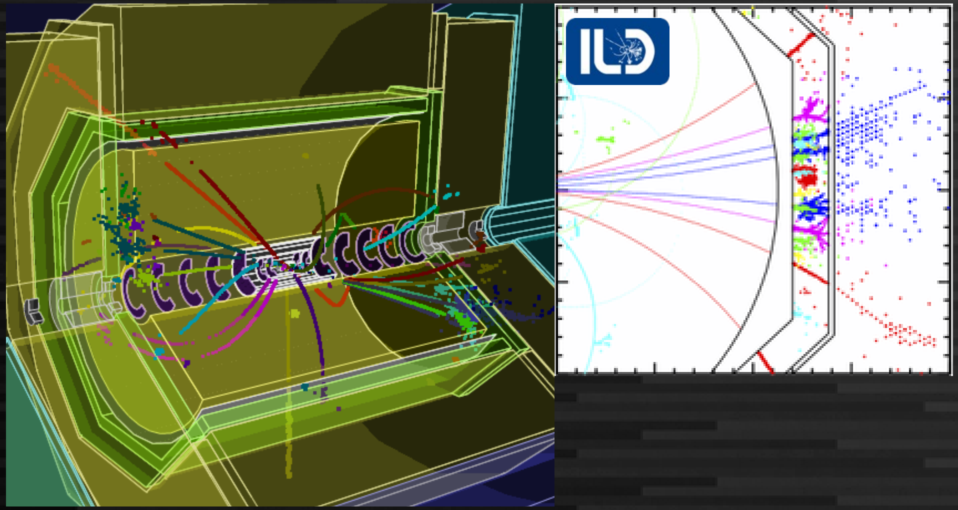
Particle flow with GNN: progress report

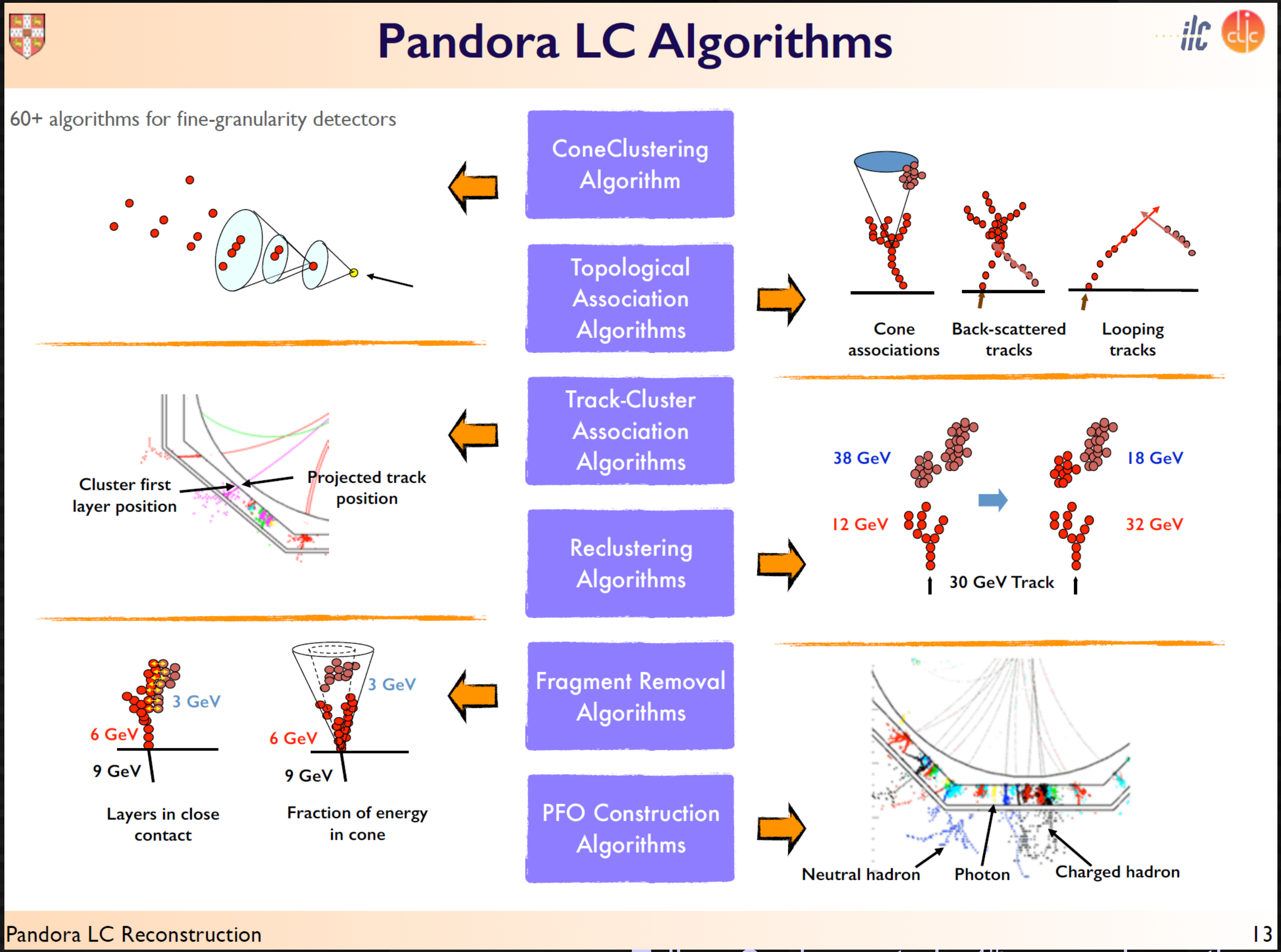
T. Suehara (Kyushu U.), S. Tsumura (Kyushu U. → KEK),
H. Nagahara, Y. Nakashima (Osaka U.), N. Takemura (Kyushu Tech.),
L. Gray, T. Klijnsma (Fermilab), T. Tanabe (MI-6 Ltd.)

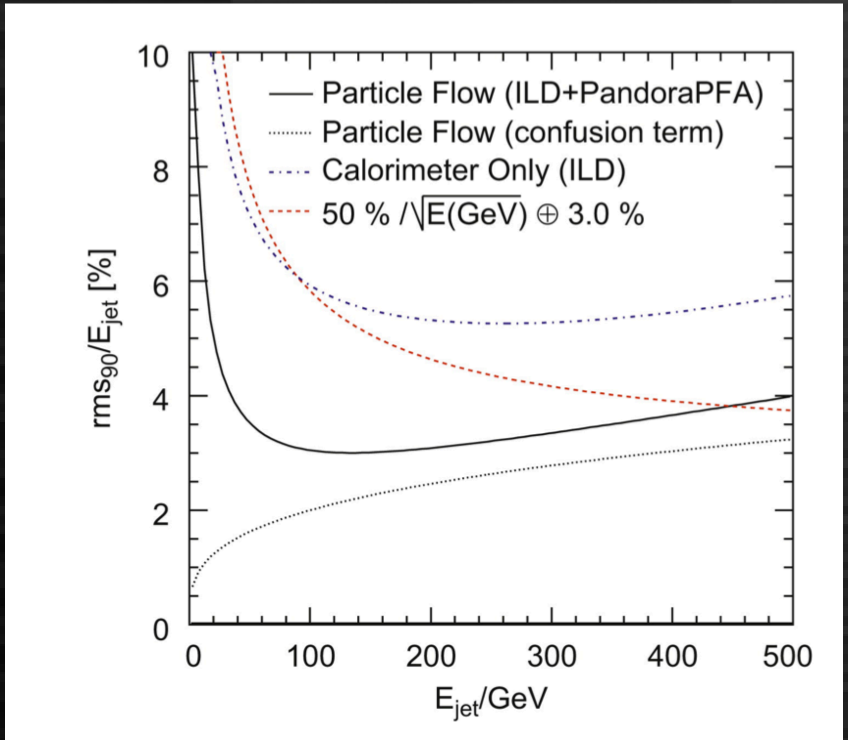
Particle flow for Higgs factories



- High granular calorimetry
  - 3D pixels for imaging EM/hadron showers at calorimeters
    - eg. 10^8 channels for ILD ECAL
  - Separation of particles inside jets
    → ~2x better energy resolution by separation of contribution from charged particles
    - Software algorithm essential (as well as hardware design)
- Particle Flow algorithm
  - Essential algorithm for high granular calorimetry
  - Complicated pattern recognition

Pandora ParticleFlow algorithm





Widely used since 2008

Reasonably good performance up to ~50 GeV jets

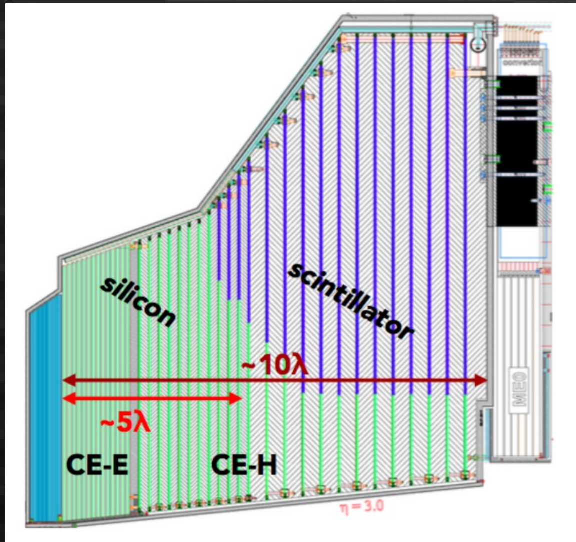
Confusion dominates at higher energies

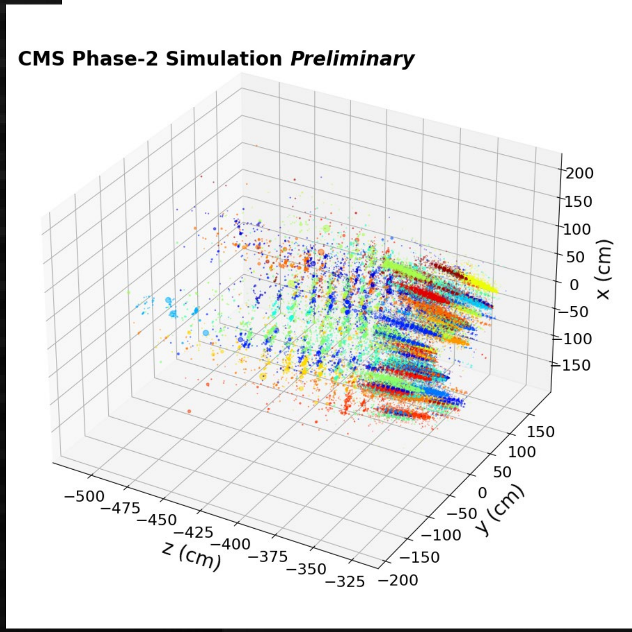
Motivations for DNN particle flow

- Performance improvement
  - Confusion dominant at jet energy > 100 GeV
  - More efficient way to separate cluster from charged particles should be investigated
- Integrate other functions
  - Software compensation, particle ID etc. closely related to PFA
- Detector optimization
  - Comparison with different detector settings
    - PandoraPFA too much depends on internal parameters
  - Effect of timing information to be investigated
    - With different timing resolution (1 ns, 100 ps, 10 ps, …)

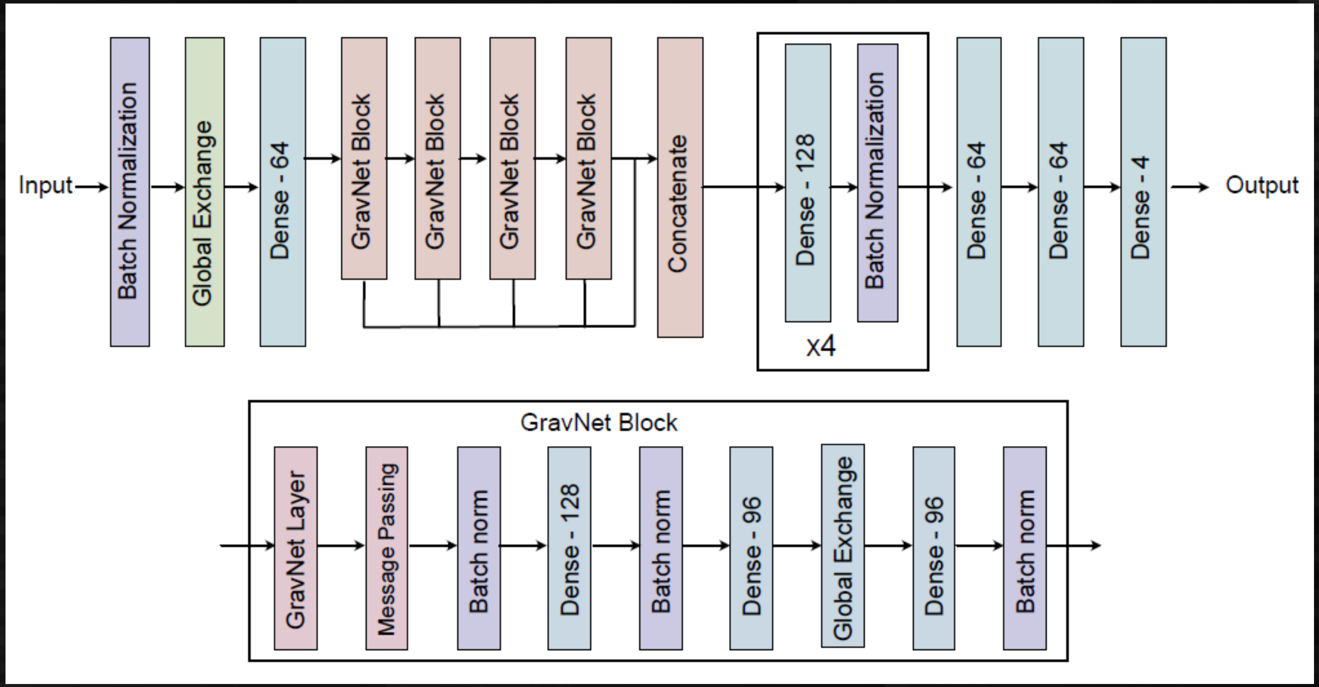
GravNet for CMS HGCAL

- CMS HGCAL
  - High granular forward calorimeter for HL-LHC upgrade at CMS
  - Similar to ILD calorimeter (silicon pixel + scintillator)
    - Inspired by CALICE development
- Reconstruction at HGCAL
  - Big noise to be separated by software
  - Numerous particles from ~200 pileups
    - Difficult to handle by current software
    - DNN reconstruction being investigated
      - Reasonable performance obtained up to ~50 pileups?





The network



Rather complicated network with ~30 hidden layers

"Object condensation" loss function is applied (shown in next page)

Input/output obtained for each hit at calorimeter

Input: Features at each hit (position, energy deposit, timing)

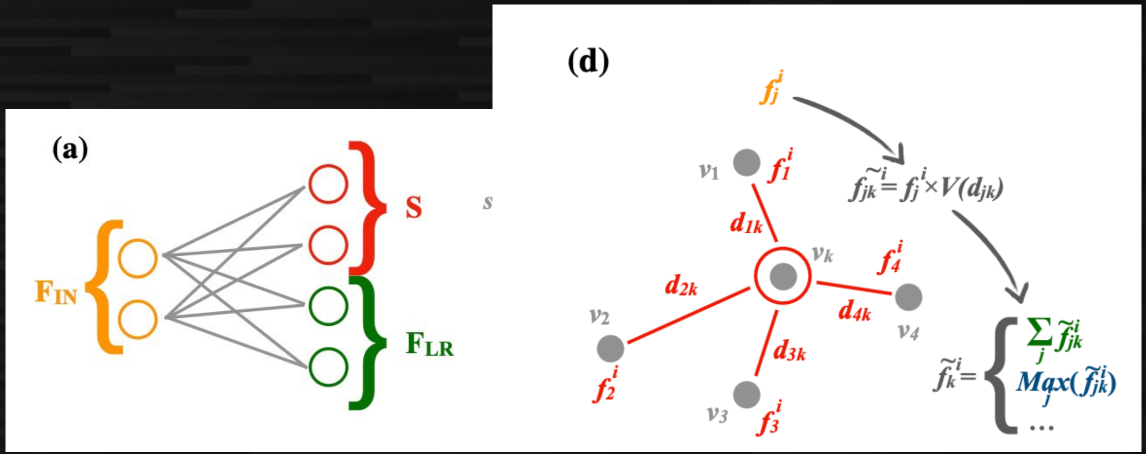
Output: "condensation coefficient" β, position at virtual coordinate (2-dim)
optional output of features such as energy, PID (not used now)

Dense (fully-connected layer) inside each hit, GravNet connects hits

GravNet and Object Condensation

GravNet arXiv:1902.07987

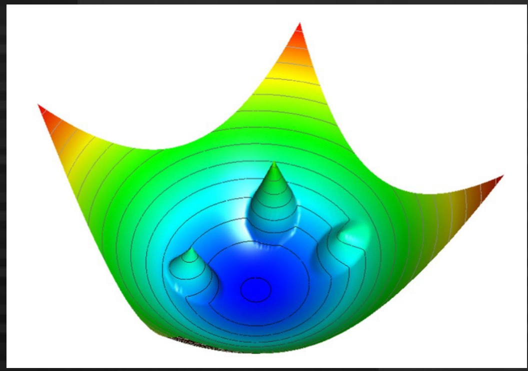
- The virtual coordinate (S) is derived from input variables with simple MLP
- Convolution using "distance" at S (bigger convolution with nearer hits)
- Repeat 2 times and concatenate the output with simple MLP



Object Condensation (loss function)

arXiv:2002.03605

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

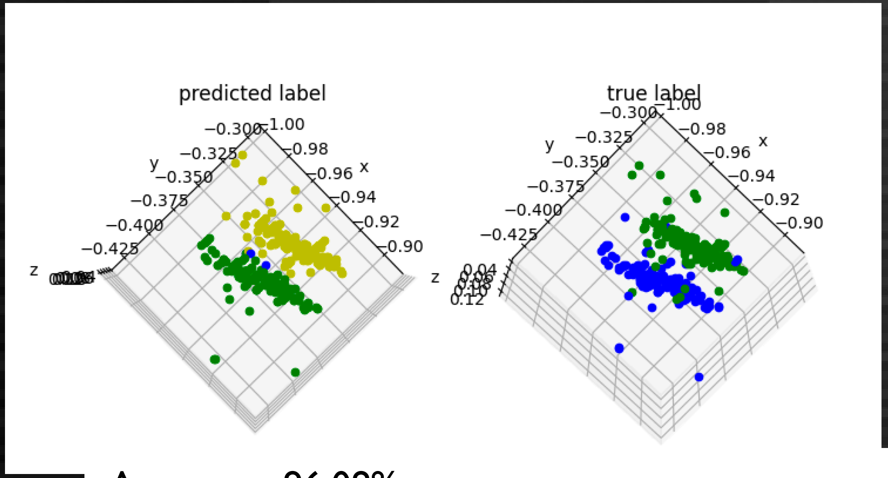


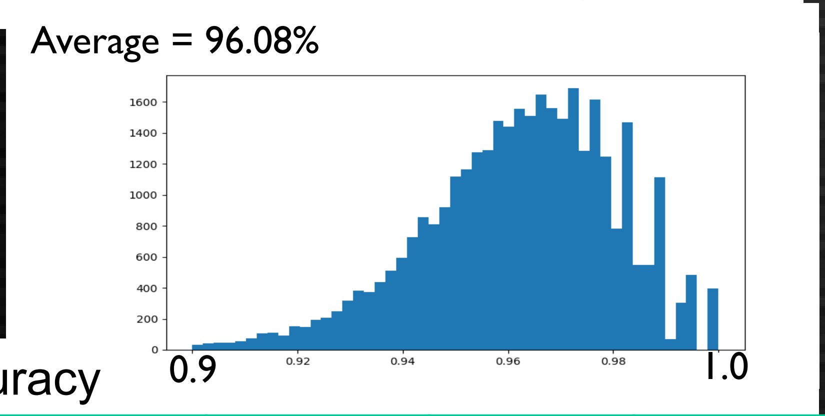
- Condensation point: The hit with largest β at each (MC) cluster
- L_V: Attractive potential to the condensation point of the same cluster and repulsive potential to the condensation point of different clusters
- L_β: Pulling up β of the condensation point
- L_p: Regression to output features (energy etc.) → currently not used

Importing to ILD full simulation

- Prepare features from ILD full simulation
  - With recent versions (> v02-02)
- Input features: (x, y, z, edep)
- True cluster info
  from MCParticle and LCRelation
- Produced events
  - Two photons
    (5/10 GeV, fixed opening angles)
  - (n x) taus (5/10 GeV)
- Evaluation
  - Fraction of hits associated to the correct cluster (accuracy)

Example of a two-photon event
(5 GeV, 30 mrad)





Reasonable
performance seen

Angle[mrad]	30	60	90	120	150
Accuracy[%]	96.08	98.64	99.30	99.68	99.56

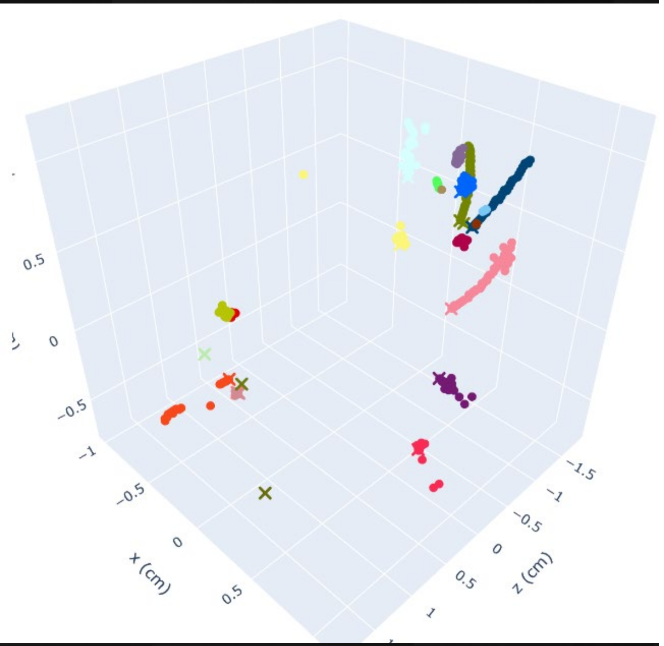
For details, refer eg. https://indico.slac.stanford.edu/event/7467/contributions/5948/attachments/2887/8032/230517-lcws2023-hlreco-suehara.pdf

Work in Progress: track-cluster matching

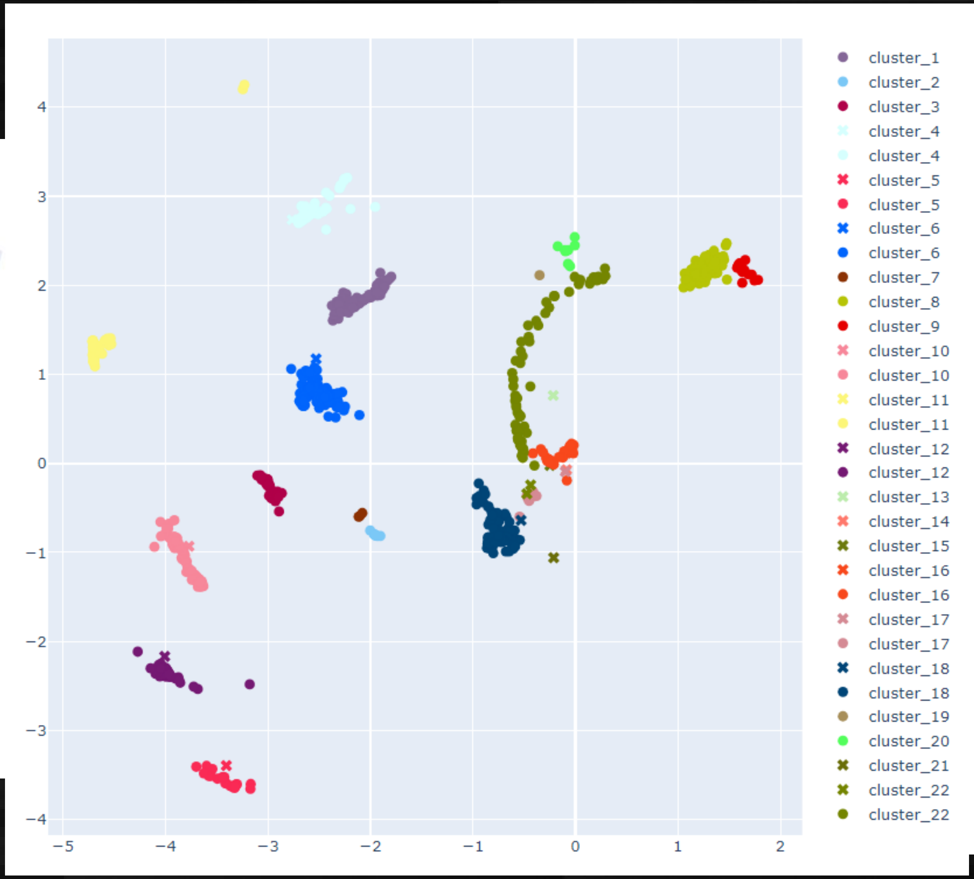
- PFA is essentially a problem "to subtract hits from tracks"
- HGCAL algorithm does not utilize track information
  - Only calorimeter clustering exists
- Simple extension to include track information
  - Adding "virtual hits" derived from track information
    - Hits at position where the track enters the calorimeter (from LCIO StackState)
  - Add a term to the object condensation loss function
    - Pulling up β of tracks (virtual hits) to promote them to condensation points (in addition to the usual beta-term, called beta-track term)
  - Evaluate fraction of (MC) charged clusters to be correctly assigned to clusters with tracks (virtual hits)

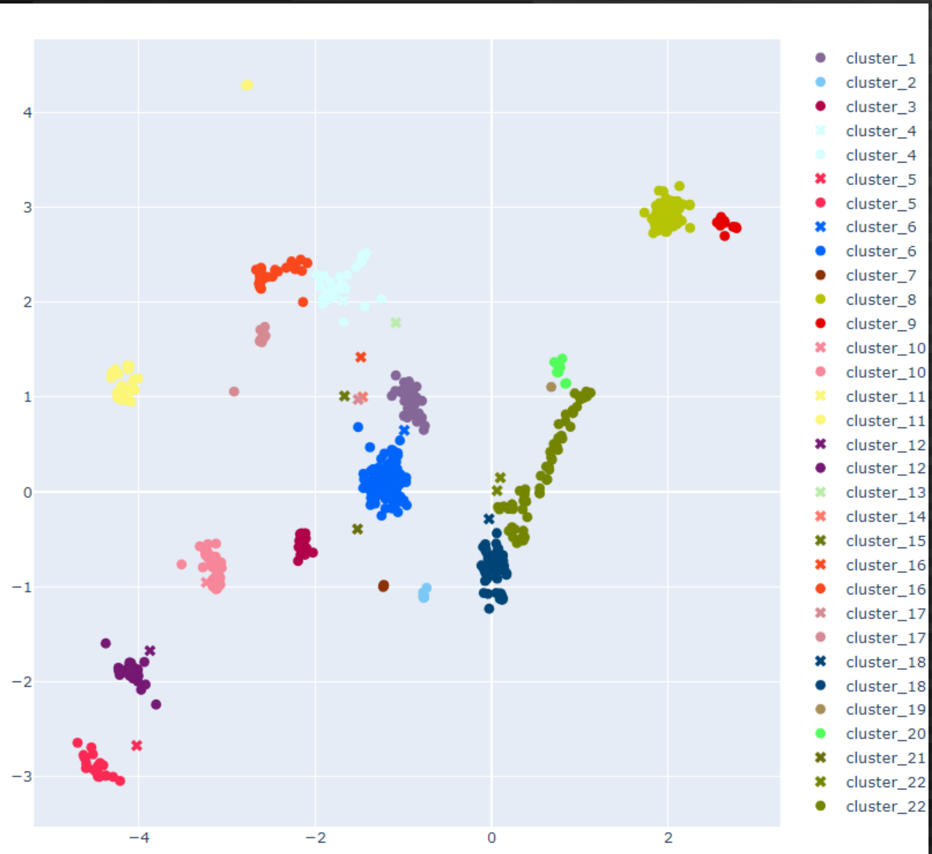
Preliminary results – event sample

10 Taus @ 10 GeV each



Real 3D coordinate





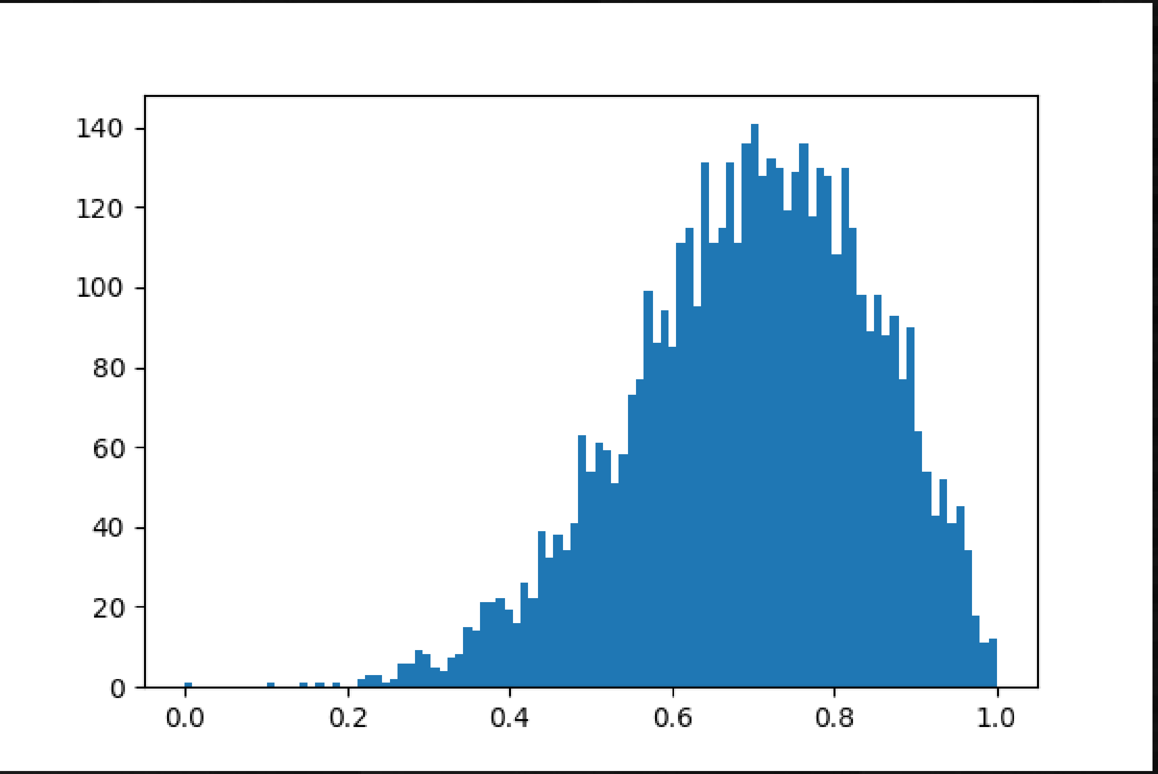
Hits on the virtual coordinate – colored by MC truth clusters
x refers virtual hits from tracks
left with beta-track term, right without beta-track term

Situations

- Position of virtual hit (from track) is distant from other hits from the same cluster even if we put beta_track term
  - Better in initial 3D coordinate
  - Maybe mapping 0/1 condition (track=1, non-track=0) to the output coordinates then getting different position
  - Beta_term only from epoch 8: is this the problem?
  - Better representation of beta_track term
    - Being tried

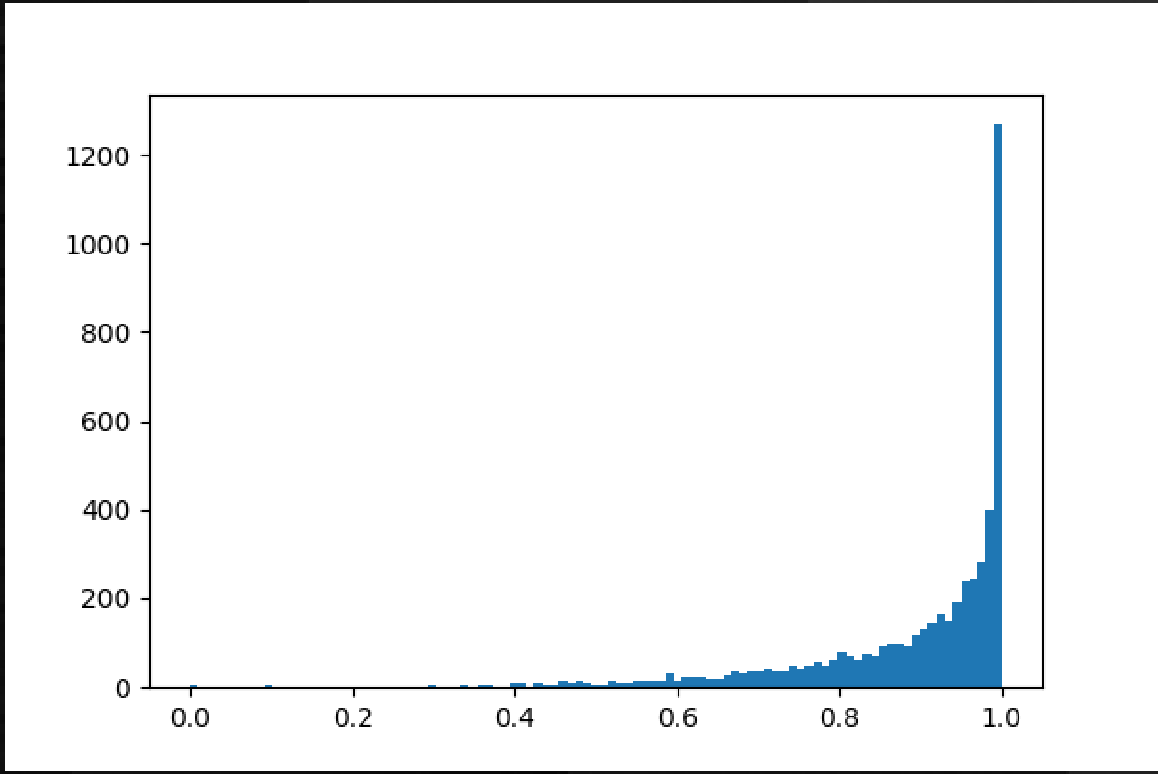
Evaluation of performance

Fraction of charged hit predicted correctly



Average: 0.698 (somehow low)

Fraction of neutral hit predicted correctly



Average: 0.890

Pred charged hit: associated cluster having at least one track

Ideas of improvement(short term / ongoing)

- Adding theta/phi → not significant
- Utilizing track momentum
  - At clustering (additional conditional association, being tried)
  - Regression of momentum at the network (not done yet)
- Adding classification of track hit or not (being done)
  - Interaction/tuning with other terms
- Tuning of learning (partially done, not significant so far)

- Fully-DNN clustering (currently clustering is done with traditional way)

Summary / long-term plans

- New DNN-based particle flow algorithm is under development based on clustering at CMS HGCAL study
- Track-cluster matching is being implemented, statistical results will come soon
  - Energy regression with track momentum information will be the next step of implementation
- Medium/long term plans (or just hopes)
  - Can be extended to any analyses using cluster/jet information using the PFA as "a foundation model"
    - Such as Particle ID, Jet clustering, even physics analyses directly
  - "Differentiate" detector parameters/designs for optimization