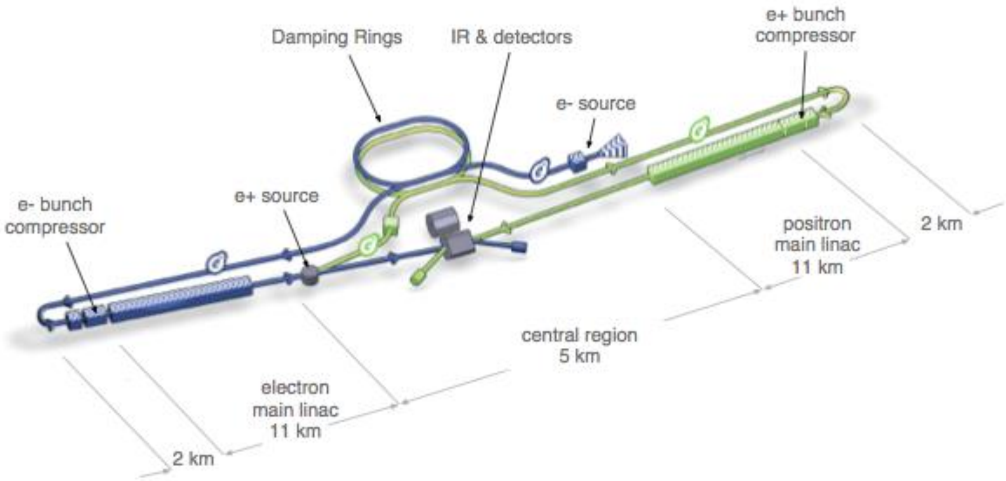
# Particle Flow Reconstruction for the International Linear Collider

## Using Graph Neural Networks and Object Condensation

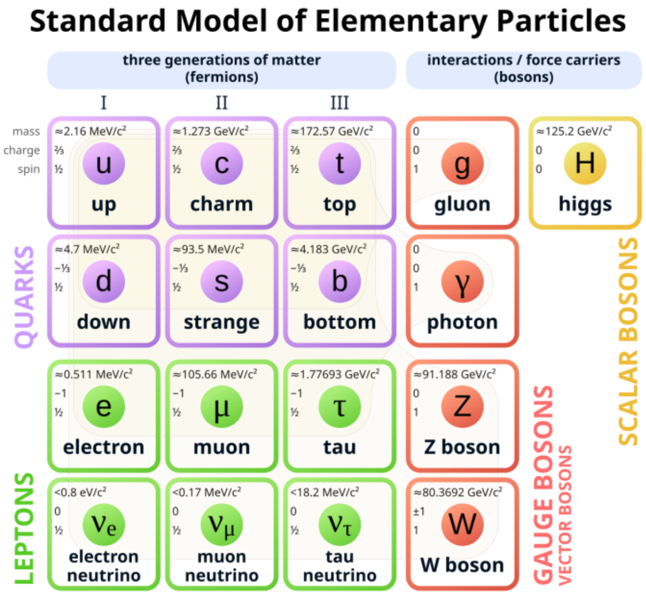
Trinh Cédric
Suehara Taikan, Schaeffer Arpad, Tatsuki Murata
University of Tokyo, ICEPP, École Polytechnique

# International Linear Collider (ILC)

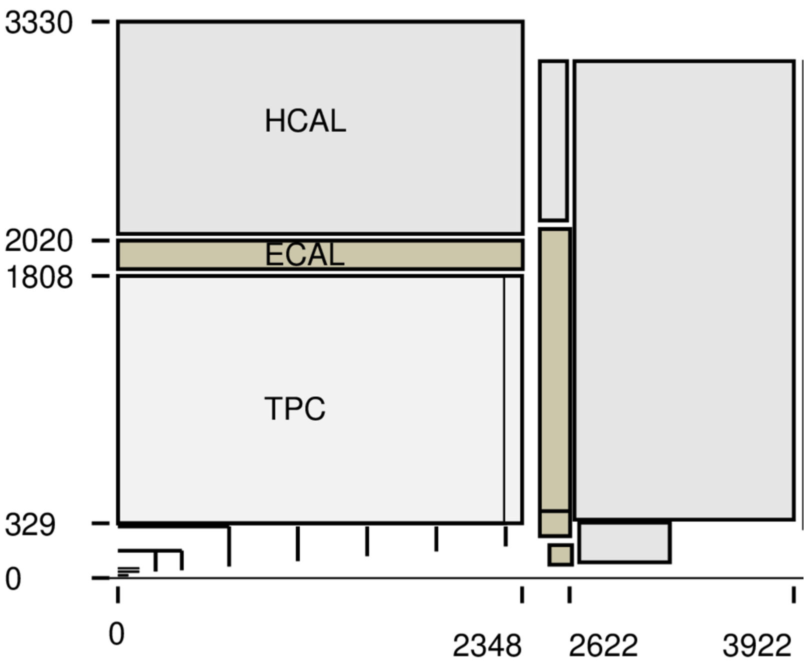
Physics goal: precise measurements of Higgs, Z, W, and top.



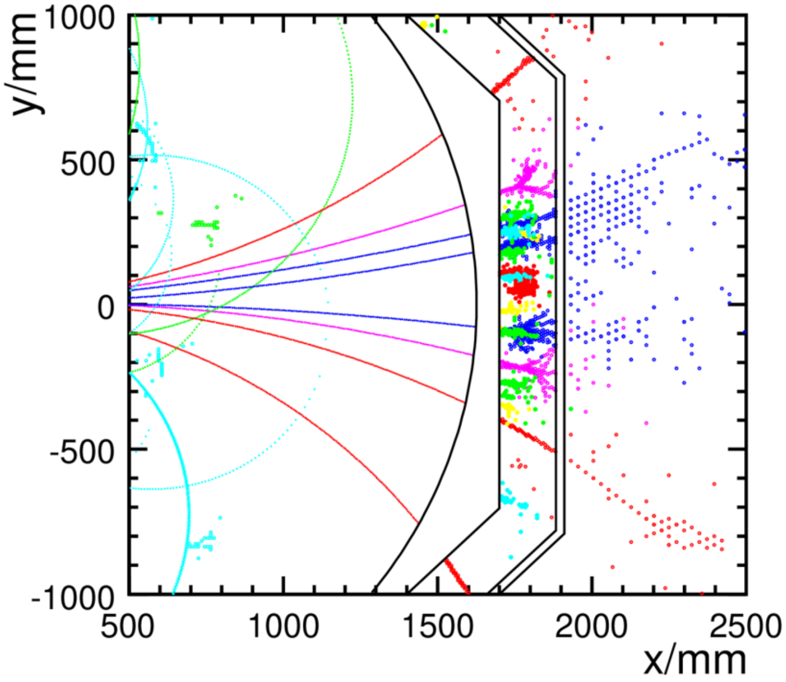

ILC layout



# International Large Detector (ILD)

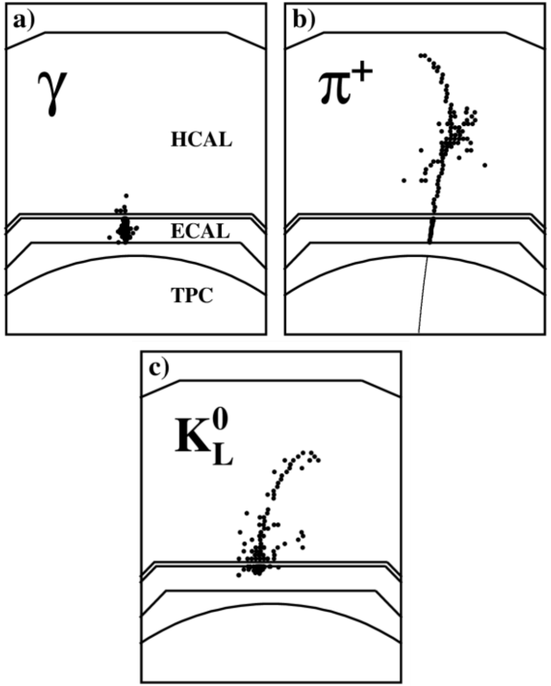



ILD subdetectors

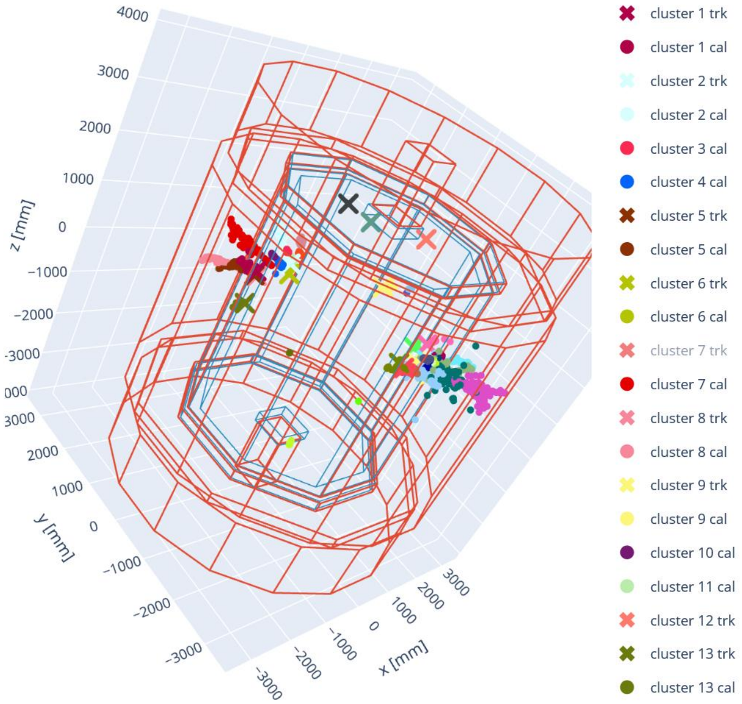



Jet showers in the ILD

# Hit clustering




Detector signature of single particles




Jet event clusters inside the ILD

# PandoraPFA and GNN

## Pandora PFA

- Physical model
- Tracks and calorimeter topology
- Detector-specific rules

↓

Reference baseline for clustering and energy regression

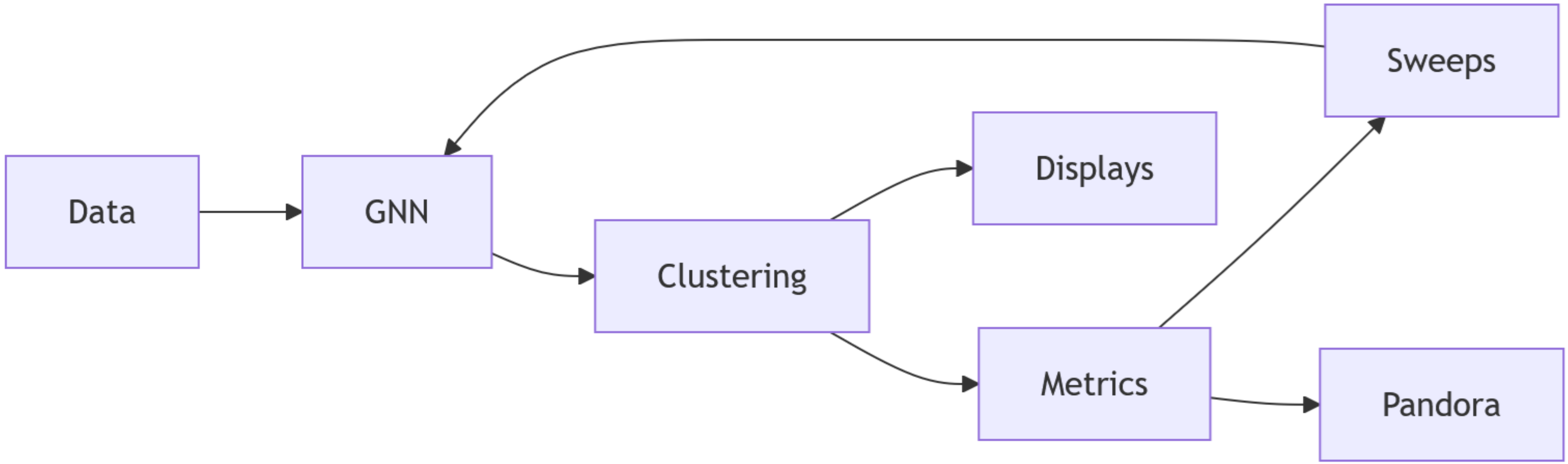
## GNN

- Deep Neural Network
- Reconstruction from detector hits
- Graph representation

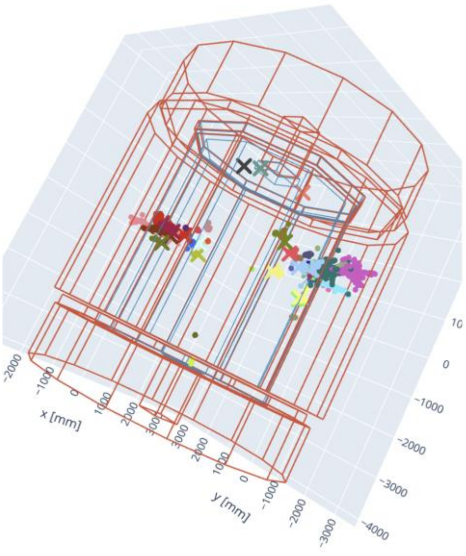
↓

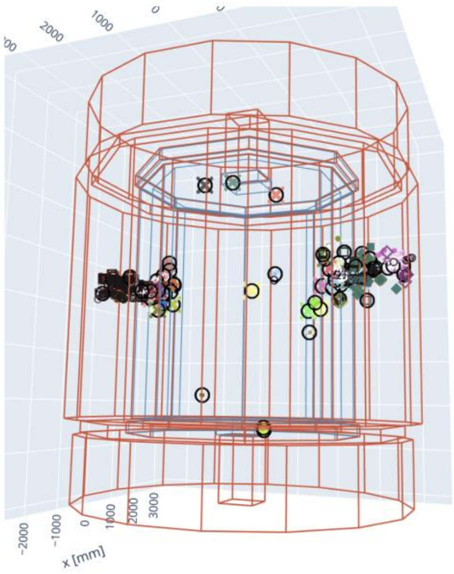
Alternative tested in this work for clustering

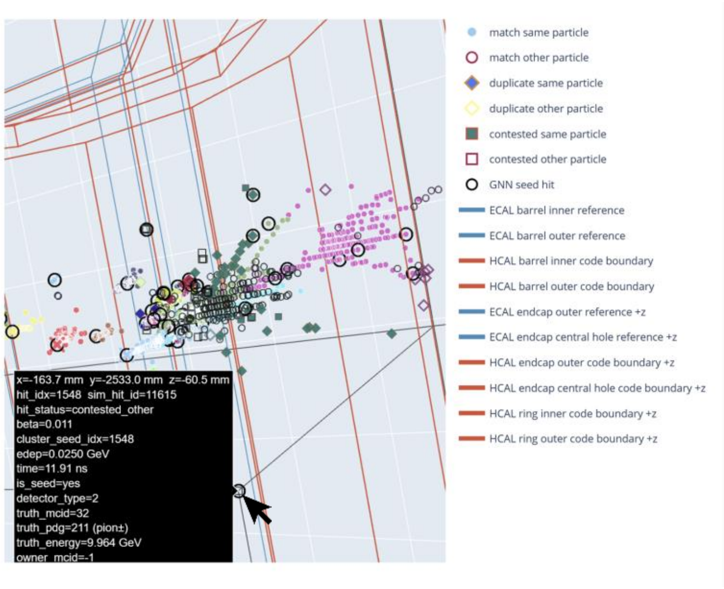
# Objective and Personal Contribution



# Visualization



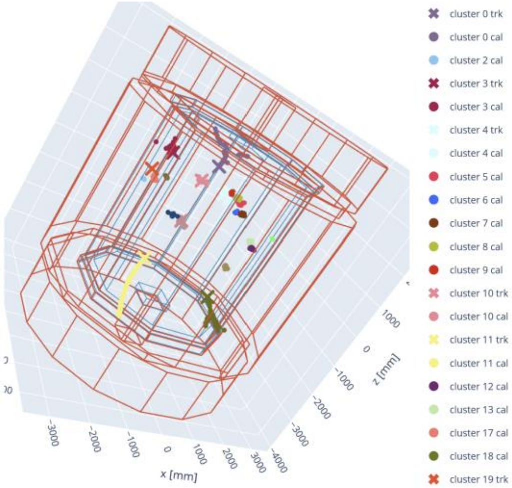

Interactive event display

# Datasets

Split of Train/Validation/Test by 80/10/10

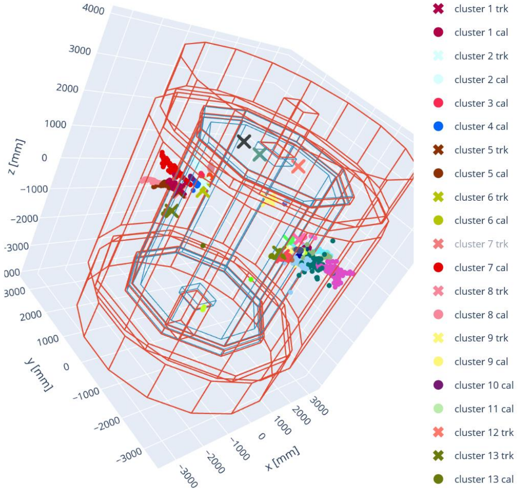
**tau dataset**

- $\tau$ events
- 100.000 events
- ~25 particles and ~1.5k per event

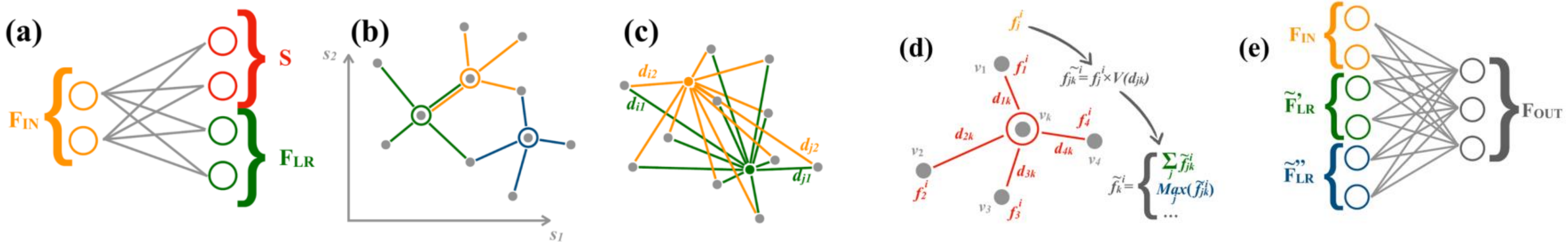



**nnqq dataset**

- e+e- -> $\nu\nu$ qq dijet events
- 120.000 events
- ~50 particles and ~3k hits per event

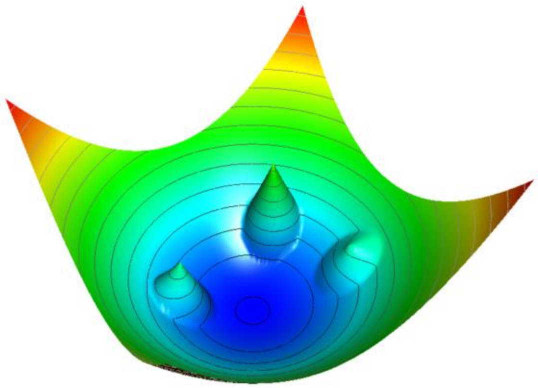


# GNN: Gravnet




(a) Input features [N, D_in] → (b) Learned neighbour space [N, D_space] → (c) k-NN graph → (d) Message propagation [N, D_prop] → (e) Aggregation + output [N, D_out]

# GNN: Object Condensation



**Hit** → (seed score beta, condensation-space coordinate)

**Beta control** = high beta on seeds + low beta on non-seeds

**Seed charge** = increasing function of beta + minimum charge

**Attraction** = Σ(hit charge × own-seed charge × distance to own seed$^2$) / number of hits

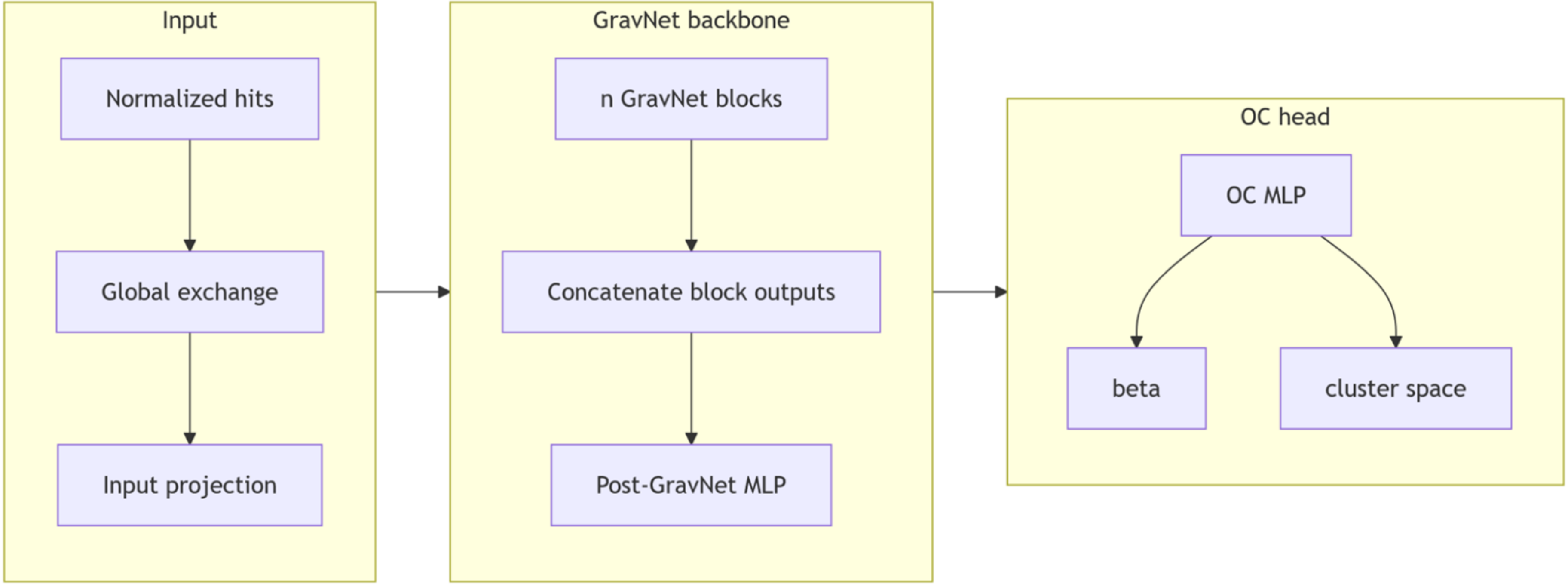
**Repulsion** = Σ(hit charge × other-seed charge × penalty if too close) / number of hits

**Potential** = Attraction + Repulsion

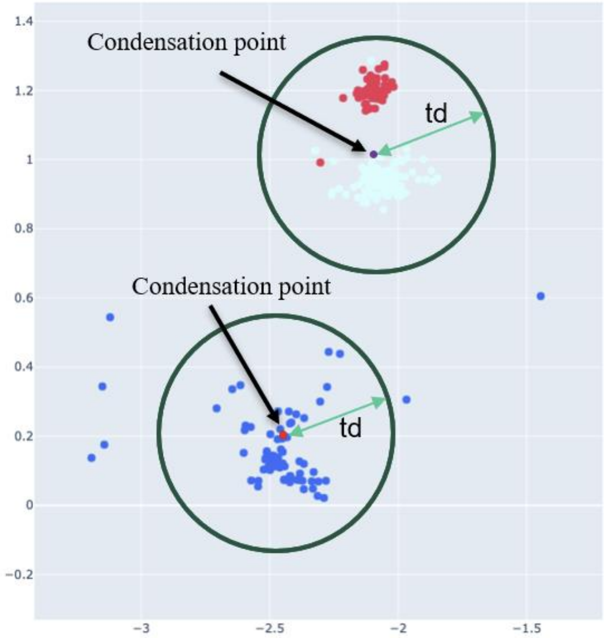
**OC loss** = Potential + Beta control

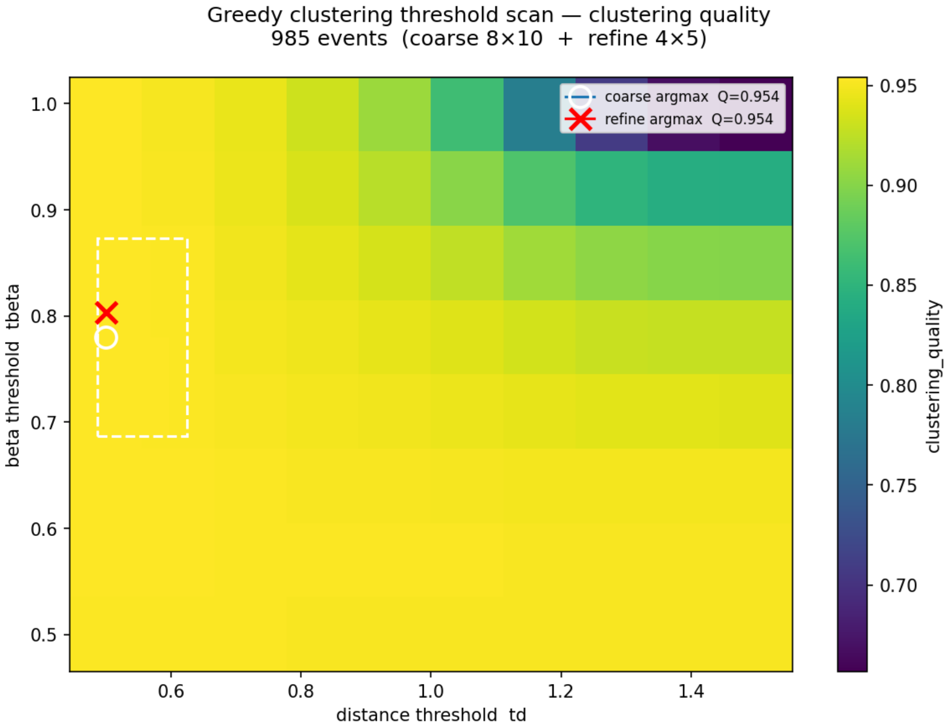
*Note: Annex with detailed formulas available*

# GNN: Pipeline



# Greedy Clustering



# Metrics: OCQuality, ClusteringQuality, EnergyQuality

**SeedQuality** = sqrt(MeanSeedBeta × (1 - MeanNonSeedBeta))

**SpaceQuality** = SeedSeparation / (SeedSeparation + HitCompactness + ε)

**OCQuality** = sqrt(SeedQuality × SpaceQuality)

---

**Efficiency** = RecoveredTrueEnergy / TotalTrueEnergy

**Purity** = CorrectEnergyInMatchedClusters / TotalEnergyInMatchedClusters

**ConfusionQuality** = sqrt((1 - ChargedToNeutralConfusion) × (1 - NeutralToChargedConfusion))

**ClusteringQuality** = $\text{Efficiency}^{0.40} \times \text{Purity}^{0.40} \times \text{ConfusionQuality}^{0.20}$

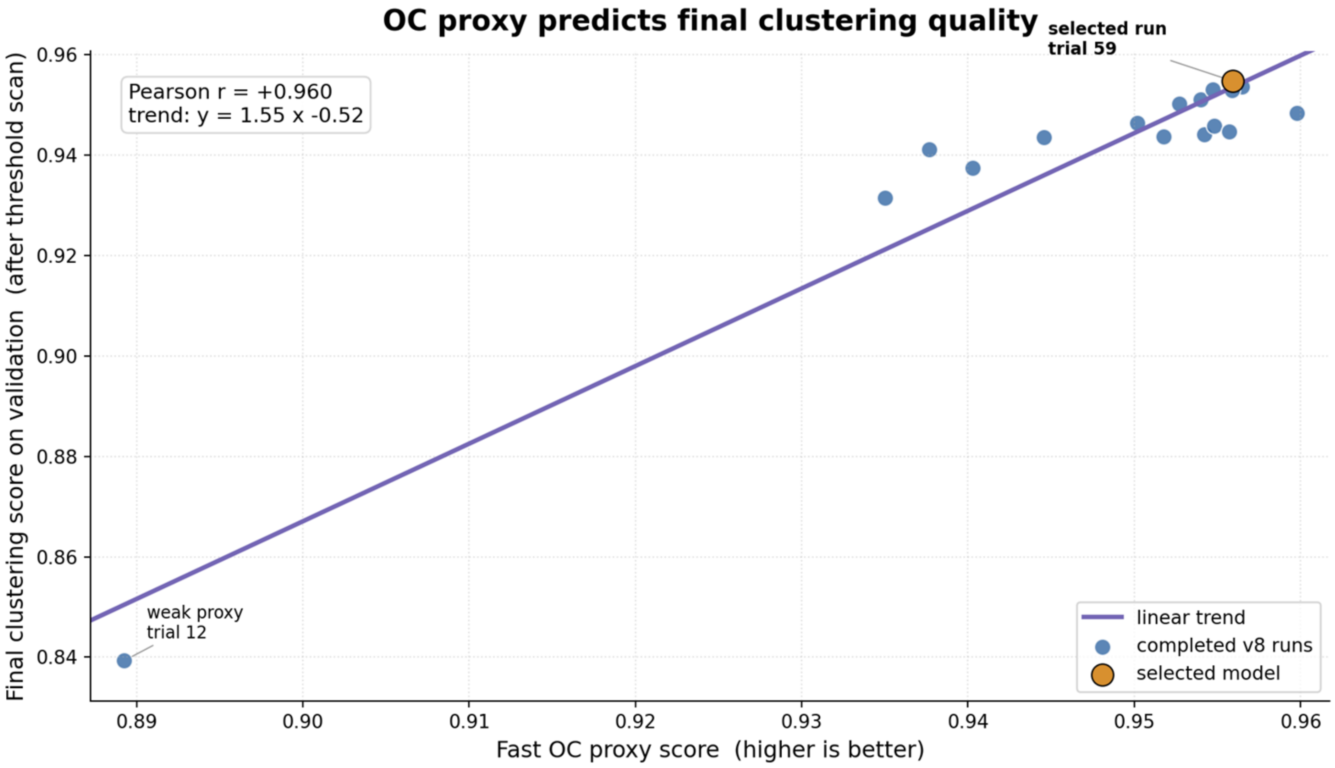
---

**RelativeEnergyRMSE** = $\text{sqrt}(\text{mean}(((\text{PredictedEnergy} - \text{TrueEnergy}) / \max(\text{TrueEnergy}, \varepsilon))^2))$

**AbsoluteEnergyRMSE** = $\text{sqrt}(\text{mean}((\text{PredictedEnergy} - \text{TrueEnergy})^2))$

*Note: Annex with detailed formulas available*

# Proxy Metrics



# Bayesian Hyperparameters Tuning

| Sweep | GPU* time | Data | Main swept axes | Best result | Conclusion |
|---|---|---|---|---|---|
| v1 | ~308h | tau | OC loss modes, qmin, LR, restart period | oc_quality 0.9475 | Pipeline becomes trainable |
| v2 | ~281h | tau | More conservative OC/loss search, OC dim, LR, weight decay | oc_quality 0.9326 | Confirms stability |
| v4 | ~4,034h | nnqq | OC dim, LR, weight decay, qmin, beta stabilization, OC loss modes | clustering_quality 0.8993 | First useful dense-event result |
| v7 | ~5,578h | nnqq | GravNet blocks, k for knn, space/propagate dims, post layers, OC head | clustering_quality 0.9456 | Main architecture jump; better geometry and seed selection |
| v8 | ~10,912h | nnqq | Refined v7 basin: blocks, k for knn, compression, post dim, OC dim, beta/track/background coefficients | clustering_quality 0.9531 | Final selected model |

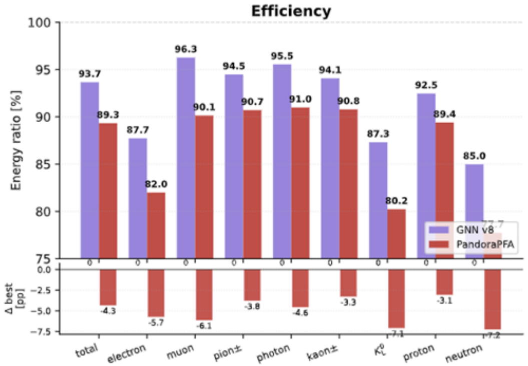
**Note:* *A100-40GB-equivalent order of magnitude*

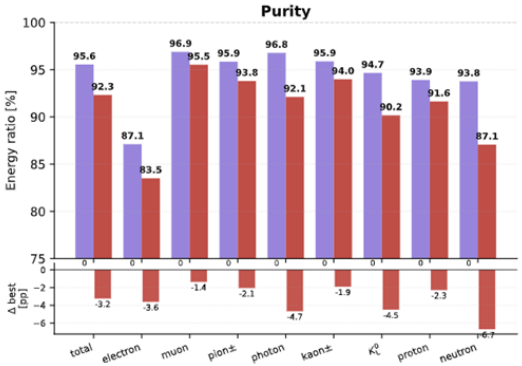
# Main Result on nnqq dataset

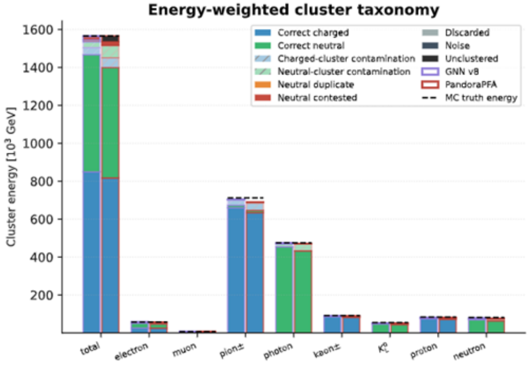
Clustering Quality

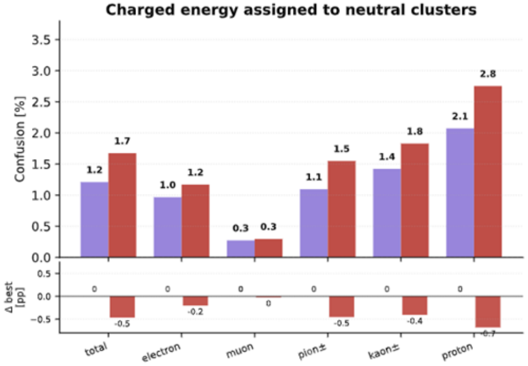
GNN v8: 0.95309

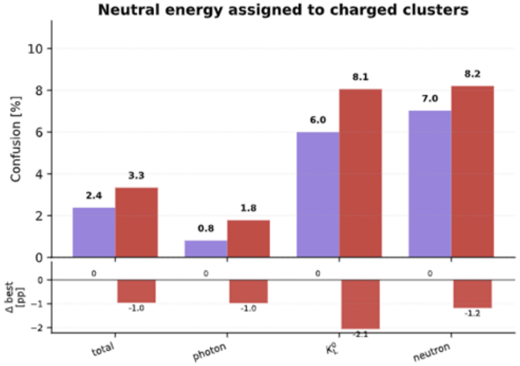
PandoraPFA: 0.92109



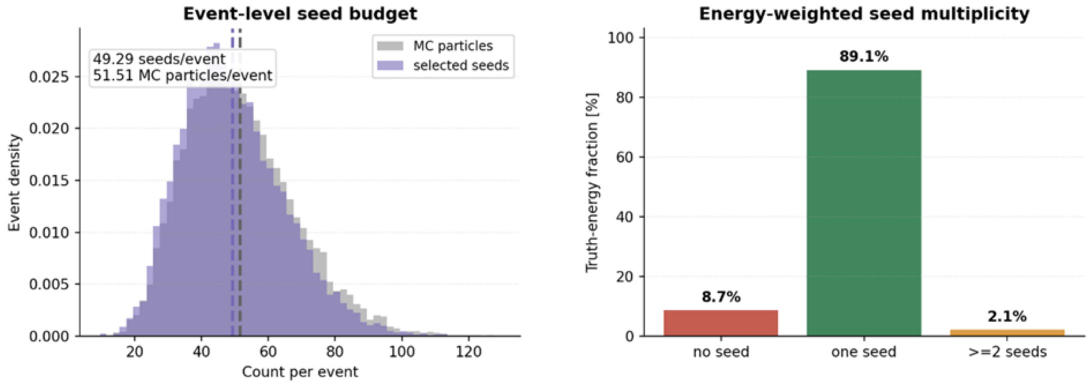



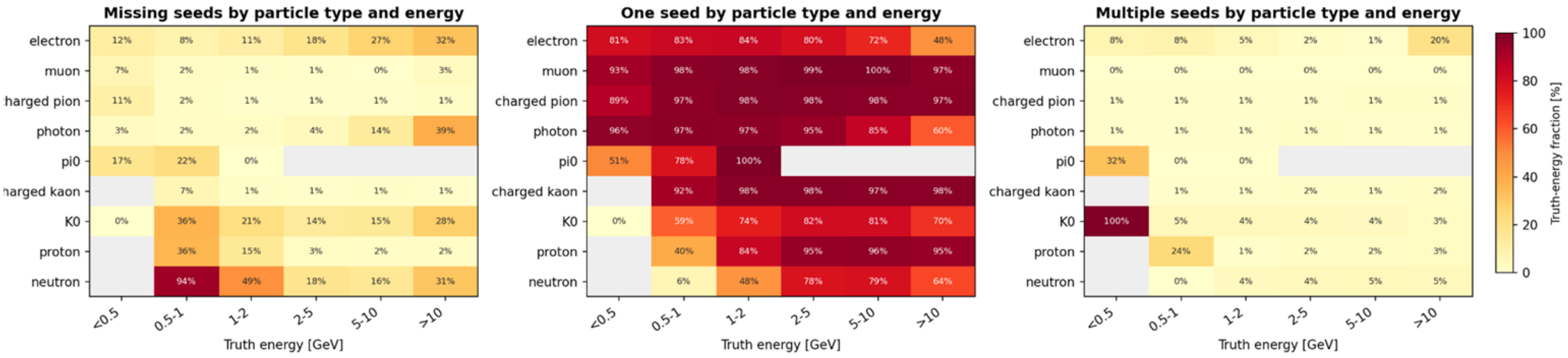




*Note: Annex with detailed sweep analysis available*

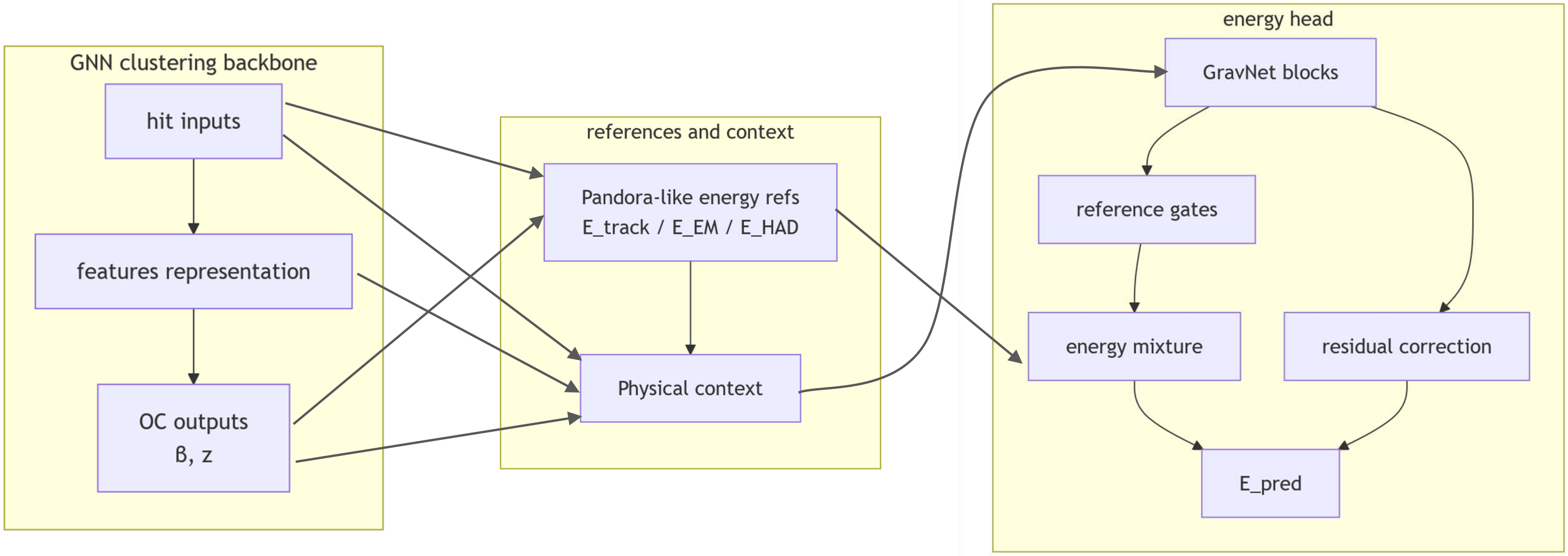
# Seeding



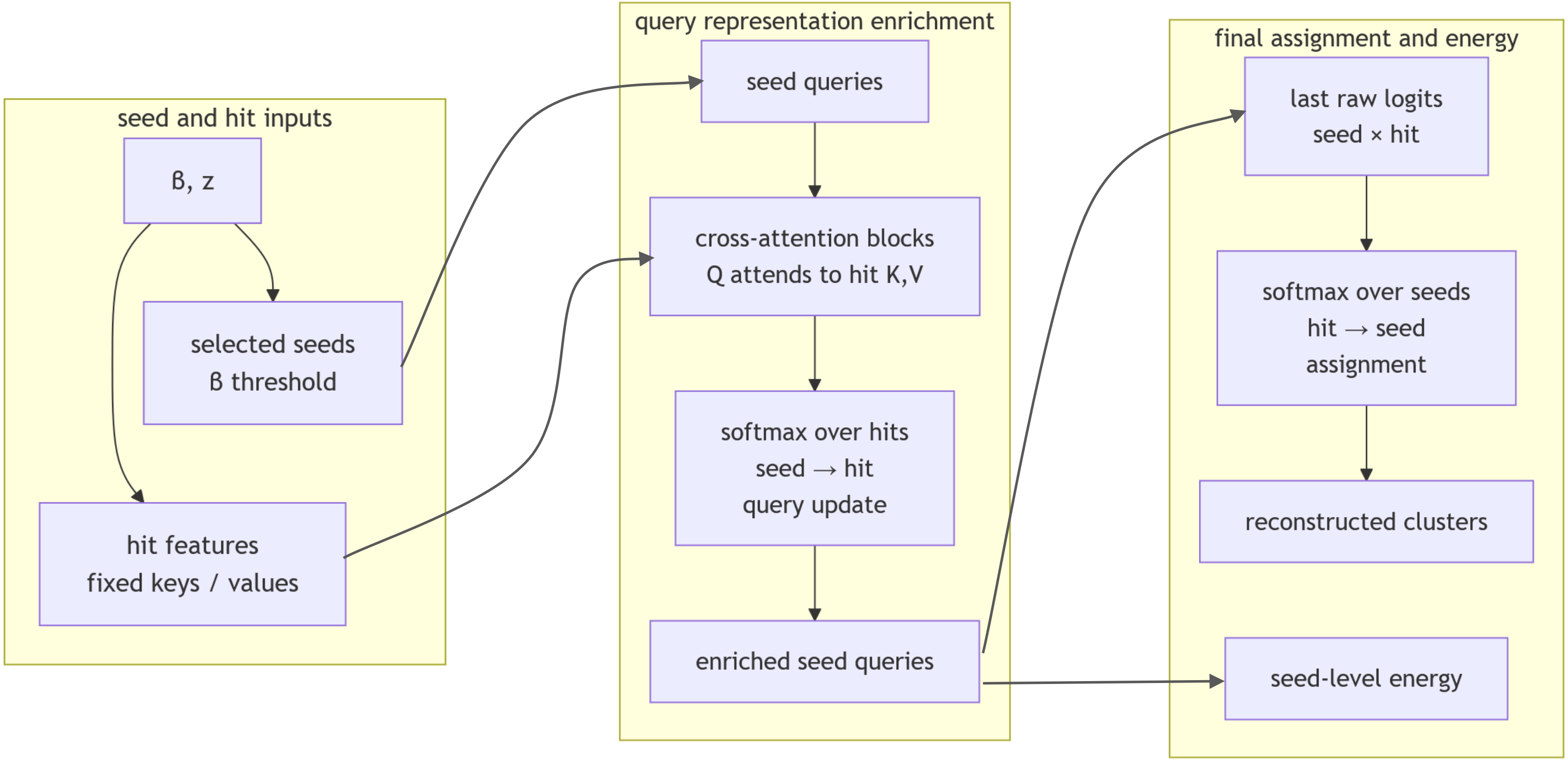




*Note: Seed analysis on the nnqq test dataset for GNN v8*

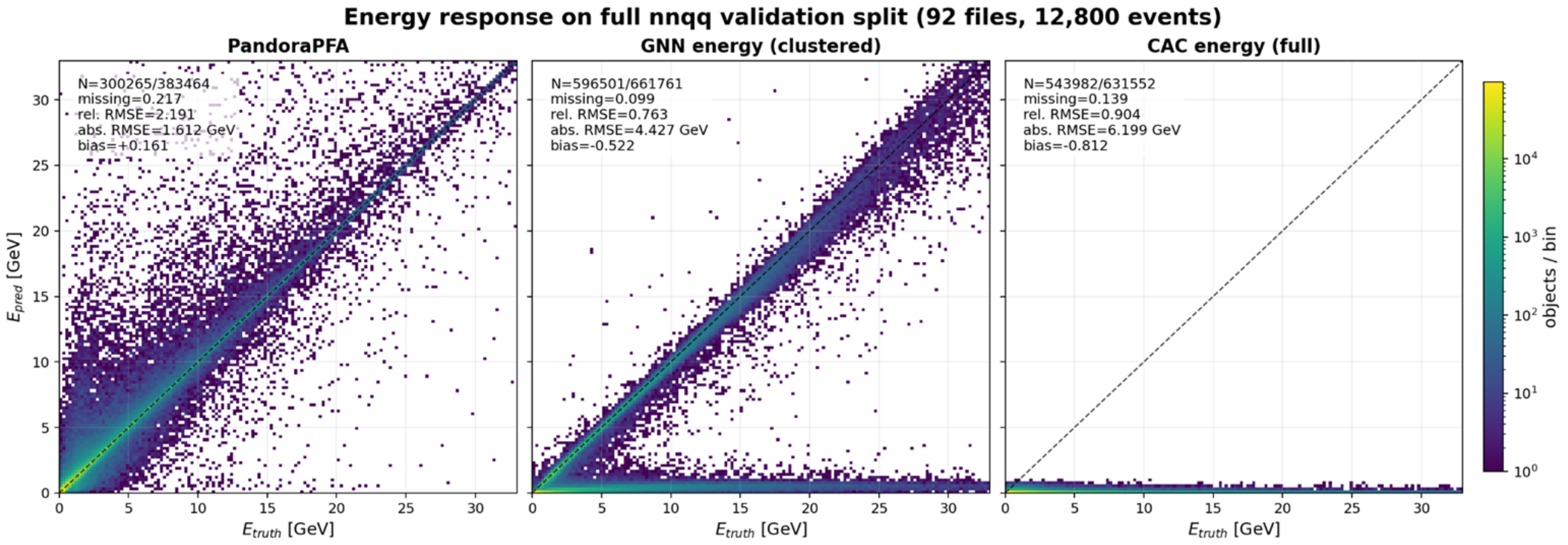
# GNN Energy Regression Extension



# Cross-Attention Clustering (CAC) and Energy Regression



# Energy Regression Preliminary Results



# Conclusion and Perspectives

Contributions

- End-to-end GNN + Object Condensation clustering pipeline.
- Metrics, threshold scans, and interactive event review.
- Robust sweeps with restart, pruning, logs, and Pandora comparison.

Results

- Full nnqq test: clustering quality 0.9531 vs 0.9211 for PandoraPFA.
- Higher efficiency and purity, lower charged-neutral confusion.
- Energy and CAC chains validated technically, but still preliminary.

Perspectives

- Finish energy comparison and seed-variance checks.
- Couple seed selection with CAC-style assignment.
- Move from particle-level diagnostics to jet-energy resolution.

# Annex Object Condensation

| Symbol | Meaning |
|---|---|
| $i$ | hit index |
| $b$ | event index |
| $k$,$\ell$ | truth object / particle indices inside one event |
| $H_b$ | hits in event $b$ |
| $H_k$ | valid hits belonging to object $k$ |
| $O_b$ | valid truth objects in event $b$ |
| $N_b$ | number of hits in event $b$ |
| $K_b$ | number of valid objects in event $b$ |
| $\beta_i$ | seed score predicted for hit $i$ |
| $z_i$ | Object Condensation coordinate of hit $i$ |
| $q_i$ | condensation charge derived from $\beta i$ |
| $\alpha_k$ | selected condensation point for object $k$ |
| $e(i)$ | event containing hit $i$ |
| $t(i)$ | truth object of hit $i$ if it is not noise |

# Annex Object Condensation

$$q_i = \operatorname{arctanh}^2\left(\frac{\operatorname{clip}(\beta_i, 0, 1-10^{-4})}{1.002}\right) + q_{\min} \qquad \alpha_k = \arg\max_{i\in H_k} q_i \qquad L_V = L_{\text{attr}} + L_{\text{rep}}$$

$$L_{\text{attr}} = \frac{1}{|\mathcal{B}|}\sum_{b\in\mathcal{B}}\frac{1}{N_b}\sum_{k\in O_b}\sum_{i\in H_k} q_i q_{\alpha_k}\left\|z_i - z_{\alpha_k}\right\|_2^2 \qquad L_{\text{rep}} = \frac{1}{|\mathcal{B}|}\sum_{b\in\mathcal{B}}\frac{1}{N_b}\sum_{i\in H_b}\sum_{\substack{\ell\in O_b\\ \ell\neq t(i)}} q_i q_{\alpha_\ell}\max\left(0, 1-\left\|z_i - z_{\alpha_\ell}\right\|_2\right)^2$$

$$L_\beta = L_{\beta,\text{seed}} + c_{\text{bg}}L_{\beta,\text{bg}} + c_{\text{track}}L_{\beta,\text{track}}$$

$$L_{\beta,\text{bg}} = \frac{1}{|\mathcal{B}|}\sum_{b\in\mathcal{B}}\frac{1}{N_b^{\text{bg}}}\sum_{\substack{i\in H_b\\ i \text{ valid}\\ i\notin\{\alpha_k\}}}\beta_i \qquad L_{\beta,\text{track}} = \frac{1}{|\mathcal{B}|}\sum_{b\in\mathcal{B}}\frac{1}{K_b}\sum_{\substack{k\in O_b\\ H_k\cap T\neq\varnothing}}\left(1-\beta_{\alpha_k^T}\right), \qquad \alpha_k^T = \arg\max_{i\in H_k\cap T} q_i$$

$$L_{\text{OC}} = c_{\text{OC}}\left(c_V L_V + c_\beta L_\beta\right)$$

# Annex Metrics

| Symbol | Meaning |
|---|---|
| $h$ | one detector hit |
| $E_{\text{dep}}(h)$ | energy deposited by hit $h$ |
| $P$ | one Monte Carlo truth particle |
| $C$ | one reconstructed cluster |
| $H(P)$,$H(C)$ | hits belonging to truth particle $P$ or cluster $C$ |
| $C^*(P)$ | representative reconstructed cluster matched to truth particle $P$ |
| P | set of truth particles |
| C, $\text{C}_{\text{match}}$ | set of kept reconstructed clusters, clusters matched to a truth particle |
| $\beta_i$ | Object Condensation seed score for hit $i$ |
| $z_i$ | learned Object Condensation coordinates for hit $i$ |
| $\alpha(i)$ | truth seed hit associated with hit $i$'s truth particle |
| $D$ | dimension of the learned clustering space |

# Annex Metrics

$$\mathrm{SeedQuality} = \sqrt{\overline{\beta}_{\text{true seeds}}\left(1 - \overline{\beta}_{\text{nonseed hits}}\right)}$$

$$\mathrm{ClusterCompactness} = \frac{1}{N_{\text{signal}}} \sum_{i \in \text{signal}} \frac{\|z_i - z_{\alpha(i)}\|_2}{\sqrt{D}}$$

$$\mathrm{SeedSeparation} = \frac{1}{K} \sum_{k=1}^{K} \min_{\ell \neq k,\ \text{same event}} \frac{\|z_{\alpha_k} - z_{\alpha_\ell}\|_2}{\sqrt{D}}$$

$$\mathrm{SpaceQuality} = \frac{\mathrm{SeedSeparation}}{\mathrm{SeedSeparation} + \mathrm{ClusterCompactness} + \varepsilon}$$

$$\mathrm{OCQuality} = \sqrt{\mathrm{SeedQuality} \times \mathrm{SpaceQuality}}$$

# Annex Metrics

$$W(P,C) = \sum_{h \in H(P) \cap H(C)} E_{\mathrm{dep}}(h), \qquad W(C) = \sum_{h \in H(C)} E_{\mathrm{dep}}(h)$$

$$\mathrm{Efficiency} = \frac{\sum_{P \in \mathcal{P}} W(P, C^*(P))}{\sum_{P \in \mathcal{P}} \sum_{h \in H(P)} E_{\mathrm{dep}}(h)}$$

$$\mathrm{Purity} = \frac{\sum_{P : C^*(P) \neq \emptyset} W(P, C^*(P))}{\sum_{P : C^*(P) \neq \emptyset} W(C^*(P))}$$

$$\mathrm{Confusion}_{\mathrm{charged} \to \mathrm{neutral}} = \frac{\sum_{P \in \mathcal{P}_{\mathrm{charged}}} \sum_{C \in \mathcal{C}_{\mathrm{neutral}}} W(P,C)}{\sum_{P \in \mathcal{P}_{\mathrm{charged}}} \sum_{h \in H(P)} E_{\mathrm{dep}}(h)}$$

$$\mathrm{Confusion}_{\mathrm{neutral} \to \mathrm{charged}} = \frac{\sum_{P \in \mathcal{P}_{\mathrm{neutral}}} \sum_{C \in \mathcal{C}_{\mathrm{charged}}} W(P,C)}{\sum_{P \in \mathcal{P}_{\mathrm{neutral}}} \sum_{h \in H(P)} E_{\mathrm{dep}}(h)}$$

$$\mathrm{ConfusionQuality} = \sqrt{(1 - \mathrm{Confusion}_{\mathrm{charged} \to \mathrm{neutral}})\,(1 - \mathrm{Confusion}_{\mathrm{neutral} \to \mathrm{charged}})}$$

$$\mathrm{ClusteringQuality} = \mathrm{Efficiency}^{0.40} \times \mathrm{Purity}^{0.40} \times \mathrm{ConfusionQuality}^{0.20}$$

$$\mathrm{RelativeEnergyRMSE} = \sqrt{\frac{1}{M} \sum_{j=1}^{M} \left( \frac{E_j^{\mathrm{pred}} - E_j^{\mathrm{true}}}{\max(E_j^{\mathrm{true}}, \varepsilon)} \right)^2}$$

$$\mathrm{AbsoluteEnergyRMSE} = \sqrt{\frac{1}{M} \sum_{j=1}^{M} \left( E_j^{\mathrm{pred}} - E_j^{\mathrm{true}} \right)^2}$$

# Annex Sweep

| Family | Hyperparameters | Viable* values / range | Practical reading |
|---|---|---|---|
| Input | energy mode, projection | edep dominant; mip_response possible; projection 96 - 160 | edep safest; capacity is flexible. |
| GravNet depth | num_blocks | 6-8 | Useful, not monotonic. |
| GravNet graph | k, space_dim, propagate_dim | k 32 - 96; space_dim 6 - 8; propagate_dim 48 - 64 | Dense events need larger neighbourhoods. |
| GravNet output | compression, post hidden, post layers | compression 32 - 96; hidden 128 - 192; layers 4 - 6 | Moderate width is enough. |
| OC head | MLP depth, clustering-space dim | 4 - 6 layers; D_z 9 - 14 | Flexible after a strong backbone. |
| OC beta scale | qmin, beta.coef | qmin 0.1364 - 0.2368<br>beta coef 0.608 - 1.130 | Broad stable basin. |
| OC seed control | background suppression, track seed, compactness | background 0.5 - 2.0; track 0.25/0.5/1.0; compactness 0.75 - 1.5 | Seed control is critical. |
| Optimisation | learning_rate.max, weight_decay | LR $1.52 \cdot 10^{-4}$ - $6.40 \cdot 10^{-4}$<br>WD $1.63 \cdot 10^{-4}$ - $8.89 \cdot 10^{-4}$ | Avoid extreme LR. |

**Note: values observed nnqq sweep that reached validation clustering_quality >= 0.94*

# Annex Sweep

| Family | Field | Exact value |
|---|---|---|
| Data | dataset | nnqq |
| Inputs | enabled features | position, deposited energy, detector type, time-of-flight |
| Inputs | energy mode, normalization, global exchange | edep, batchnorm, mean/min/max |
| Input projection | input_projection_dim | 160 |
| GravNet | blocks, k, space_dim, propagate_dim | 7, 96, 6, 64 |
| GravNet | output dim, compression, post hidden | 128, 96, 128 |
| GravNet / post MLP | normalization and activations | GravNet batchnorm + tanh; post MLP batchnorm + relu |
| Post-GravNet | layers, hidden dim | 5, 128 |
| OC head | layers, hidden dim, output dim | 4, 64, $D_z$=10 |
| Physics branch | status | disabled for the clustering model |

# Annex Sweep

| Family | Field | Exact value |
|---|---|---|
| Training | epochs, batch size, early stopping | 500, 48, patience 60 |
| Optimisation | max LR, min LR ratio, weight decay | $3.9829 \cdot 10^{-4}$; $10^{-4}$: $8.2405 \cdot 10^{-4}$ |
| Optimisation | dropout, gradient clip | 0.0, 1.0 |
| OC loss | global coefficient, potential coefficient | 1.0, 1.0 |
| OC potential | attractive, repulsive | paper, paper |
| OC beta | coefficient, epoch start, qmin | 0.782359, 6, 0.167863 |
| OC beta | stabilization, forced track point | soft_q_scaling, true |
| OC seed | mode, kernel, compactness, log-beta | positive_short_range, 20.0, 1.0, 0.2 |
| OC seed control | background suppression, track seed | 2.0, 0.5 from epoch 6 |
| Scheduler | cyclic restarts | cosineReduce, period 30, t_mult=1.1, decay 0.7 |