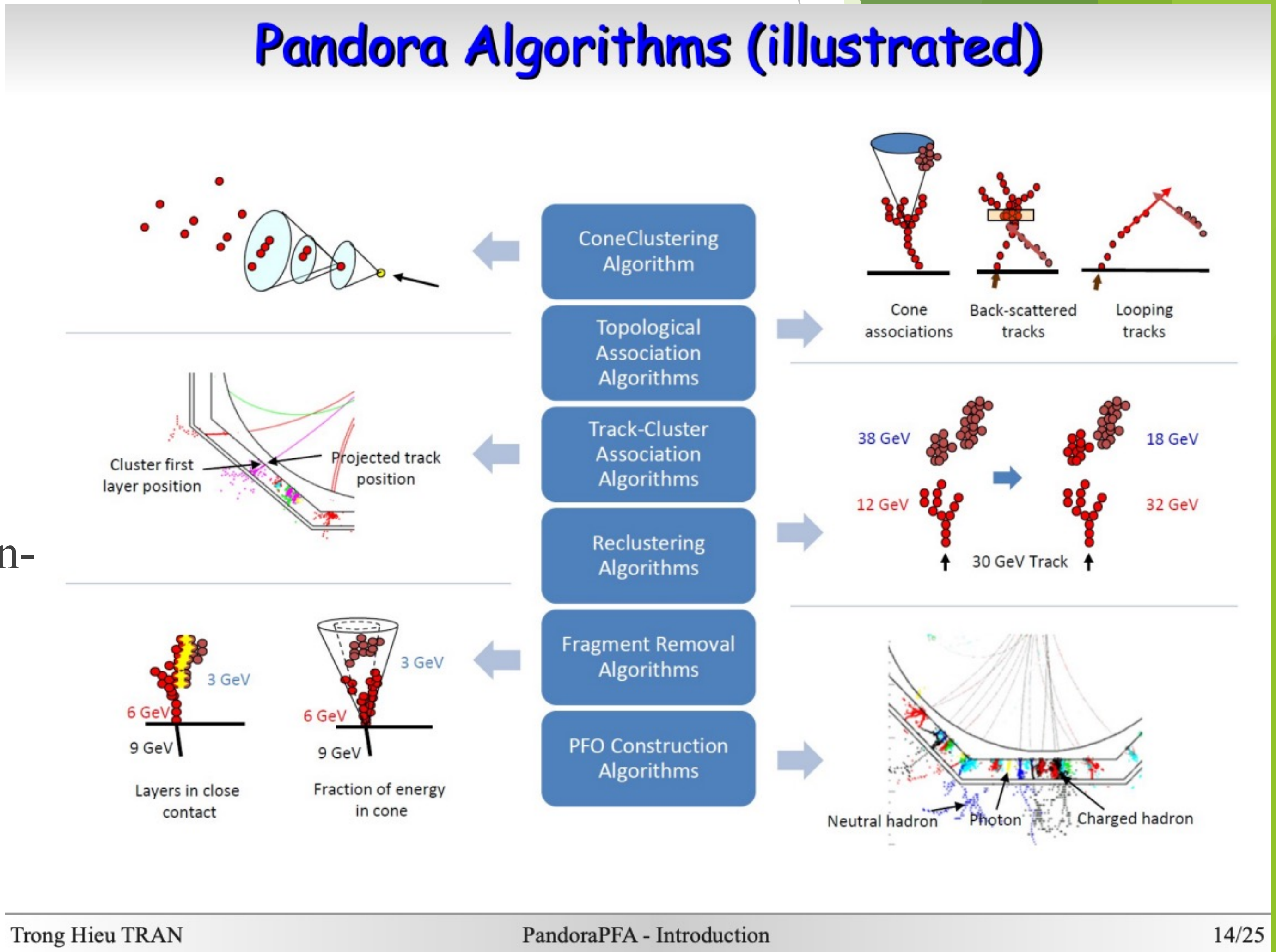
LCWS 2024

Collaborators: T. Suehara (ICEPP, U. Tokyo), T. Tanabe (MI-6 Co.),
L. Gray (Fermilab), P. Wahlen (IP Paris & ETHZ / internship at Tokyo)

Particle flow

- Many detectors are PFA-oriented designed
  - ILD, SiD, etc …
- Separation of cluster at calorimeter
  - Charged or neutral cluster
- Essential for jet energy resolution
- Current algorithm : PandoraPFA
  - Pattern recognition based on the human-tuned parameters
  - Combination of various process
  - Not easy to optimize/add more info.
    - Timing information, etc…

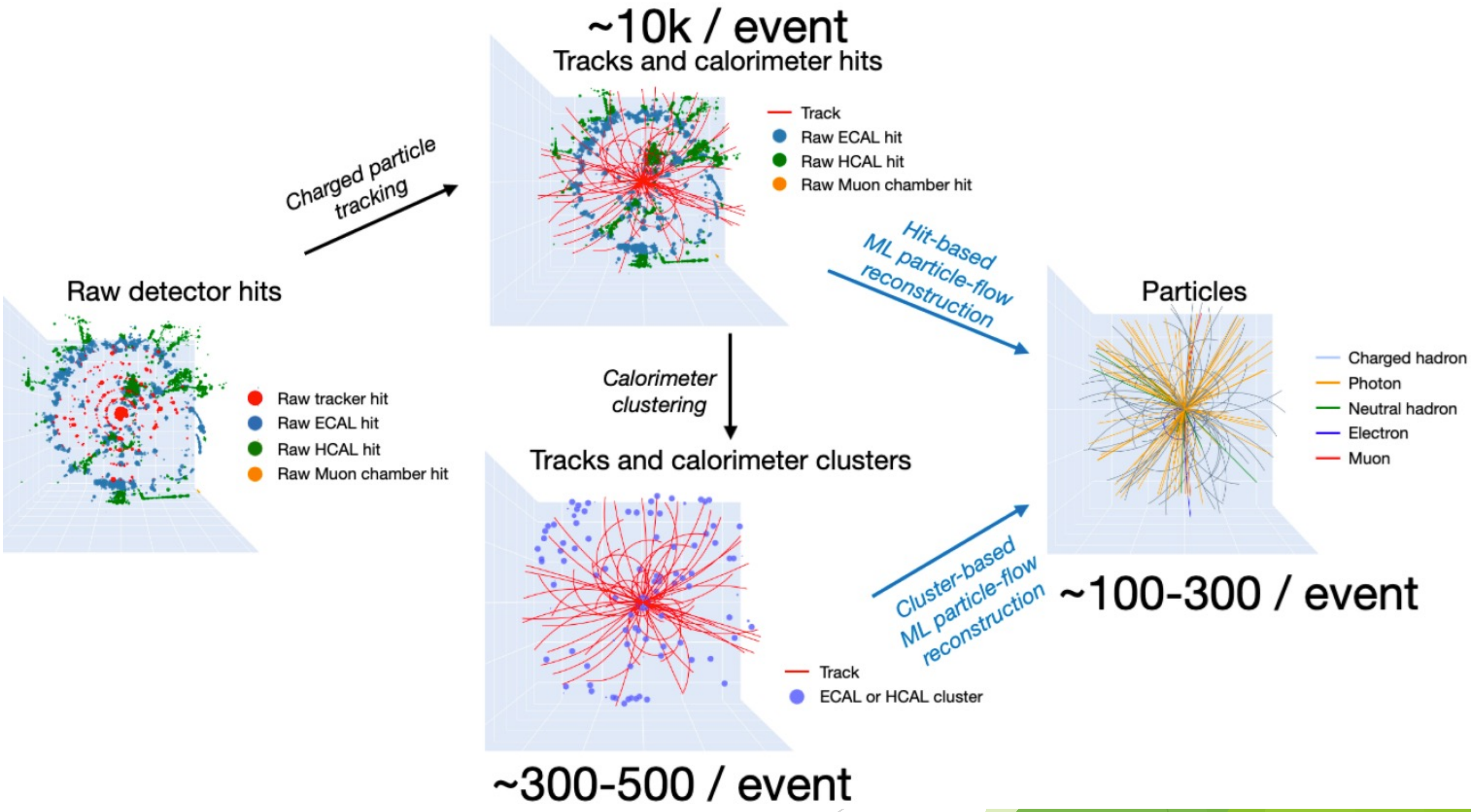




Two ways for particle flow

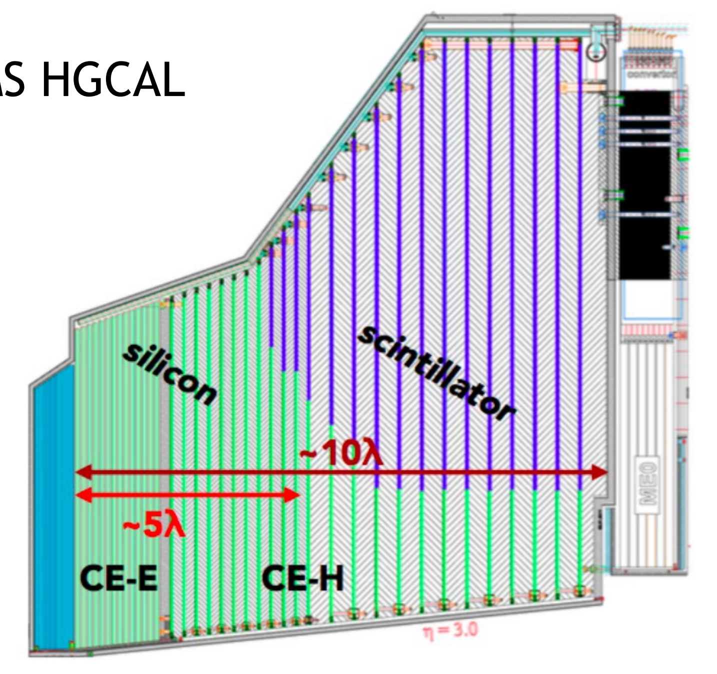
- Track-cluster matching from calorimeter hits
  - More freedom
  - Distance-based connection
  - More efficient

- Track-cluster matching from subclusters
  - Less input
  - Additional clustering is needed



GravNet for CMS HGCAL

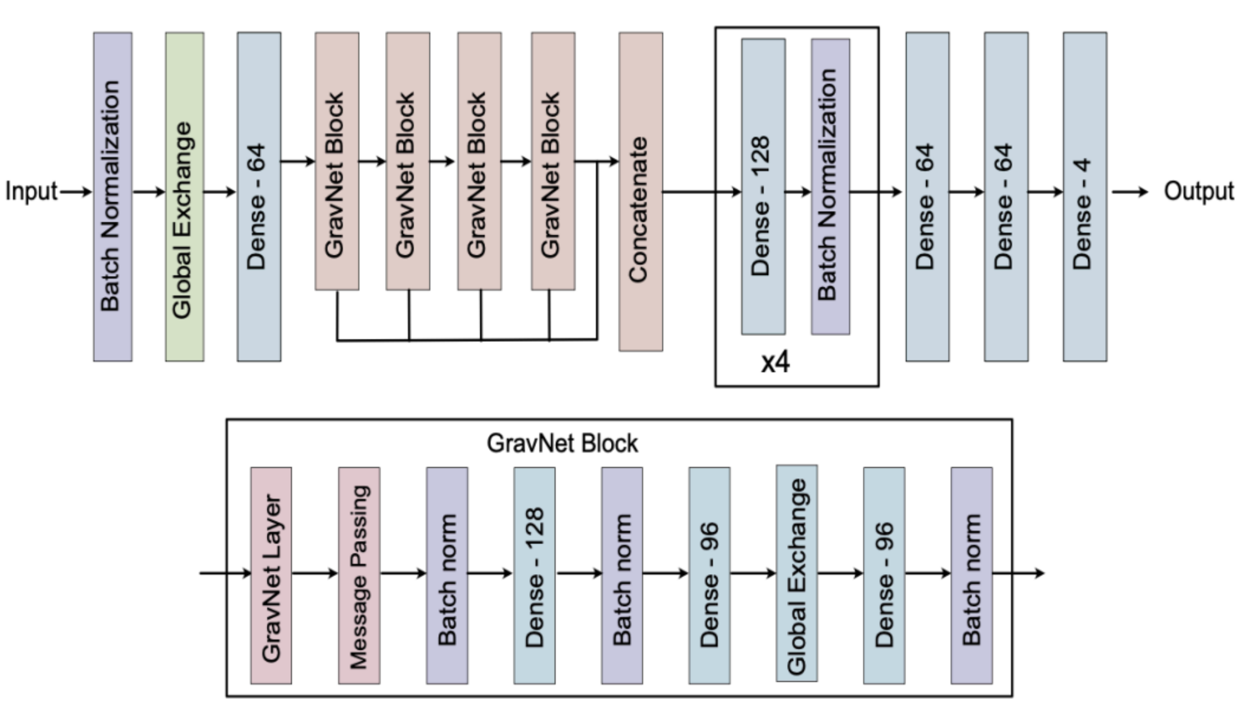
- CMS HGCAL
  - High granular forward calorimeter for HL-LHC upgrade at CMS
  - Similar to ILD calorimeter (Silicon pixel + scintillator)
    - Inspired by CALICE development

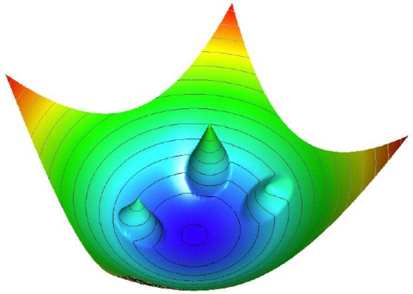


- Reconstruction at HGCAL
  - Pileup/noise separation by software
  - Numerous particles from ~200 pileups
    - Difficult to handle
    - DNN reconstruction is investigated

Network

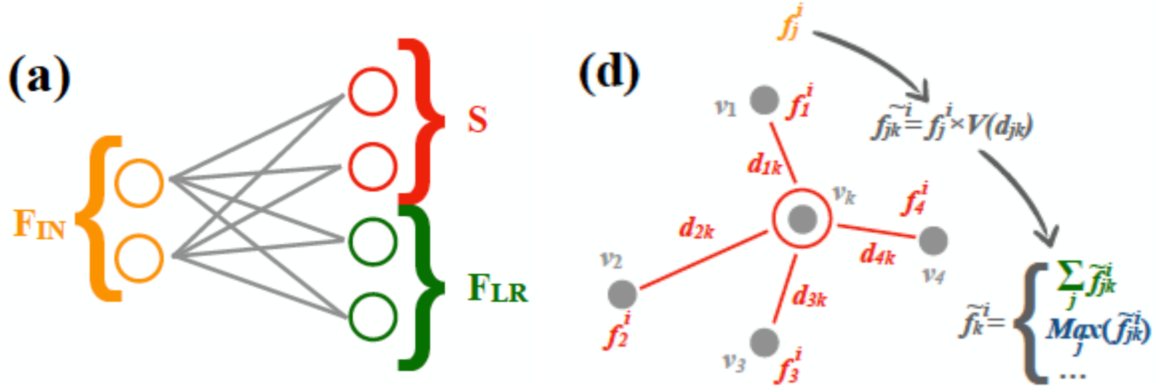
- Input/output are obtained for each hit at calorimeter
  - Input :features at each hit (position, energy deposit, timing)
  - Output:
    “condensation coefficient” β
    position of virtual coordinate (2-dimention)
    (optional)
    energy, PID





GravNet and object condensation

- GravNet arXiv:1902.07987
- The virtual coordinate (S) is derived from inputs with simple multilayer-perceptron(MLP)
- Convolution using "distance" at S (bigger convolution with nearer hits)
- Concatenate the output with MLP



arXiv:2002.03605

Object condensation (loss function)

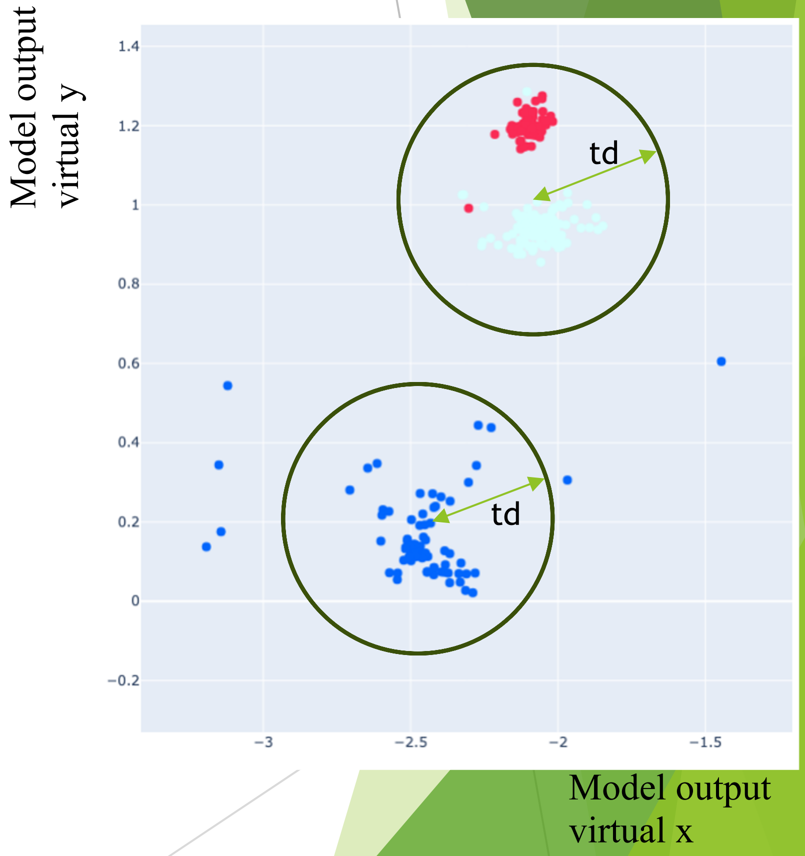
$$L = L_p + s_C(L_\beta + L_V)$$

- <u>Condensation point : the hit with largest β at each MC cluster</u>
- L_V: attractive potential to the condensation point of the same cluster and repulsive potential to the condensation point of different clusters
- L_β: pulling up β of the condensation point (up to 1)
- (L_p: regression to output features)

Modifications from CMS HGCAL algorithm

- Putting tracks as “virtual hits”
  - Locate at entry point of calorimeter (have “track” flag)
  - Energy deposit = 0
  - Forcibly treat tracks as condensation points regardless of β
    - β of tracks become spontaneously close to 1due to L_β term in loss function

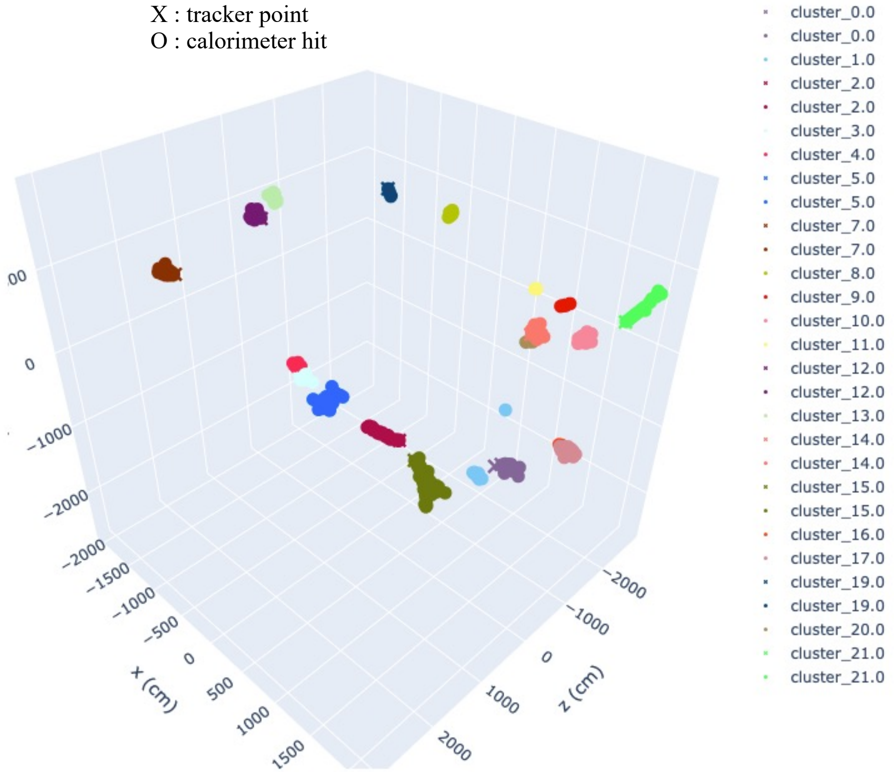
- Clustering algorithm
  - Hits that are within a certain distance (td) from the highest β point assume as a cluster
  - Continues clustering until all hits are clustered or β of remaining hits are below threshold (tbeta)



Samples for performance evaluation

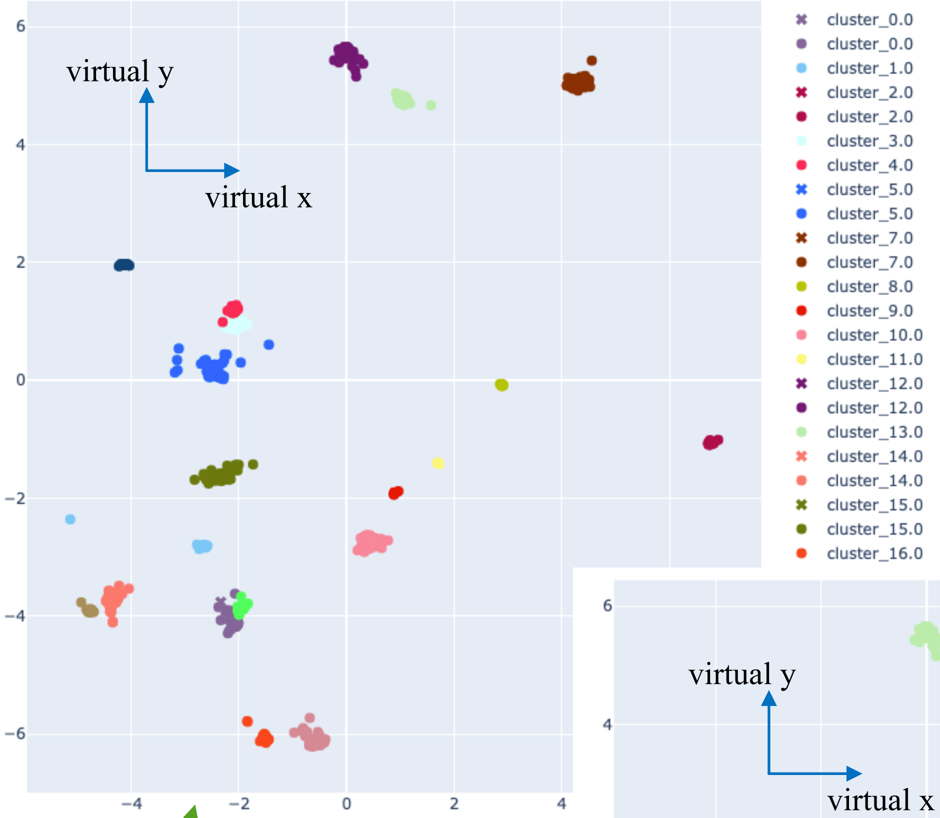
- ILD full simulation with SiW-ECAL and AHCAL (ILD_l5_o1_v02, 020301)
  - ECAL : 5×5 mm^2, 30 layers   HCAL : 30×30 mm^2, 48 layers
  - Two types of samples : τ, jets(u, d, s)
    - τ^-(10 GeV)
      - τ has many decay modes, hadrons, leptons, and photons
      - Good for training
    - qq (q= u, d, s) (91 GeV)
      - Official sample for PFA calibrations
  - Converted to awkward array stored in HDF5 format

Event display



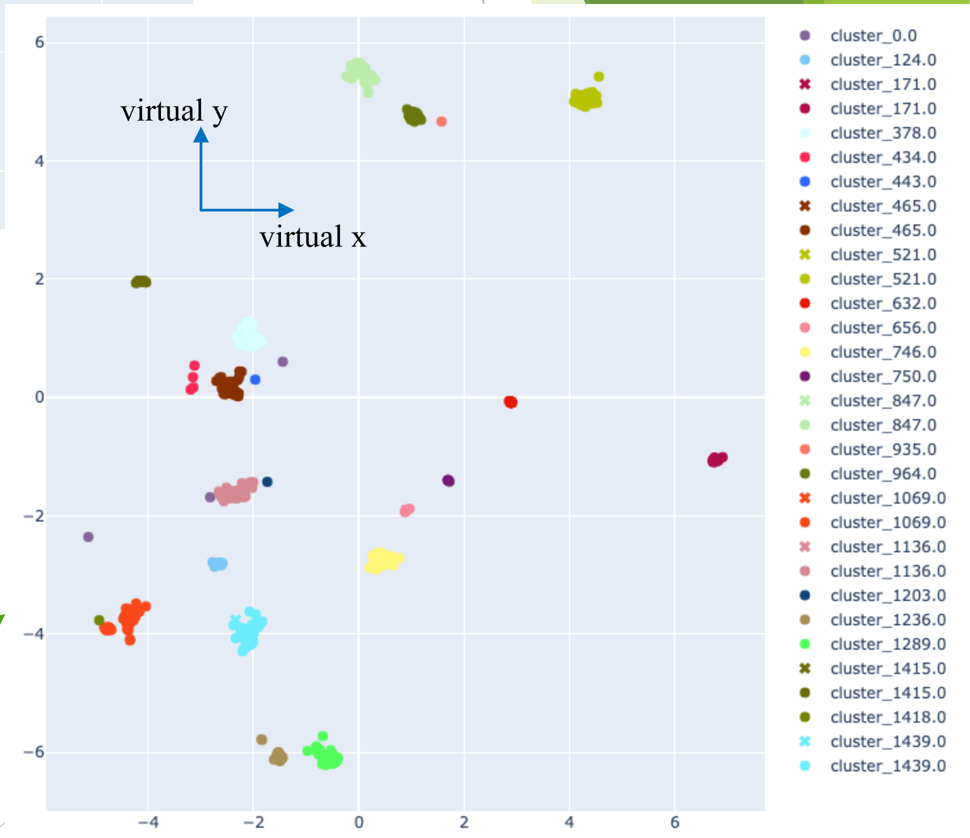
Input features
Real coordinate in detector

Colored by true clusters



Output features
Virtual coordinate

Colored by true clusters



Colored by reconstructed clusters

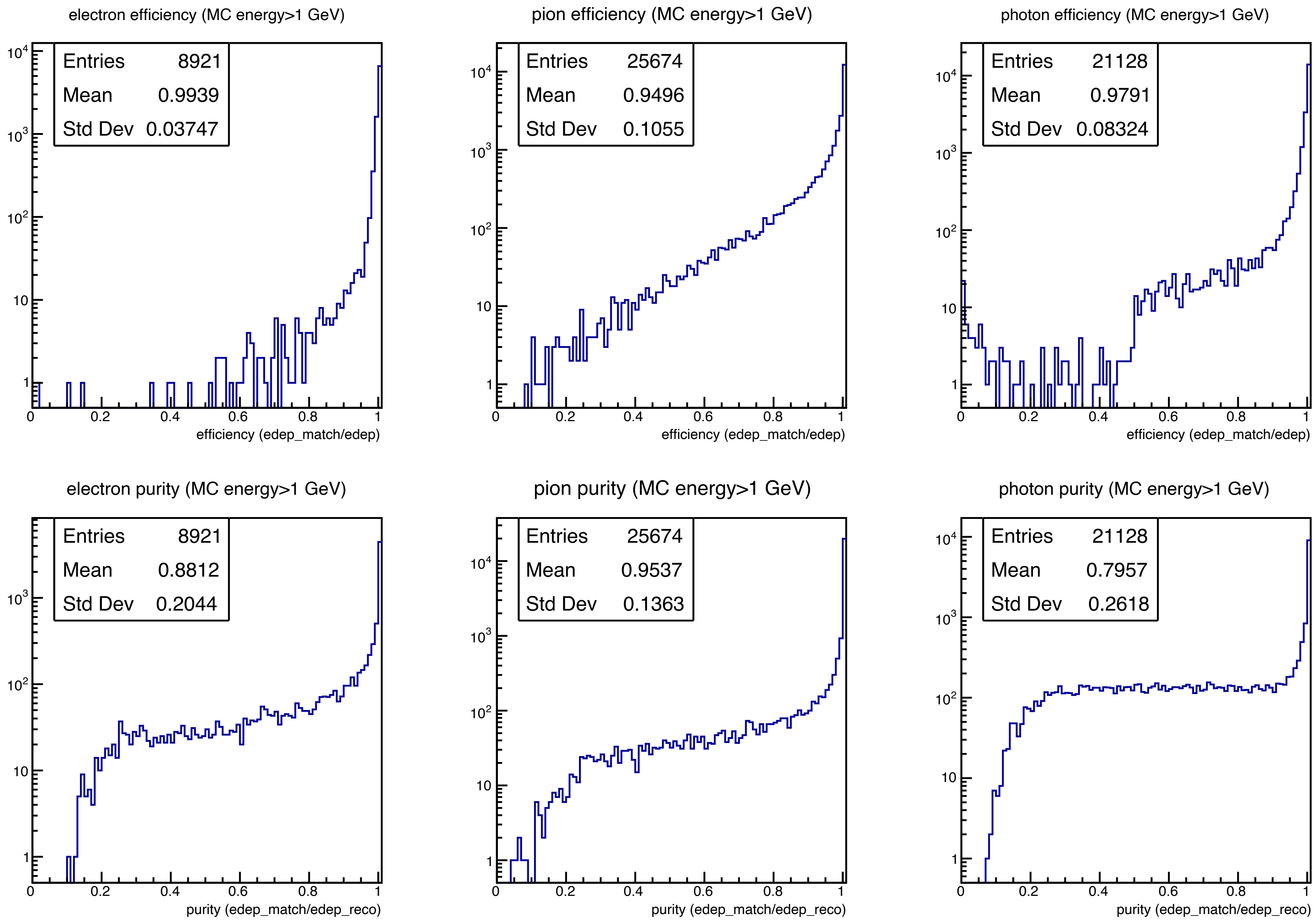
Quantitative evaluation

- Make 1-by-1 connection of MC and reconstructed cluster
  - Reconstructed cluster with highest fraction of hits from the MC is taken
  - Multiple reconstructed cluster may connect to one MC cluster

- Quantitative comparison with PandoraPFA
  - Compared “efficiency” and “purity” of particle flow
    - Efficiency : (reconstructed cluster energy that matches the MC cluster) / (MC cluster energy)
    - Purity : (reconstructed cluster energy that matches the MC cluster) / (reconstructed cluster energy)

Efficiency and purity for GNN, tau train / tau prediction

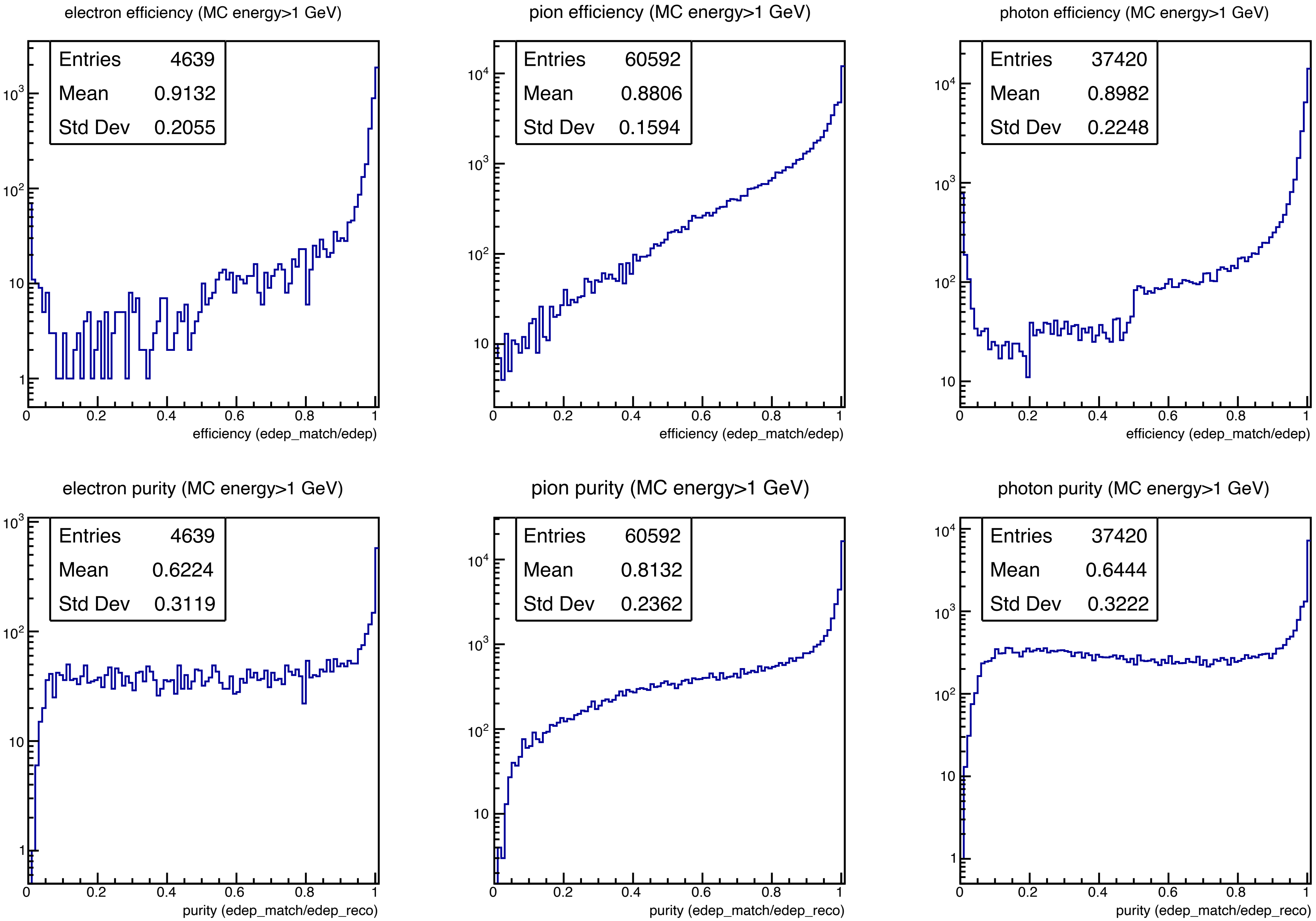
- Efficiency :
  over 90% for all particles
  slightly low in pions
- Purity :
  over 88% for all tracks
  79% for photons
  merged photons?
- Reasonably well reconstructed



Efficiency and purity for GNN, tau train / qq prediction

- **Efficiency :**
  over 88% for all particles
  slightly worse than tau pred.

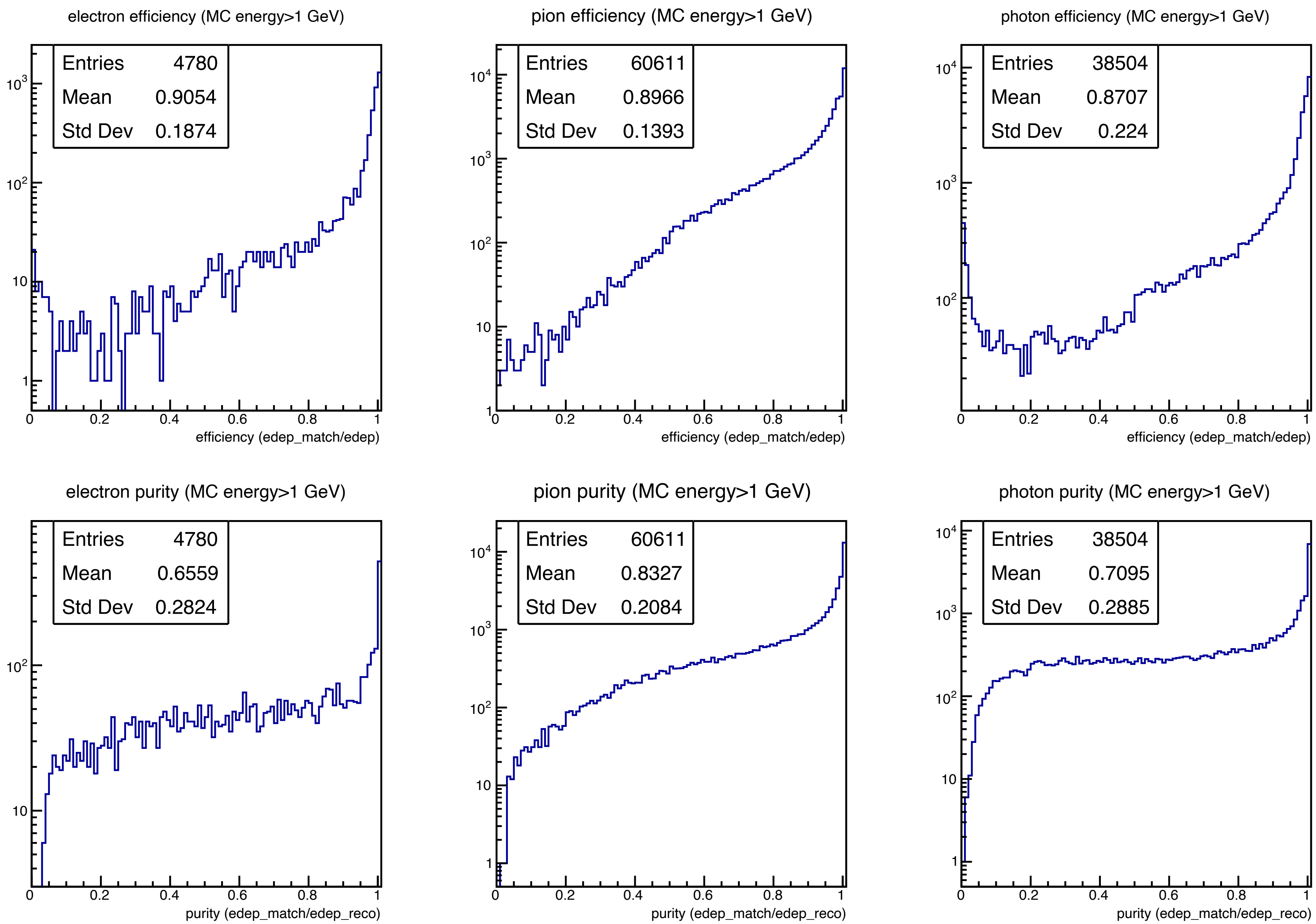
- **Purity :**
  slightly worse in pions
  significantly worse in electrons/photons



Efficiency and purity for GNN, qq train / qq prediction

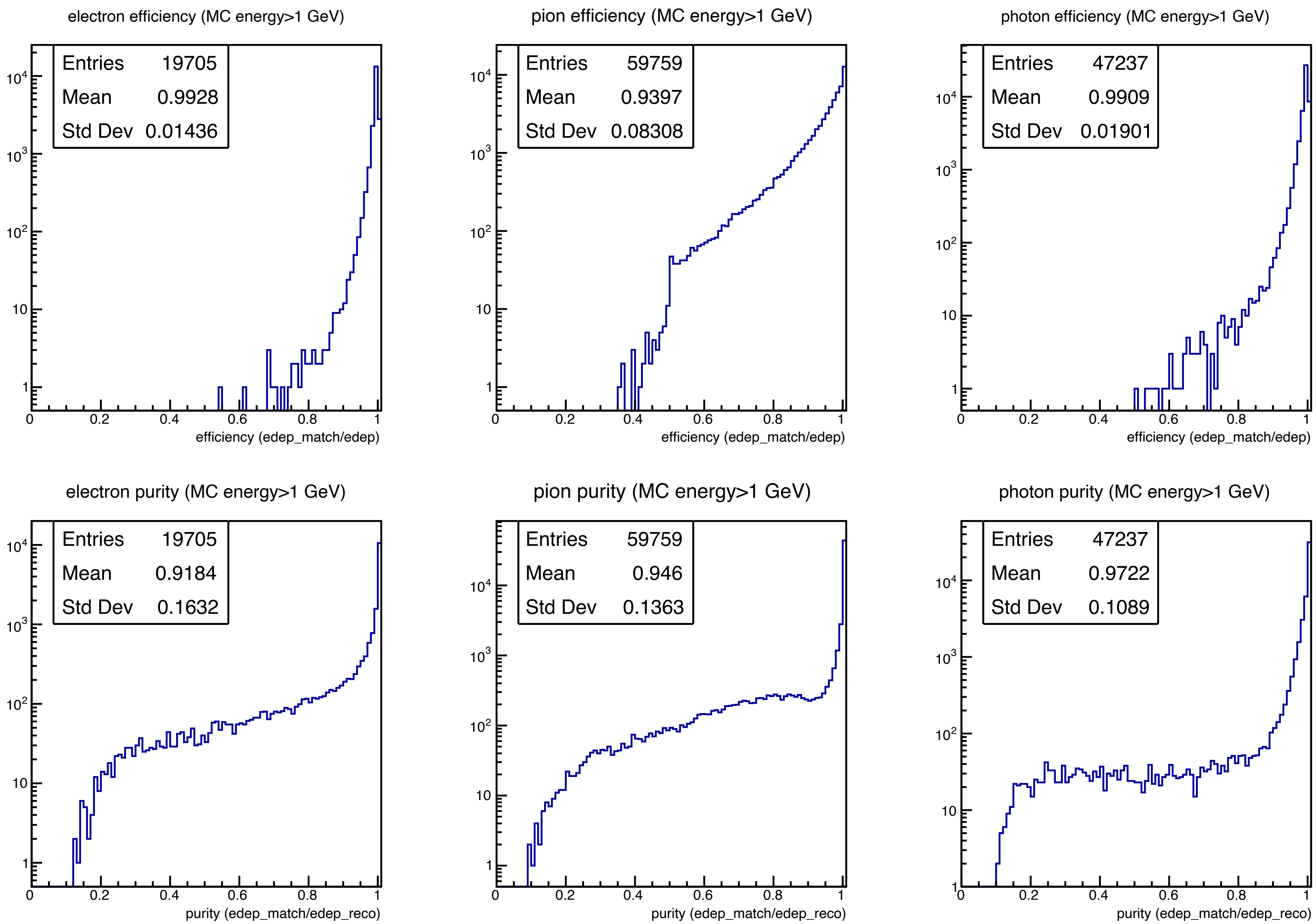
- Efficiency :
  similar to tau training
  strong to different type of events

- purity :
  slightly better than tau training



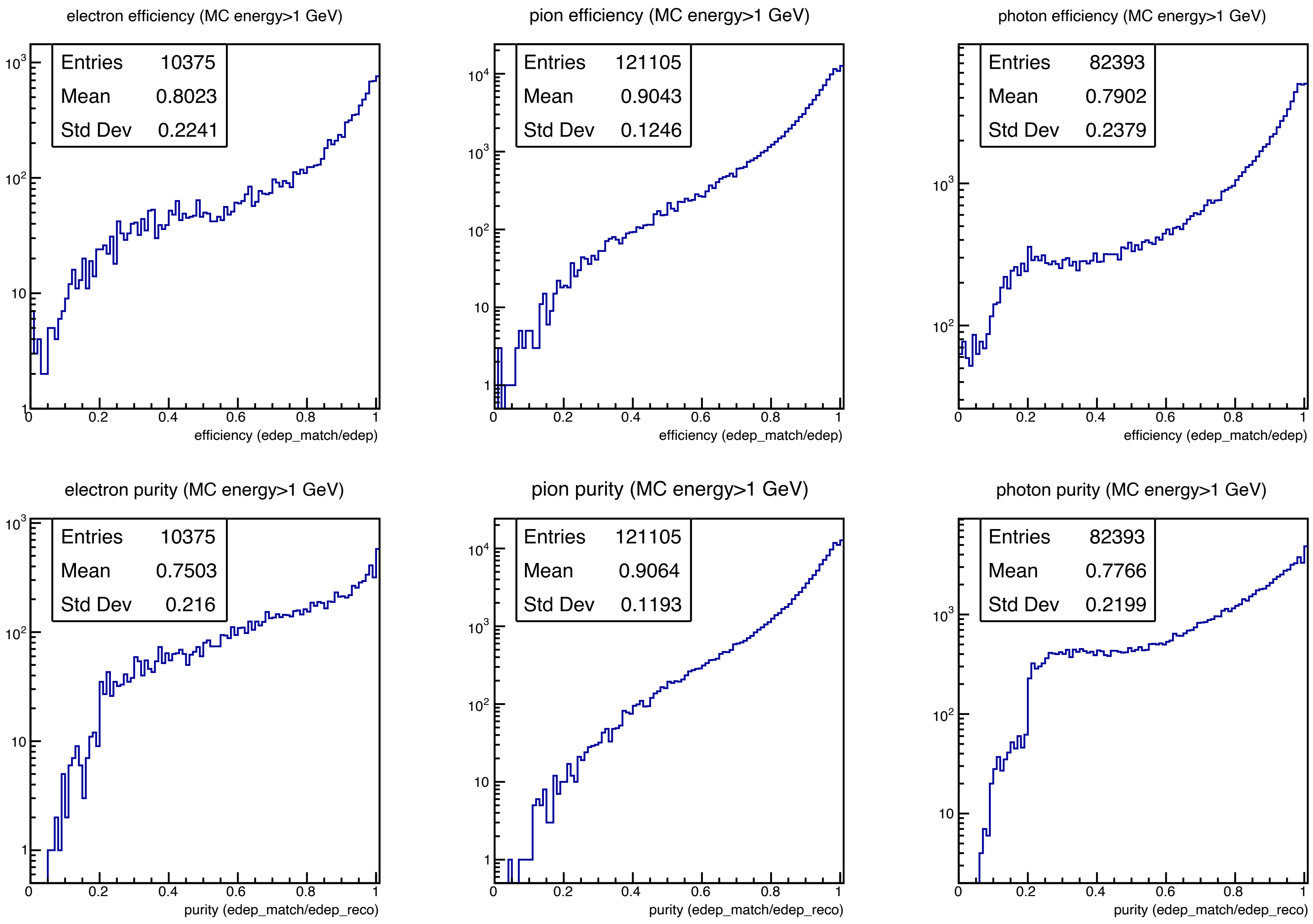
Efficiency and purity with PandoraPFA, tau events

- Efficiency and purity for pion is similar to GNN
- Pandora is better in photon reconstruction (especially in purity)



Efficiency and purity with PandoraPFA, qq events

- Similar performance with GNN method obtained
- Inconsistency with analysis using MC-cluster matching in official software (ILCSoft)



Comparison of results (>1 GeV for MC truth)

Algorithm train / test	Electron eff.	Pion eff.	Photon eff.	Electron pur.	Pion pur.	Photon pur.
taus / taus	99.4	95.0	97.9	88.1	95.4	79.6
taus / jets	91.3	88.1	89.8	62.2	81.3	64.4
jets / jets	90.5	89.7	87.1	65.6	83.3	70.9
PandoraPFA taus	99.3	94.0	99.1	91.8	94.6	97.2
PandoraPFA jets	80.2	90.4	79.0	75.0	90.6	77.7
PandoraPFA jets (ILCSoft)	96.7	95.5	96.4	97.1	90.4	97.7

- The performance of GNN is comparable to PandoraPFA at least on pions, which have less uncertainty related to MC truth definitions
- There is no tunings of hyperparameters
  - The performance of GNN could be improved by hyperparameter tuning

Comparison of results (>1 GeV for MC truth)

Algorithm train / test	Electron eff.	Pion eff.	Photon eff.	Electron pur.	Pion pur.	Photon pur.
taus / taus	99.4	95.0	97.9	88.1	95.4	79.6
taus / jets	91.3	88.1	89.8	62.2	81.3	64.4
jets / jets	90.5	89.7	87.1	65.6	83.3	70.9
PandoraPFA taus	99.3	94.0	99.1	91.8	94.6	97.2
PandoraPFA jets	80.2	90.4	79.0	75.0	90.6	77.7
PandoraPFA jets (ILCSoft)	96.7	95.5	96.4	97.1	90.4	97.7

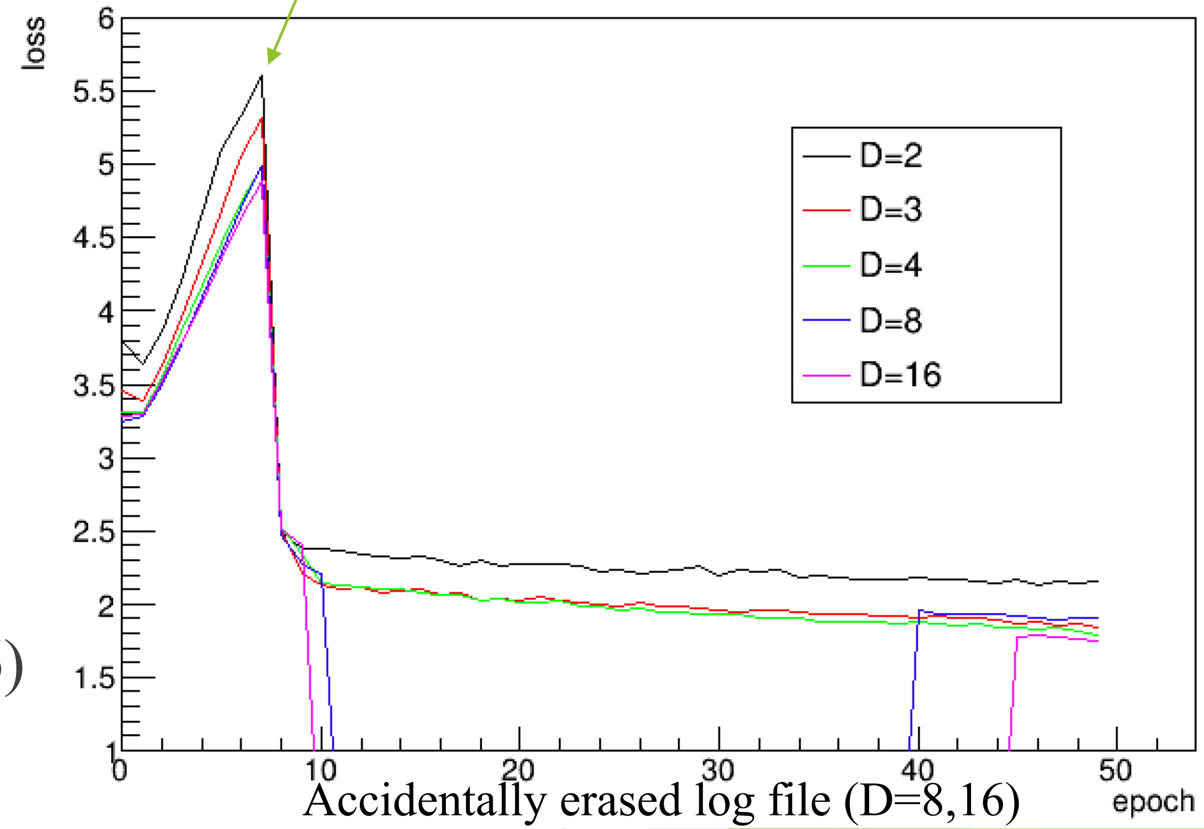
- The performance of GNN is comparable to PandoraPFA at least on pions, which have less uncertainty related to MC truth definitions
- There is no tunings of hyperparameters
    - The performance of GNN could be improved by hyperparameter tuning

$$L = L_p + s_C(L_\beta + L_V)$$

Hyperparameter tuning

- Hyper parameters
  - Training model's output dimensions ($D = 2$, two for virtual coordinate)
  - Epochs to turn on beta term in loss function
  - Clustering parameters
    - Beta threshold ($thre_\beta = 0.2$)
    - Diameter threshold ($thre_{diameter} = 0.5$)
  - etc…
- Checked output dimensions (D=2,3,4,8,16)
  - Train : 10 taus (10 GeV)
  - pred : 10 taus (10 GeV) / jets (91 GeV)

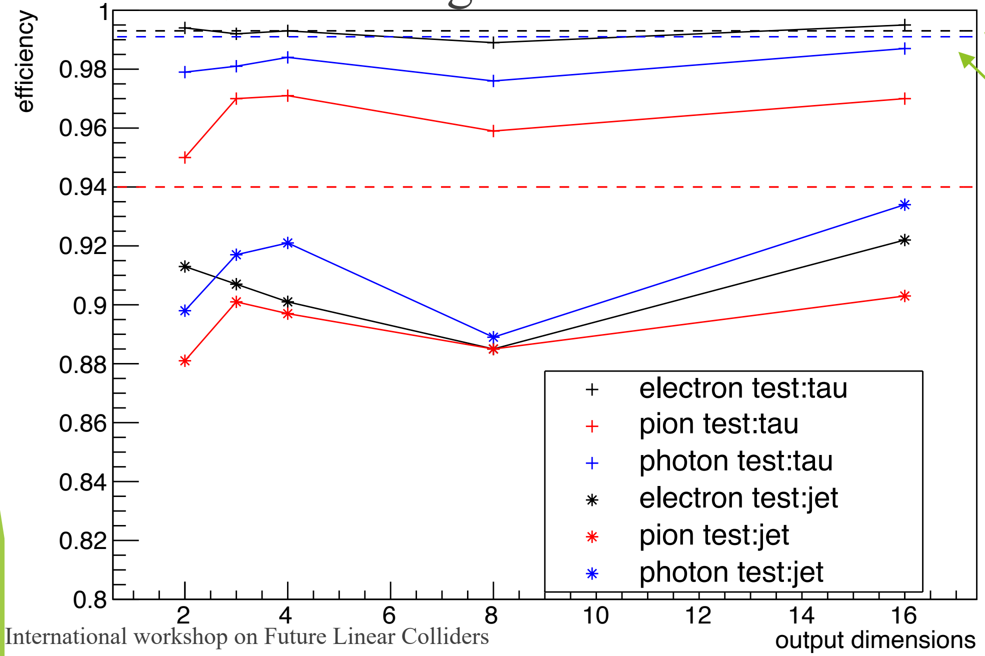
Beta term in loss function is not activated for first few epochs
Resulting increasing loss



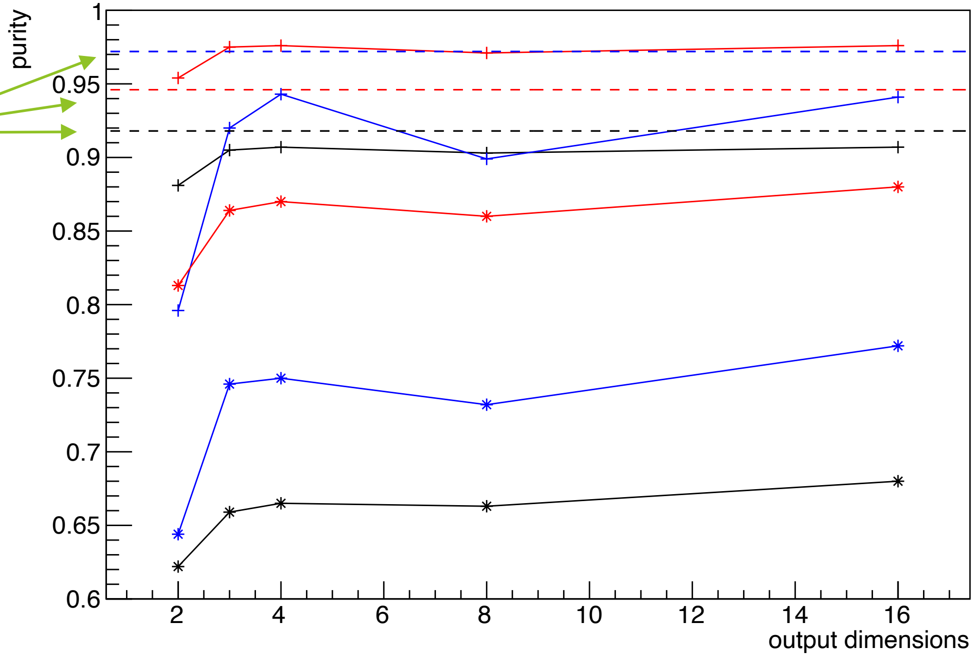
Accidentally erased log file (D=8,16)
Re-calculating loss

Comparison of results (>1 GeV for MC truth)

- For all output dimensions,
  - Pion GNN results are overcoming the PandoraPFA
  - Electron GNN results are comparable to the PandoraPFA
  - Photon GNN results are below the PandoraPFA
    - For output dimension of 4, GNN results are reaching the PandoraPFA
- Both efficiency and purity are low at D=8
  - This might be due to lack of statistics

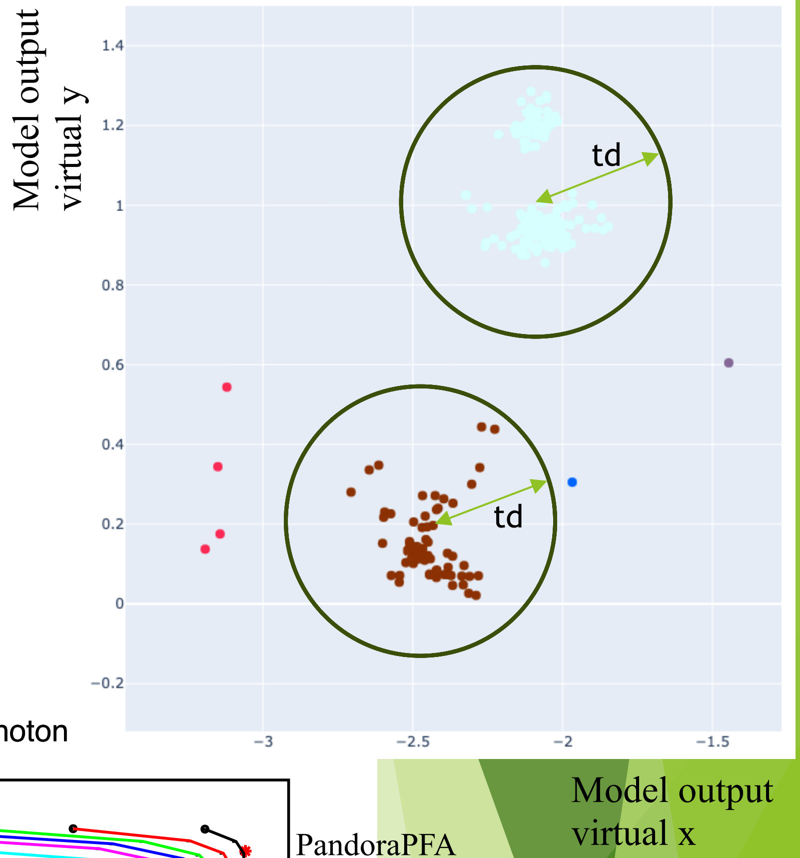


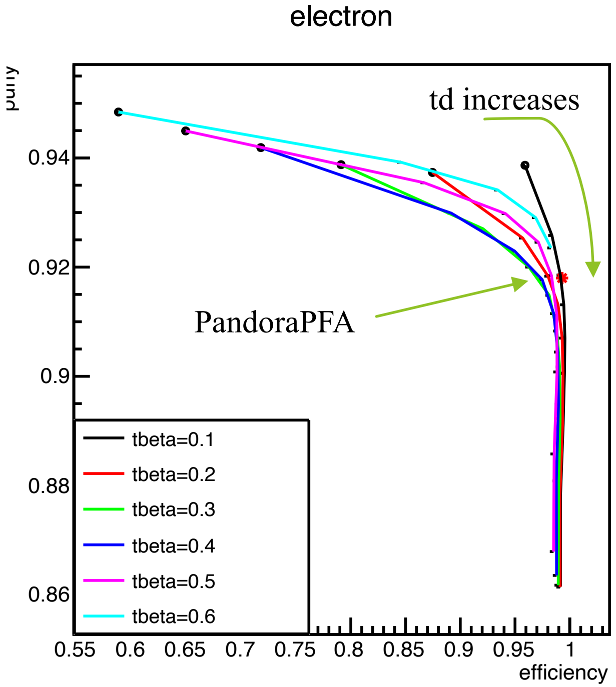
PandoraPFA
Tau events

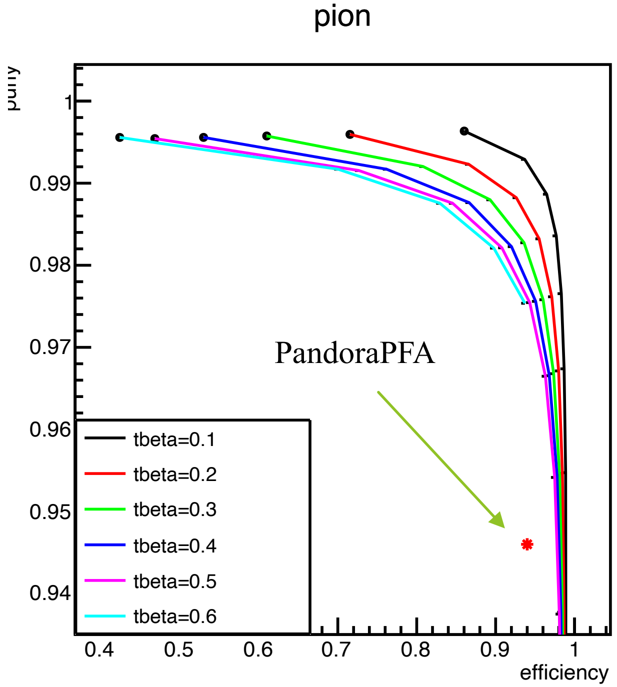


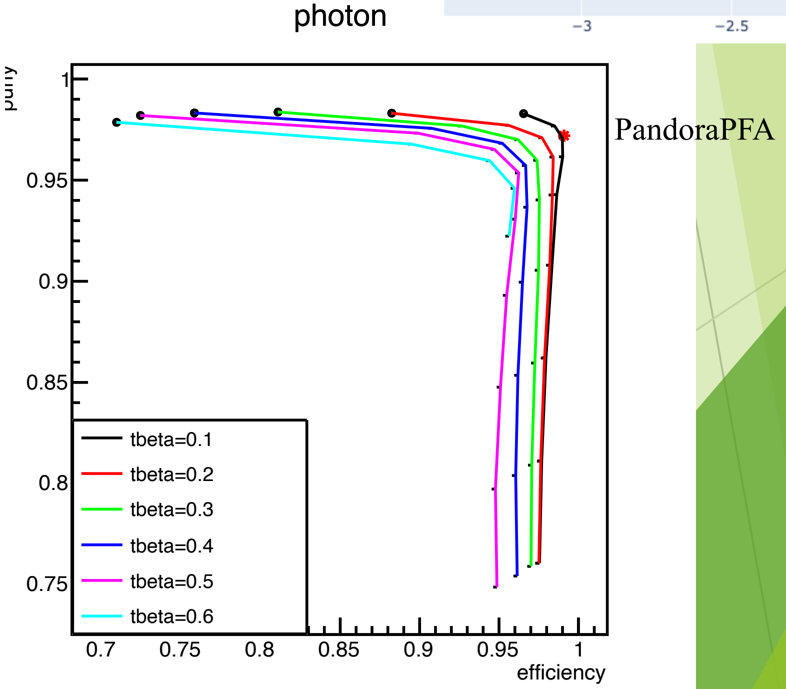
Hyper parameter tuning (clustering)

- Grid search of parameter "tbeta" and "td" (output dimension=4)
  - tbeta : points whose beta > tbeta , candidates of a condensation point
  - td : points which distance from the highest beta point < td is considered as the cluster
- Lower beta threshold tend to have higher purity and efficiency









Summary and prospect

- Applied CMS HGCal clustering algorithm to ILD simulation
  - GravNet and object condensation
  - Virtual hit form tracker
  - Simulation samples
    - 10 τ (10 GeV) and 10 u, d, s (91 GeV)
- Quantitative comparison with PandoraPFA is ongoing
  - GNN method showed a bit worse performance in clustering calorimeter hits than PandoraPFA
- Prospect
  - Optimizing network/input
    - Improvement of MC truth matching
    - Input sample particle/size
    - Other hyperparameters
  - Clustering method : replace by NN?
  - Further comparison with PandoraPFA in terms of jet energy resolutions

Back up