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

Tatsuki Murata[1]
Arpad Schaeffer[2], Cedric Trinh[2], Taikan Suehara[3]

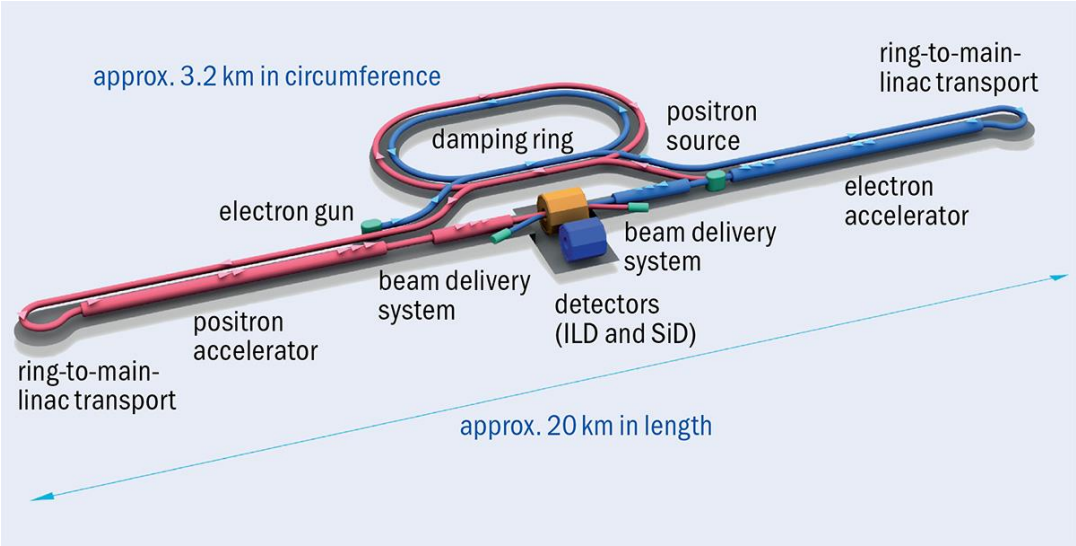
[1] The University of Tokyo
[2] École Polytechnique / ILANCE, [3] ICEPP, The University of Tokyo

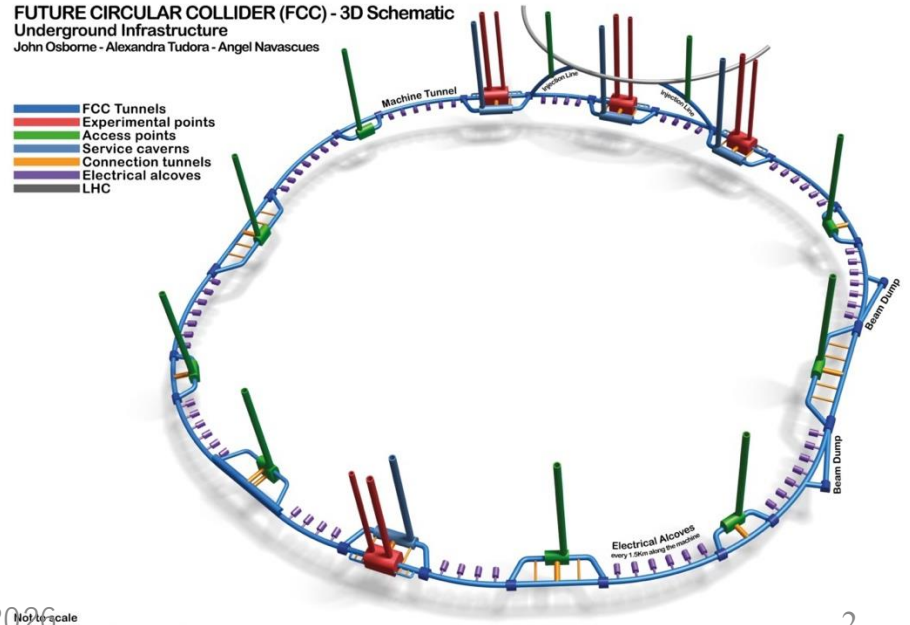
Higgs factories

- Precise measurements of Higgs boson properties provide sensitive probes of physics beyond the Standard Model

- Future e+e- colliders are being developed as Higgs factories.
  - ILC, FCC, etc…
  - Jet energy resolution ~30%/√E

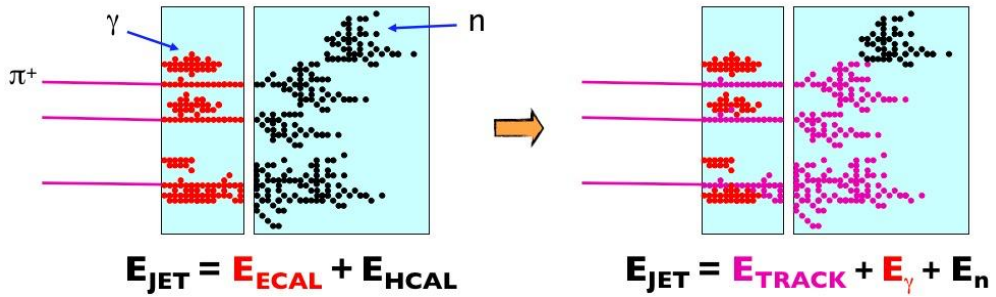
- To achieve the requirement, Particle Flow Algorithm (PFA) oriented highly granular detectors are considered
  - ILD, SiD, CLD, etc…
  - (e.g., silicon ECAL: $5\times5mm^2$, 30 layers)





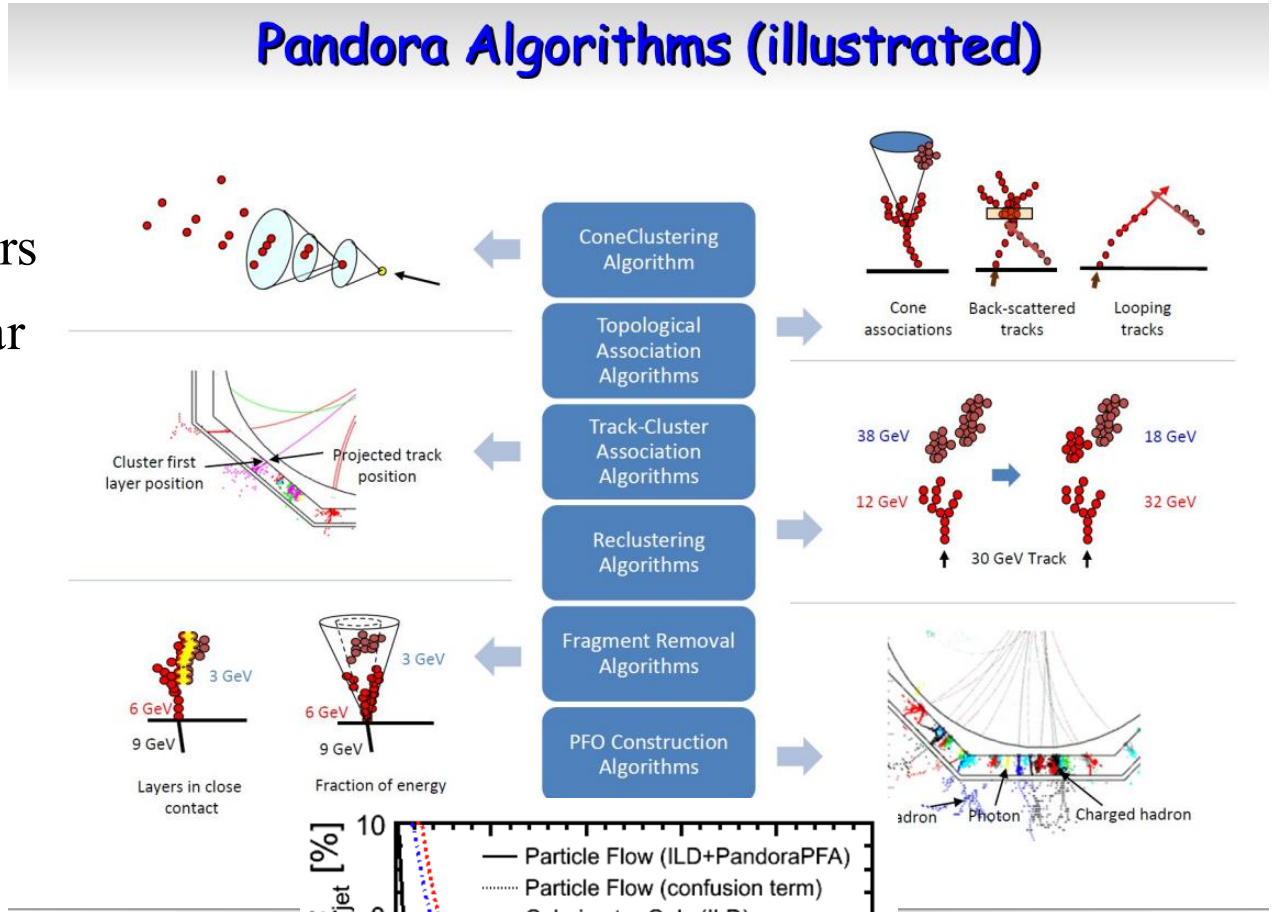
Particle flow

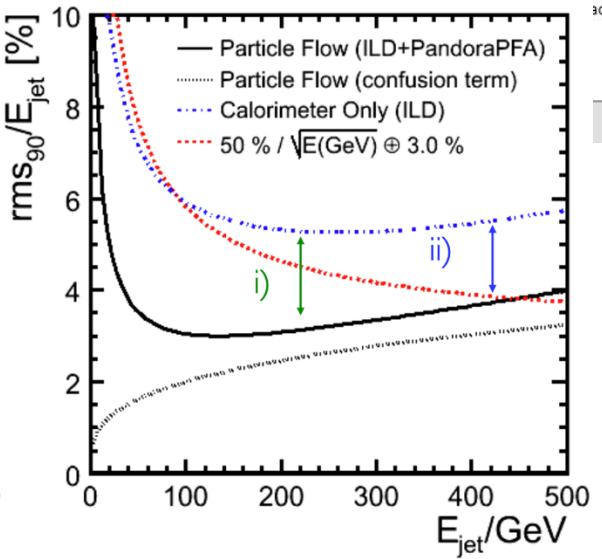
- Reconstruct particle energy using suitable sub-detectors
- Separation of charged/neutral cluster at highly granular calorimeter is crucial
- Essential for high jet energy resolution



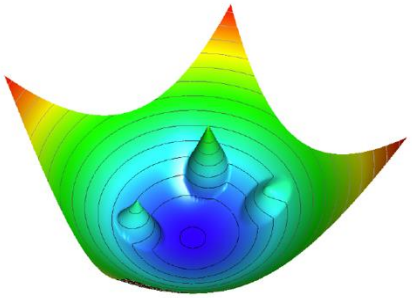
- Current algorithm : **PandoraPFA**
  - Pattern recognition based on the human-tuned parameters
  - Limitations: difficult to optimize or incorporate new information (e.g., timing)

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





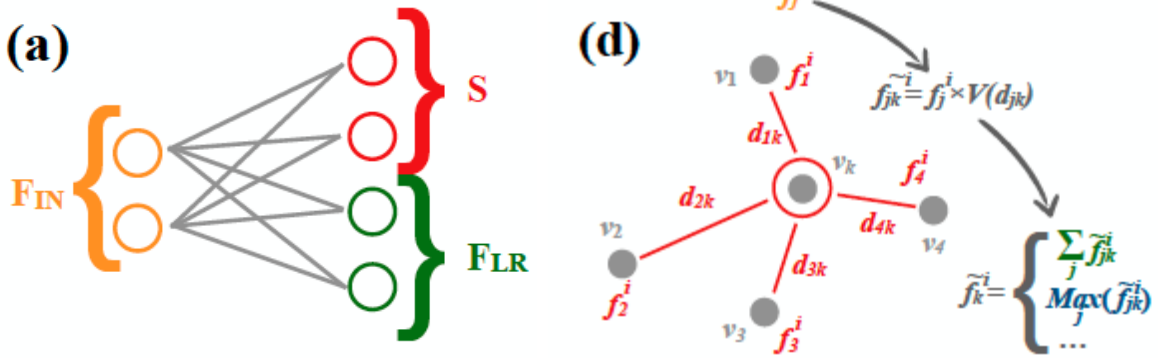
GravNet and object condensation



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 (distance-dependent weights)
- Concatenate the output with MLP



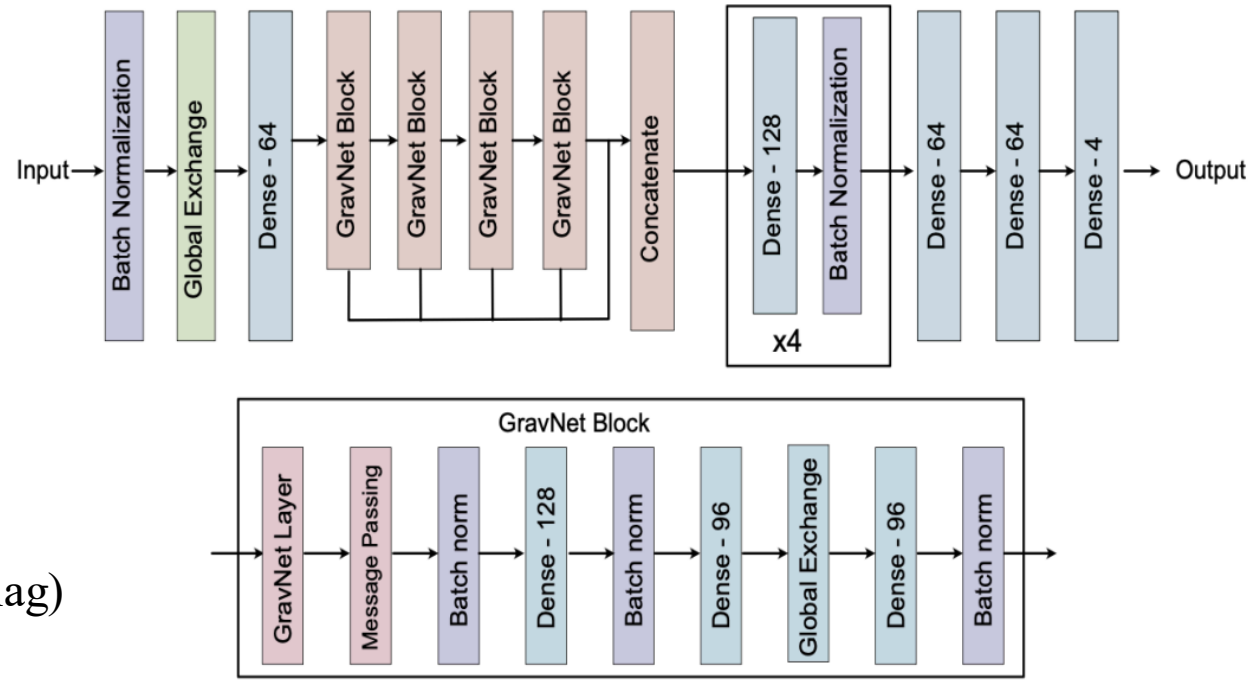
Object condensation (loss function)

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

- 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

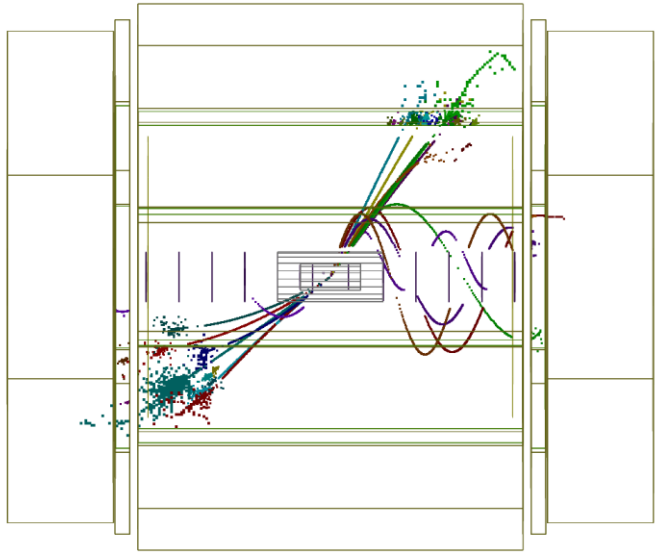
- Originally developed by CMS HGCal group
- Per-node inputs and outputs
  - Input: features at each hit
    - Position
    - energy deposit
    - Tracker momentum (only for track)
  - Output:
    - "condensation coefficient" β
    - position of virtual coordinate for clustering (4D)
    - cluster energy
- 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

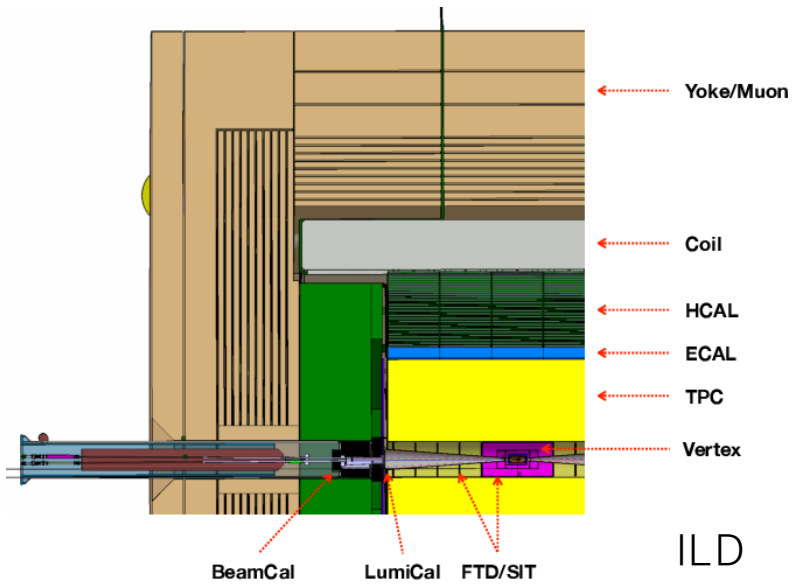


# of parameters: ~1.1M

Training an evaluation samples

- ILD full simulation with SiW-ECAL and AHCAL (ILD_l5_o1_v02, 020301)
  - Training sample
    - di-jets (ee→Zh→qq$\nu\nu$,q=ud, √s=250 GeV)
    - 2,000,000 events
  - Evaluation sample
    - Fixed-energy light-quark, q=uds
    - √s=40, 91, 200, 350, 500 GeV
    - 150,000 events

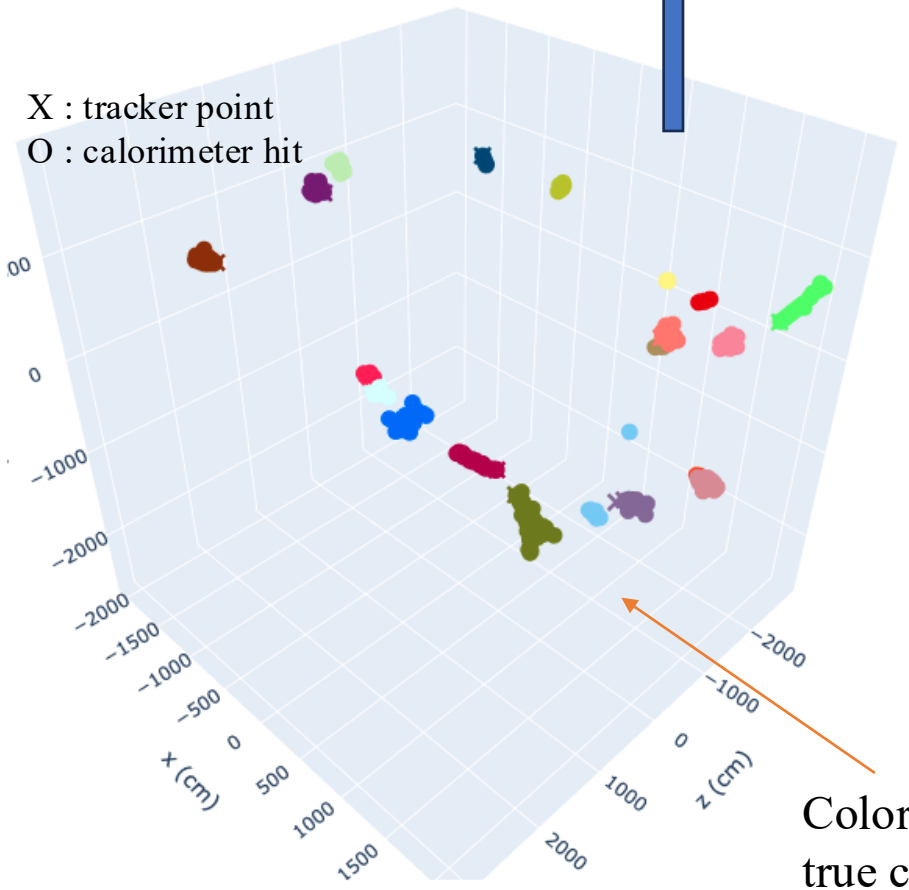




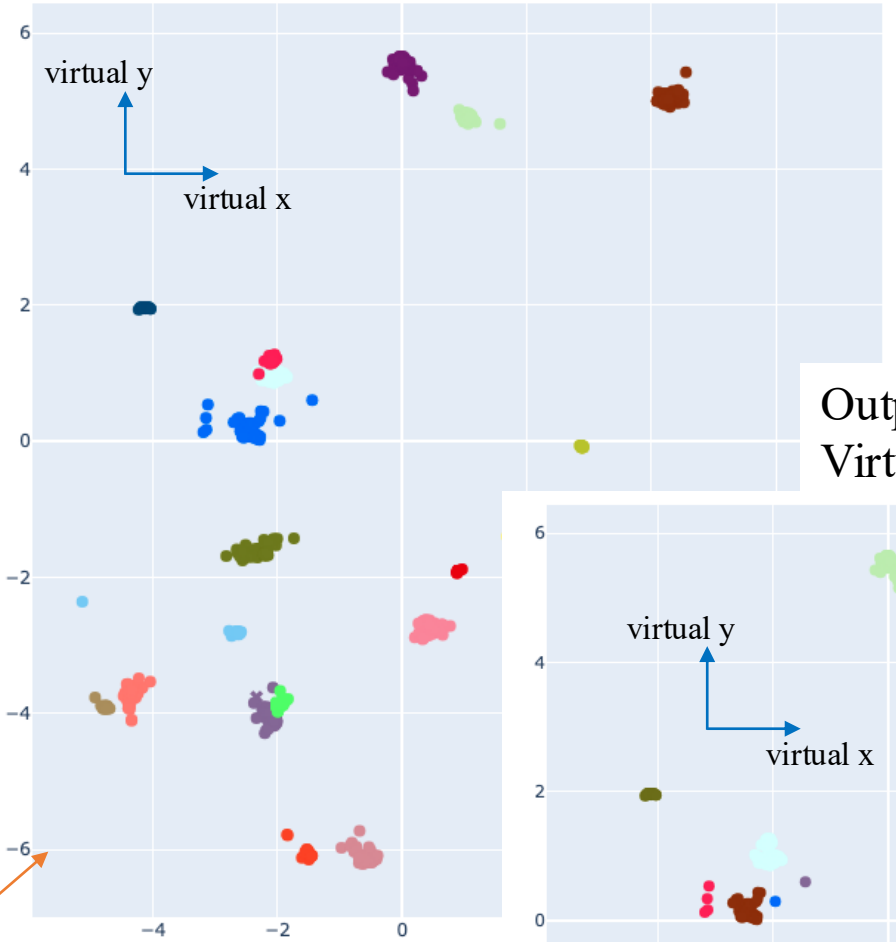
Event display

Network

X : tracker point
O : calorimeter hit

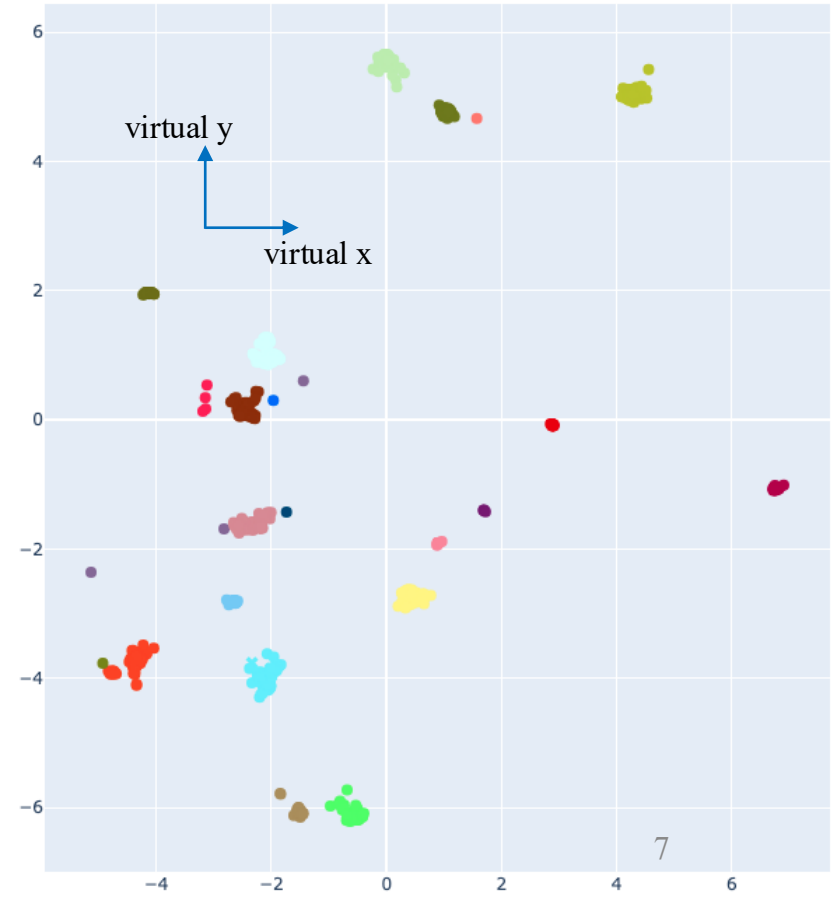


Input features
Real coordinate in ILD



Colored by
true clusters

Output features
Virtual coordinate



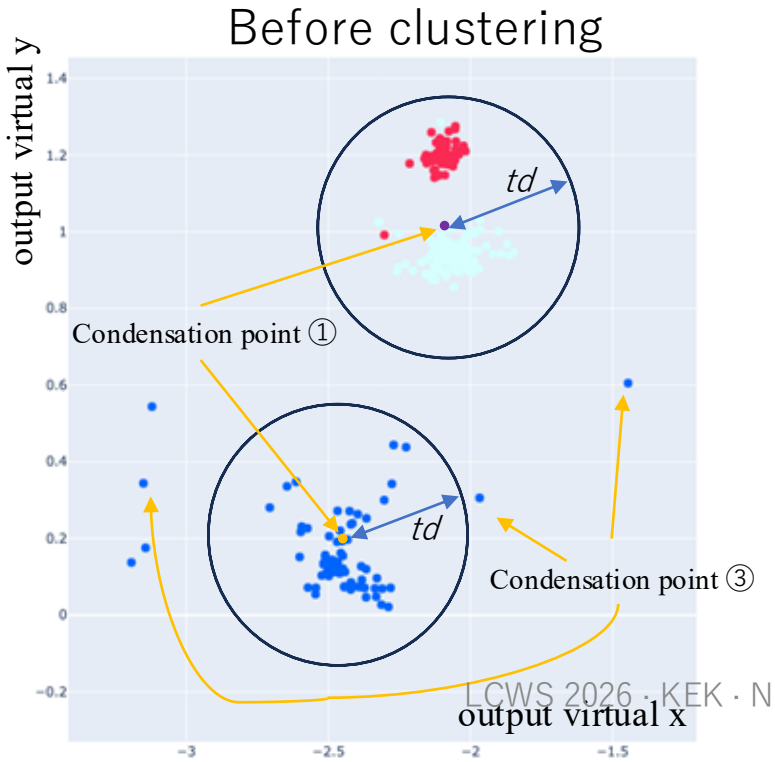
Colored by reconstructed
clusters

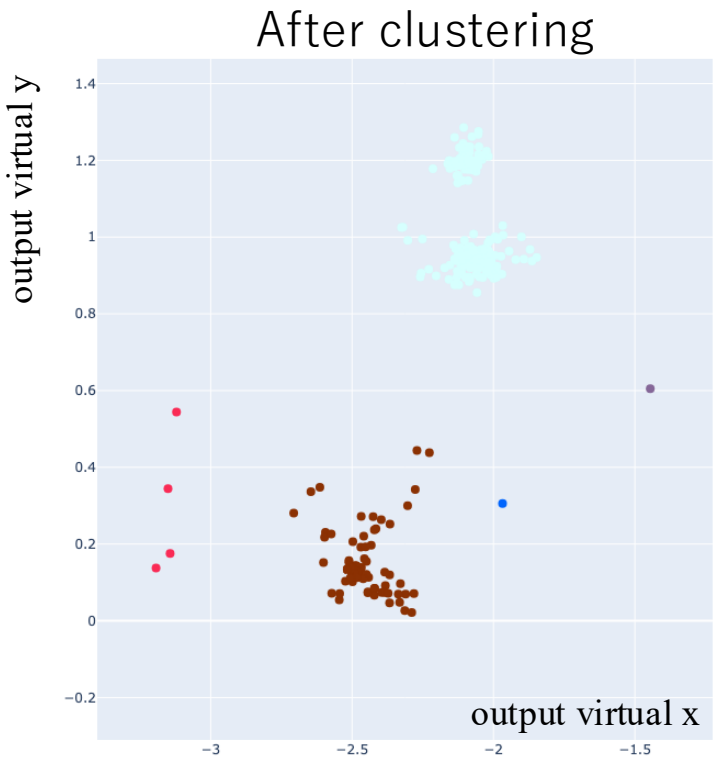
Clustering algorithm

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

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

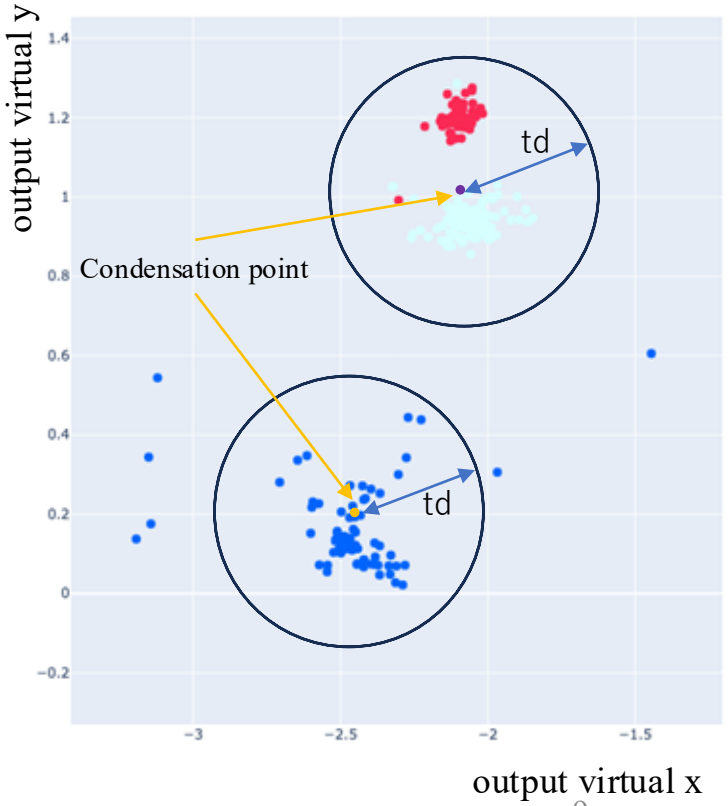
③ Repeat with next-highest β hit until all hits are clustered





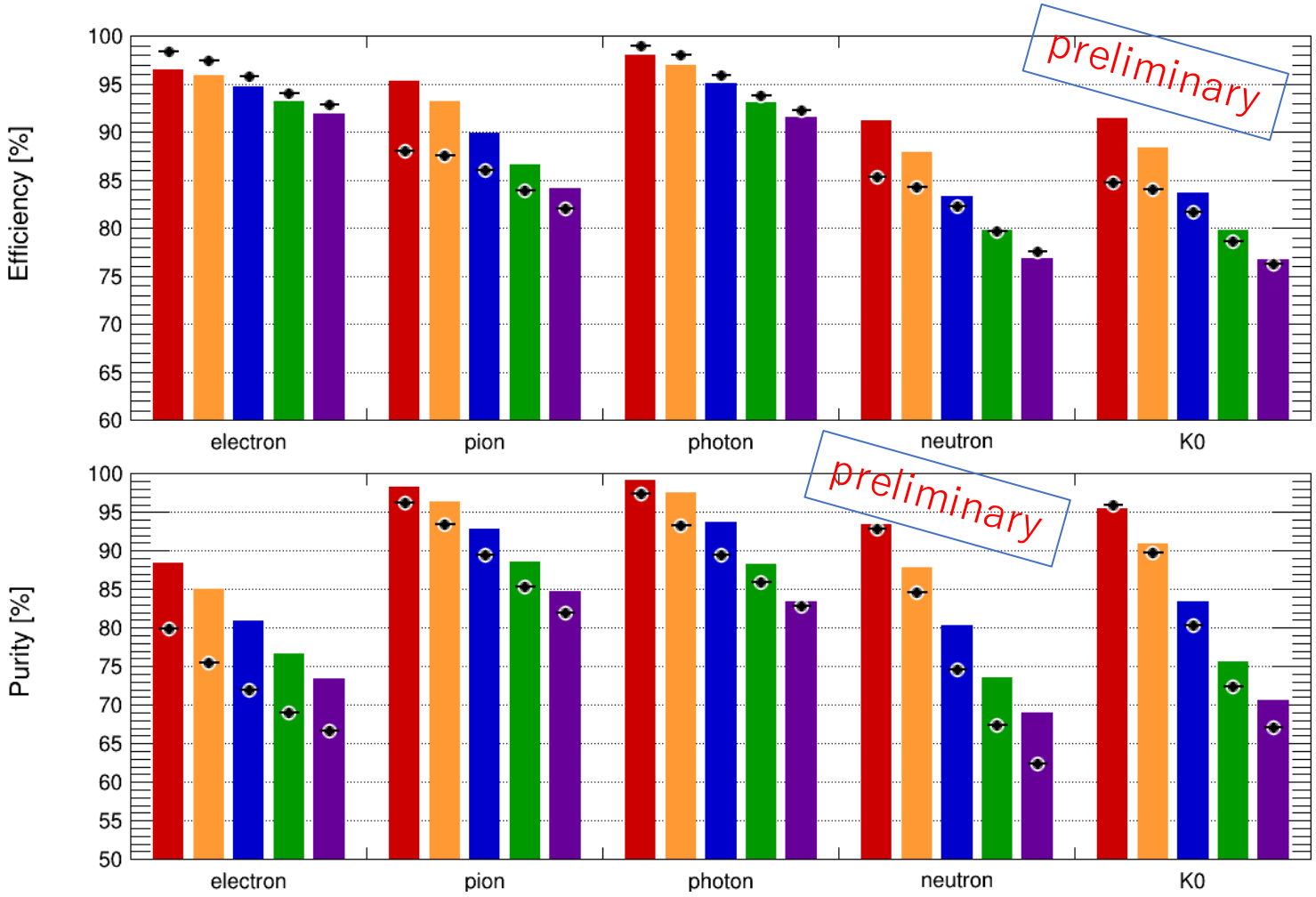
Clustering performance evaluation

- 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* $= \dfrac{E_{match}}{E_{dep}}$
    - *Purity* $= \dfrac{E_{match}}{E_{reco}}$
  - Compared efficiency and purity for
    - Charged particles: <u>electrons, pions,</u>
    - Neutral particles: <u>photons, kaons, neutrons</u>

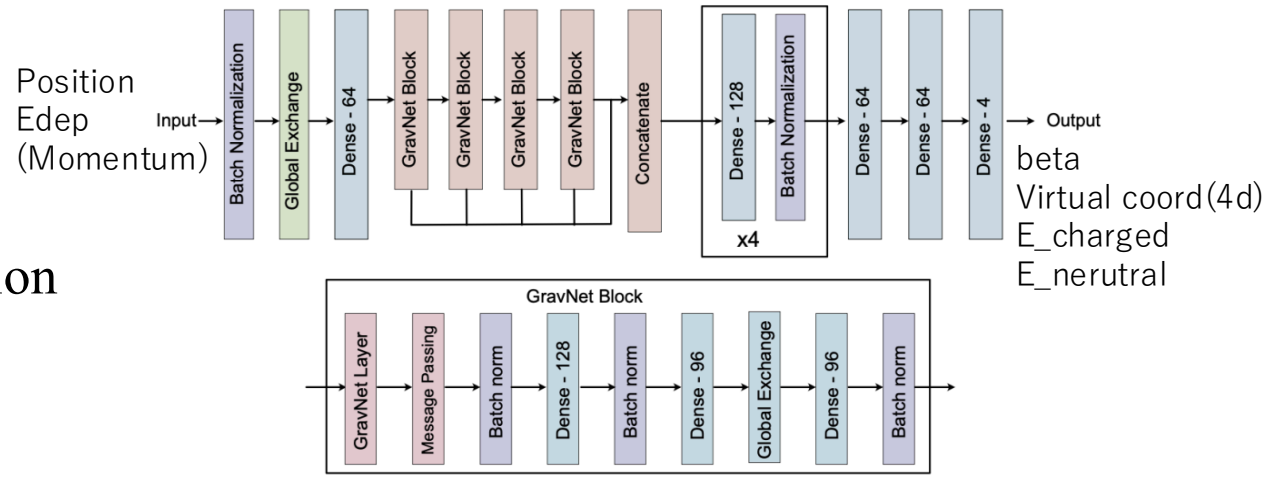


Clustering performance across event energies

- Electron and photon efficiency remains above 90% across the displayed energies.
- Achieved higher pion efficiency compared to PandoraPFA.
- Cluster purity is generally higher than PandoraPFA in this comparison.
- The efficiency degrades at higher energies, which is likely due to the lack of high-energy samples in the train dataset.



Energy regression



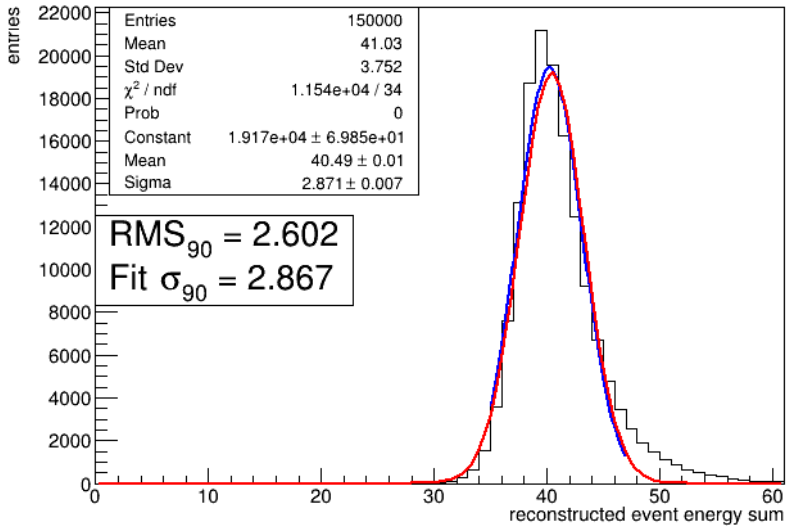
- PandoraPFA aims to minimize jet energy resolution
  →regression is critical

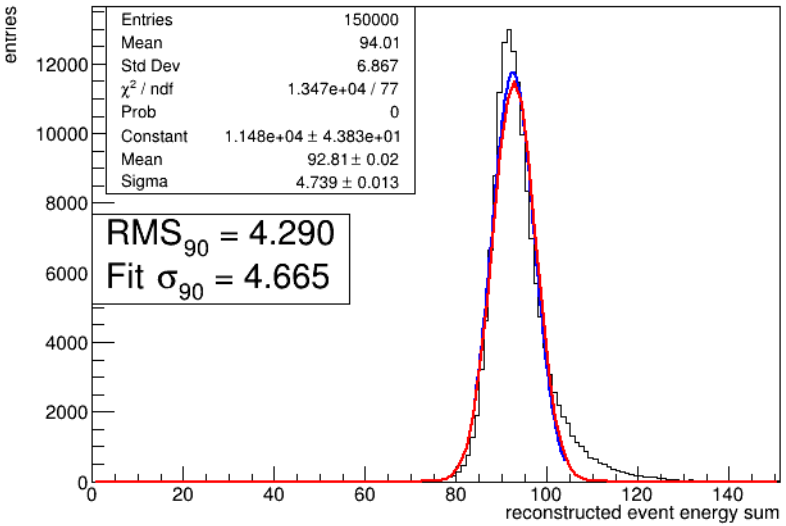
- Added two energy regression terms to model output and loss function
  - $L_{E,charged} = \sum_i \log\left(\left|E_{truth,i} - E_{pred,charg,i}\right| + 1\right)$ i:summation over only condensation points
    - **For charged particles**
    - To fully use track information
    - Estimate cluster energy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

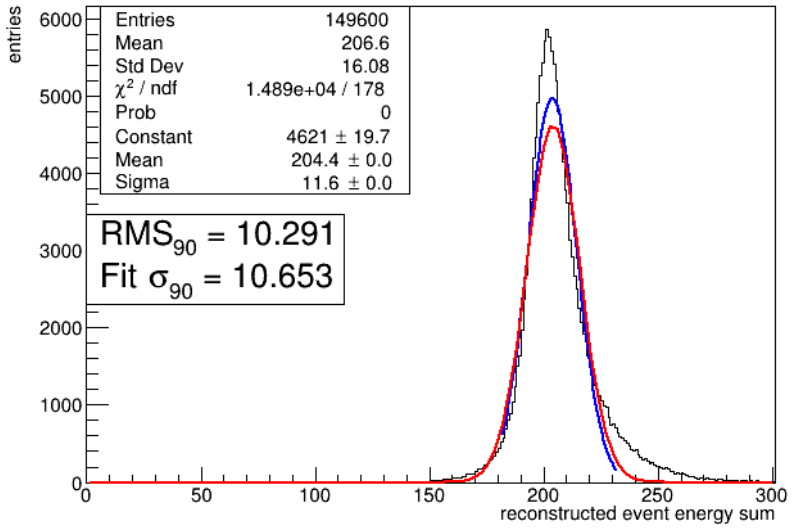
Event-energy reconstruction

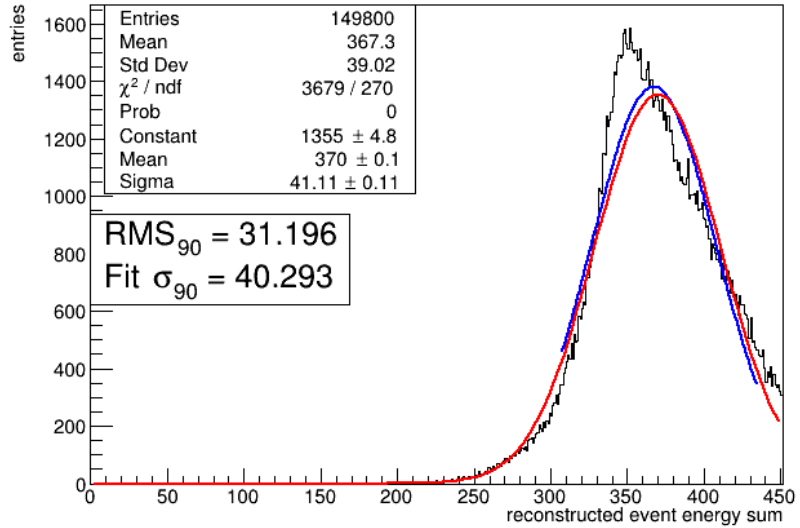
- Reconstruct the event energy by summing the selected cluster energies.
  - Added the ”**charged**” energy for clusters with tracks.
  - Added the ”**neutral**” energy for clusters **without** tracks.

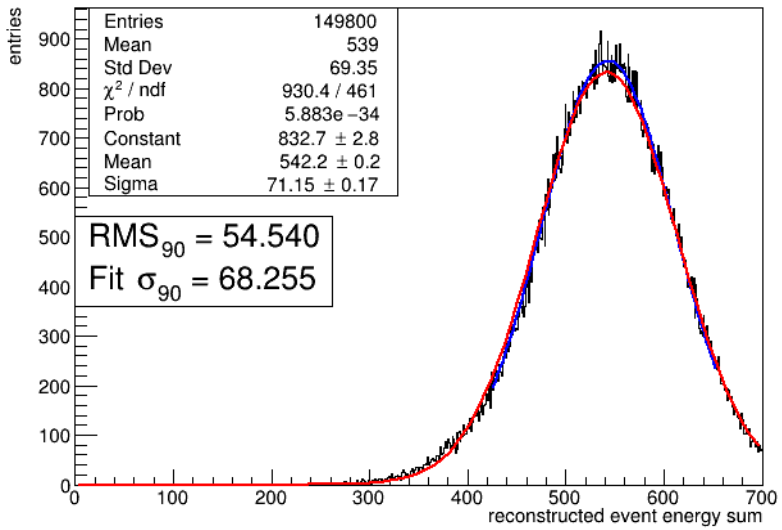
- The GNN systematically under-estimates energy of high-energy particles
  - Due to lack of high energy particles in training samples





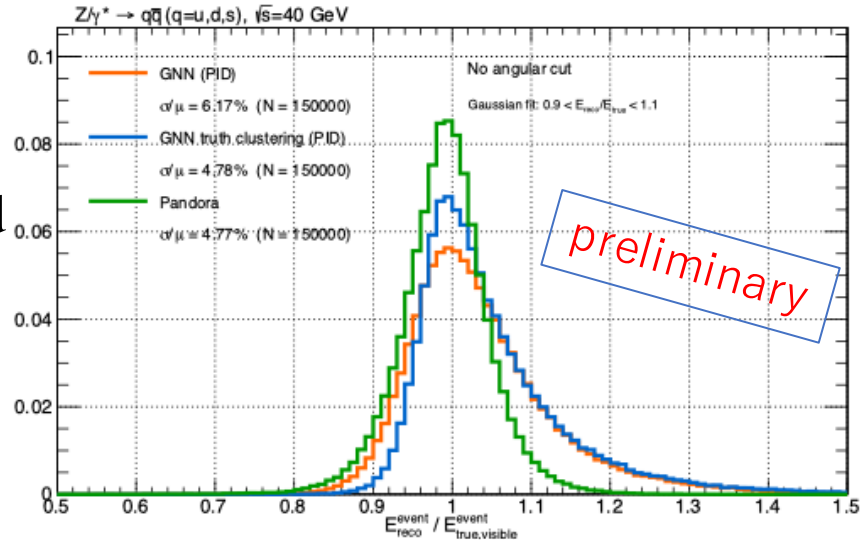


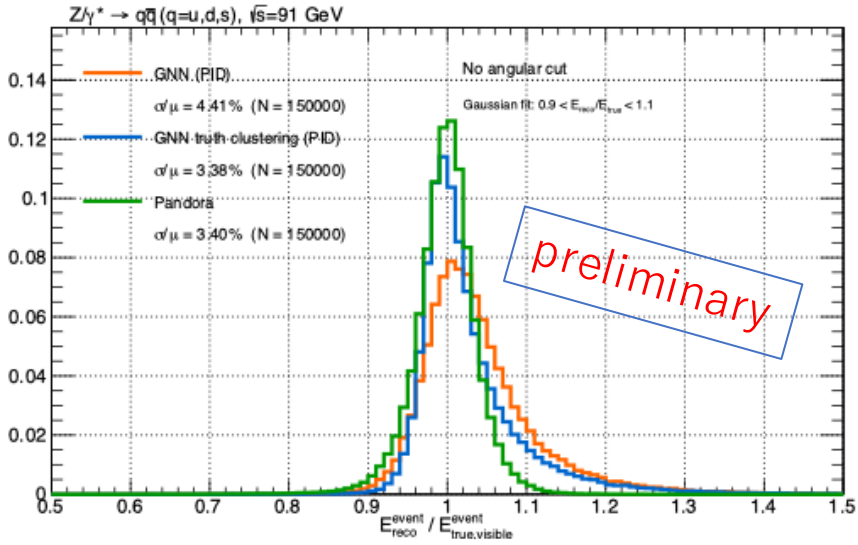


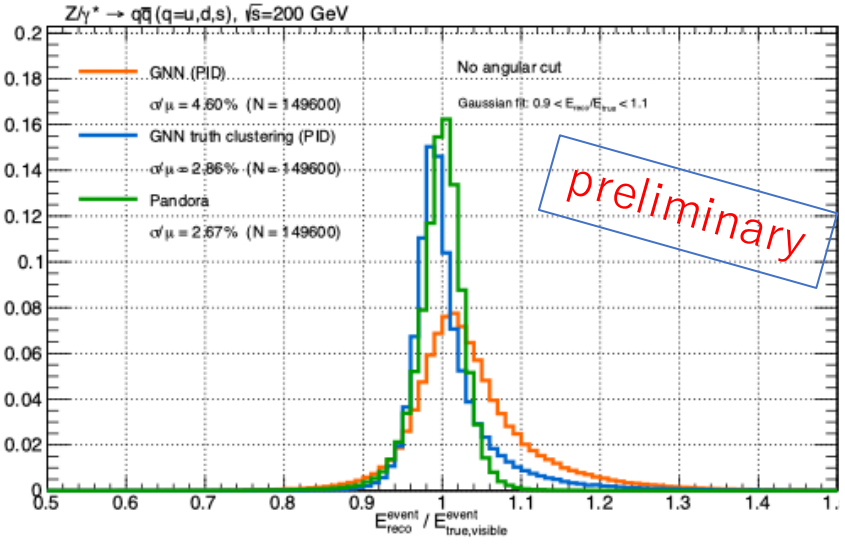


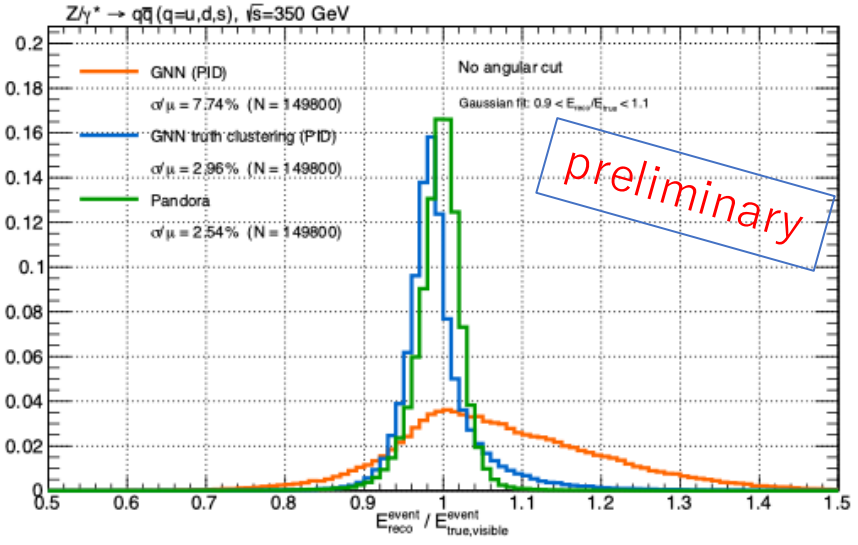
Event energy responce

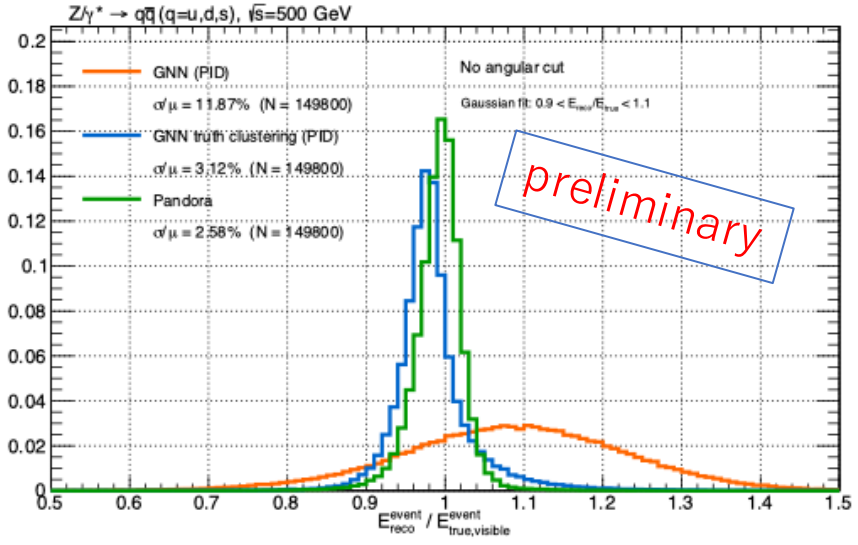
- When evaluating the central peak resolution (σ/μ), GNN performs competitively with PandoraPFA across low-to-mid energies
  - The quoted core widths closely match PandoraPFA at 40 and 91 GeV.
- At 200–500 GeV, the truth-assisted core widths remain approximately 2.9–3.1%.





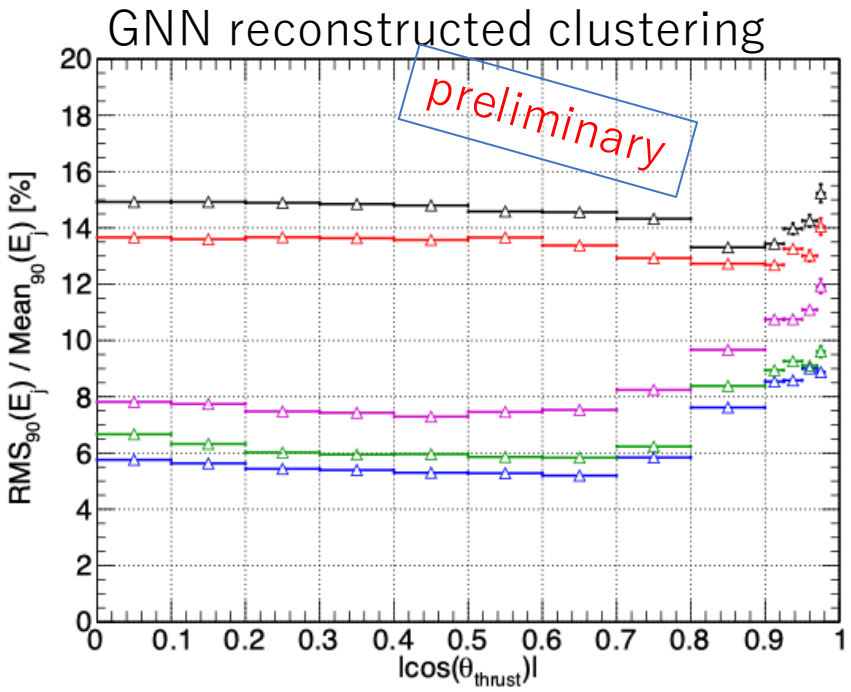


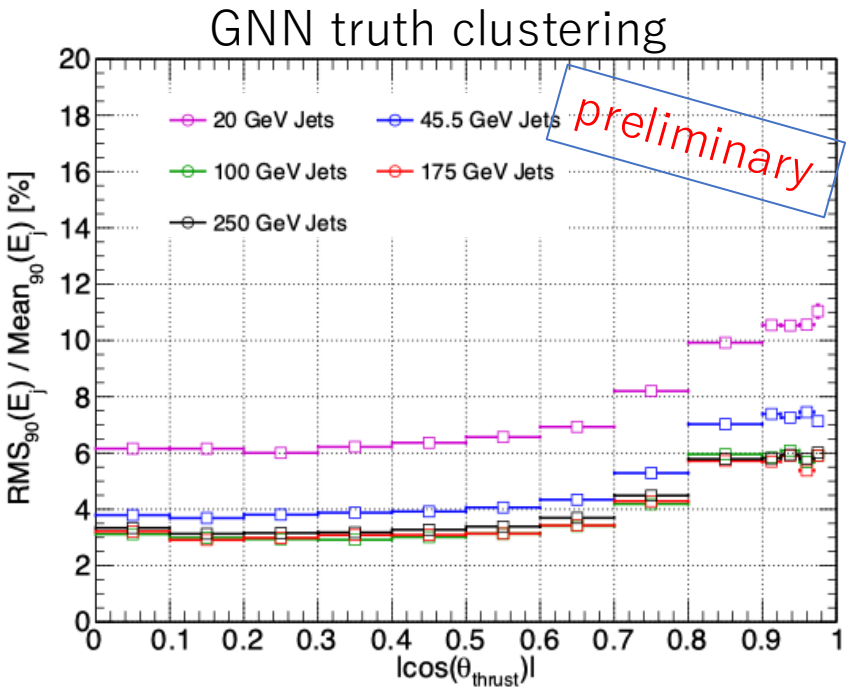


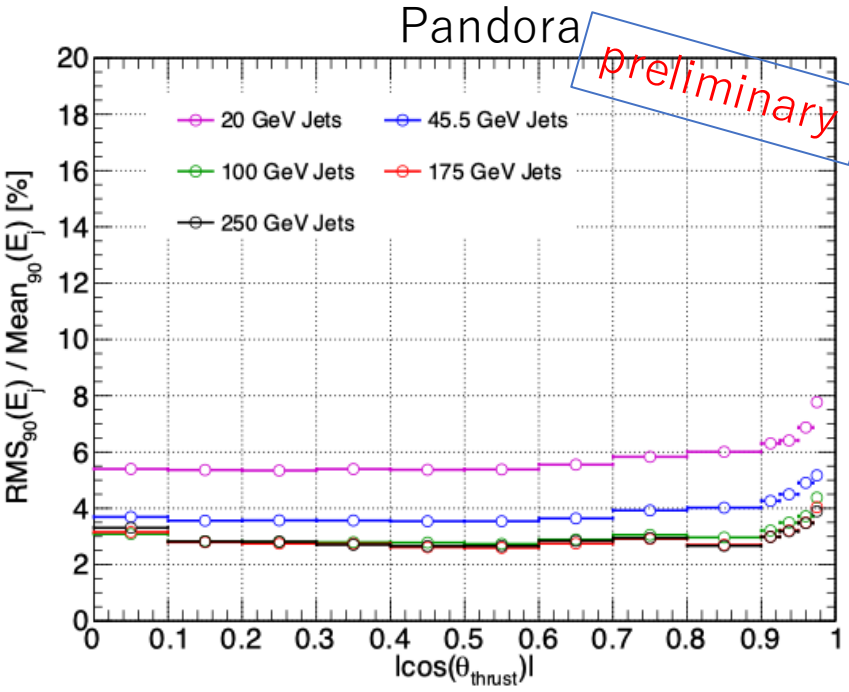


Angular Dependence of Jet Energy Resolution

- Evaluated using standard resolution metric $\frac{\mathrm{RMS_{90}}}{\mathrm{Mean90}}$ vs. jet polar angle $|\cos\theta|$ for direct comparison with PandoraPFA.
- At high energies, truth-assisted clustering gives substantially narrower response widths across the angular range.

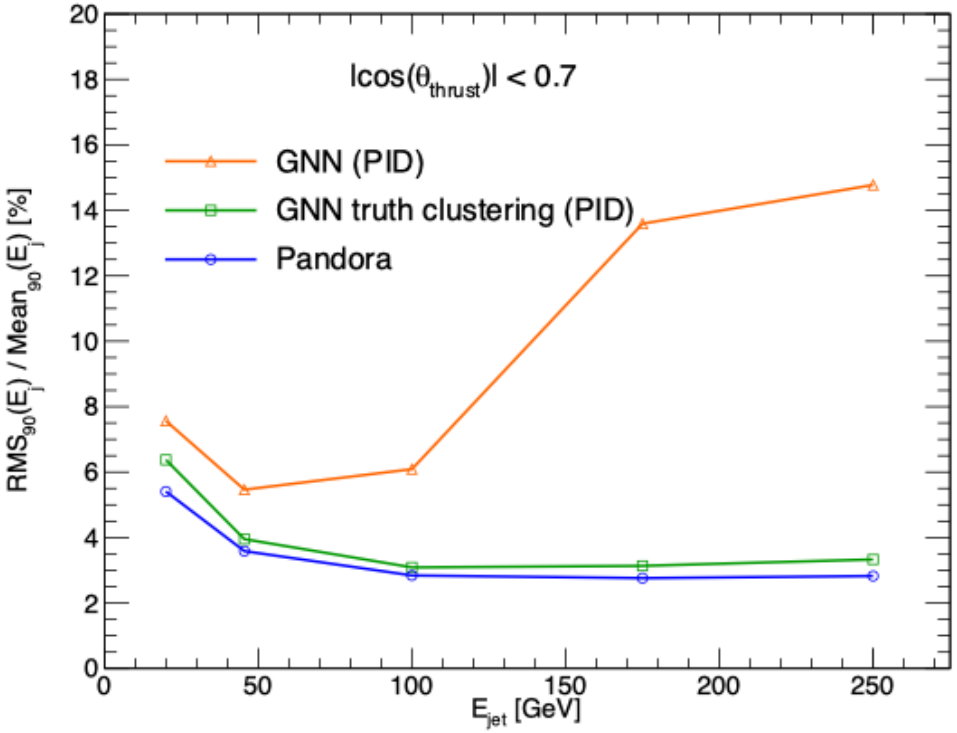


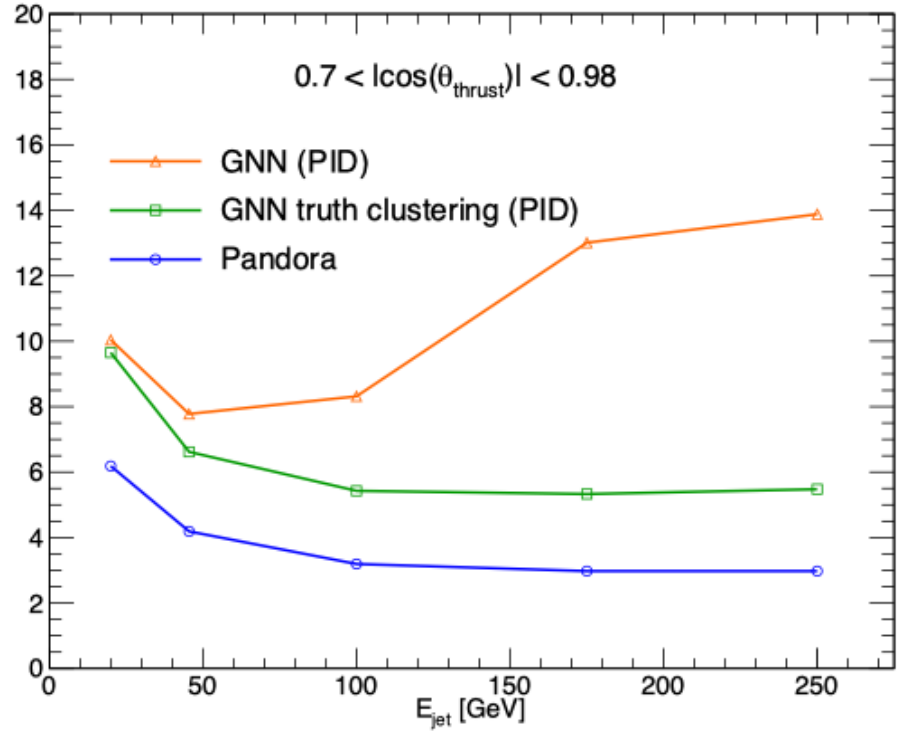


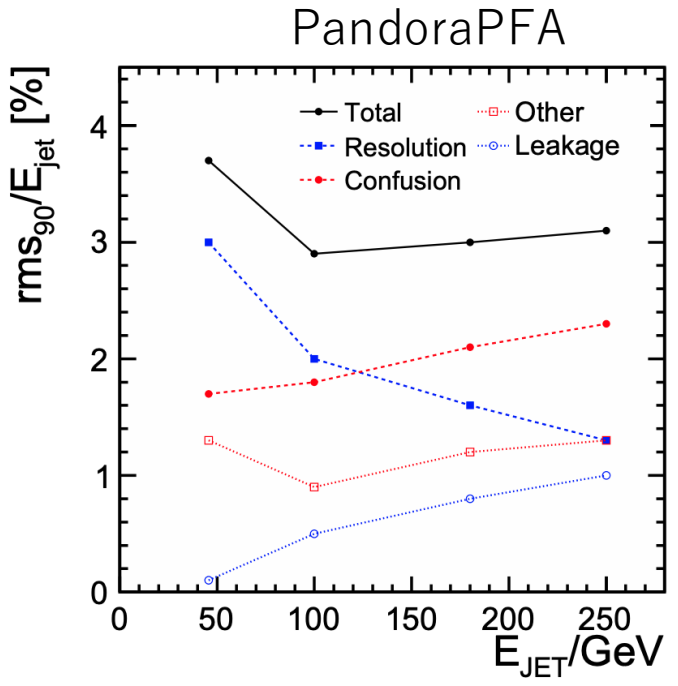


Jet energy resolution

- With truth-assisted clustering, the GNN resolution is comparable to PandoraPFA for central high-energy events.
- Truth-assisted clustering also substantially narrows the high-energy response in the forward region.
- These results provide concrete targets for optimizing cluster assignment and track–cluster association.







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

Summary

- Developing GNN based PFA
  - Demonstrated GNN based clustering and event-energy reconstruction in full ILD simulation over 40–500 GeV.
  - Achieved high clustering efficiency and generally higher purity than PandoraPFA in the displayed comparison.
  - With truth-assisted clustering, the GNN achieves resolution comparable to PandoraPFA for central high-energy events.

- Ongoing work
  - Optimize cluster assignment and track–cluster association, and refine PID-guided energy readout.
  - 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.