Overview: Deep learning based calorimeter clustering and PFA

Fangyi Guo, IHEP

CEPC International Workshop, 2023 EU Edition

University of Edinburgh

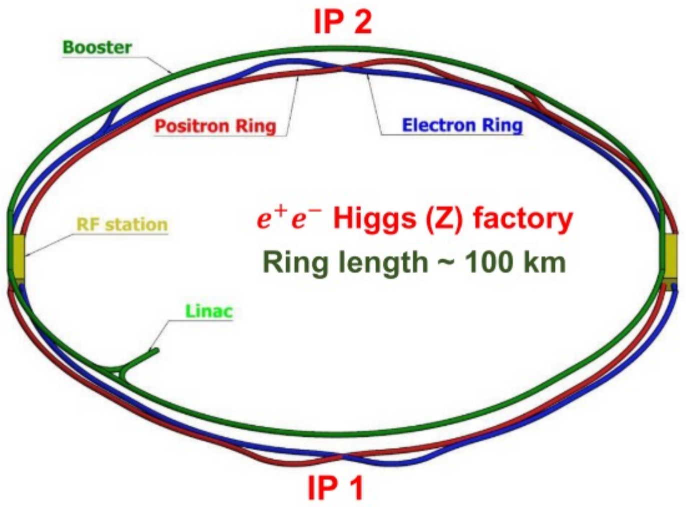
3-6 July, 2023

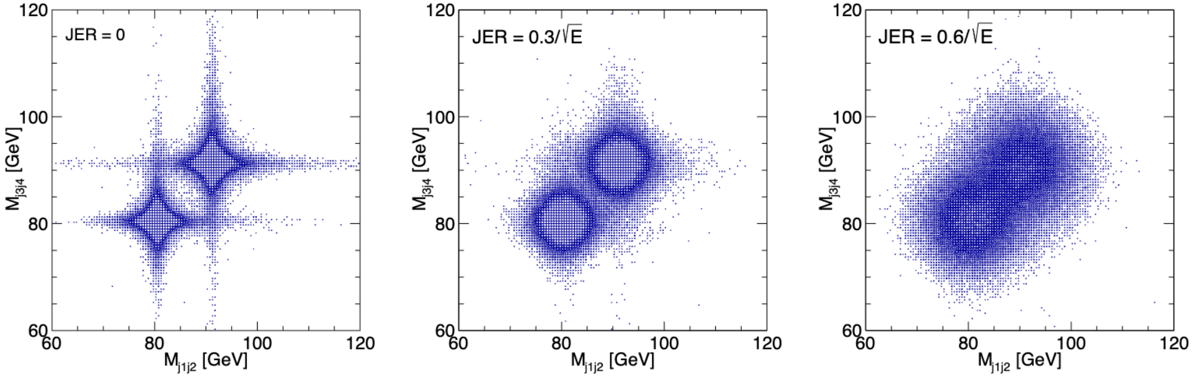
中国科学院高能物理研究所
Institute of High Energy Physics Chinese Academy of Sciences

Introduction

- **CEPC: future lepton collision experiment**
  - e^+e^- collider, $\sqrt{s} = 91\sim360$ GeV within 100 km turnel.
  - Physics target:
    - Z mode: EW, flavor factory for b, c, τ, QCD.
    - W threshold: EW, W mass …
    - Higgs mode: Higgs precise measurement, new physics …
    - top mode: $t\bar{t}$ physics e.g. top mass, $\alpha_S(m_{top})$…
  - Detector requirement:
    - Jet energy resolution $\sigma/E < 30\%/\sqrt{E}$ (For separation of $W/Z \rightarrow q\bar{q}$ process.)

    ⇨ Particle flow approach.

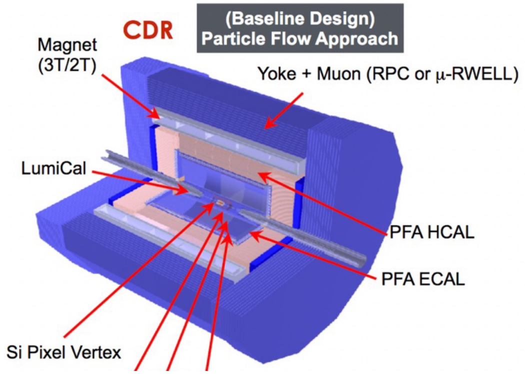


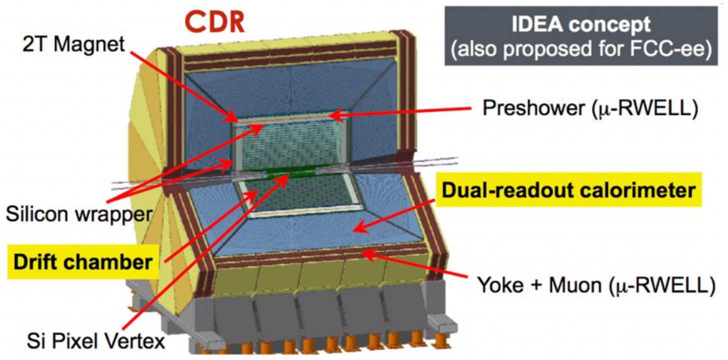


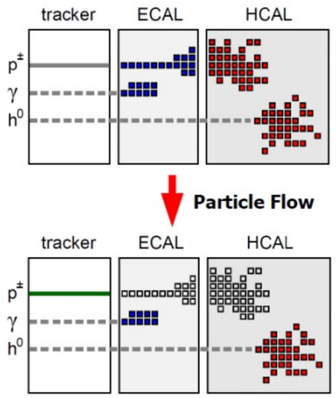
Jet E res.	W/Z sep
perfect	3.1 σ
2%	2.9 σ
3%	2.6 σ
4%	2.3 σ
5%	2.0 σ
10%	1.1 σ

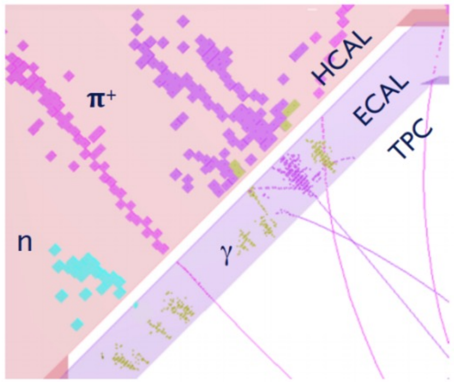
Introduction

- **Particle flow approach: principle**
  - Measure the jet by it’s components: 60% charged particles, 30% photons, 10% neutral hadrons.
  - Final resolution: $\sigma_{Jet} = \sqrt{\sigma_{track}^2 + \sigma_{EM}^2 + \sigma_{Had}^2 + \sigma_{confusion}^2}$.
  - Requirement: Hardware + software
    - Distinguish showers in calorimeter ➡ high granularity ECAL/HCAL.
    - Minimize transverse spread of EM shower ➡ small Moliere radius R_M ➡ SiW sampling ECAL in ILD.
    - Separate EM and Hadronic showers longitudinally ➡ large λ_I / X_0 ratio.
    - Software: a novel pattern recognition algorithm.



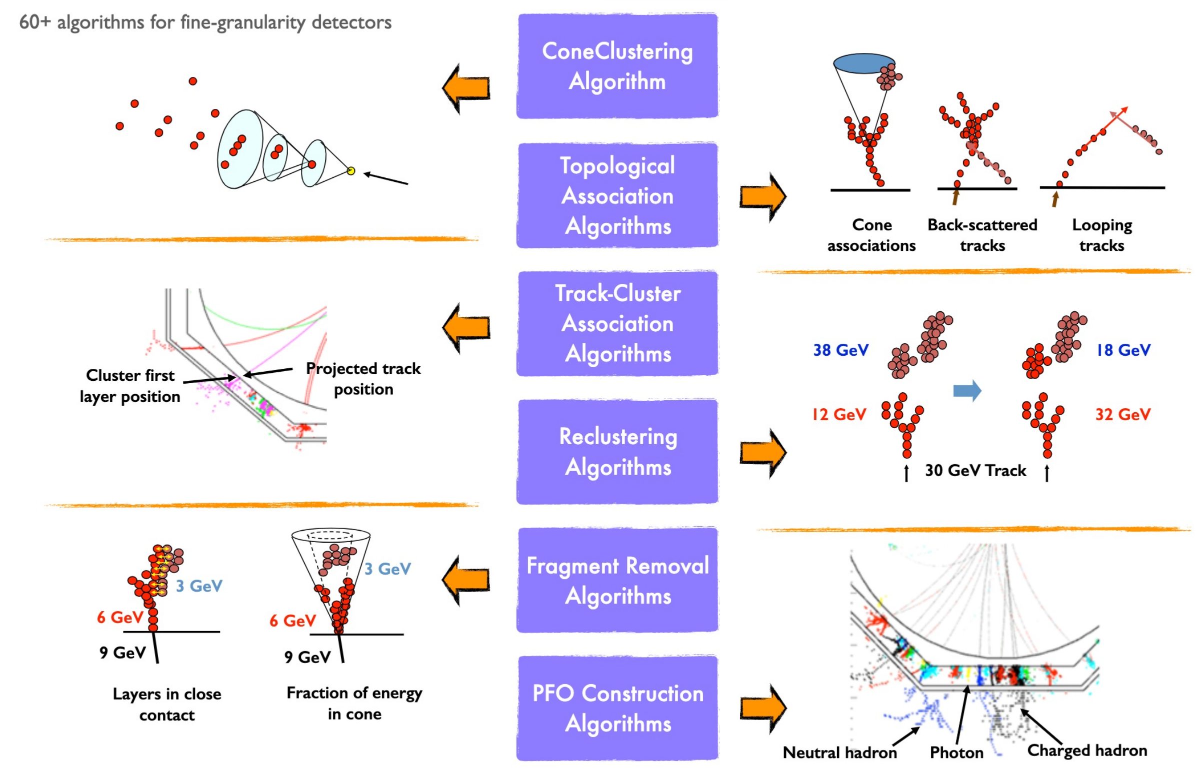


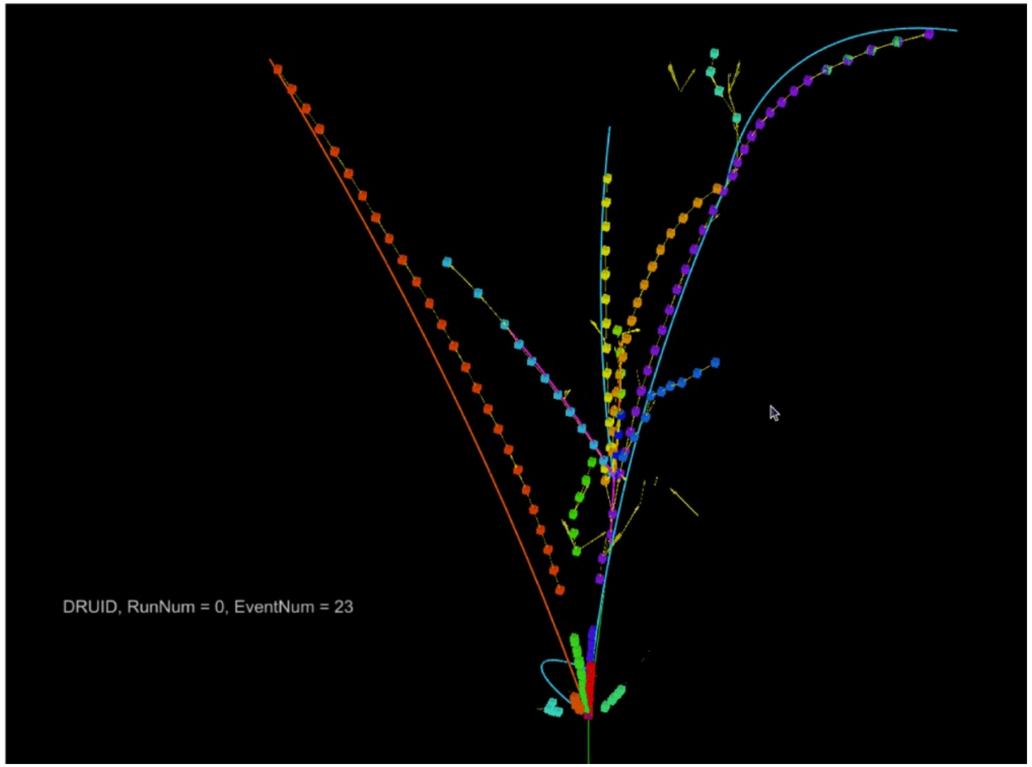




Introduction

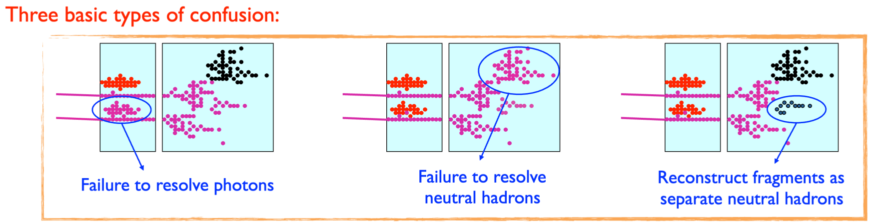
- **Particle flow approach: PandoraPFA and ArborPFA**
  - PandoraPFA: hand-tuned algorithms for clustering. **Artificial recognition.**
  - ArborPFA: arbor-structure of shower development.

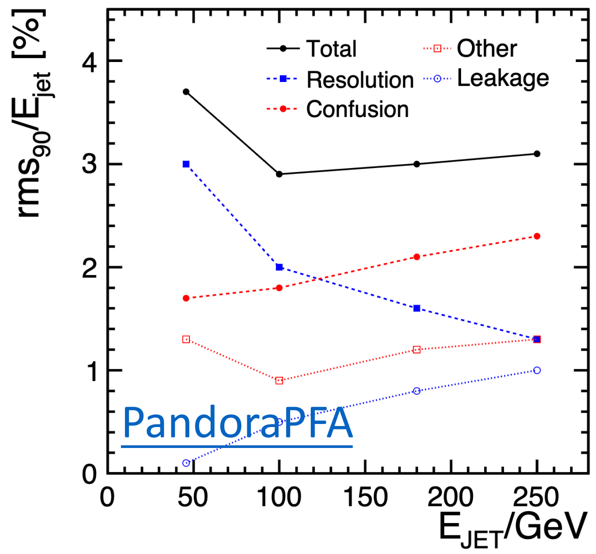


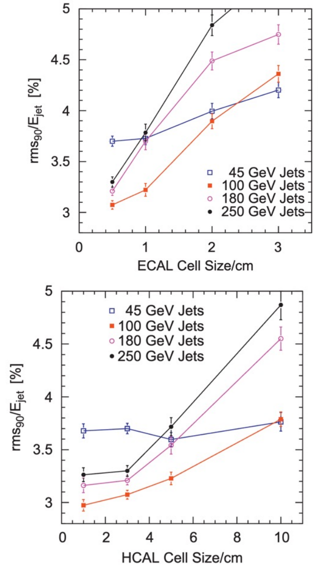


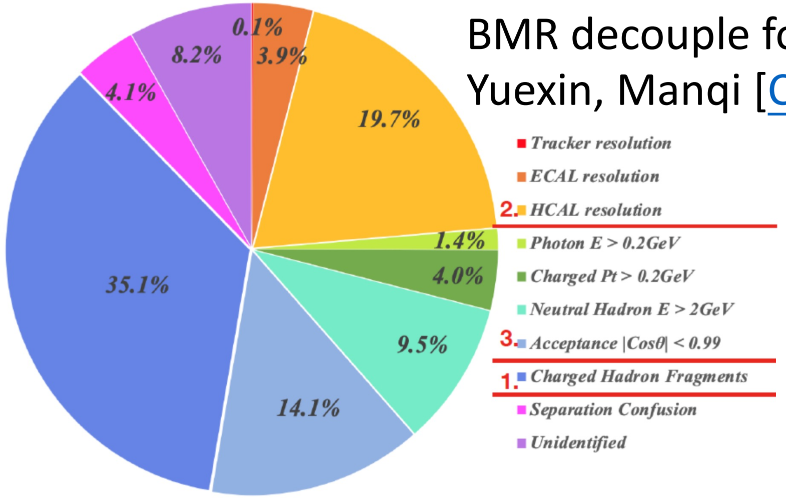
Introduction

- **Particle flow approach: PandoraPFA and ArborPFA**
  - PandoraPFA and ArborPFA are trying to demonstrate the dominant contribution to σ_{jet};
  - And optimize the performance w.r.t. the detector.



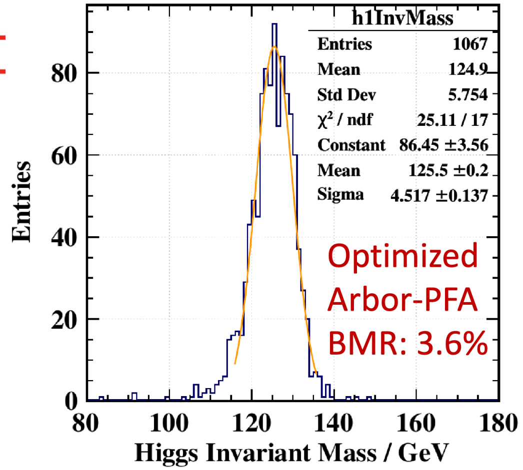






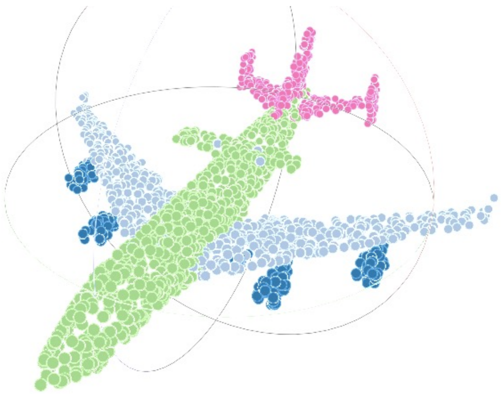
BMR decouple for ArborPFA, Yuexin, Manqi [CEPCWS 2019]

Optimized Arbor for crystal ECAL Baohua et.al [ICHEP 2022]

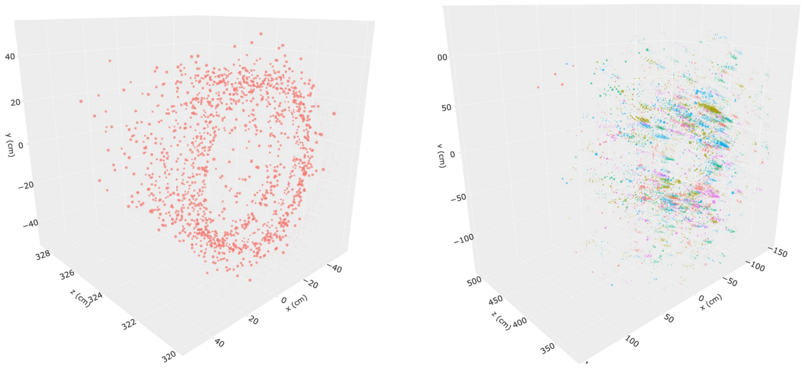


Introduction

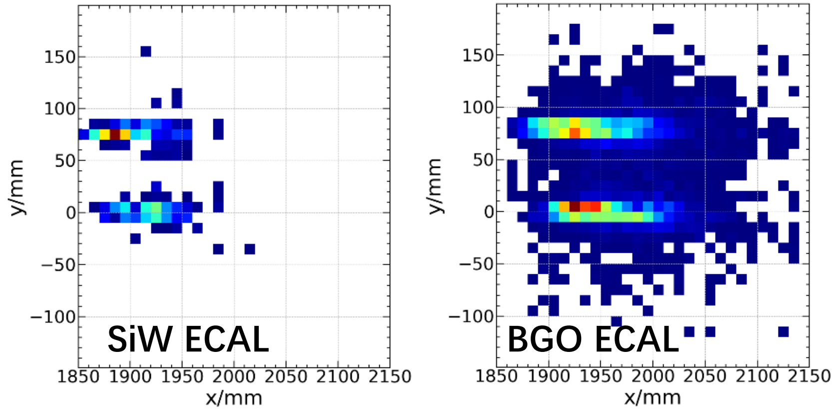
- **Particle flow for the future collider:**
  - 2020s is the AI era: **point cloud**
    - HG calorimeter hits are naturally point clouds: spatial + features (energy + time).
    - Several mature models can be used: PointNet, DGCNN, GravNet, etc.
    - Facing the challenge of large dataset: O(10 k) hits / event.
  - What we want from AI:
    - A general PFA: lower confusion, better performance.
    - Easy migrated model: fast reconstruction for detector design & optimization. (e.g. CEPC 4th crystal ECAL)



Point cloud segmentation in DGCNN



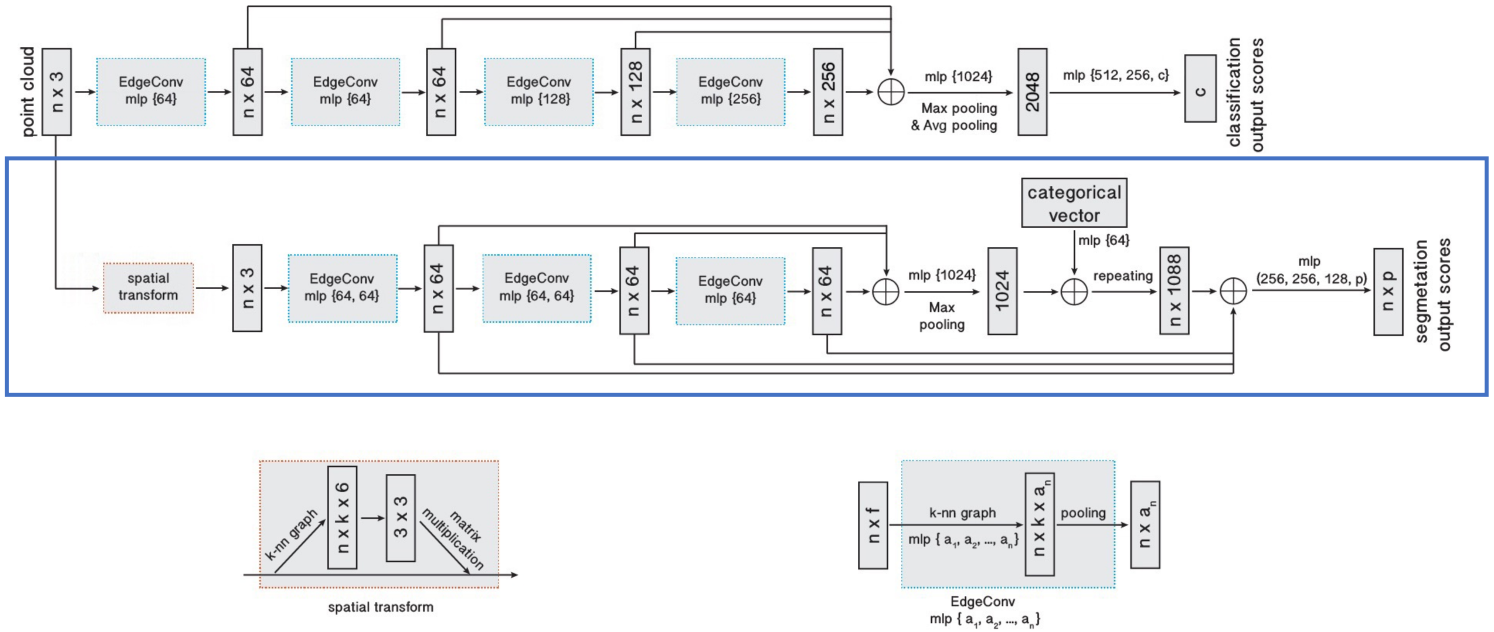
CMS HGCAL clustering with GravNet

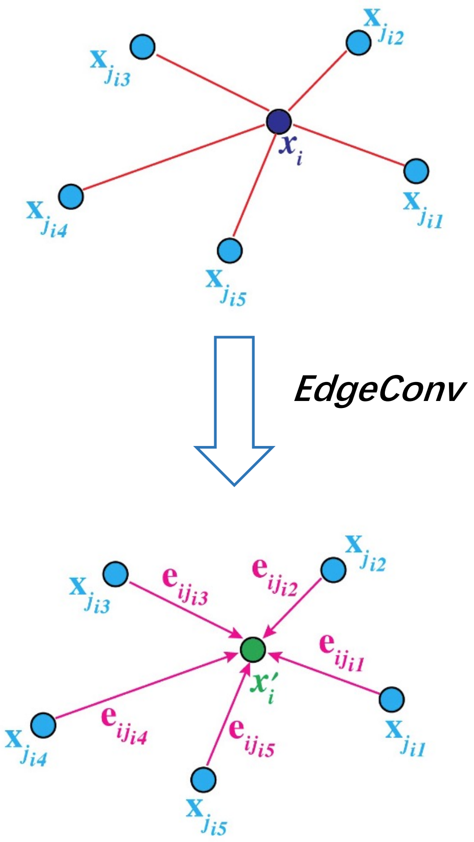


EM showers in SiW and crystal ECAL

Deep learning clustering

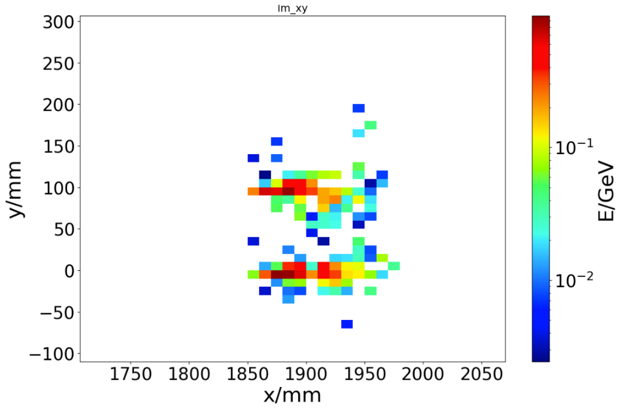
- **Model: Dynamic graphic CNN (DGCNN)**
  - CNN-based model for graphic dataset.
  - Proposed the ***EdgeConv*** to handle the graph structure.
  - Is able to handle classification and **segmentation** tasks.
  - Application: ParticleNet for c-tagging in CMS.

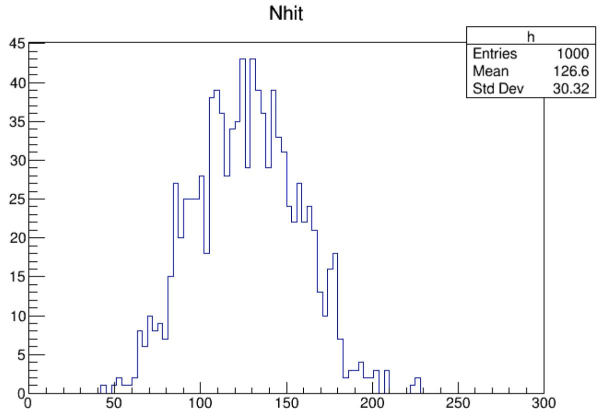


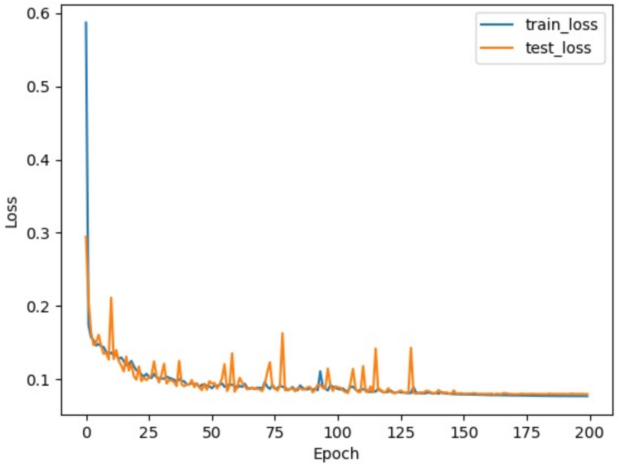


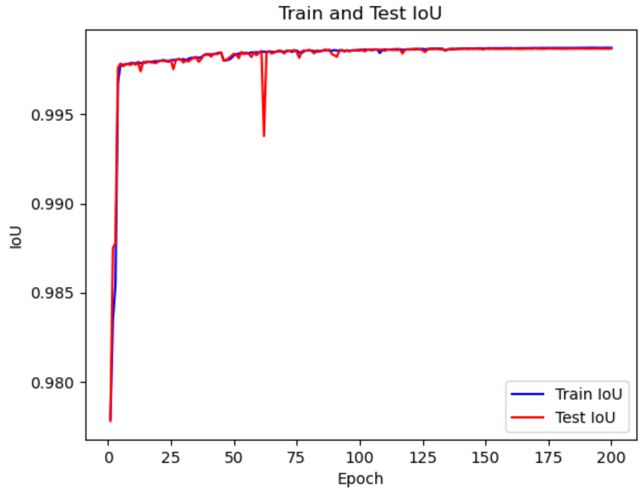
Deep learning clustering

- **Application: from simplest case**
  - Diphoton separation:
    - Event simulated within CEPC-v4 (SiW ECAL) geometry for training
      - γ_1: $E_\gamma \in [3,\ 7]$ GeV, $\theta_\gamma \in [88°, 90°]$, $\phi \in [0°, 3°]$
      - γ_2: $E_\gamma \in [3,\ 7]$ GeV, $\theta_\gamma \in [90°, 92°]$, $\phi \in [0°, 3°]$
      - ~200 hits / event.
    - Input feature: spatial coordinates only.
    - Loss function: cross entropy.



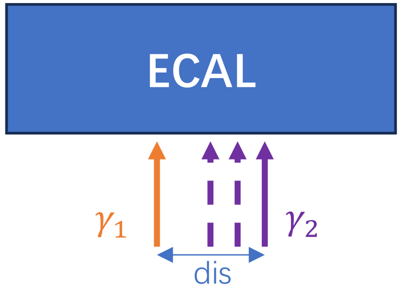


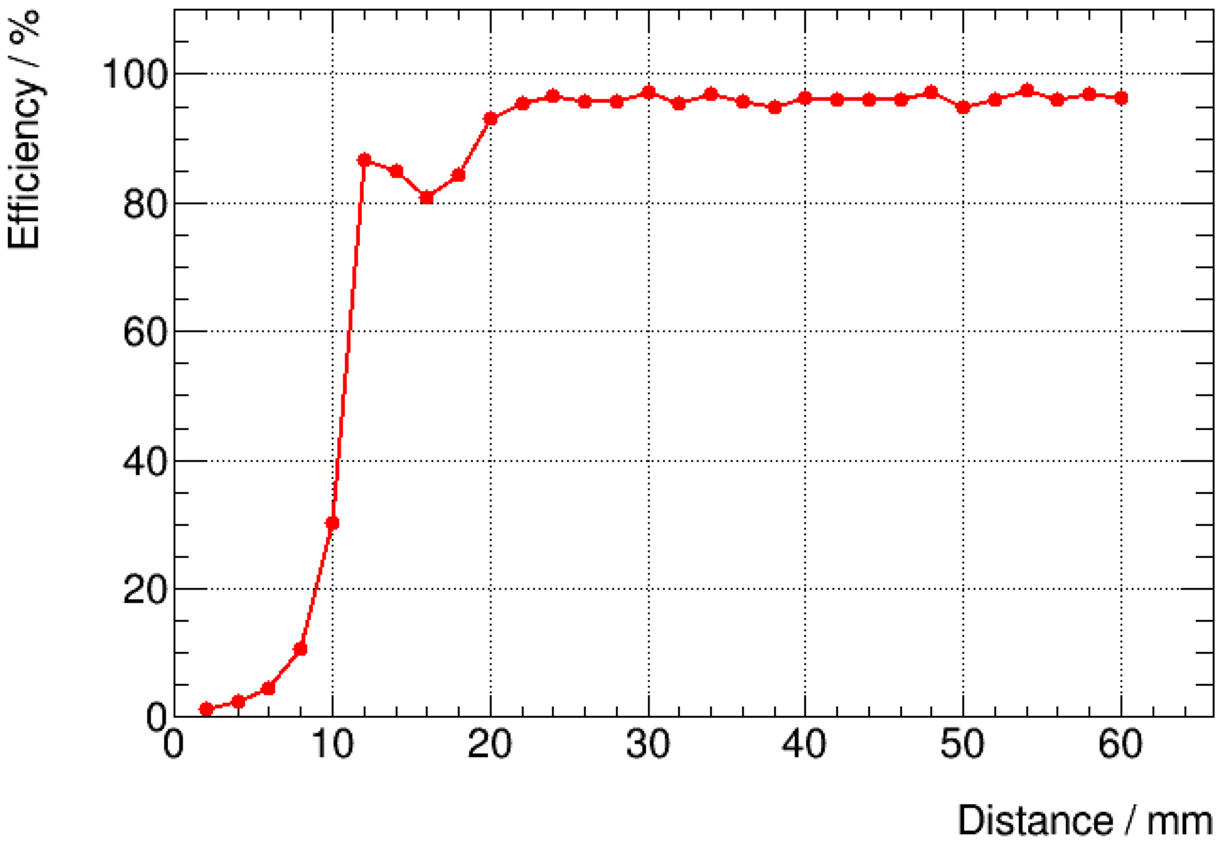


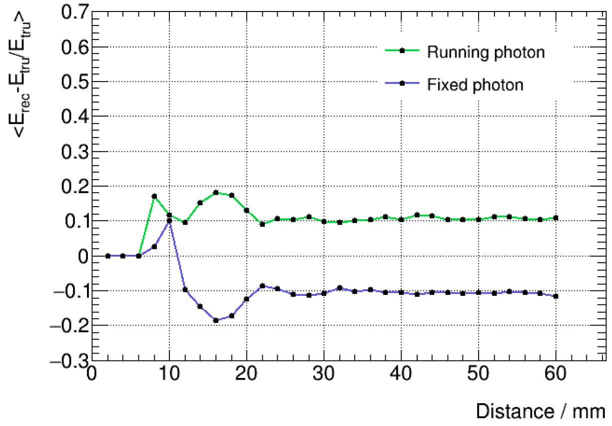


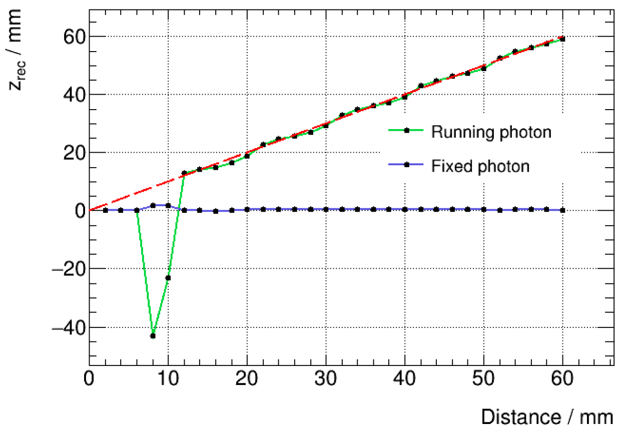
Deep learning clustering

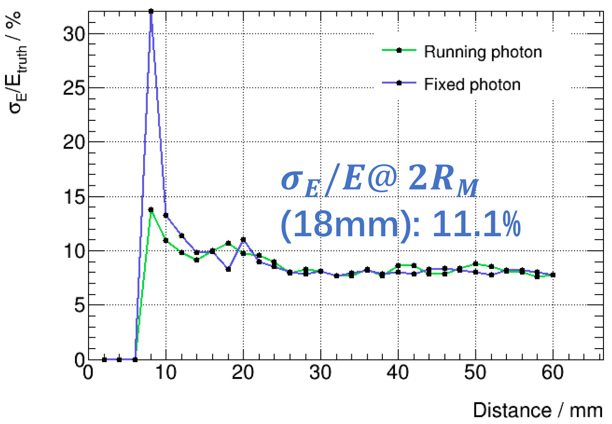
- **Preliminary performance: separation efficiency with distance.**
  - Applied into a set of di-photon events

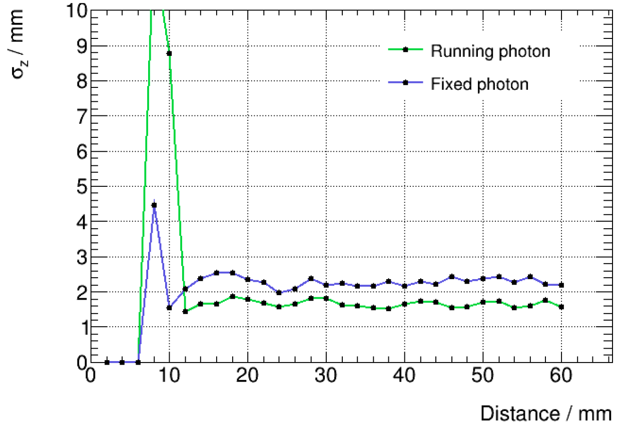






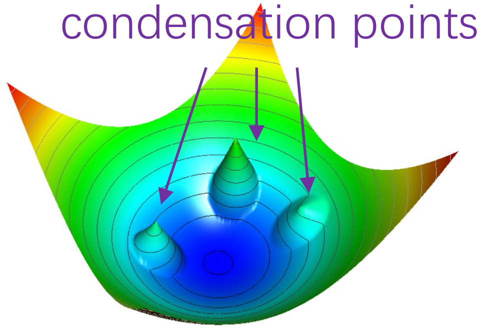






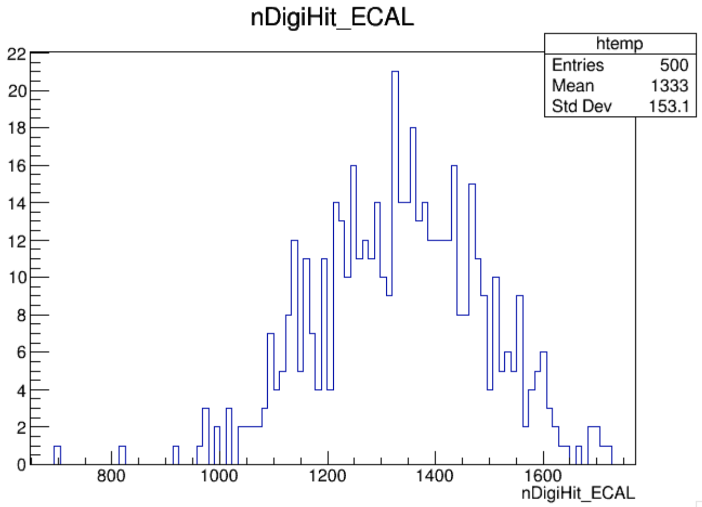
Discussion

- **Facing challenges:**
  - Loss function: cross entropy can not handle the clustering problem.
    - Condensation loss: $L = L_V + L_\beta + L_P$ [2002.03605]
      - L_V: potential loss, make hits closer to the condensation points.
      - L_β: condensation score, $\beta \rightarrow 1$ is the cluster center.
      - L_P: features can be trained in the model: e.g. $E_{cluster}$.
    - Can deal with multiple objects, but need to design for overlapping.



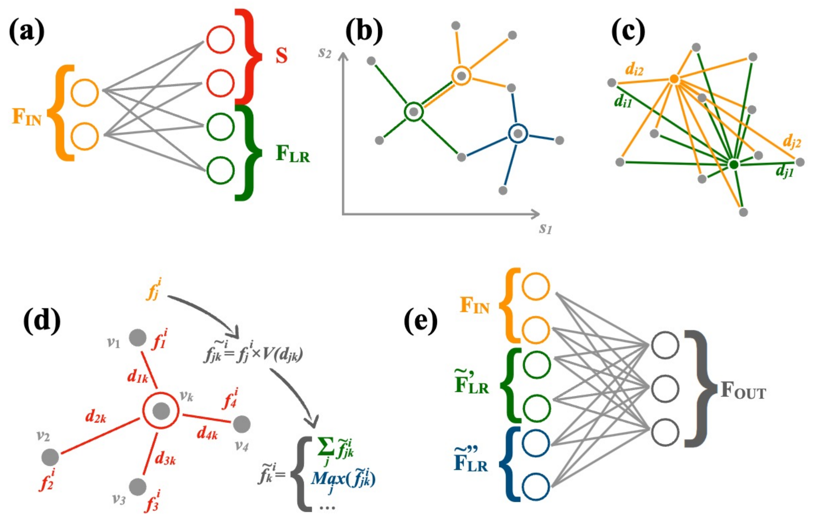
(Expected) potential from 2002.03605

  - Hit size: O(10) more in crystal ECAL than SiW ECAL.
    - Harder to converge.
    - Overlapping issue would be more critical.



Discussion

- **Other models:**
  - GravNet: distance weighted graph network architectures.
    - Commonly used in CMS machine learning studies: MLPF, End-to-End reconstruction, etc.



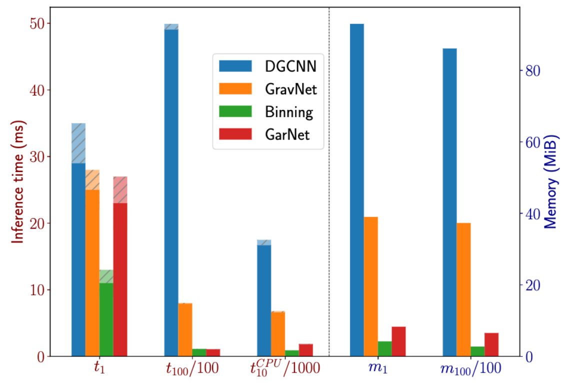
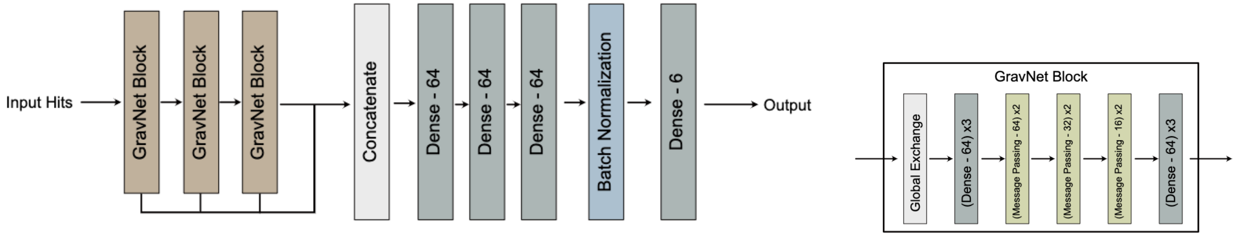


Fig. 5: Comparison of inference time for the network architectures described in the text, evaluated on CPUs and GPUs with different choices of batch size. The shaded area represents the $+1\sigma$ statistical uncertainty band.



Model architecture in MLPF

Summary and outlook

- **Calorimeter clustering with deep-learning:**
  - Showed the attempt of 2 clusters with DGCNN segmentation task.
  - Is a promising way to the detector reconstruction in AI's era.

- **Next step:**
  - Tune the models and loss functions.
  - Add features (energy, time) into the model.
  - Solve the data size issue in crystal ECAL.

- **Future:**
  - Add track info and pre-trained shower shape info as bias.
  - Final target: a general deep learning based PFA.