Development of a Graph Neural Network-Based Particle Flow Algorithm for Future Higgs Factories

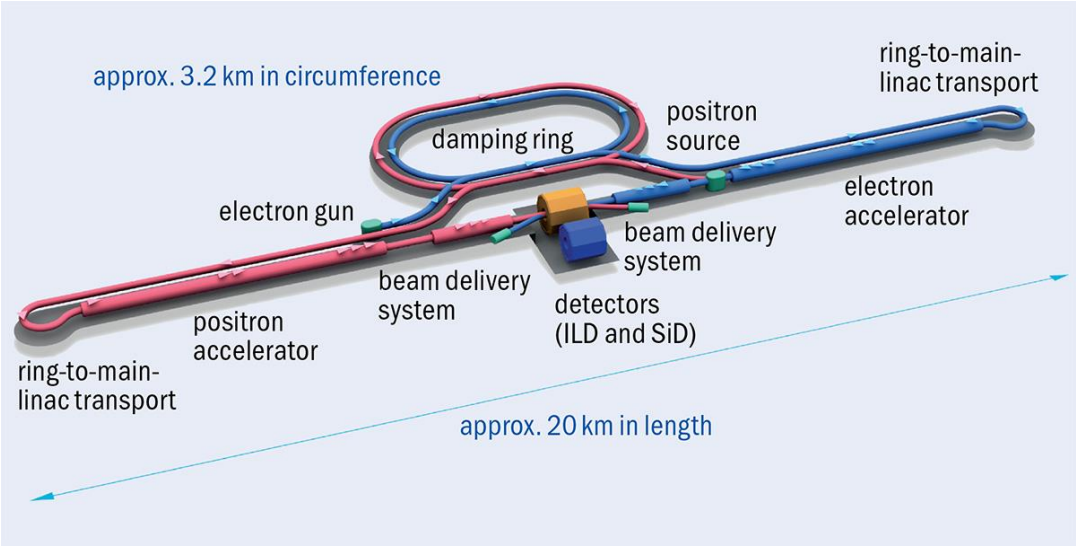
Tatsuki Murata

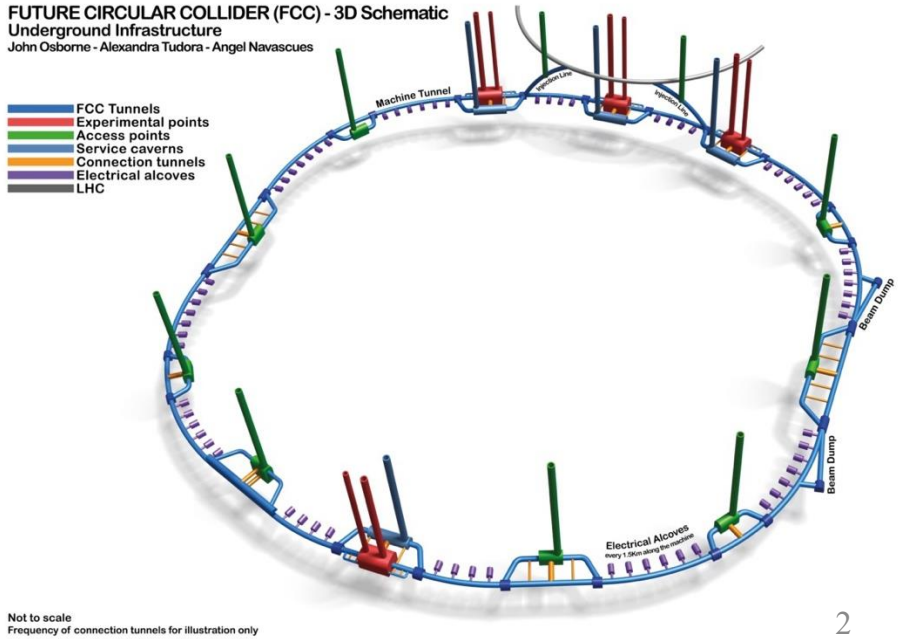
Higgs factory

- 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

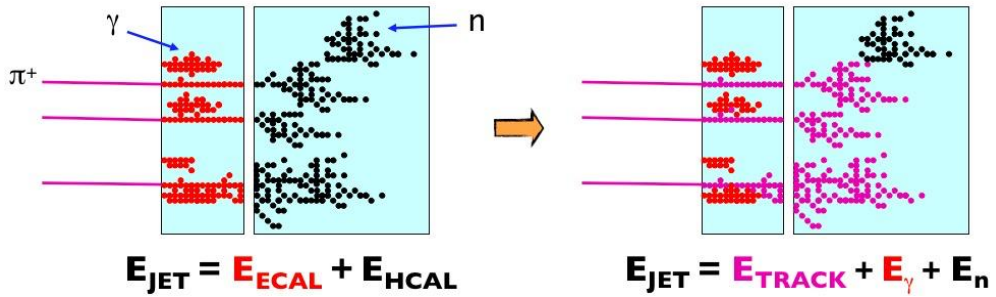
- To achieve the requirement, Particle Flow Algorithm (PFA) oriented high granular detectors are considered
  - ILD, SiD, CLD, etc…
  - (e.g., silicon ECAL: 5×5mm², 30 layers)





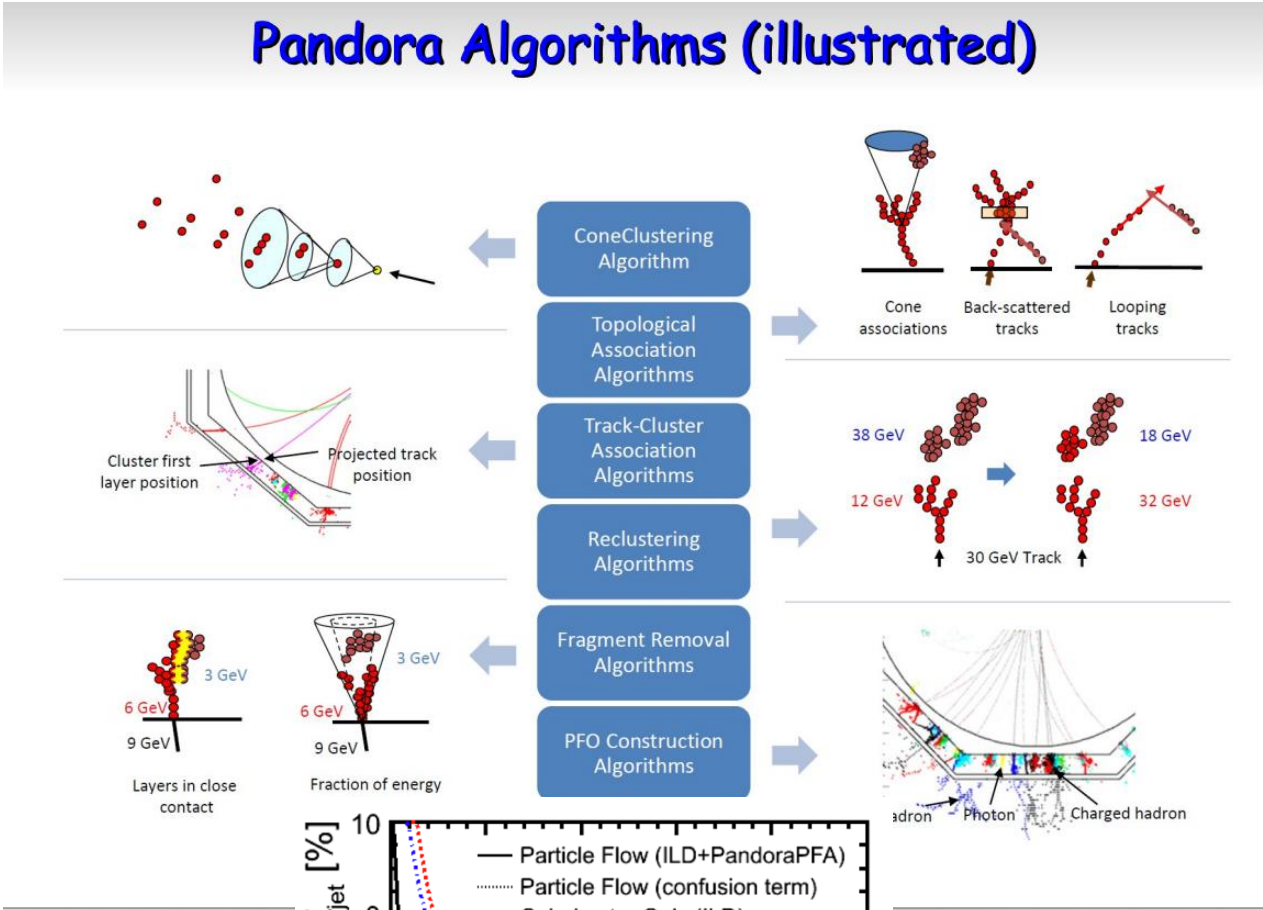
Particle flow

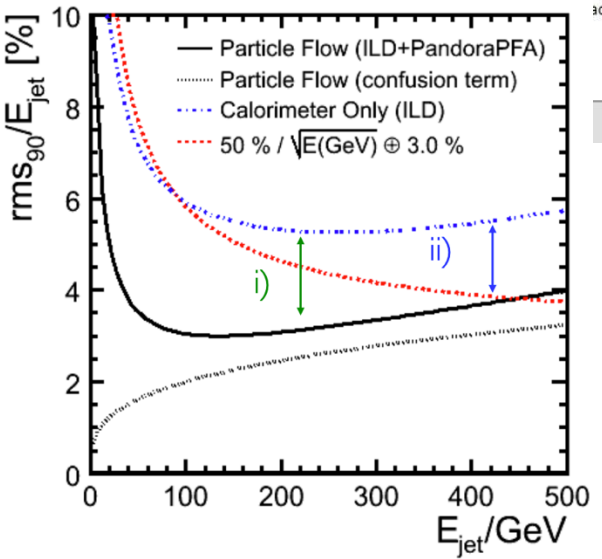
- Detect particle energy by suitable sub-detectors
- Separation of charged/neutral cluster at high granular calorimeter is crucial
- Essential for high jet energy resolution



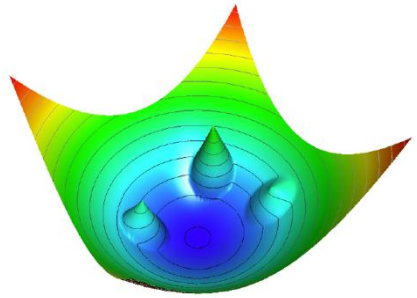
- Current algorithm : **PandoraPFA**
  - Pattern recognition based on the human-tuned parameters
  - Limitations: difficult to optimize or incorporate new information (e.g., timing)

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





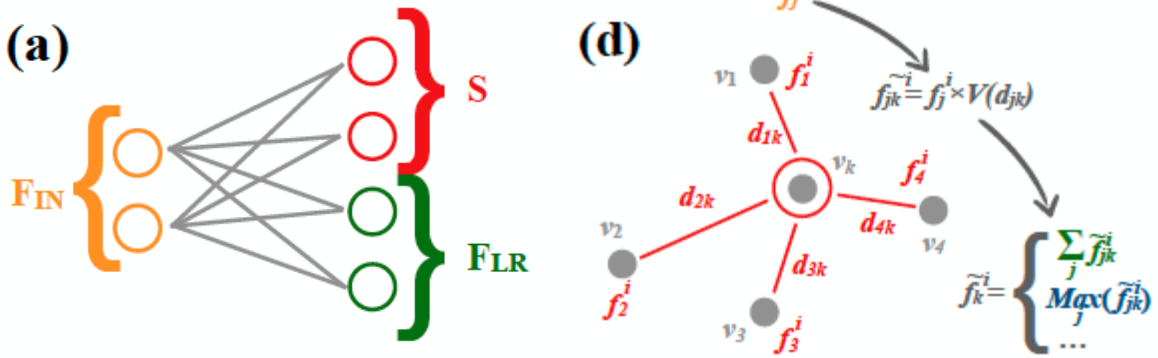
GravNet and object condensation



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



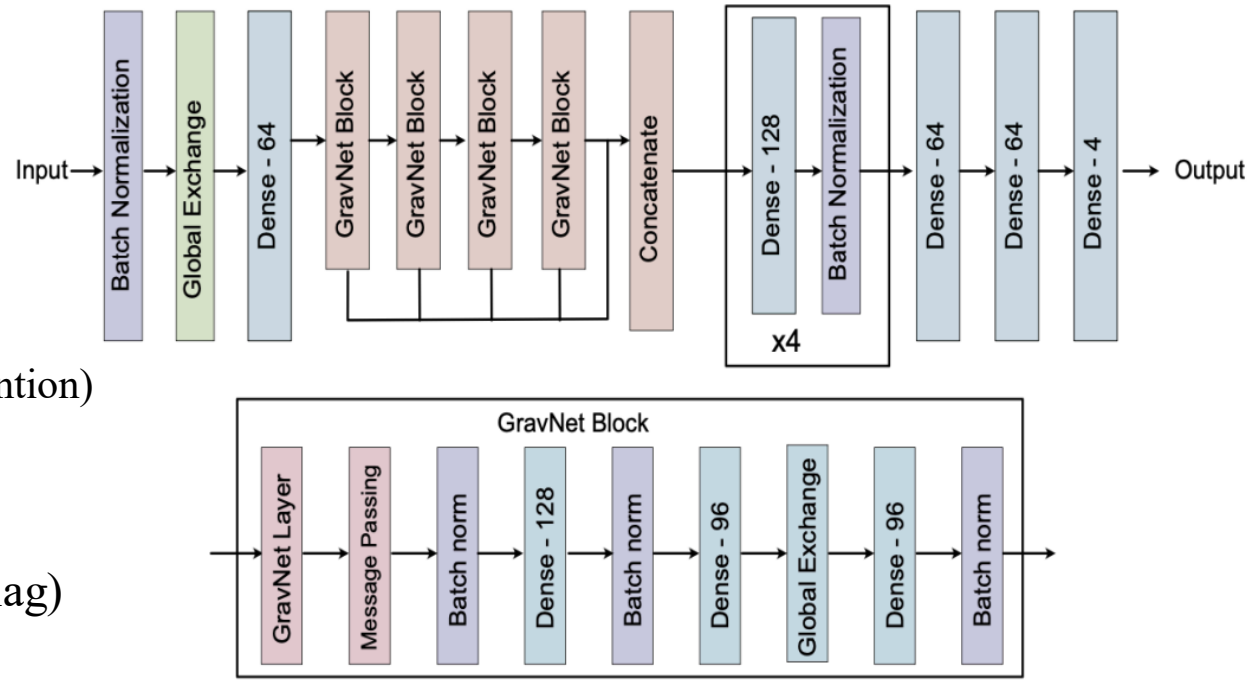
Object condensation (loss function)

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

- 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

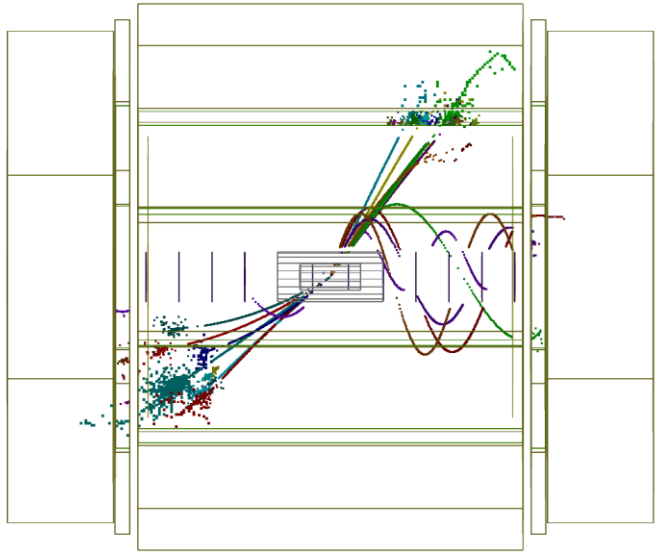
- Originally developed by CMS HGCal group
- Input/output are obtained for each hit at calorimeters
  - Input: features at each hit
    - Position
    - energy deposit
    - Tracker momentum (only for track)
  - Output:
    - "condensation coefficient" β
    - position of virtual coordinate for clustering (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

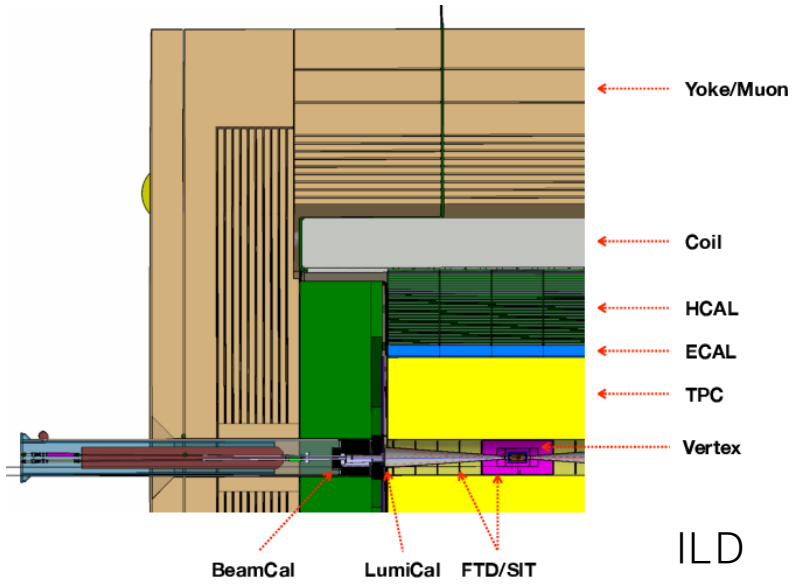


# of parameter: ~1.1M

Samples for performance evaluation

- ILD full simulation with SiW-ECAL and AHCAL (ILD_l5_o1_v02, 020301)
  - Two types of samples : di-jets, fixed energy jets
    - di-jets (ee→Zh→qq$\nu\nu$,q=uds, √s=250 GeV)
      - 2,000,000 events
      - Training sample
    - fixed energy jets (√s=40, 91, 200, 350, 500 GeV)
      - 150,000 events
      - Validation sample

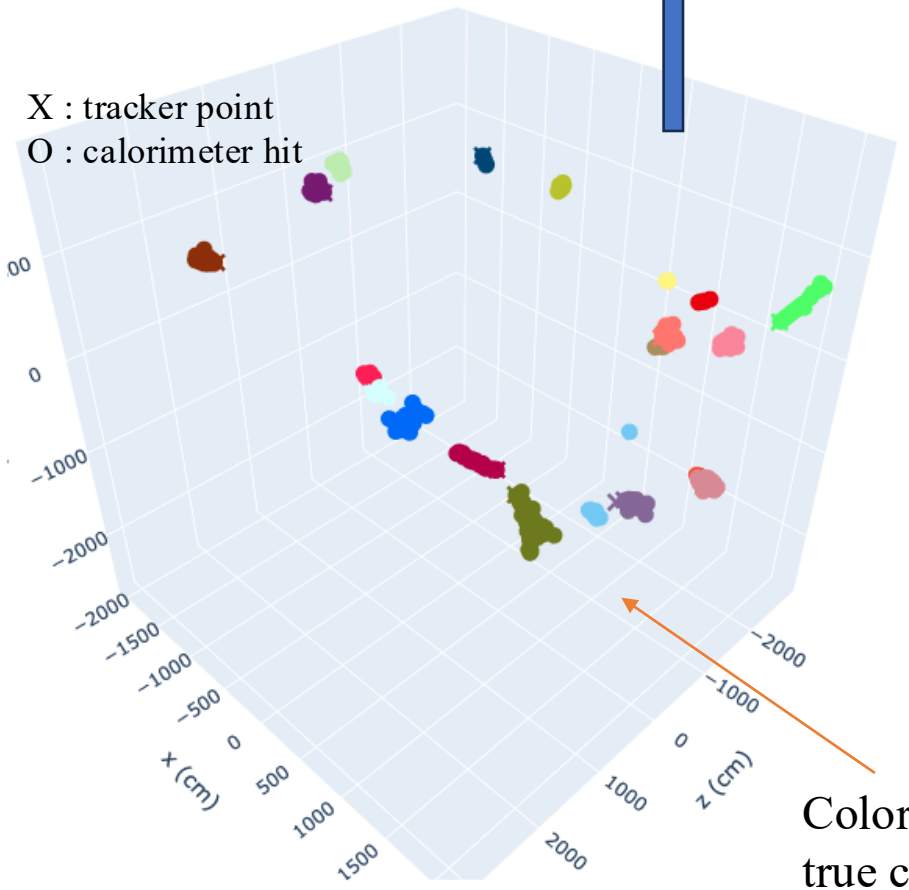




Event display
tau event

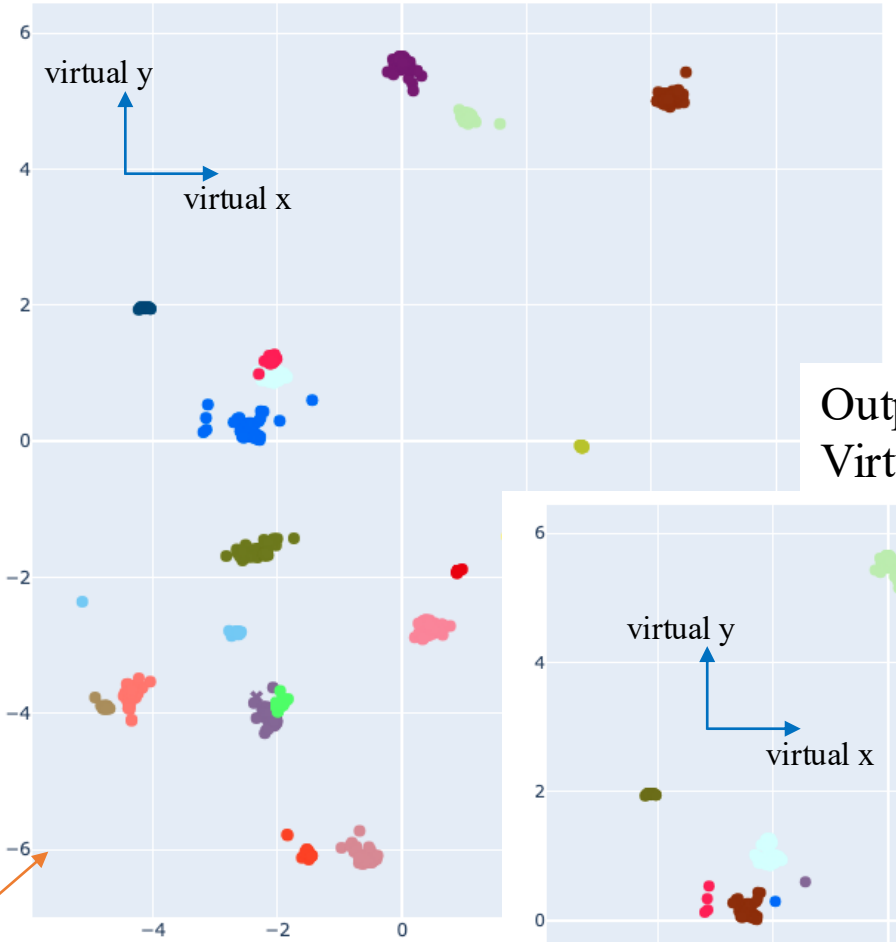
Network

X : tracker point
O : calorimeter hit

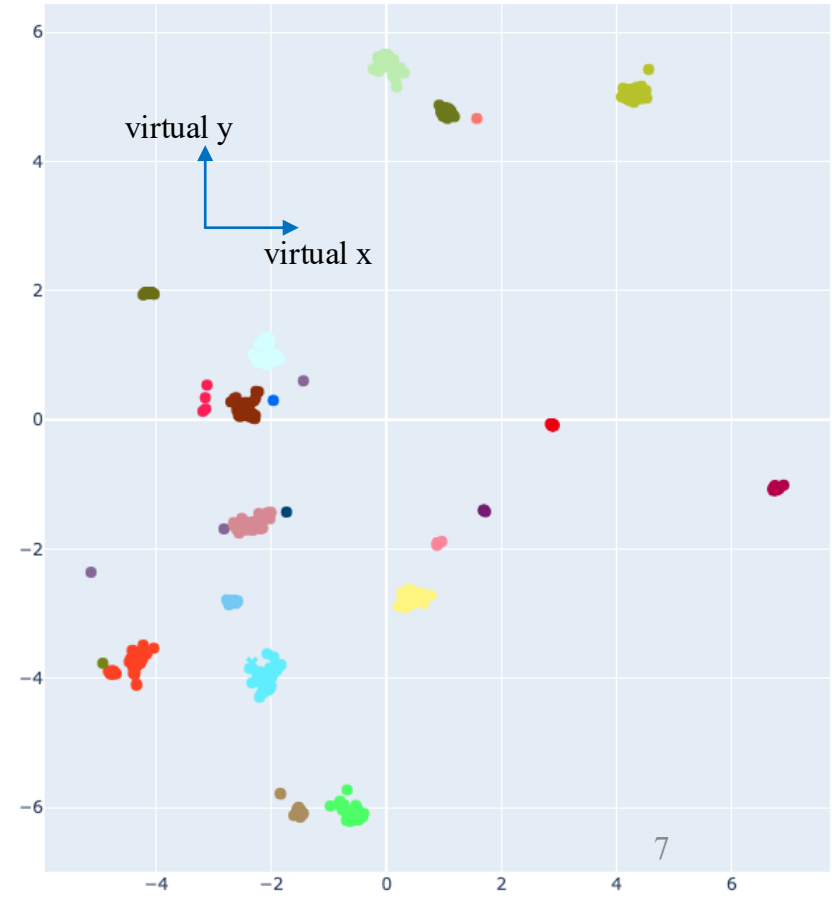


Input features
Real coordinate in ILD

Colored by
true clusters



Output features
Virtual coordinate



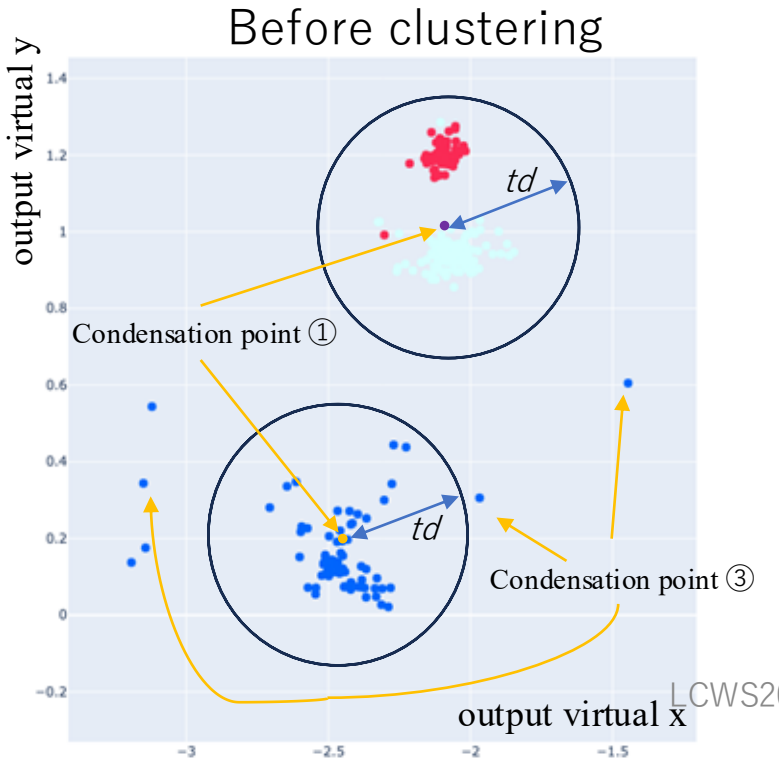
Colored by reconstructed
clusters

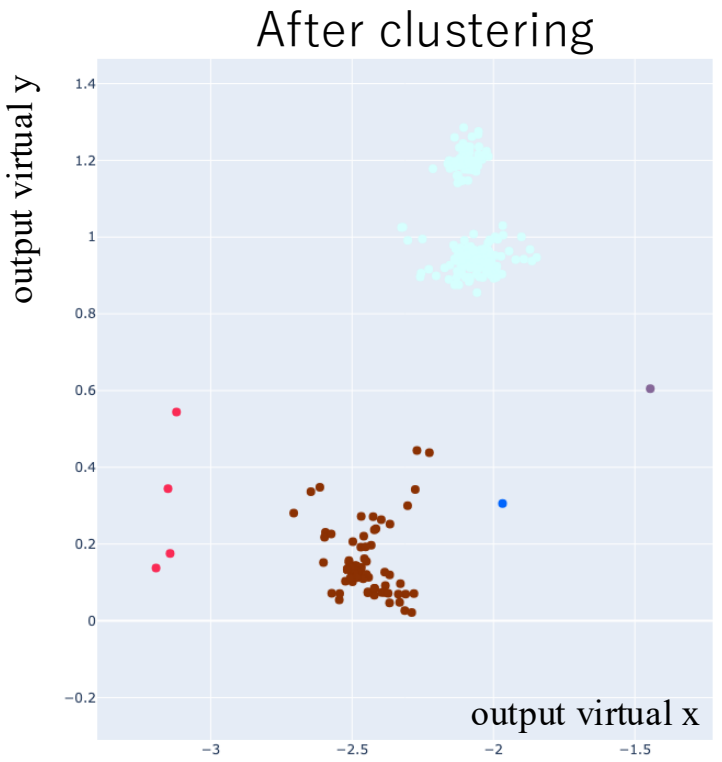
Clustering algorithm

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

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

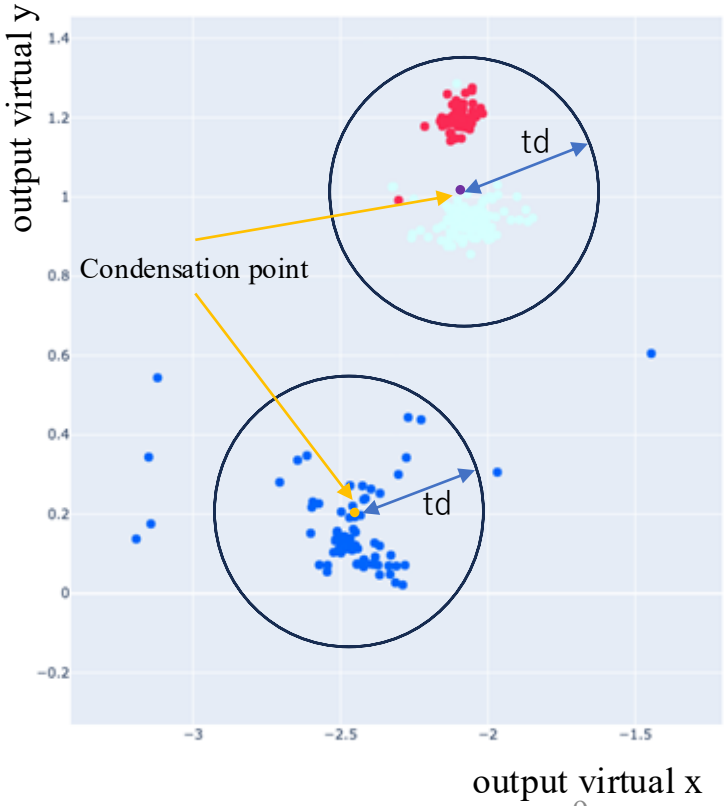
③ Repeat with next-highest β hit until all hits are clustered





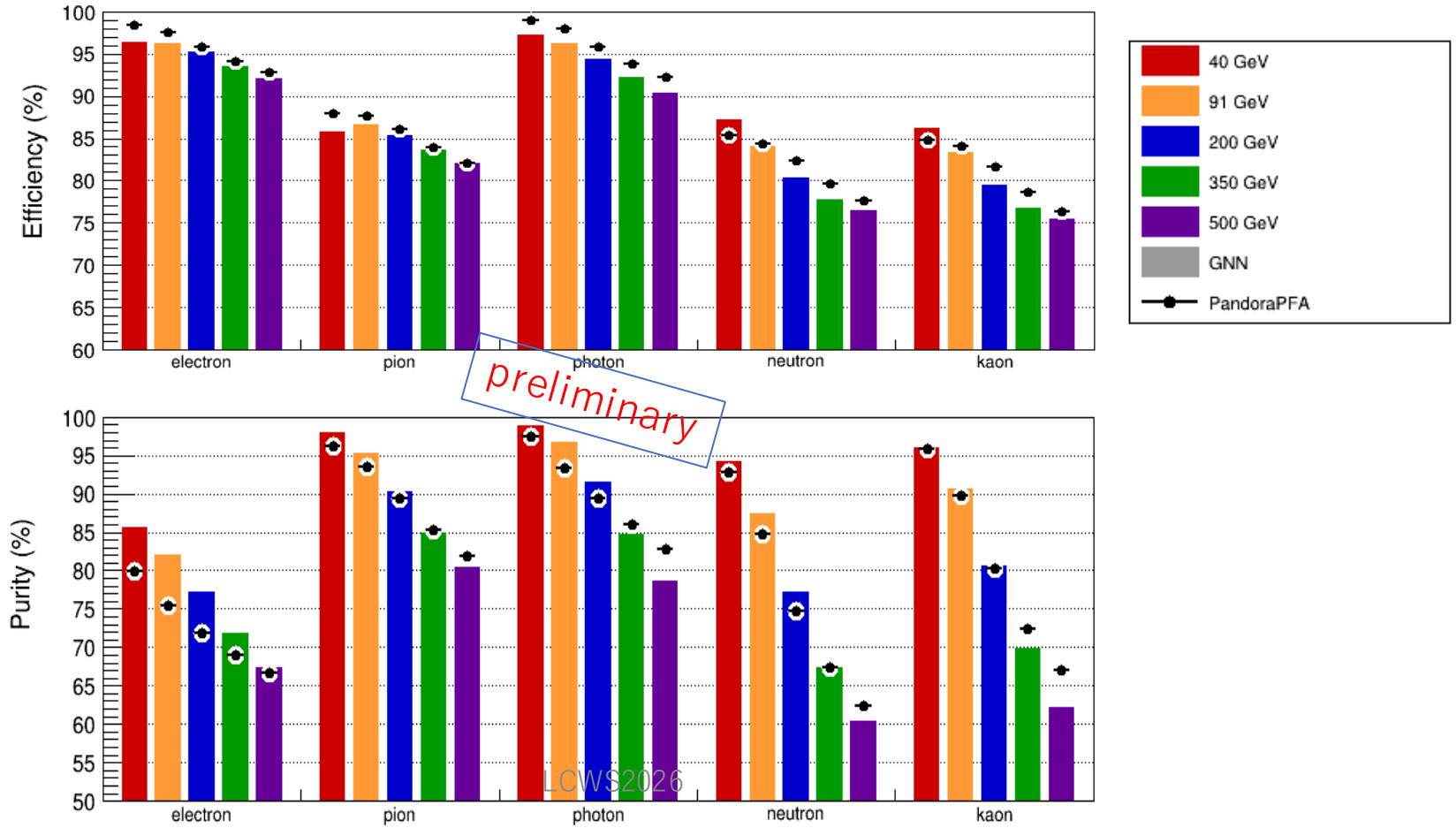
Clustering performance evaluation

- 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>

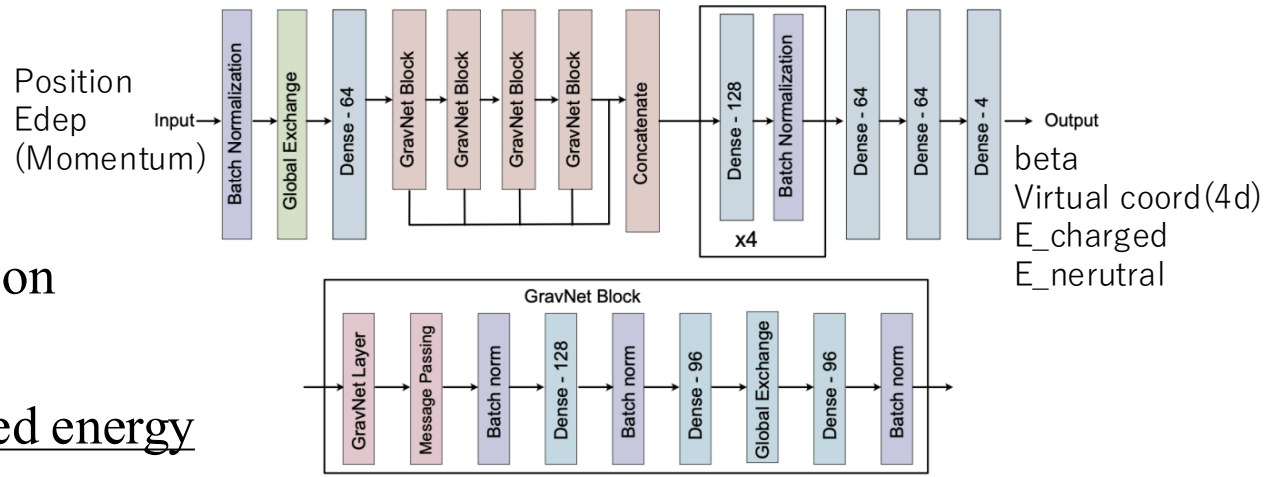


Fixed energy samples

- Used fixed energy di-jet ($q = u, d, s$) samples ($\sqrt{E_{jj}}$ = 40, 91, 200, 350, 500 GeV)
- The efficiency degrades at higher energies, which is likely due to the lack of high-energy samples in the train dataset.



Energy regression



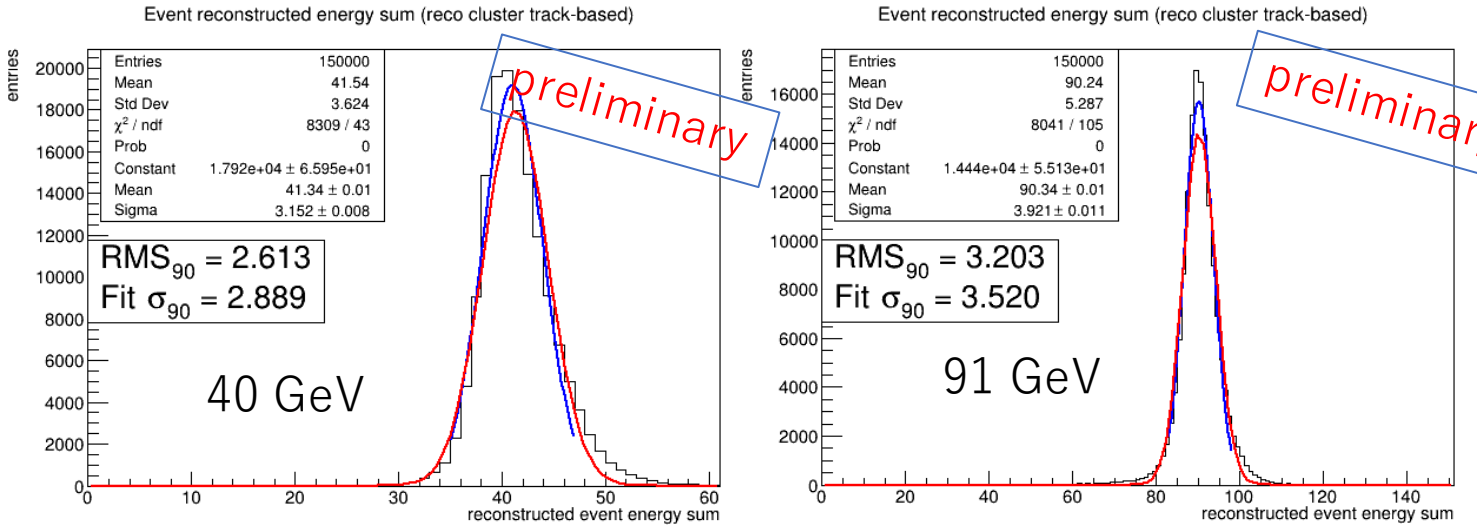
- PandoraPFA aims to minimize jet energy resolution
  →regression is critical
- Before examining jet energy resolution, performed energy regression on decay particle level

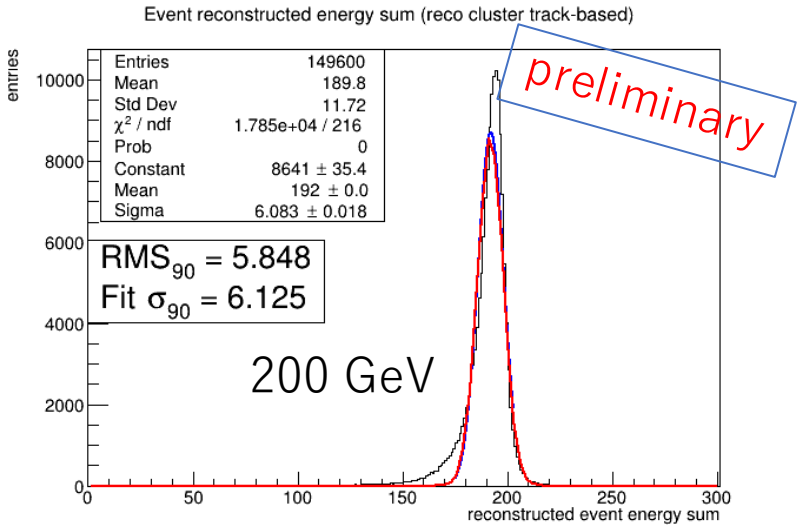
- Added two energy regression terms to model output and loss function
  - $L_{E,charged} = \sum_i \log\left(\left|E_{truth,i} - E_{pred,charg,i}\right| + 1\right)$ i:summation over only condensation points
    - **For charged particles**
    - To fully use track information
    - Estimate cluster energy only at the condensation point
  - $L_{E,nerutral} = \sum_i \log\left(\left|E_{truth,i} - \sum_j E_{pred,calo,i,j}\right| + 1\right)$ 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

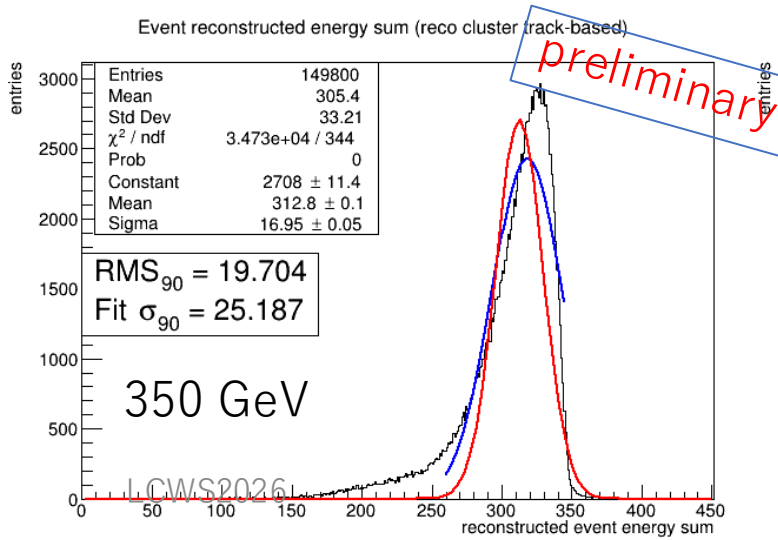
Jet energy reconstruction

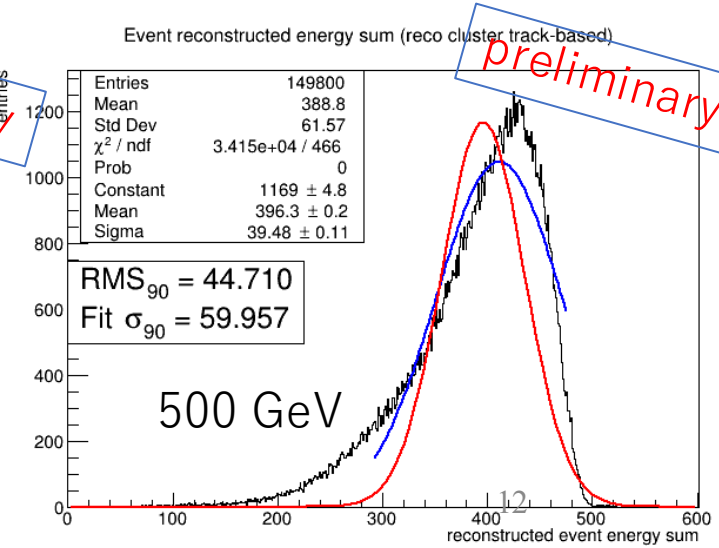
- Di-jet energy was estimated by summing the predicted energies of reconstructed clusters.
  - Added the ”**charged**” energy for clusters with tracks.
  - Added the ”**neutral**” energy for clusters **without** tracks.

- The GNN systematically under-estimates energy of high-energy particles
  - Due to lack of high energy particles in training samples



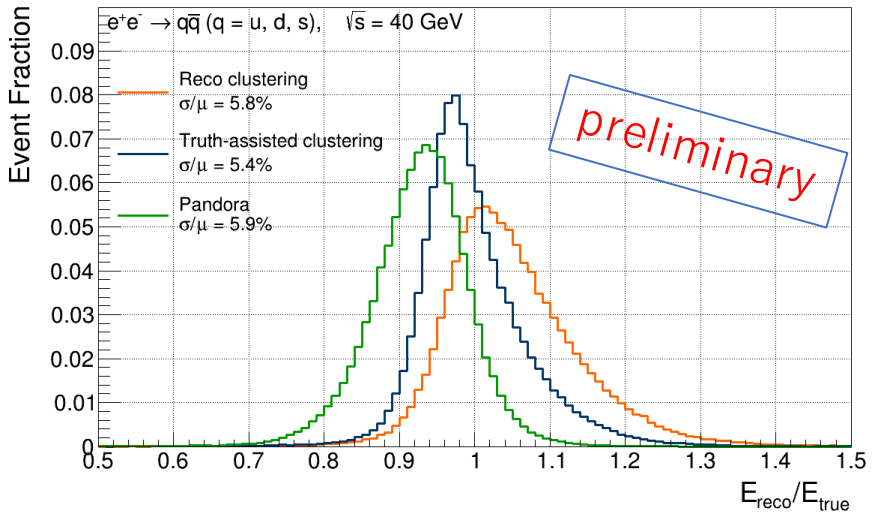


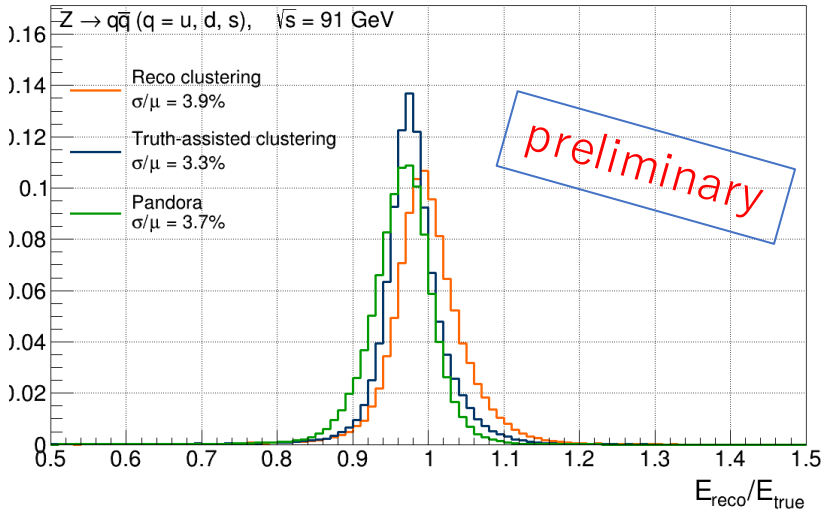


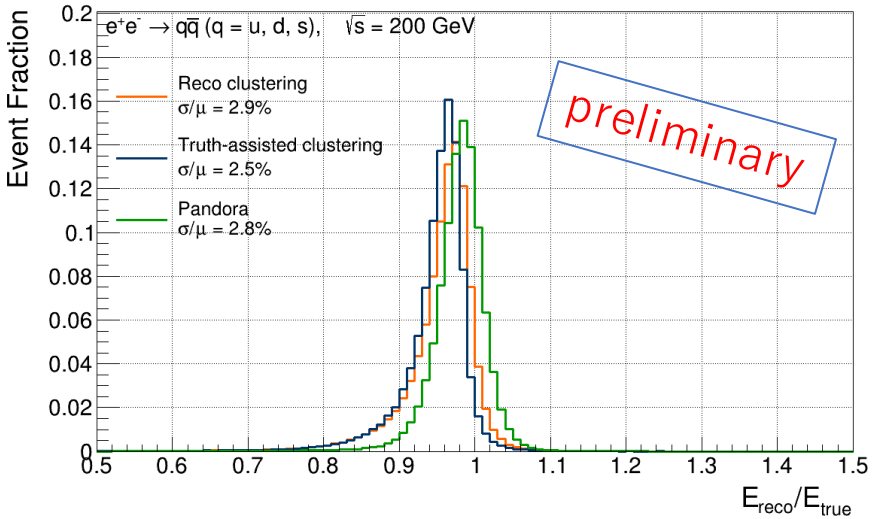


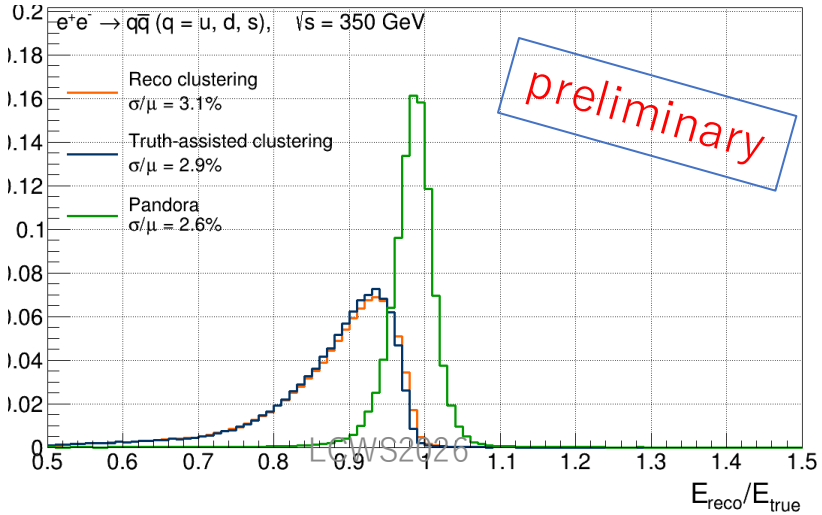
Jet energy resolution

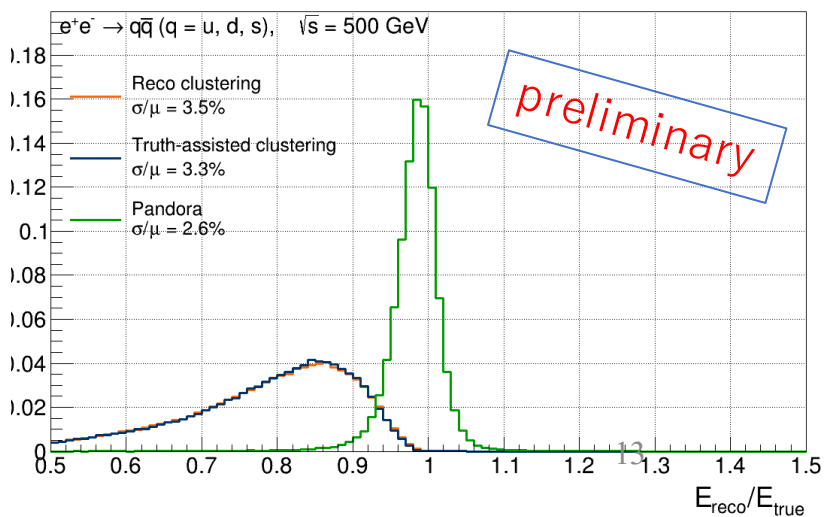
- When evaluating the central peak resolution (σ/μ), GNN performs competitively with PandoraPFA across low-to-mid energies
- Insufficient energy regression for high-energy samples creates a pronounced long tail on the lower-energy side of the distribution.





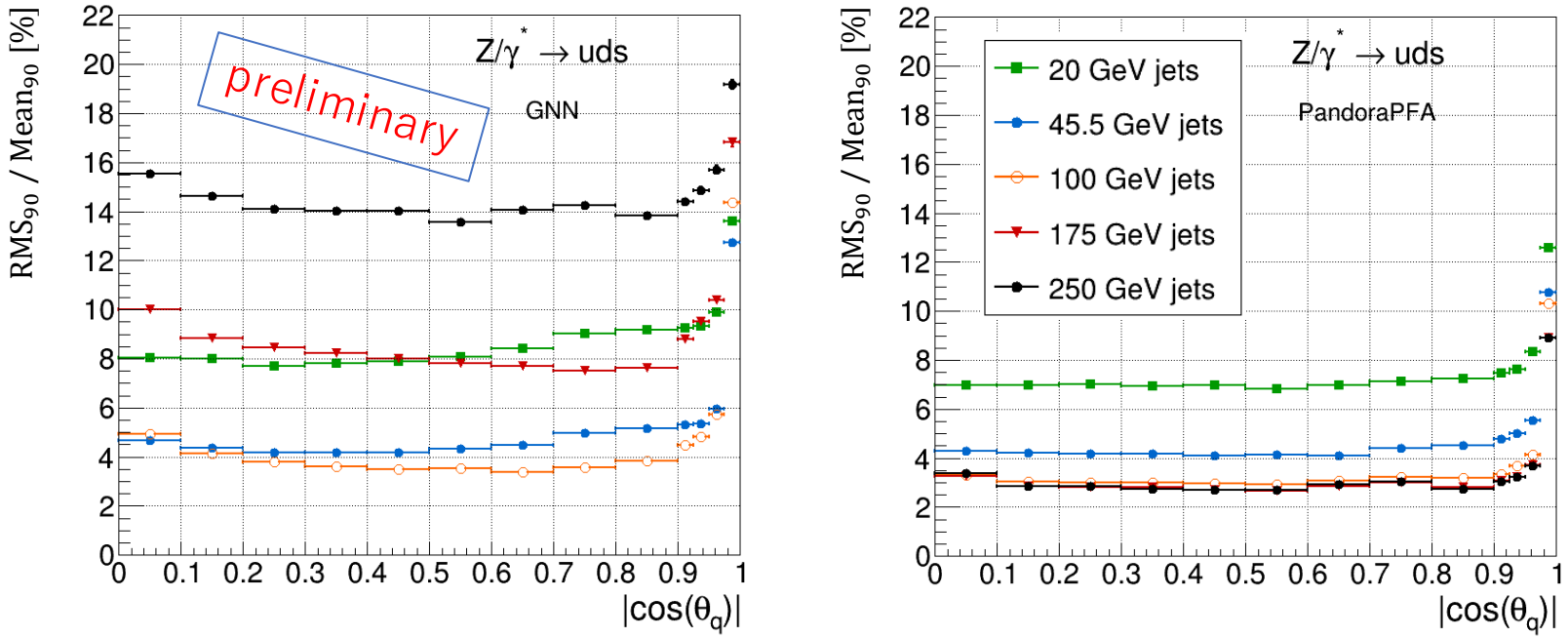






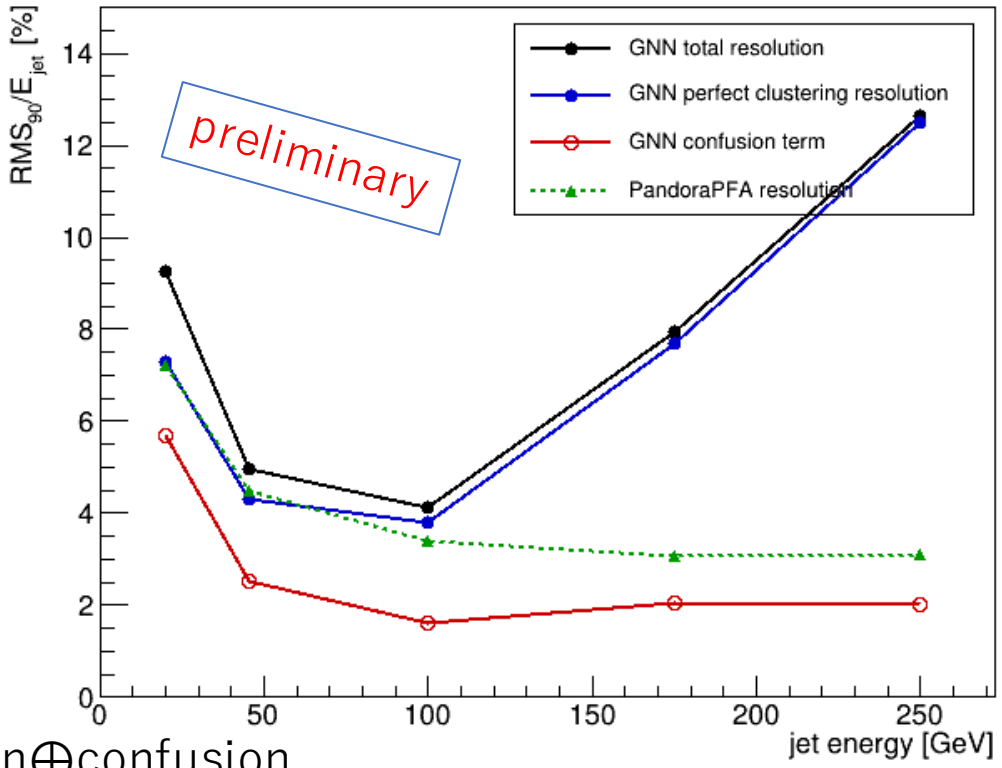
Angular Dependence of Jet Energy Resolution

- Evaluated using standard resolution metric $\frac{\mathrm{RMS_{90}}}{\mathrm{Mean90}}$ vs. jet polar angle $|\cos\theta|$ for direct comparison with PandoraPFA.
- GNN exhibits a flat angular response across barrel and endcaps.
  - However, GNN resolution remains **consistently higher than PandoraPFA by ~0.5-1%** even at lower energies due to minor low-energy tails.

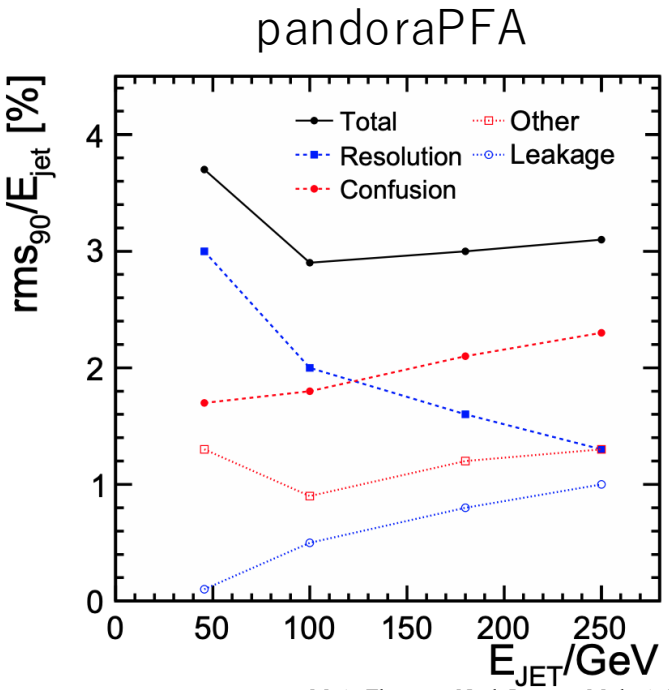


Jet energy resolution

- GNN resolution (black) remains very close to PandoraPFA (green dashed)
- Performance divergence relative to PandoraPFA at high energies is driven by GNN intrinsic regression failure, NOT by cluster confusion or merging errors.



Total = resolution⊕confusion



M. A. Thomson, Nucl. Instrum. Meth. A 611, 25 (2009)

Summary

- Developing GNN based PFA
  - Evaluated GNN based PFA across $e^{-}e^{+} \rightarrow q\bar{q}$ events from 40 GeV to 500 GeV
  - Maintained high clustering performance and purity across low-to-mid energy scales, effectively grouping calorimeter hits into distinct particle clusters.
  - Achieved performance comparable to PandoraPFA at energies $\leq$ 91 GeV
  - High-energy JER degradation energies $\leq$ 91 GeV is driven by GNN energy regression failure (systematic energy under-estimation and response loss) rather than clustering errors

- Ongoing work
  - Energy Regression Calibration: Redesigning energy loss terms and introducing energy-dependent scale factors for high-density shower hits.
  - Architecture Upgrade: Transitioning to Transformer/Attention-based models to capture long-range energy correlations in dense jets.
  - Enhanced PID Integration: Tightening coupling between tracker momentum and calorimeter energy regression.