Graph neural network-based particle flow for the future Higgs factories

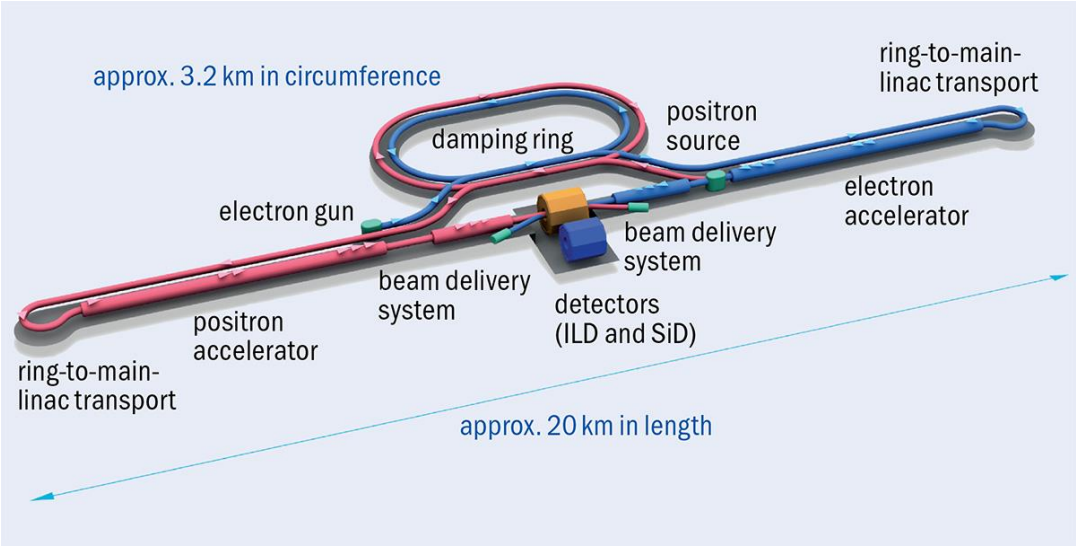
Tatsuki Murata [1]
Taikan Suehara [1,2], Stefan Barbu [1,3], Wataru Ootani [1,2]
University of Tokyo [1], ICEPP [2], École polytechnique[3]

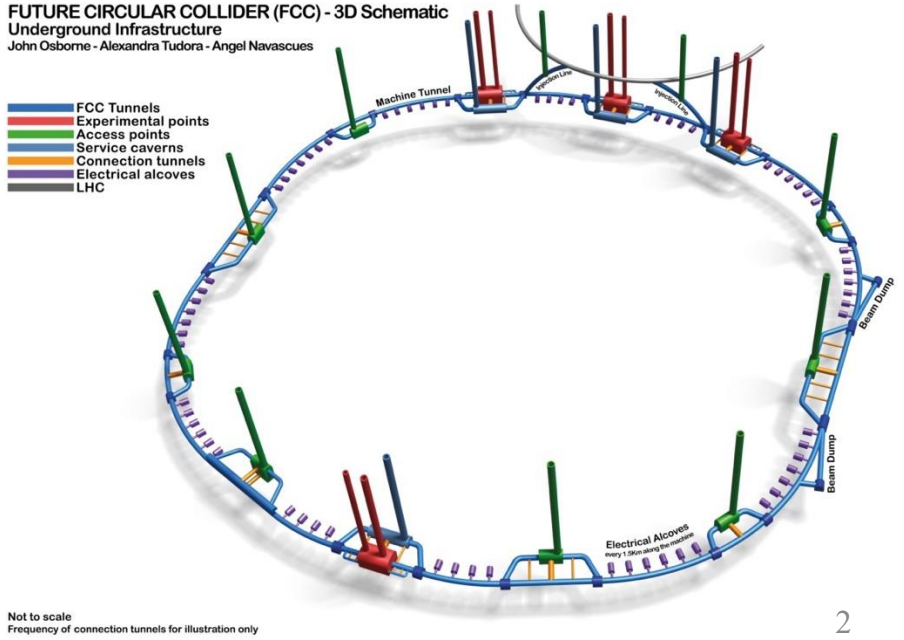
Higgs factory

- Precise measurement of Higgs bosons are necessary to search for the beyond standard model

- e+e- colliders are designed
  - ILC, FCC, etc…
  - Jet energy resolution ~30%/√E

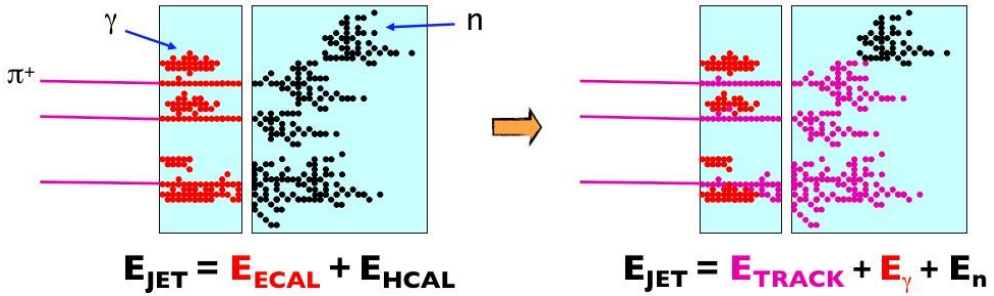
- To achieve the requirement, Particle Flow Algorithm (PFA) oriented high granular detectors are considered
  - ILD, SiD, CLD, etc…
  - (e.g., silicon ECAL: $5\times5mm^2$, 30 layers)





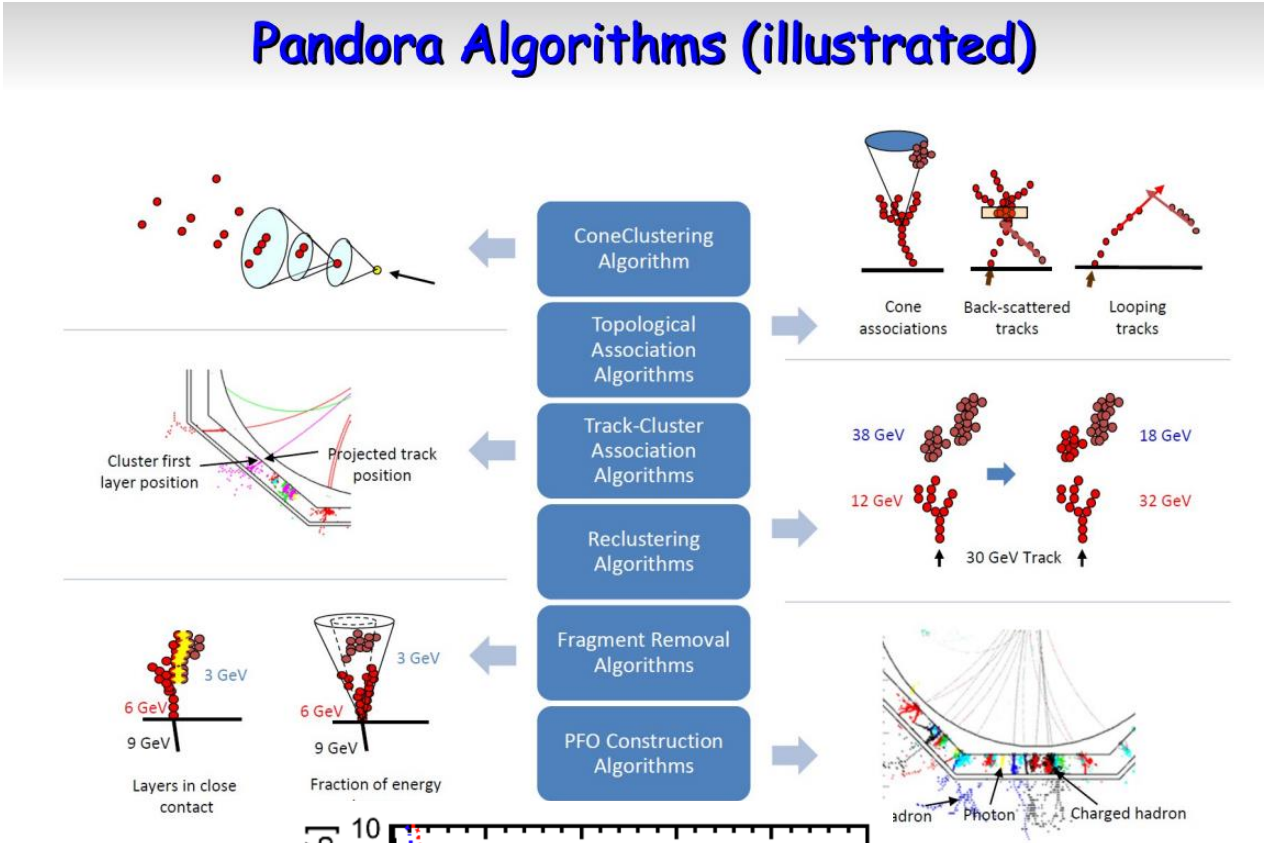
Particle flow

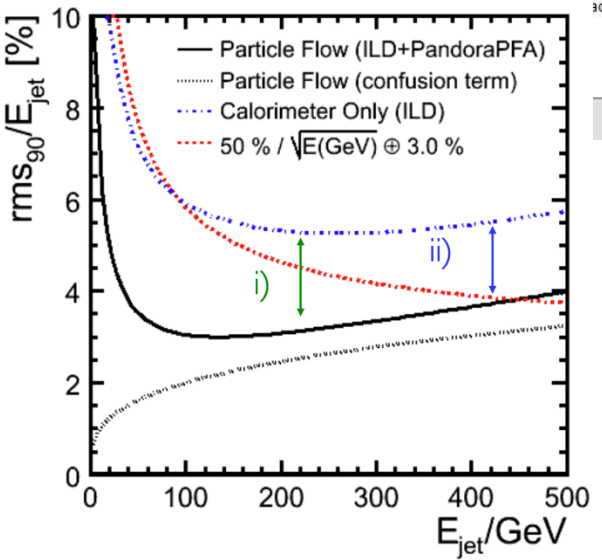
- Detect particle energy by suitable sub-detectors
- Separation of charged/neutral cluster at high granular calorimeter is crucial
- Essential for high jet energy resolution



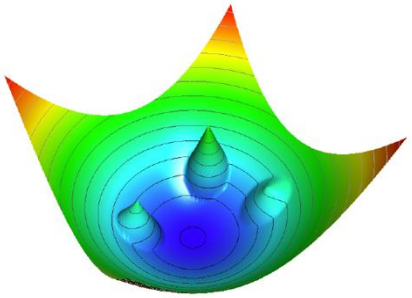
- Current algorithm : **PandoraPFA**
  - Pattern recognition based on the human-tuned parameters
  - Limitations: difficult to optimize or incorporate new information (e.g., timing)

→ <u>replace and improve the algorithm with machine leaned one</u>





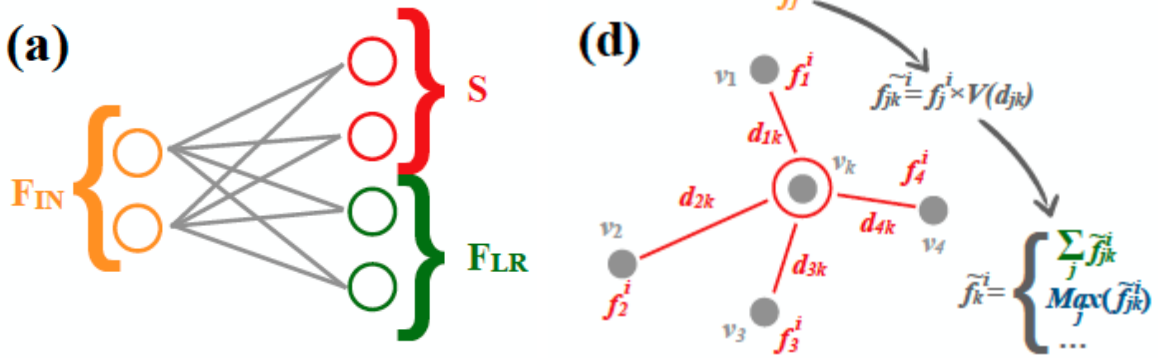
GravNet and object condensation



arXiv:2002.03605

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



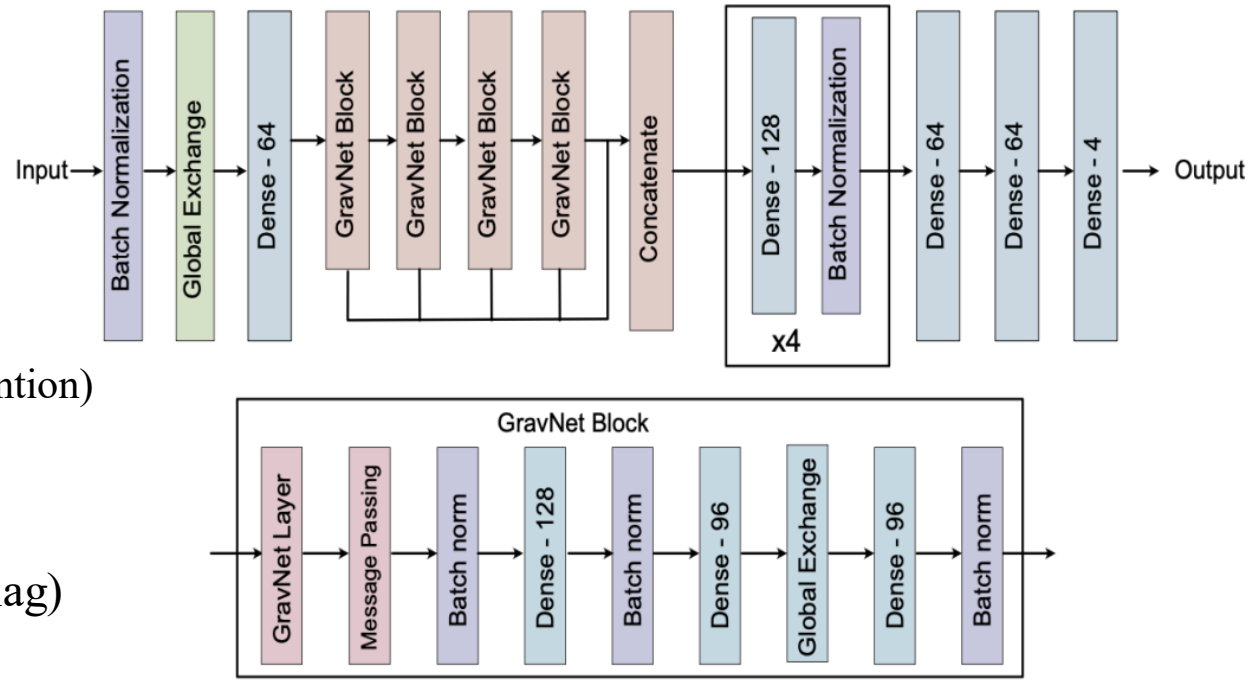
Object condensation (loss function)

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

- Condensation point: the hit with largest β at each MC cluster
- 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

Model architecture

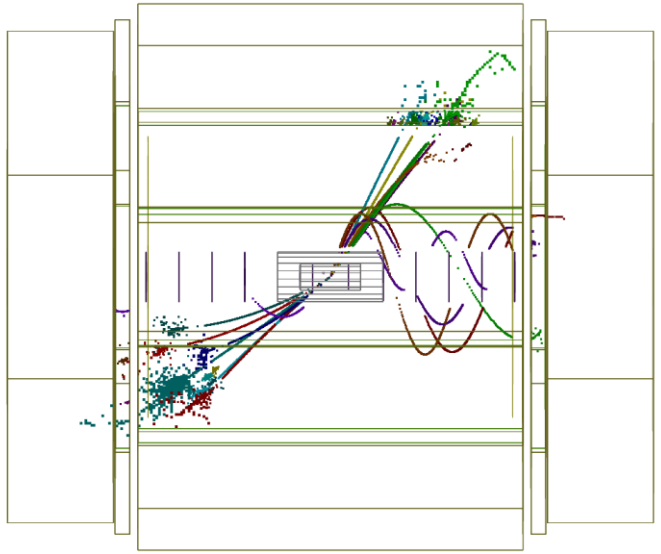
- Originally developed by CMS HGCal group
- Input/output are obtained for each hit at calorimeters
  - Input: features at each hit
    - Position
    - energy deposit
    - Tracker momentum (only for track)
  - Output:
    - "condensation coefficient" β
    - position of virtual coordinate for clustering (4-dimention)
    - cluster energy
- 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 1 due to L_β term in loss function

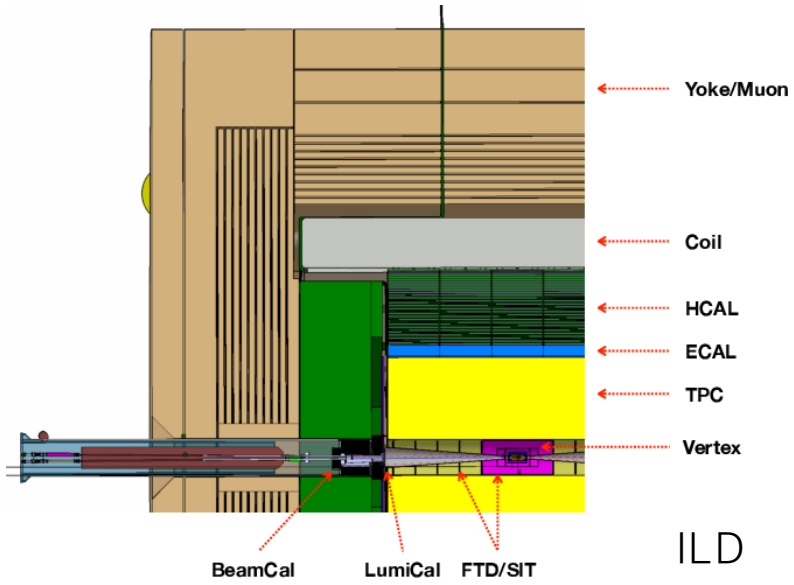


# of parameter: ~400k

Samples for performance evaluation

- ILD full simulation with SiW-ECAL and AHCAL (ILD_l5_o1_v02, 020301)
  - Two types of samples : τ, di-jets
    - τ^- (p=10 GeV)
      - 10 taus overlayed
      - Good mixture of decay products (pi, e, mu, photons)
      - 100,000 events
    - di-jet ($ee \rightarrow Zh \rightarrow qq\nu\nu$, $\sqrt{s}$=250 GeV)
      - 40,000 events
  - train : 80%
    validation : 10%
    test : 10%

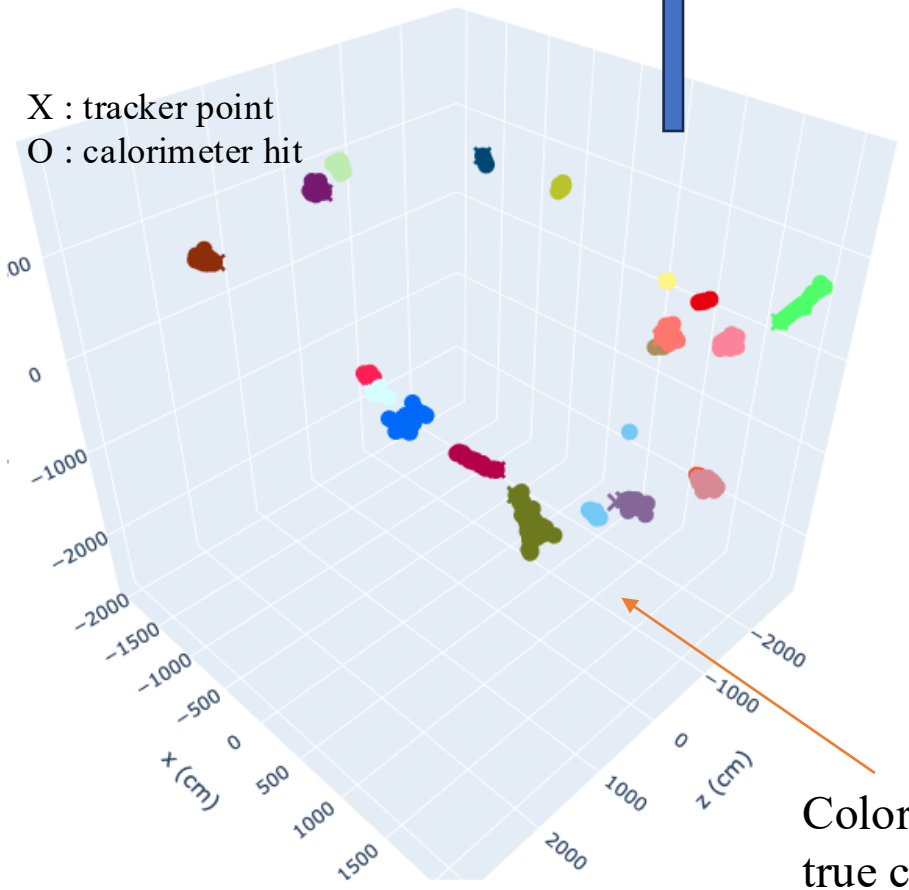




Event display
tau event

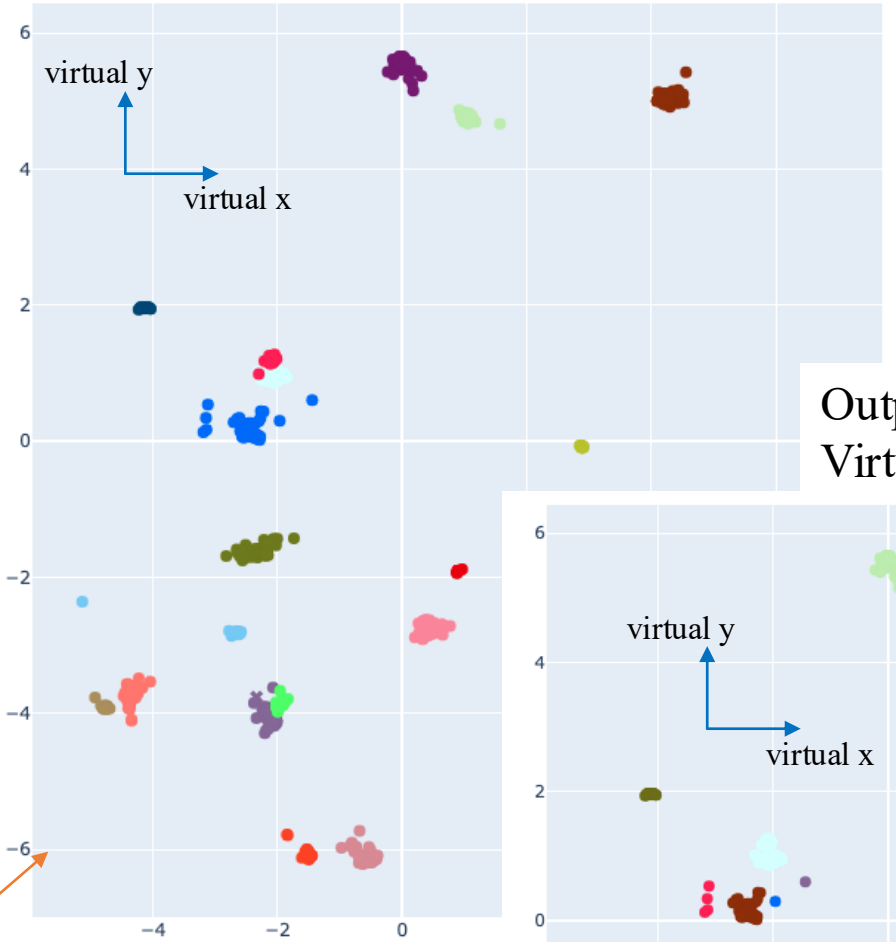
Network

X : tracker point
O : calorimeter hit

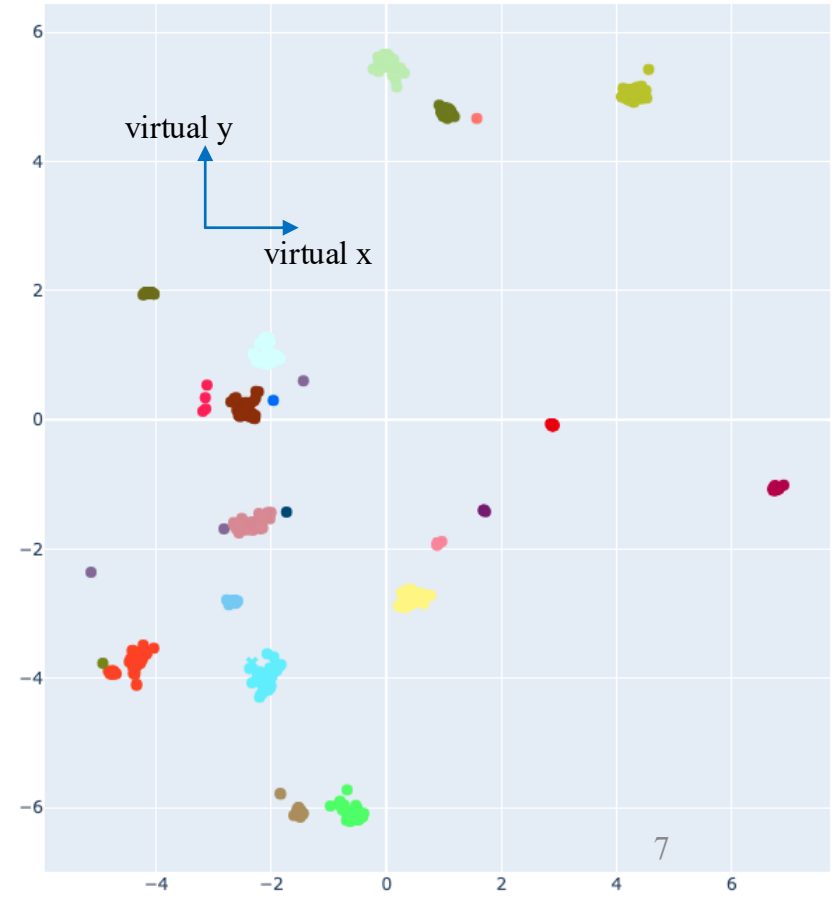


Input features
Real coordinate in ILD

Colored by
true clusters



Output features
Virtual coordinate



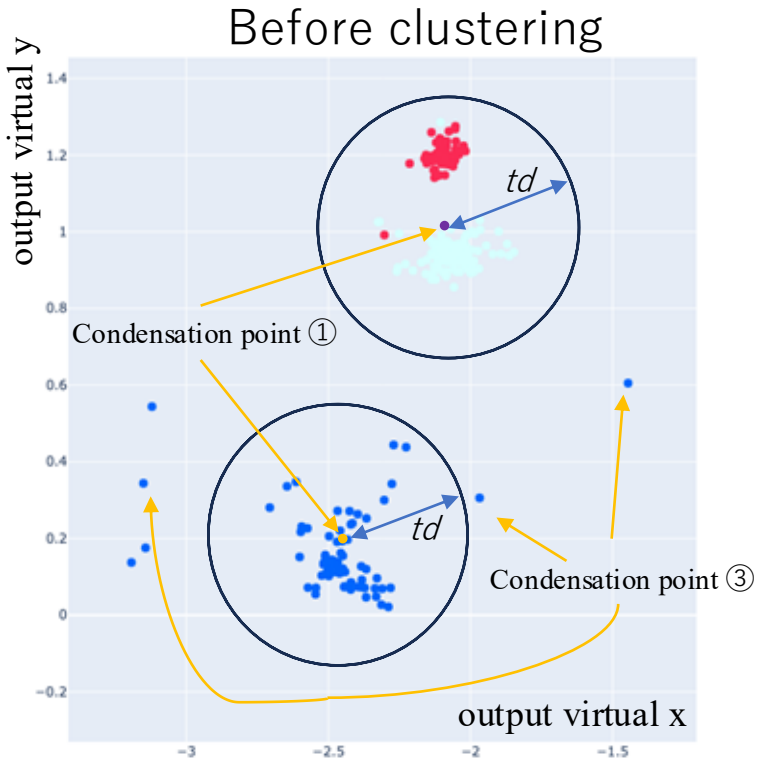
Colored by reconstructed
clusters

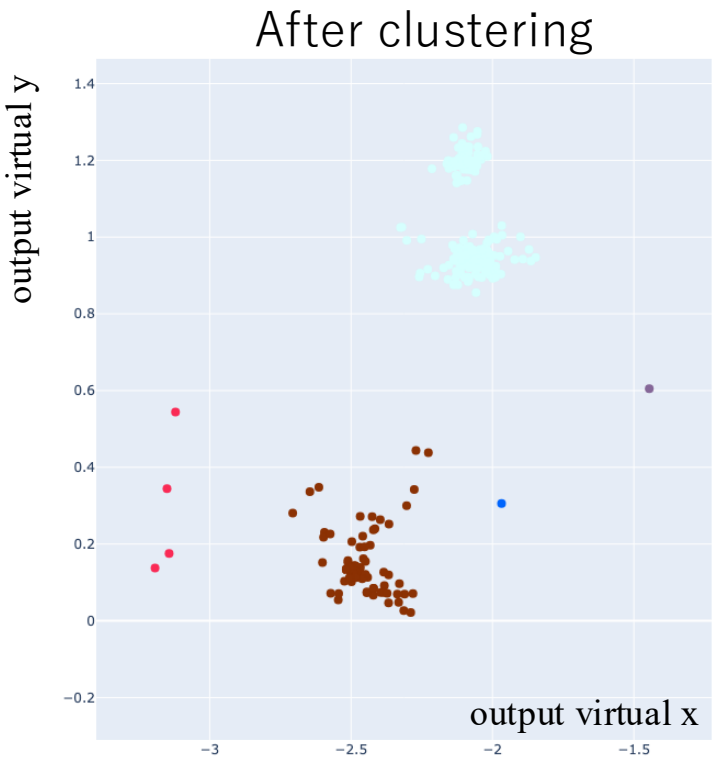
Clustering algorithm

① Select hit with β above threshold (*tbeta*) as condensation seeds

② Assign nearby hits (within distance *td*) to the nearest seed

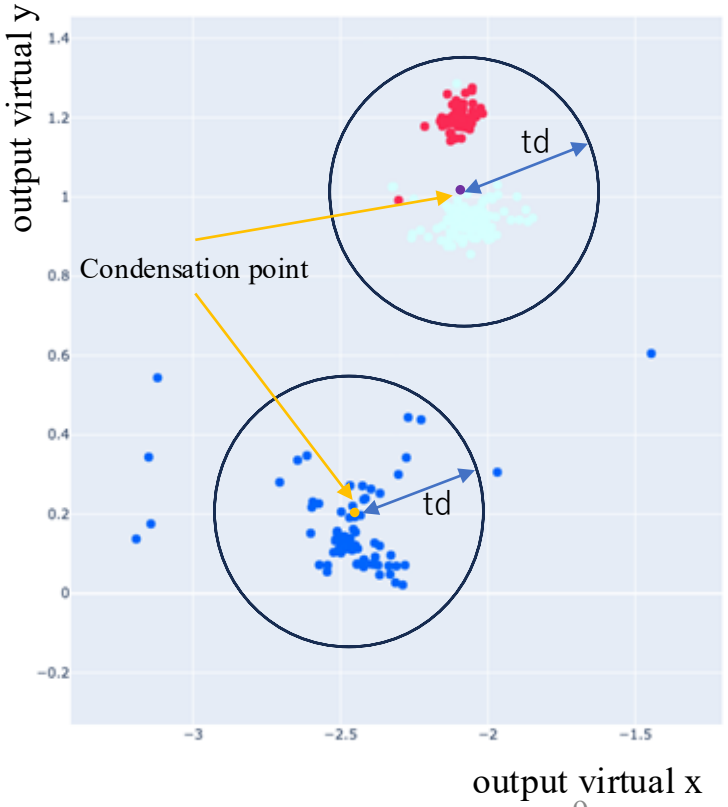
③ Repeat with next-highest β hit until all hits are clustered





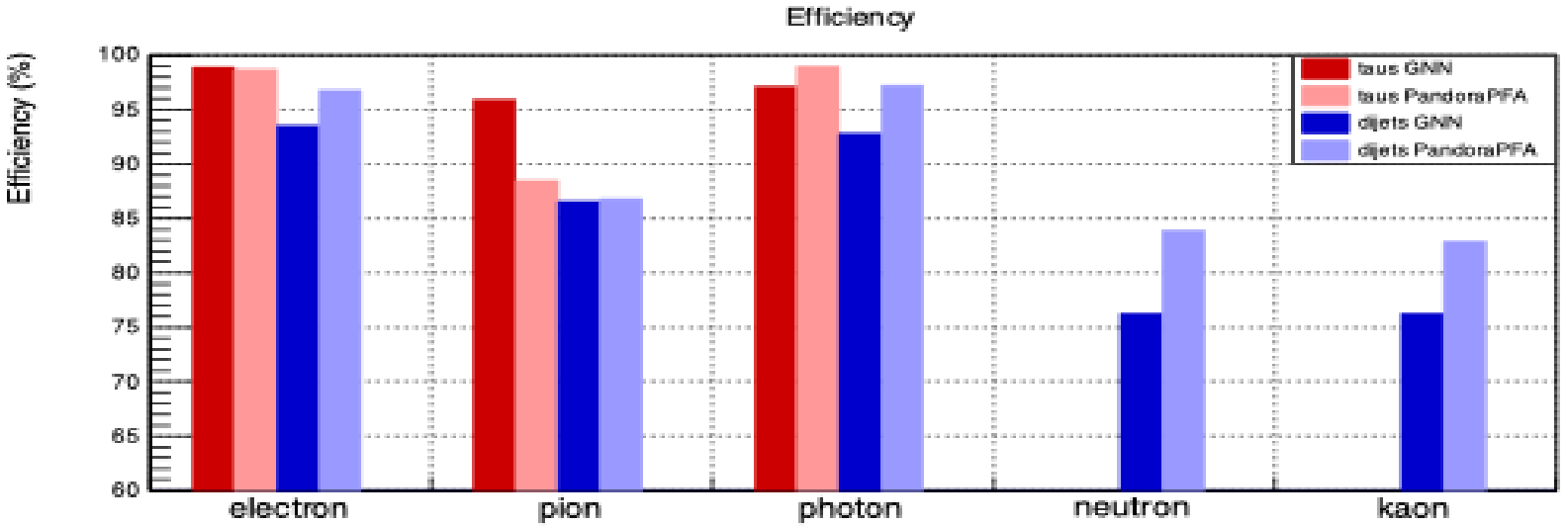
Clustering performance evaluation

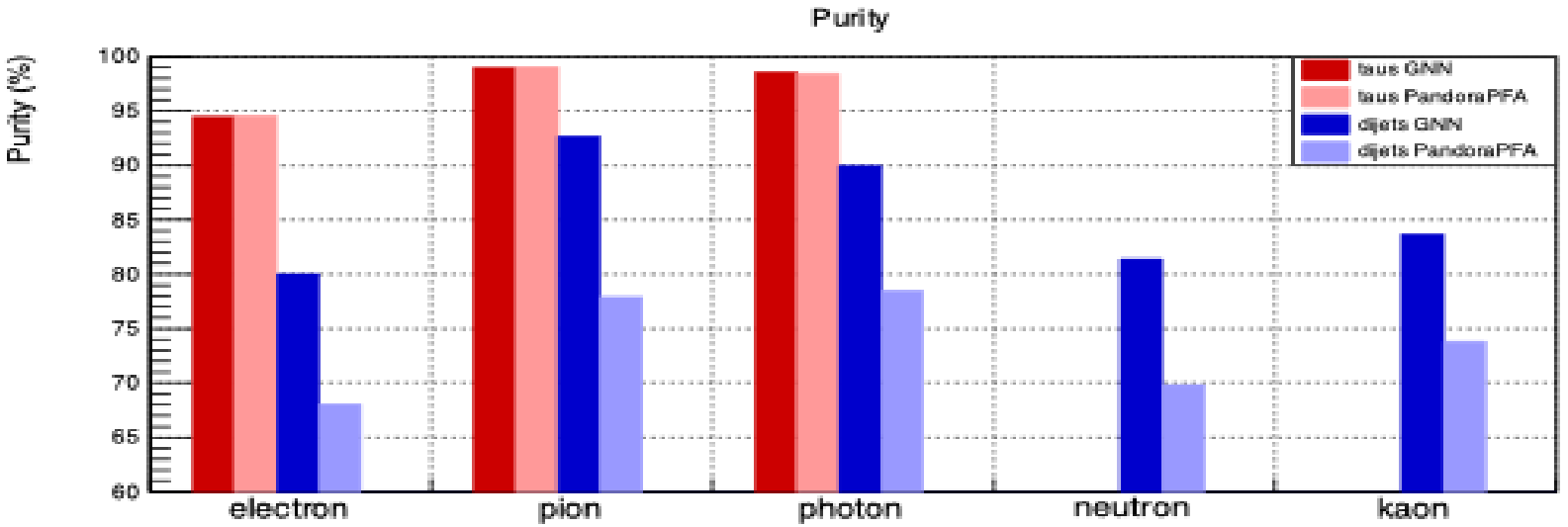
- Make 1-to-many connection of MC and reconstructed cluster
  - Reconstructed cluster with highest fraction of hits from the MC is taken
  - Multiple reconstructed cluster may connect to one MC cluster
    - Chose reconstructed cluster with the highest fraction
- Quantitative comparison with PandoraPFA
    - E_{dep} = MC truth cluster energy
    - E_{reco} = reconstructed cluster energy
    - E_{match} = correctly clustered energy
  - Compared "efficiency" and "purity" of each cluster
    - *Efficiency* $= \frac{E_{match}}{E_{dep}}$
    - *Purity* $= \frac{E_{match}}{E_{reco}}$
  - Compared efficiency and purity for
    - Charged particles: <u>electrons, pions,</u>
    - Neutral particles: <u>photons, kaons, neutrons</u>



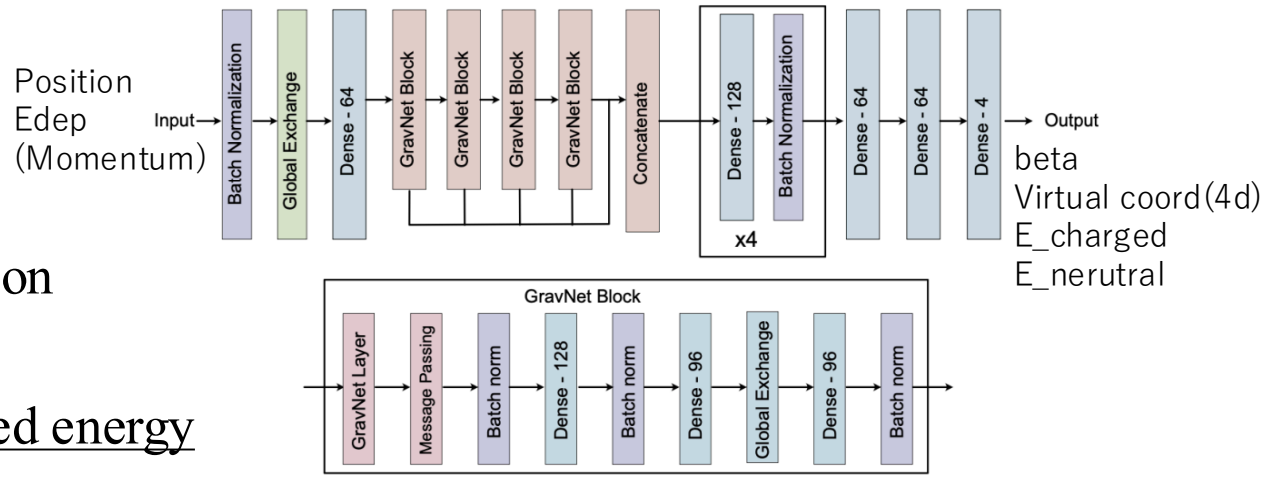
Clustering performance

- GNN achieves higher purity, while PandoraPFA has slightly higher efficiency
  - The MC assignment selects clusters with highest E_{match}, which leads to higher purity
  - Significant improvement at pinon





Energy regression

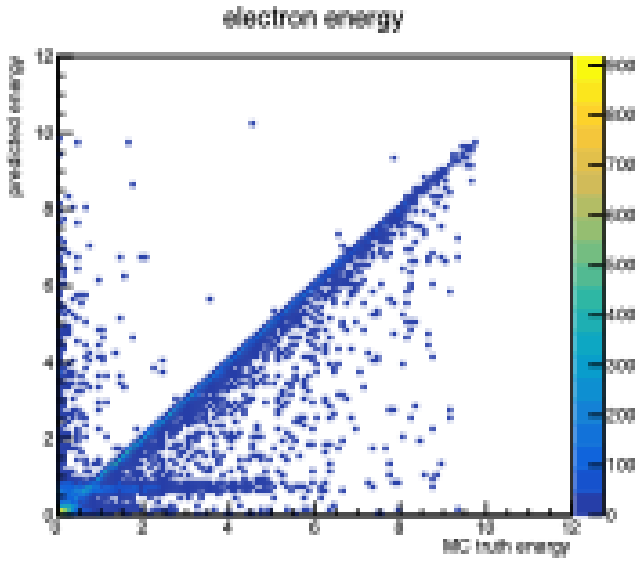


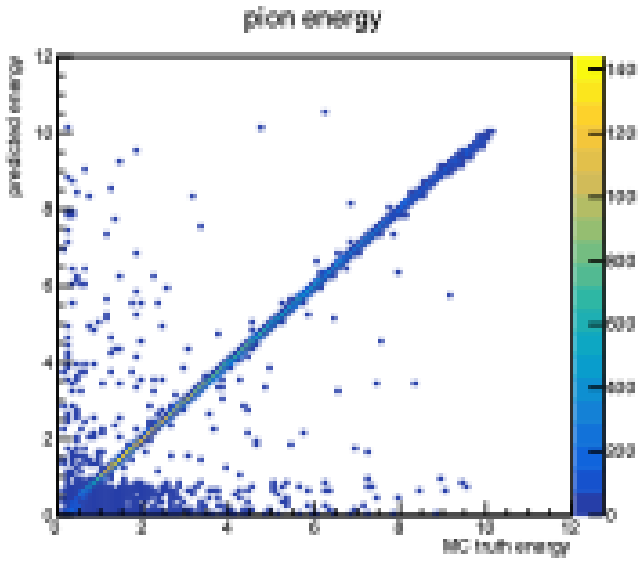
- PandoraPFA aims to minimize jet energy resolution
  →regression is critical
- Before examining jet energy resolution, performed energy regression on decay particle level

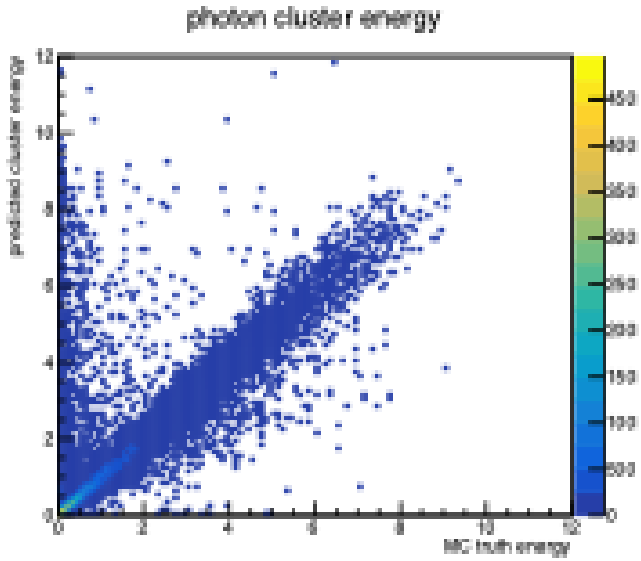
- Added two energy regression terms to model output and loss function
  - $L_{E,charged} = \sum_i \left(E_{truth,i} - E_{pred,charg,i}\right)^2$ i:summation over only condensation points
    - **For charged particles**
    - To fully use track information
    - Estimate cluster energy only at the condensation point
  - $L_{E,nerutral} = \sum_i \left(E_{truth,i} - \sum_j E_{pred,calo,i,j}\right)^2$ i:summation over clusters, j:summation over hits in a cluster
    - **For neutral particles**
    - Condensation beta of neutral particles tend to have lower values than charged particles
    - This regression strongly depends on the clustering

Energy regression -tau samples- (ongoing work)

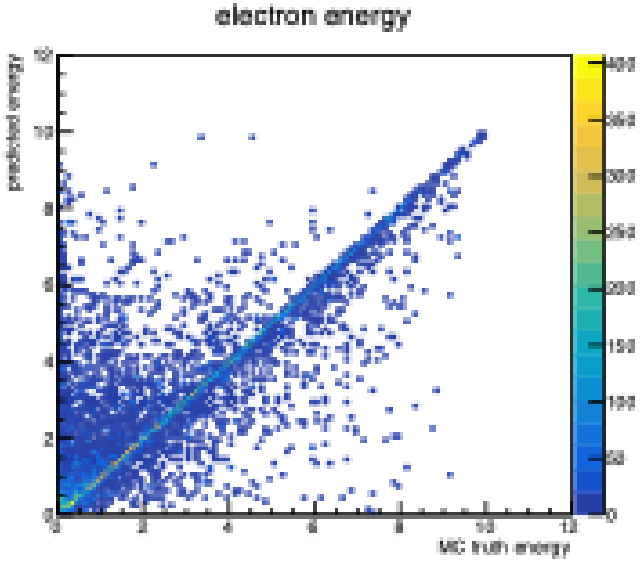
GNN energy regression

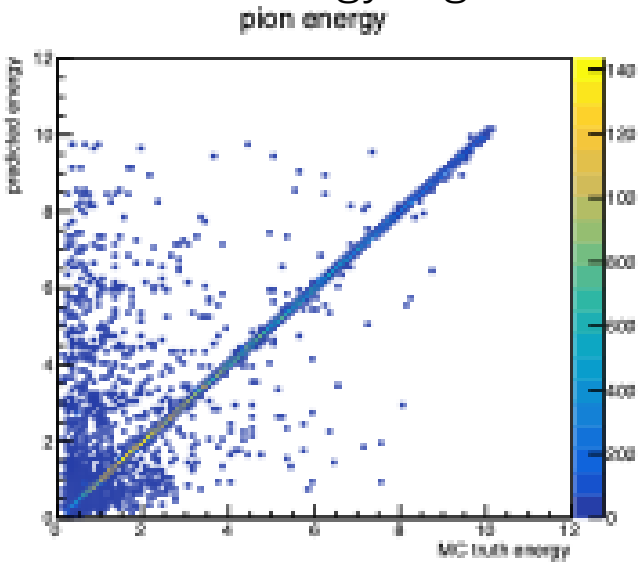


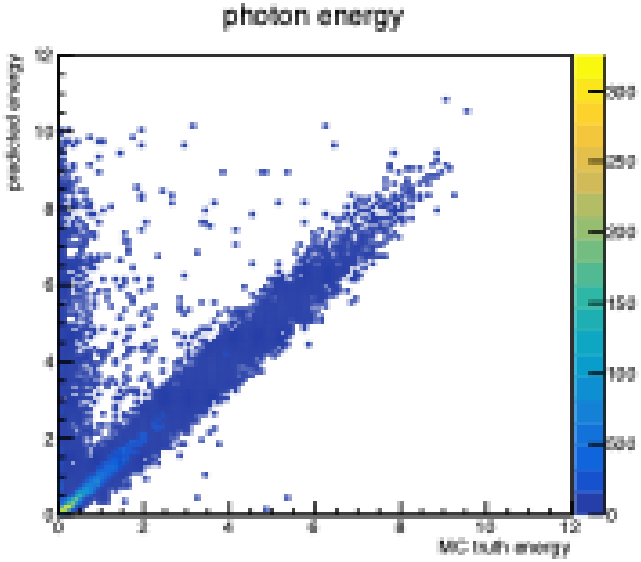




PandoraPFA energy regression

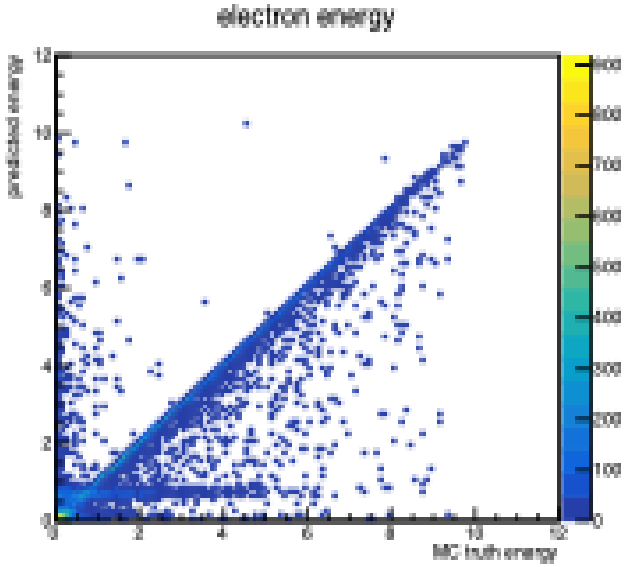


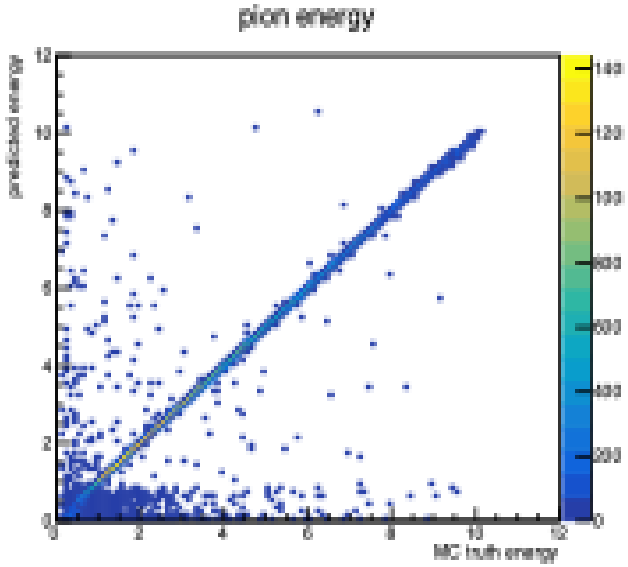


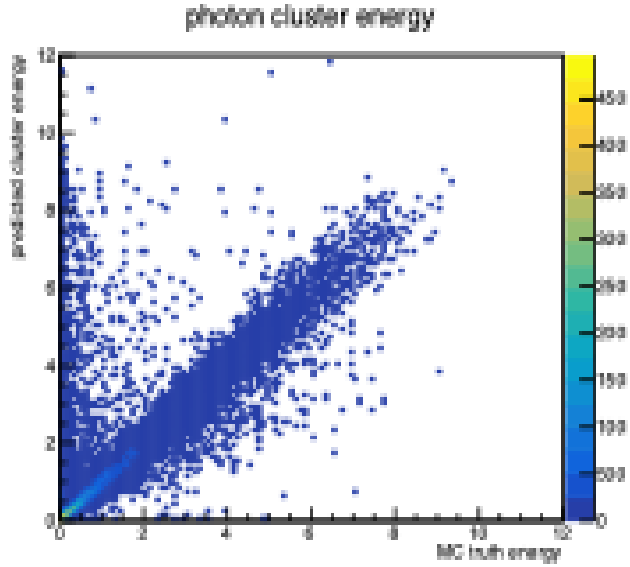


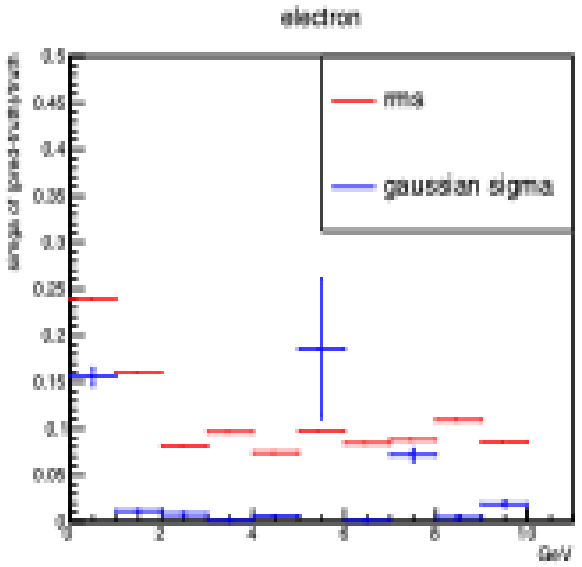
Energy regression -tau samples- (ongoing work)

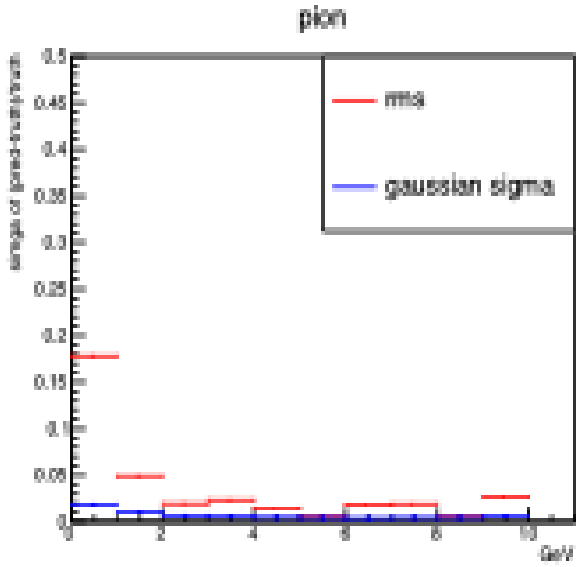
- Energy regression for charged particles are working well
- Neutral particles have worse regression
  - Depends on clustering performance

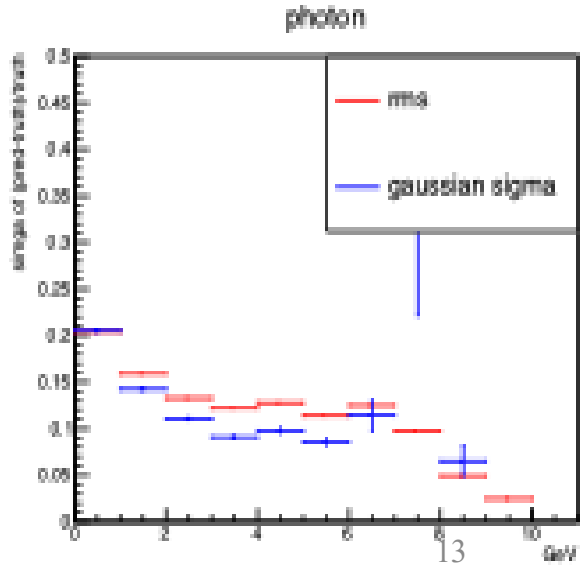






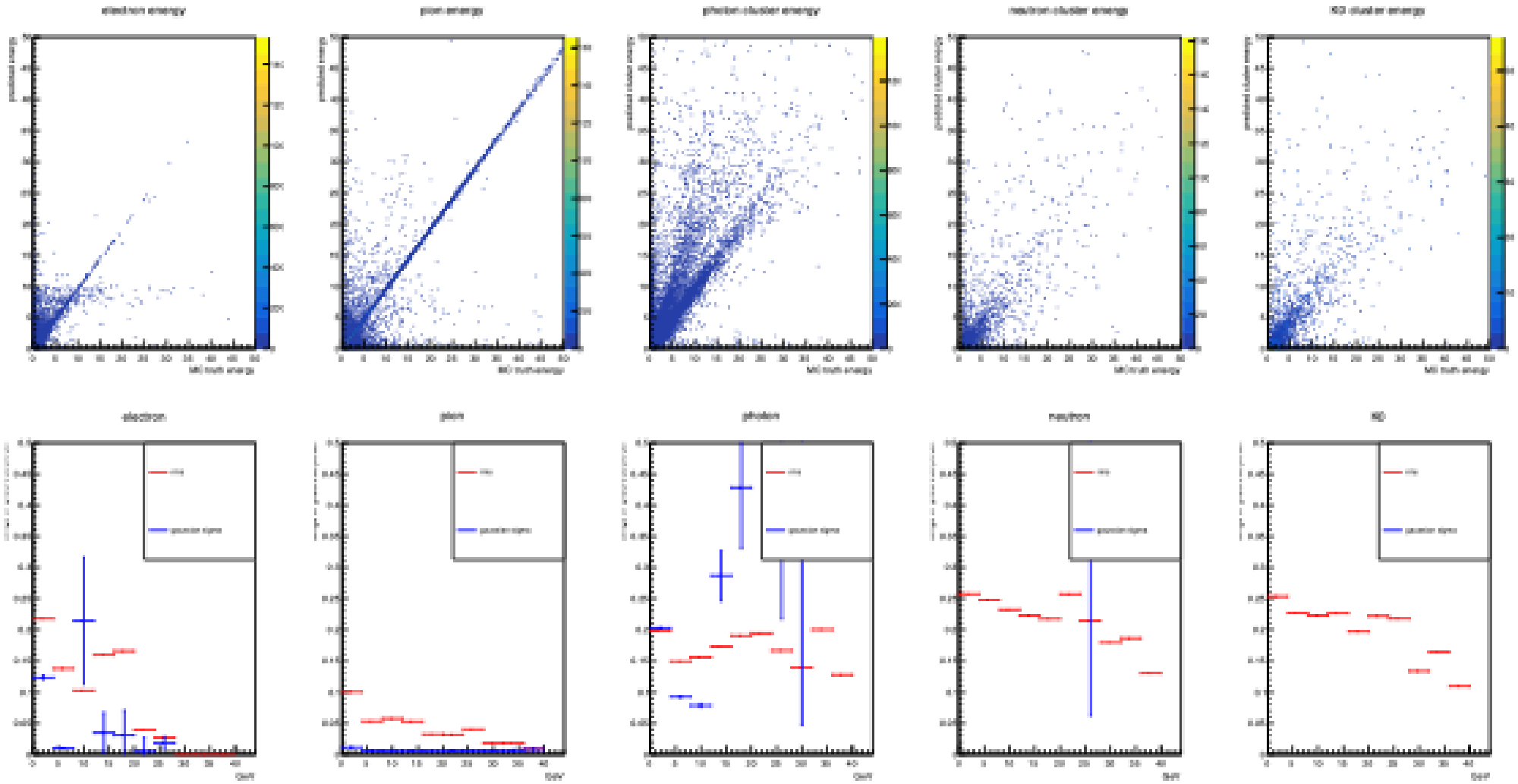






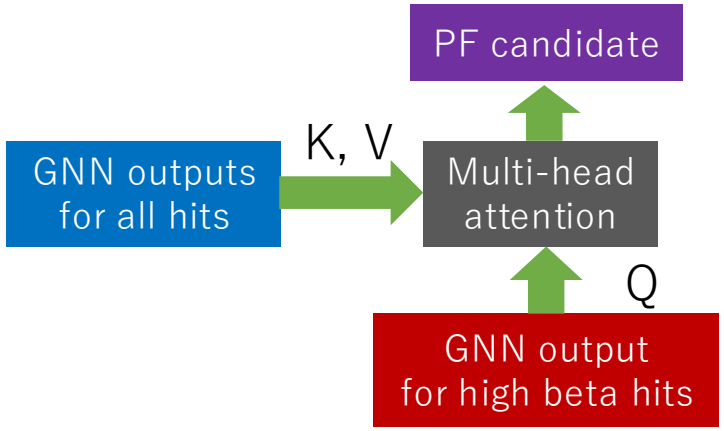
Energy regression - di-jet samples – (ongoing work)

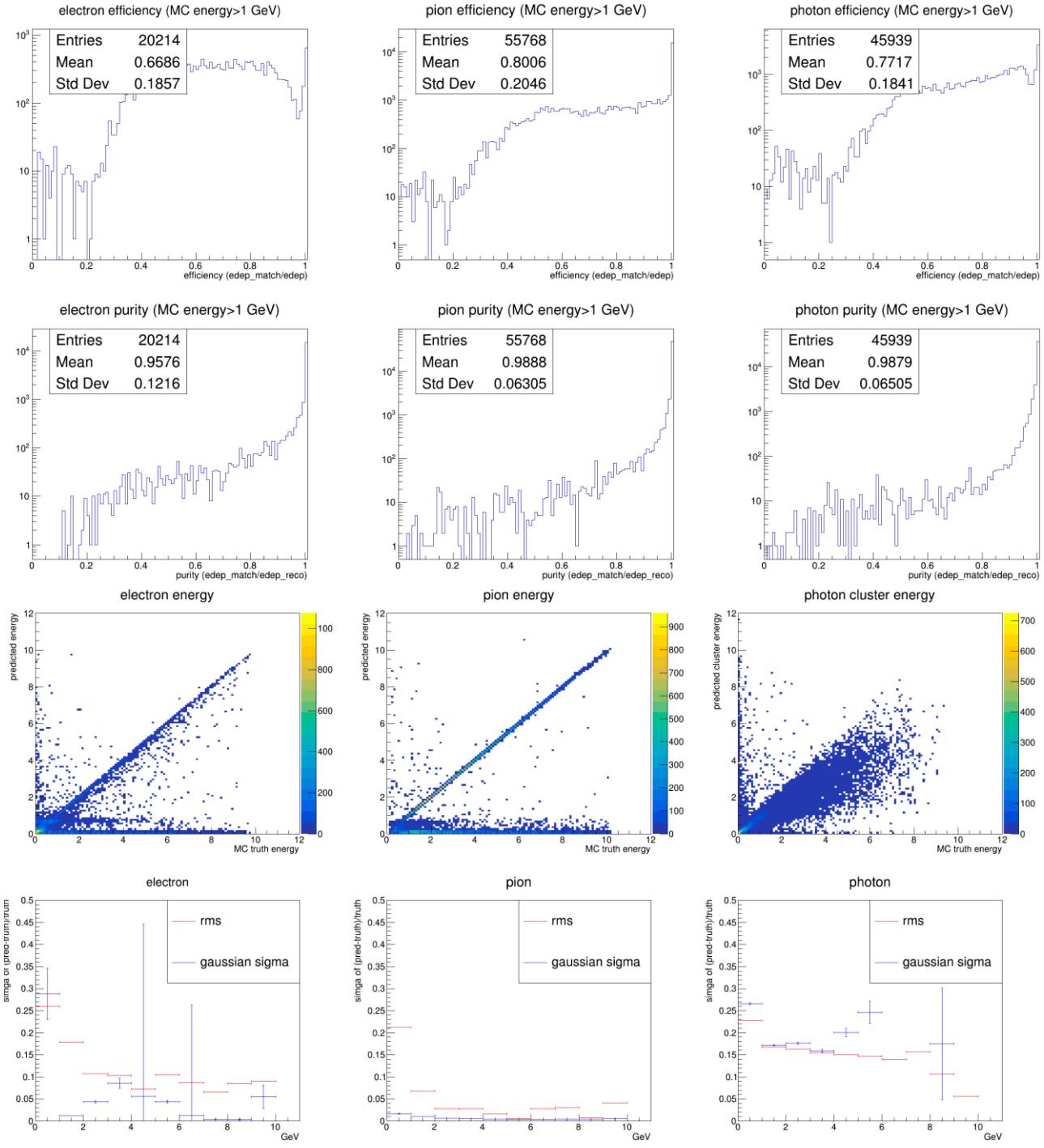
- Energy regression for pions are working well
- Electron energy regression worsens at low energies, likely due to photon misidentification.
- Regression of photon seems to have two lines
  - Top line mainly comes from $\pi^0 \rightarrow \gamma\gamma$ decay



Upgrading clustering algorithm (ongoing work)

- The current clustering simply applies a set of rules using the outputs of the GNN
- Modifying this method to cross-attentioned clustering model is ongoing
  - Estimation of number of particles and particle energy is the next task





Summary

- Developing GNN based PFA
  - Evaluated performances with tau and jets samples
  - Clustering performance of GNN have comparable or superior result to the PandoraPFA
- Evaluated energy regression performance
  - Added condensation point energy and cluster energy term for charged and neutral particles
  - Charged particle resolution is completely determined by the track
- The performance of energy regression is still not good enough, but it is working well for charged hadrons
  - Charged particle energy resolution comparable to the Pandora PFA
  - Neutral particle energy resolution is still not comparable to the Pandora PFA
    - The degraded resolution is likely due to the poor clustering performance of neutral particles
  - Evaluation of the jet energy regression is being done

- Ongoing work
  - Modifying clustering algorithm into machined learned one
  - Replacing the whole GNN with transformer