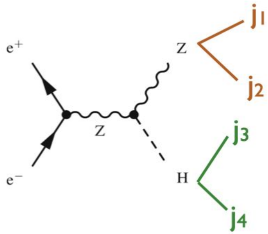
Color Singlet clustering and MLPF

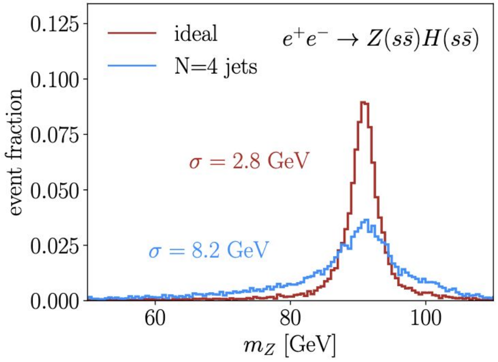
GNN based Reconstruction at FCC-ee

Dolores Garcia

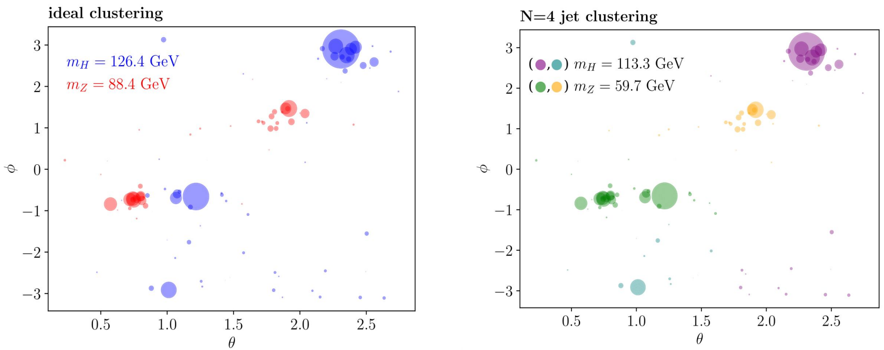
Clustering Color Singlets

- FCC-ee would serve as a Higgs factory, electroweak and top at highest luminosities
  - Measure Higgs particle properties and interactions in challenging decay modes
- Identification of color-neutral resonances relies on clustering final state into jets
- Calorimetry is expected to be much improved at future e+e− colliders, so that the 2-jet invariant mass resolution will be dominated not by detector resolution but rather by mis-clustering [1] (A)
- Jets are not well defined but color connection is physical, this may help **improve the mass estimation for color singlets** (H,Z,W) **and remove more background**





A Comparison of clustering performance vs ideal reconstruction



B Example of miss clustering

ACKS : Michele Selvaggi

[1]Fujii, K., Grojean, C., Peskin, M. E., Barklow, T., Gao, Y., Kanemura, S., ... & Murayama, H. (2020). ILC study questions for snowmass 2021. *arXiv preprint arXiv:2007.03650*..
[2] Gallicchio, J., & Schwartz, M. D. (2010). Seeing in color: jet superstructure. *Physical review letters*, *105*(2), 022001.

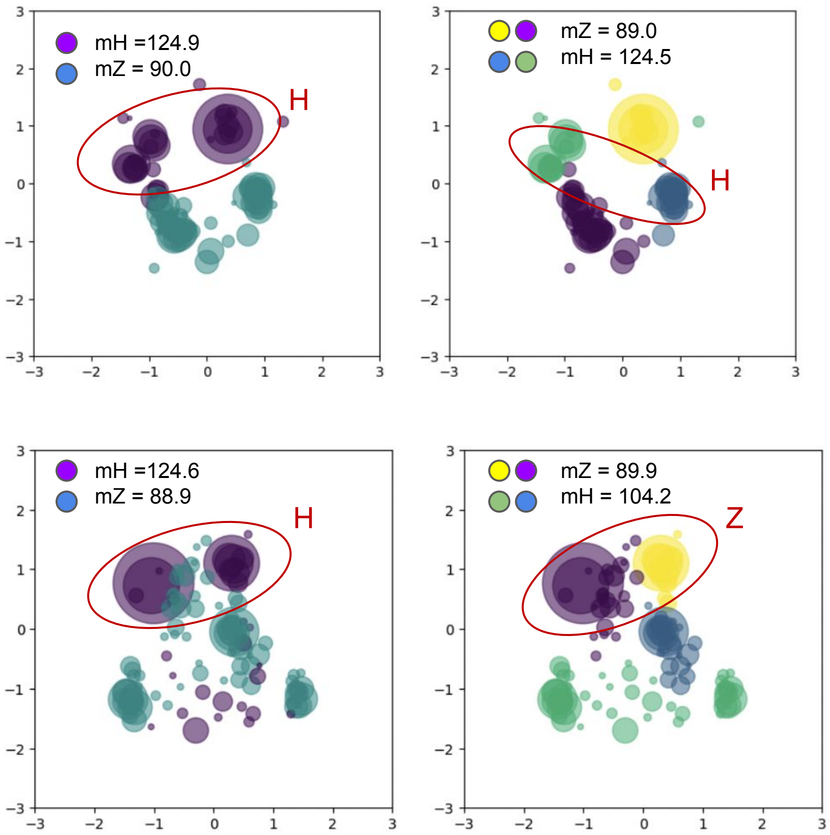
Clustering Color Singlets

Loss in performance can be due to:

- Miss matching of jets pairs
- Miss clustering of soft particles leading to degraded resolution

Possible solutions:

- Parameter tuning (BAO)
- Optimize distance metrics?: piecewise continuous function, hard optimization problem
- **End-to-end approach**



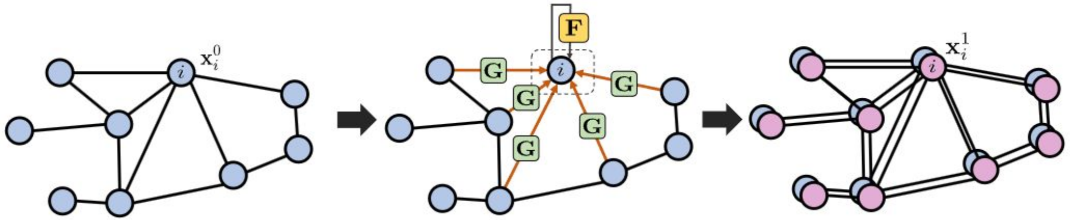
A Mismatching of jets pairs

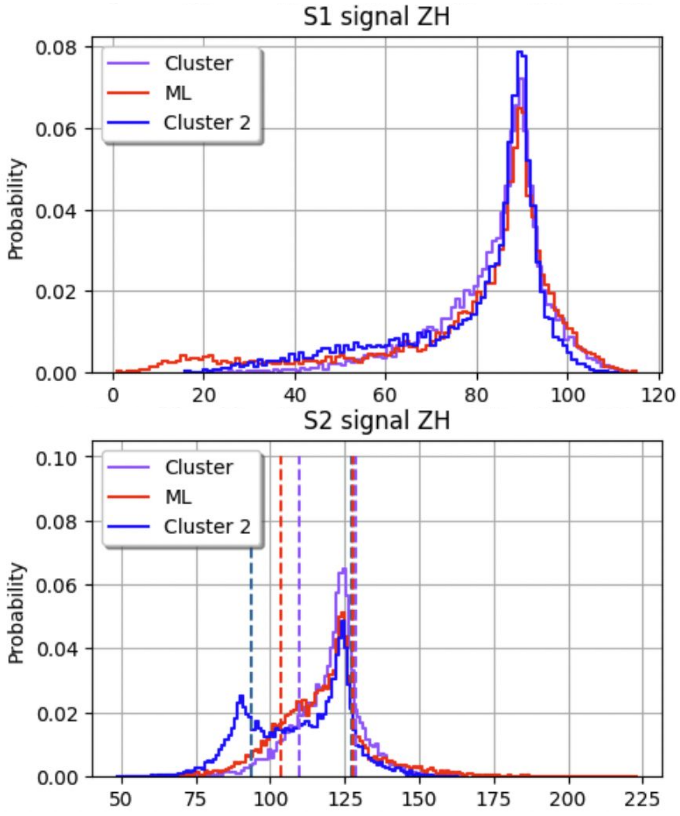
ACKS : Michele Selvaggi

[1]Fujii, K., Grojean, C., Peskin, M. E., Barklow, T., Gao, Y., Kanemura, S., ... & Murayama, H. (2020). ILC study questions for snowmass 2021. *arXiv preprint arXiv:2007.03650*..
[2] Gallicchio, J., & Schwartz, M. D. (2010). Seeing in color: jet superstructure. *Physical review letters*, *105*(2), 022001.

CSC- Approach

- **GNN** - Node classification (instantiation) problem, permutation invariant and equivariant
- Arch: FC - Graph **Transformer [1]**
- **Results:**
  - Similar performance to classical approach
  - **Baselines:**
    - **Chi-squared**
      $\chi^2 = 1/\sigma_H(M_{1/2}-M_H)^2 + 1/\sigma_z (M_{2/1}-M_Z)^2$
    - **Z only**
      $\chi^2 = (M_{1/2}-M_Z)^2$
- Can find events that reduce background by assigning a score per event





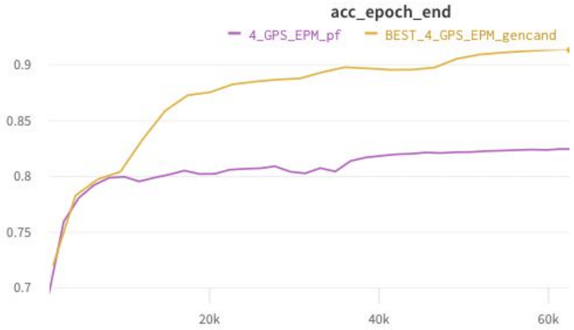
A. Mass distributions of signal

ACKS : Michele Selvaggi

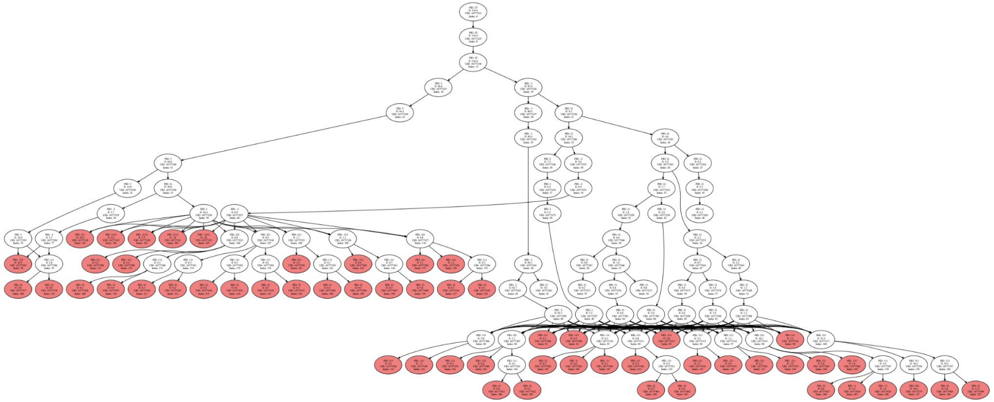
[1]Dwivedi, V. P., & Bresson, X. (2020). A generalization of transformer networks to graphs. *arXiv preprint arXiv:2012.09699*..

CSC- Approach

- **Wiring is important**
- Using information about **the ordering** (<tree structure) performance can be improved
- Efforts to obtain MLE (A*, beam search…) all for small number of leaves [1,2]



A. Accuracy increase with new wiring, ordering by tree structure



B. Example tree

ACKS : Michele Selvaggi

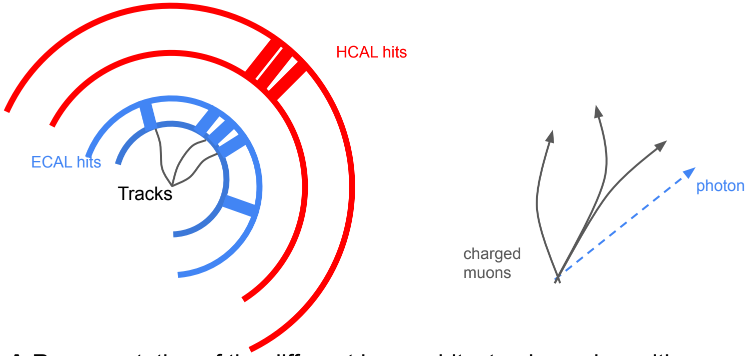
[1]Brehmer, J., Macaluso, S., Pappadopulo, D., & Cranmer, K. (2020). Hierarchical clustering in particle physics through reinforcement learning. *arXiv preprint arXiv:2011.08191*.
[2] Greenberg, C. S., Macaluso, S., Monath, N., Dubey, A., Flaherty, P., Zaheer, M., ... & McCallum, A. (2021, December). Exact and approximate hierarchical clustering using A. In *Uncertainty in Artificial Intelligence* (pp. 2061-2071). PMLR.

MLPF: Motivation

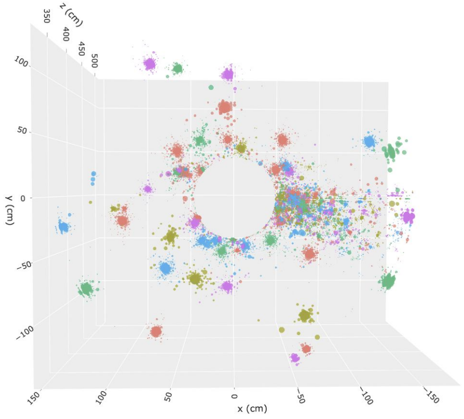
- The particle flow algorithm aims to identify the produced particles in a collision through the combination of the information from the entire detector and provide best combined energy/momentum resolution

- Hoping to achieve higher reconstruction performance: cluster merging, arbitration of track vs cluster energy

- First step: **focus on calorimeter clustering**



A Representation of the different layers, hits, tracks and resulting particles (reproduced from [1])

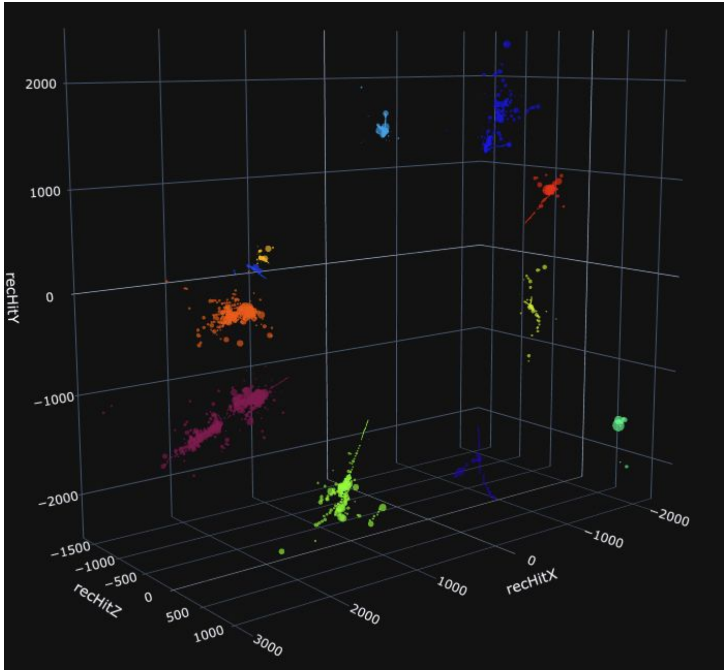


B Example of an event, the shower of secondary particles generated by an individual particle is labelled with one colour [2]

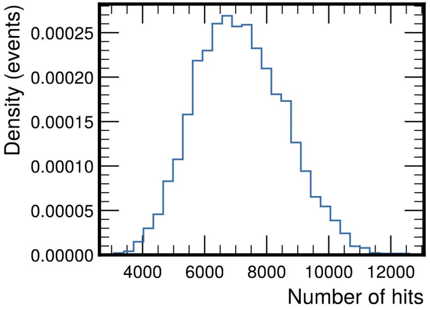
[1] Pata, J. Machine learning for particle flow reconstruction at CMS, presentation at CDS.
[2] Qasim, S. R., Chernyavskaya, N., Kieseler, J., Long, K., Viazlo, O., Pierini, M., & Nawaz, R. (2022). End-to-end multi-particle reconstruction in high occupancy imaging calorimeters with graph neural networks. *The European Physical Journal C*, *82*(8), 753.

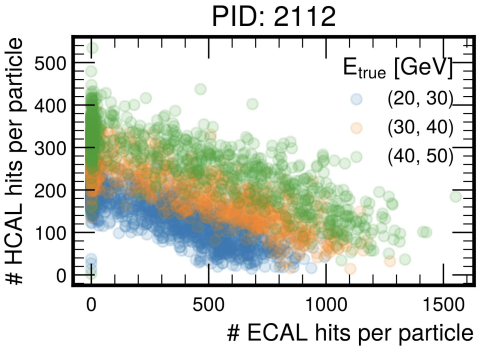
Training Data

A Example train event - 15 particles



- Event generation:
  - Use particle gun (10-15 particles)
  - $E \in [0.5, 50]$ GeV
  - p, n, K_L, π
- FCC-ee O(100)
- Simulation and reconstruction: Key4HEP turnkey + Geant4



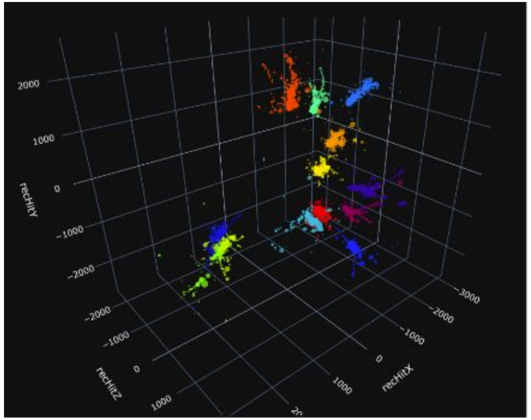


B Number of hits per event (left) and #hits ECAL vs HCAL (right)

ACKS : Michele Selvaggi, **Gregor Krzmanc,** Jan Kieseler, Philipp Zehetner

CERN

Architecture: Object condensation (End-to-End approach)



Input:

- A set of hits from different sensors (coordinates, type of hit, energy, A)
- Each one node in the graph O(600) per particle

CP

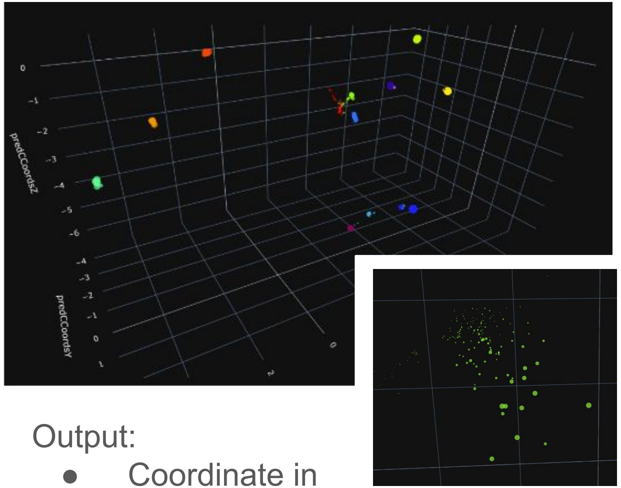
$$q_{\alpha k} = \max_i q_i M_{ik}.$$

$$\breve{V}_k(x) = \|x - x_\alpha\|^2 q_{\alpha k}, \text{ and}$$

$$\hat{V}_k(x) = \max(0, 1 - \|x - x_\alpha\|) q_{\alpha k}.$$

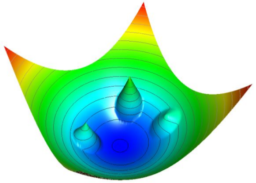
$$L_V = \frac{1}{N} \sum_{j=1}^{N} q_j \sum_{k=1}^{K} \left(M_{jk} \breve{V}_k(x_j) + (1 - M_{jk}) \hat{V}_k(x_j) \right).$$

- **Each object 1 condensation point (CP)**
- **Repulsive +Attractive potentials for each CP**



Output:

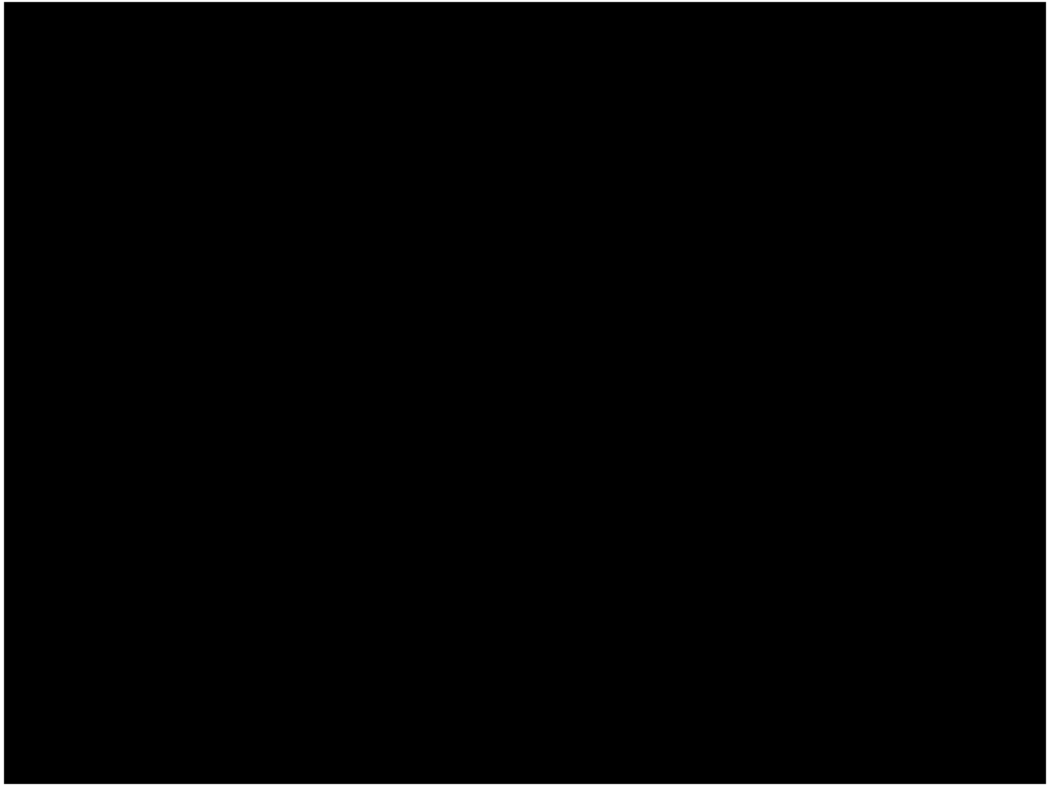
- Coordinate in embedding space (3D>)
- Beta (q)
- Use clustering space to build showers



ACKS : Michele Selvaggi, Gregor Krzmanc, Jan Kieseler, Philipp Zehetner

[1] Kieseler, J. (2020). Object condensation: one-stage grid-free multi-object reconstruction in physics detectors, graph, and image data. *The European Physical Journal C*, *80*, 1-12.
[2] Qasim, Shah Rukh, et al. "Learning representations of irregular particle-detector geometry with distance-weighted graph networks." *The European Physical Journal C* 79.7 (2019): 1-11.

Architecture: Object condensation (End-to-End approach)

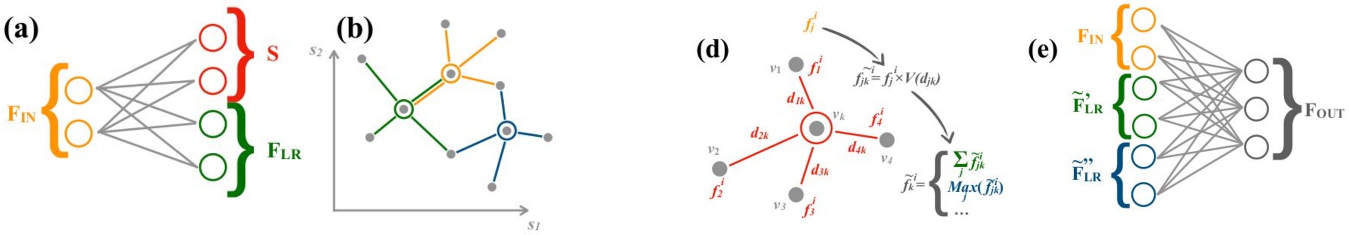


ACKS : Michele Selvaggi, Gregor Krzmanc, Jan Kieseler, Philipp Zehetner

Architecture: Gravnet Model

- Input: a set of hits from different sensors (coordinates, type of hit, energy), each one node in the graph O(600) per particle
- Graph representation with **no given graph structure**
- **Dynamically** compute edges in embedding space with knn

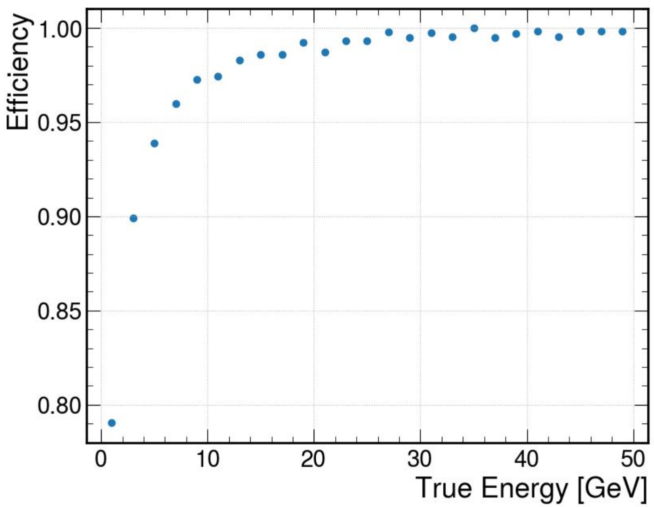
```
a)    Transform input features F_IN into
        ○    transformed features F_LR
        ○    latent coordinates S
b)    Build graph using coordinates S
d)    Aggregate weighted features
        ○    Weights depending on distance
        ○    Aggregation typically is mean or max
e)    Concatenate the new features
```



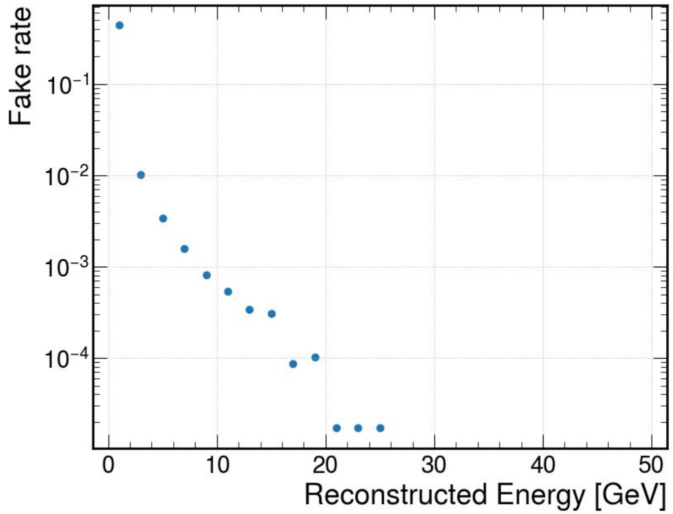
ACKS : Michele Selvaggi, Gregor Krzmanc, Jan Kieseler, Philipp Zehetner

[1] Kieseler, J. (2020). Object condensation: one-stage grid-free multi-object reconstruction in physics detectors, graph, and image data. *The European Physical Journal C*, *80*, 1-12.
[2] Qasim, Shah Rukh, et al. "Learning representations of irregular particle-detector geometry with distance-weighted graph networks." *The European Physical Journal C* 79.7 (2019): 1-11.

Efficiency and fake rate

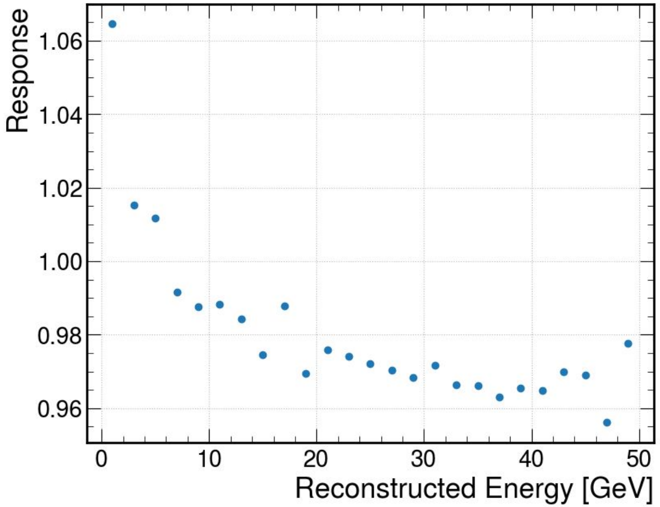


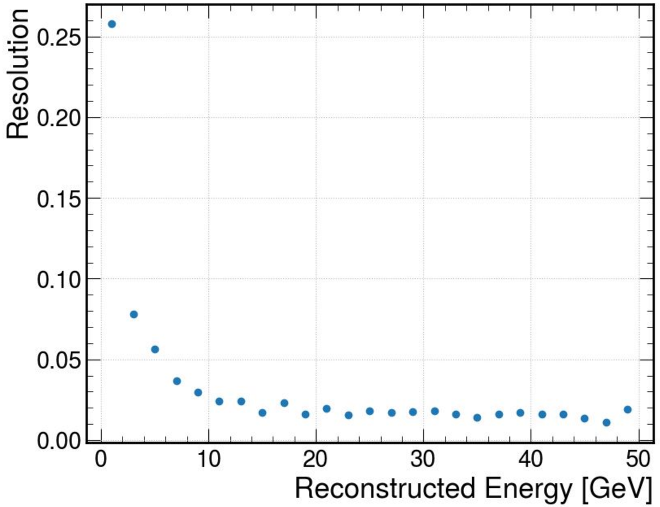
- Efficiency approaches 100% with high p_T
- Adding tracks will improve efficiency



- Most fakes with E< 1 GeV
- Other clustering methods in the embedding space can improve fakes
- Resulting from bad beta distributions

Response and Resolution (clustering metrics)





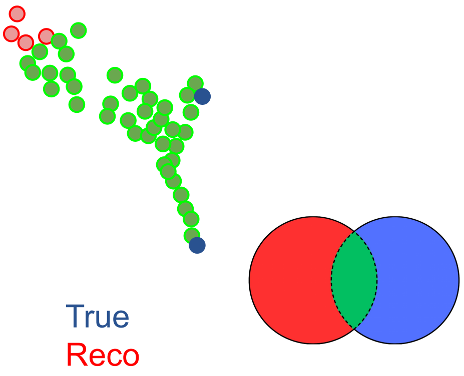
- Evaluated on reco values (for clustering evaluation)

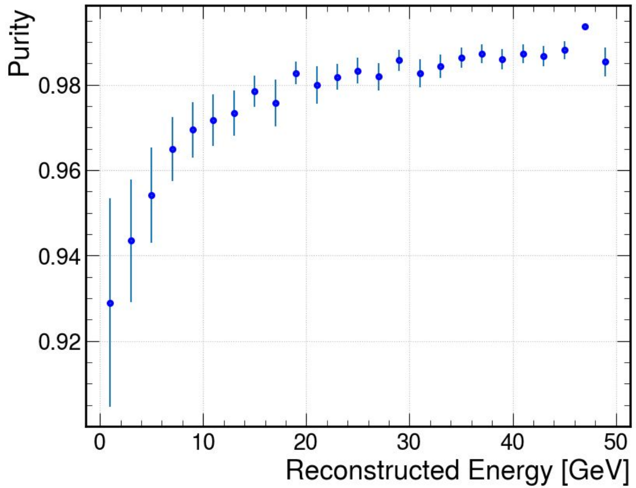
- Resolution performance must be improved for low energies
- Can be improved with better clustering in embedding space

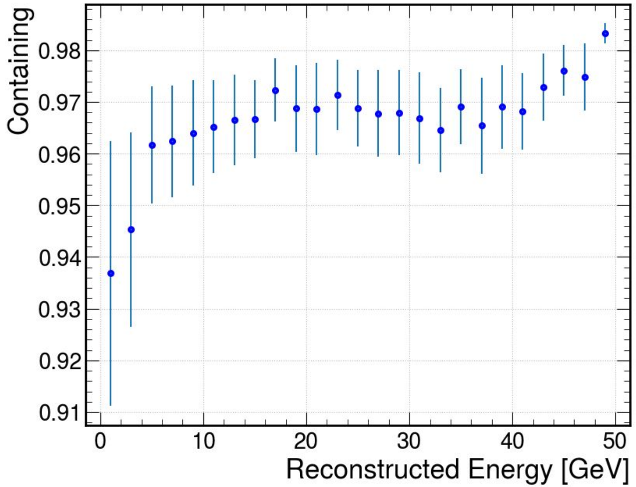
ACKS : Michele Selvaggi, Gregor Krzmanc, Jan Kieseler, Philipp Zehetner

Calorimeter clustering - Results

- Containing: percentage of reco energy that belong to the reconstructed particle **(G+R)/(G+B)**
- Purity: Percentage of reco energy **contained** in reconstructed cluster **(G)/(G+R)**







Summary and next steps MLPF

Summary:

- Color singlet clustering seems promising with ML but more research is necessary
- **MLPF**
  - Promising performance, we will soon compare to PFA (baseline for CLD)
  - Demonstrated generalization over different types of events (for now kept particle number low)
  - Fast execution time, linear scaling with number of hits

Ongoing work and next steps:

- Add tracks as inputs to the graph
- Regress particle properties
- Try heterogeneous graph architectures
- Compare to the performance of PFA

ACKS : Michele Selvaggi, Gregor Krzmanc, Jan Kieseler, Philipp Zehetner

Thank you!