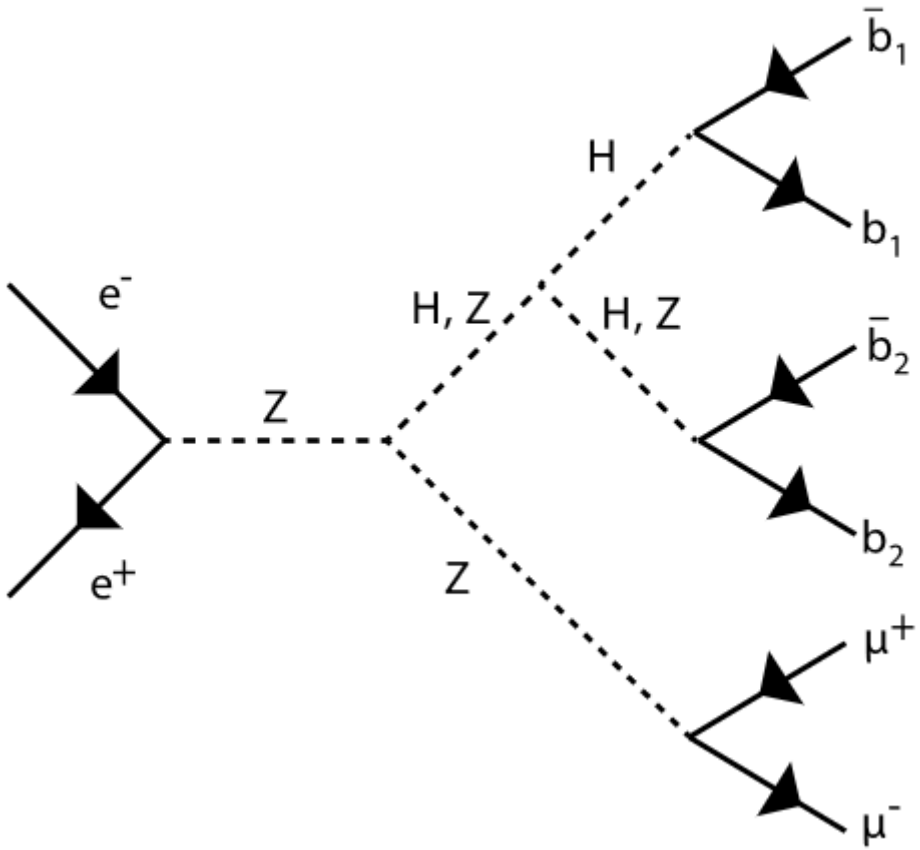
Status on Supervised Jet Clustering with GNNs and Spectral Clustering

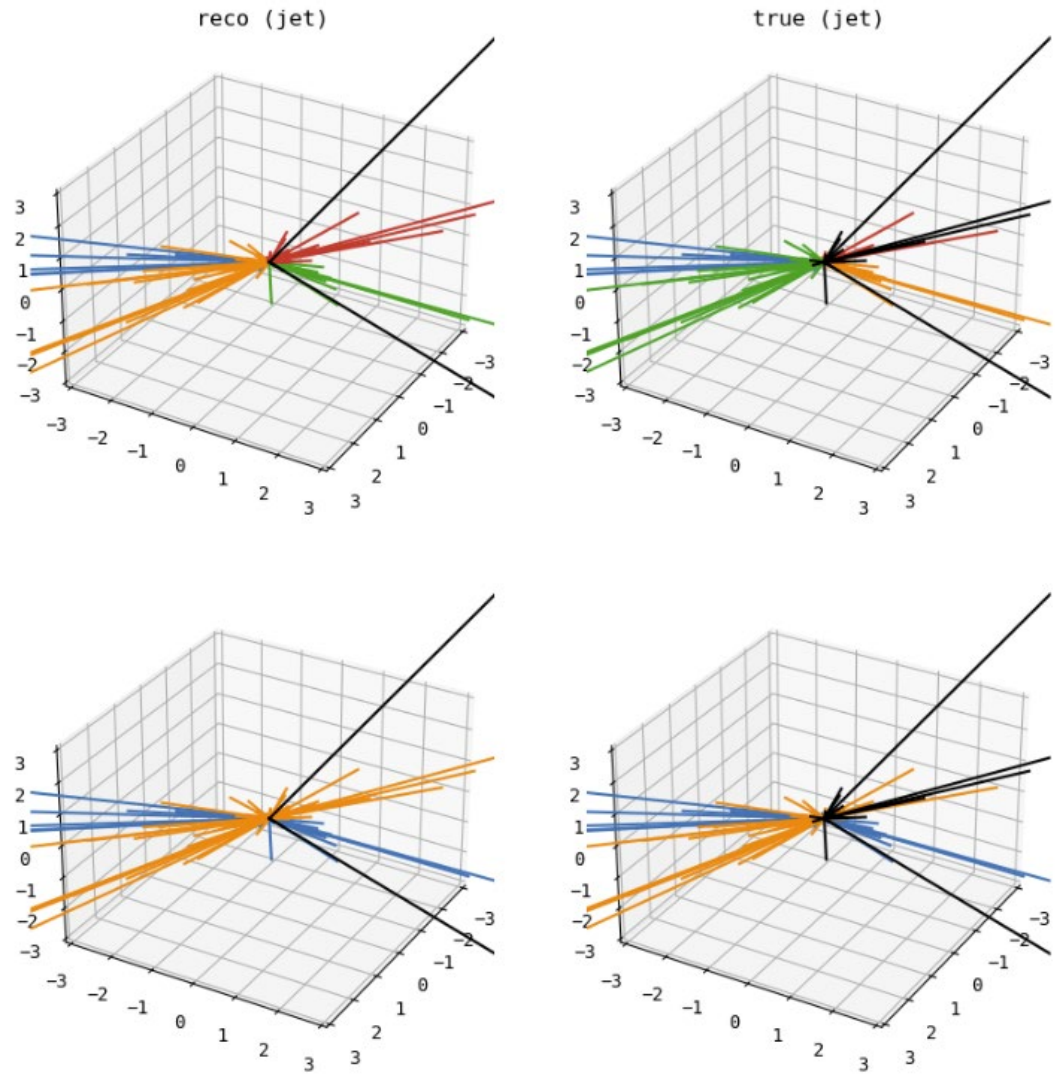
ILD Software Analysis Meeting | 2024/03/13

Bryan Bliewert
DESY & Technical University of Munich

HELMHOLTZ

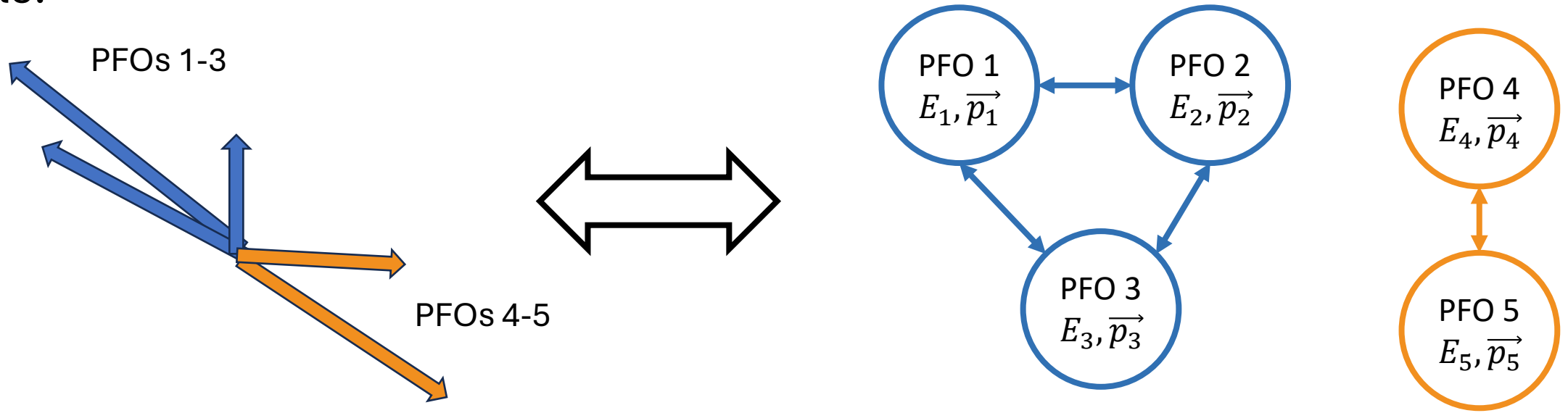
Jets in the ZHH-Analysis





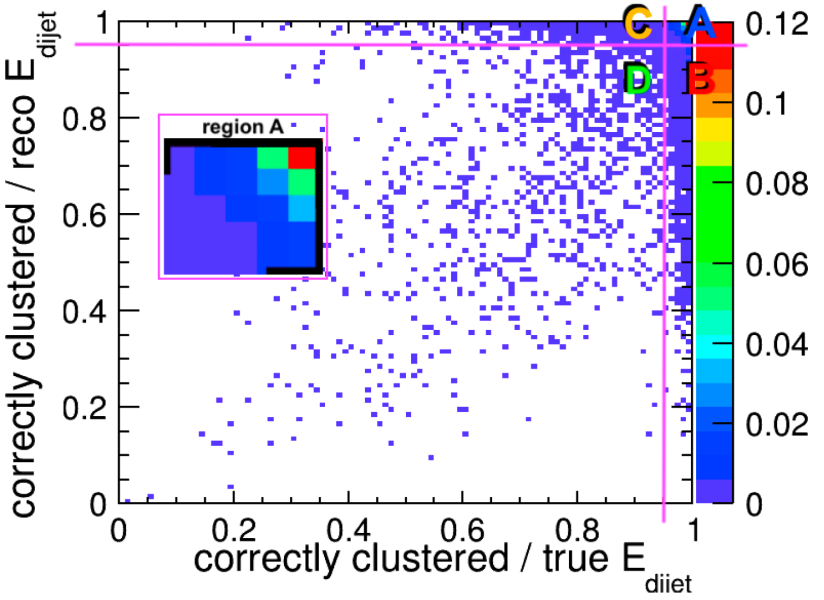
Graphs and Graph Neural Networks (GNNs)

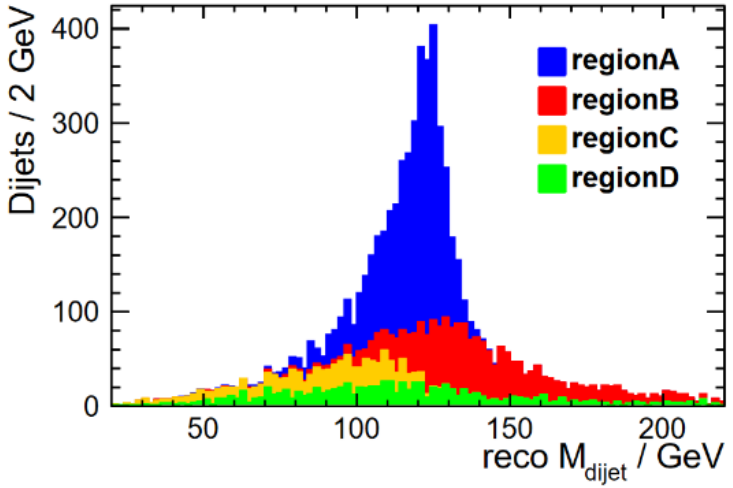
- Graph: set of
  - nodes (e.g. Particle Flow Objects (PFOs) with four-momenta as properties)
  - edges/links, i.e. connections between nodes (e.g.: two PFOs linked they belong to same cluster)
  - (global) graph attributes (e.g. is signal/background event)
- example:



Data Preparation & Comparison

- Goal: improve reconstruction by reducing misclustering
- Metric: compare what *is* and *should be* in the jet
  - ground truth given by TrueJet (uses MCRecoTruthLink)
  - **benchmark ratios** $\mathrm{r}_x = \frac{E_{intersection}}{E_{x,total}}$ (for $x = \mathrm{true}, \mathrm{reco}$)
    - intersection: PFOs which are in both the true and reconstructed dijets
  - Procedure: (for Durham and GNN)
    - each jet is matched to a *true jet* by closest angular matching
    - the jets are combined to dijets by the matching given by truejet (combined to *initial color neutrals*)

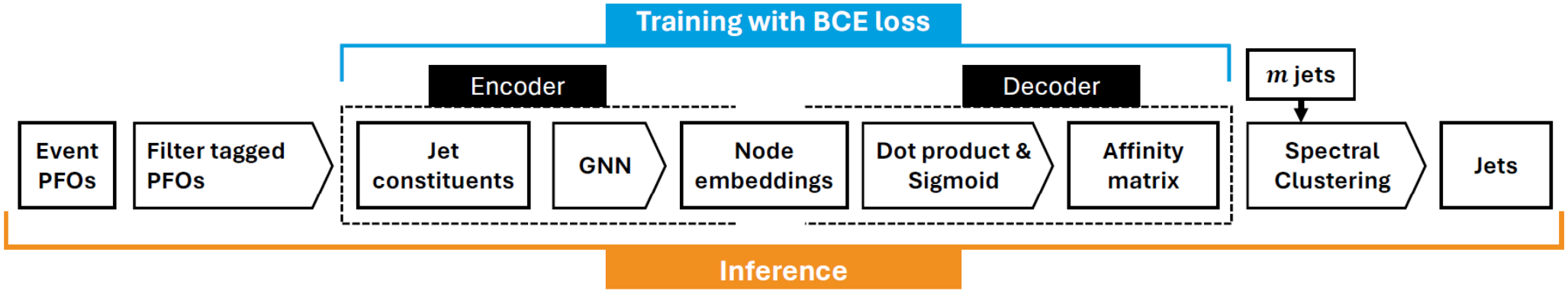




Architecture: Overall Model

Idea of the *GNN-SC* model

- *edge-scorer* based on *GNN layers* learns to calculate a *similarity* between two PFO
  here: score for whether or not they belong into the same jet (ground-truth labels given by TrueJet)
- Spectral Clustering is then used on the resulting affinity matrix to form“*jets*“



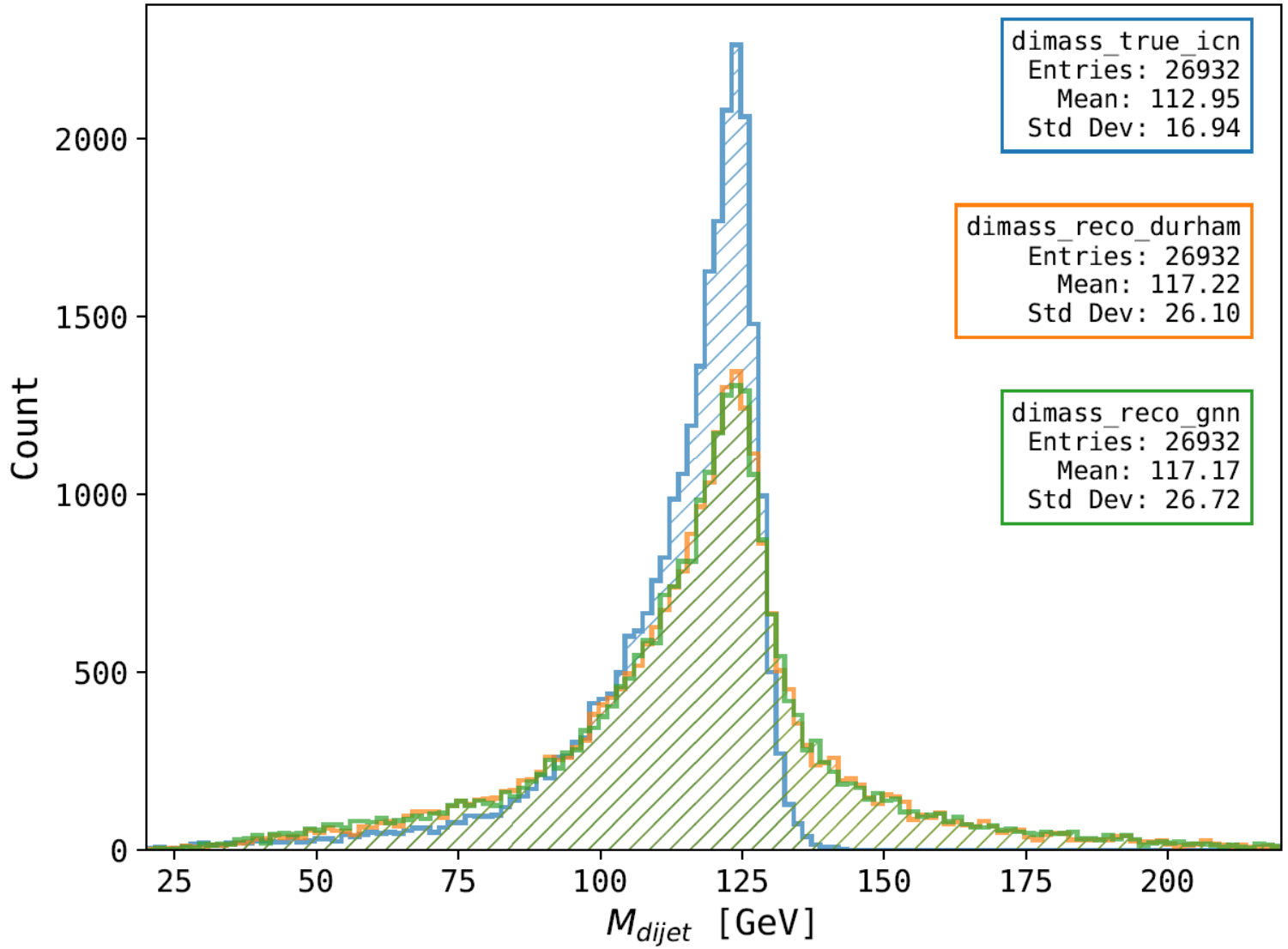
design decisions:

- **permutation invariance built in** by using pairwise dot products of node embeddings
- however, **no IR/C-safety** enforced in model
- training and hyperparam. optimization in Python, **inference possible in Marlin (JetConvProcessor)**

Data Preparation & Comparison

- Training on ZHH events $ZHH \rightarrow \mu\bar{\mu}b\bar{b}b\bar{b}$
- No gluon splittings $g \rightarrow gg$ or $g \rightarrow q\bar{q}$
  (by imposing n=4 hadronic true jets)
- filter out isolated leptons and recover ISR/BS photons *before* clustering for both Durham and the GNN (iso-leptons only if they pass cut on $M_{l\bar{l}}$)
- Input information:
  - GNN : PFO four momenta
  - Durham@LCFIPlus : PFO four momenta + vertexing information
  - Durham only : PFO four momenta

Results: Dijet mass reconstruction



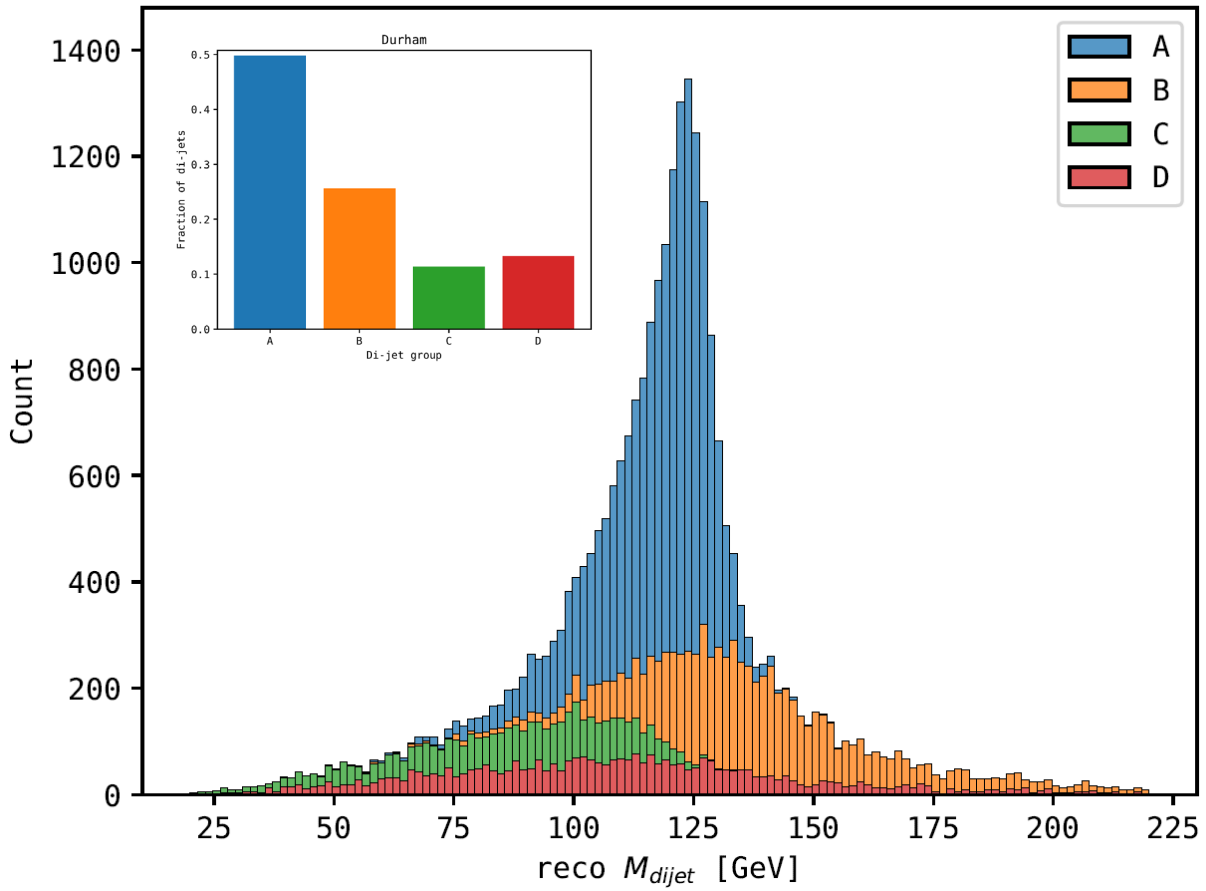
TrueJet ICNs (cheated clusters)

Durham

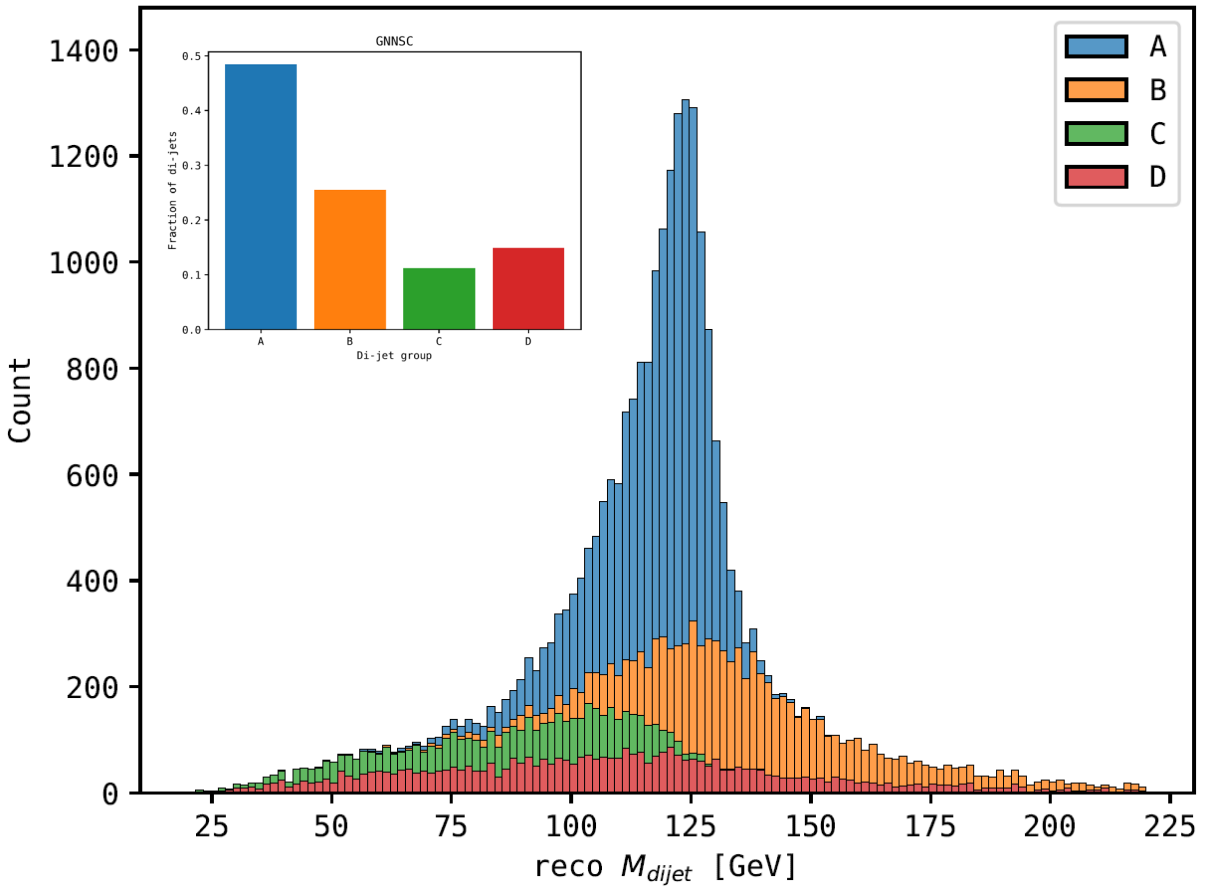
GNN-SC model NEW

Results: Dijet mass reconstruction / Misclustering

Durham

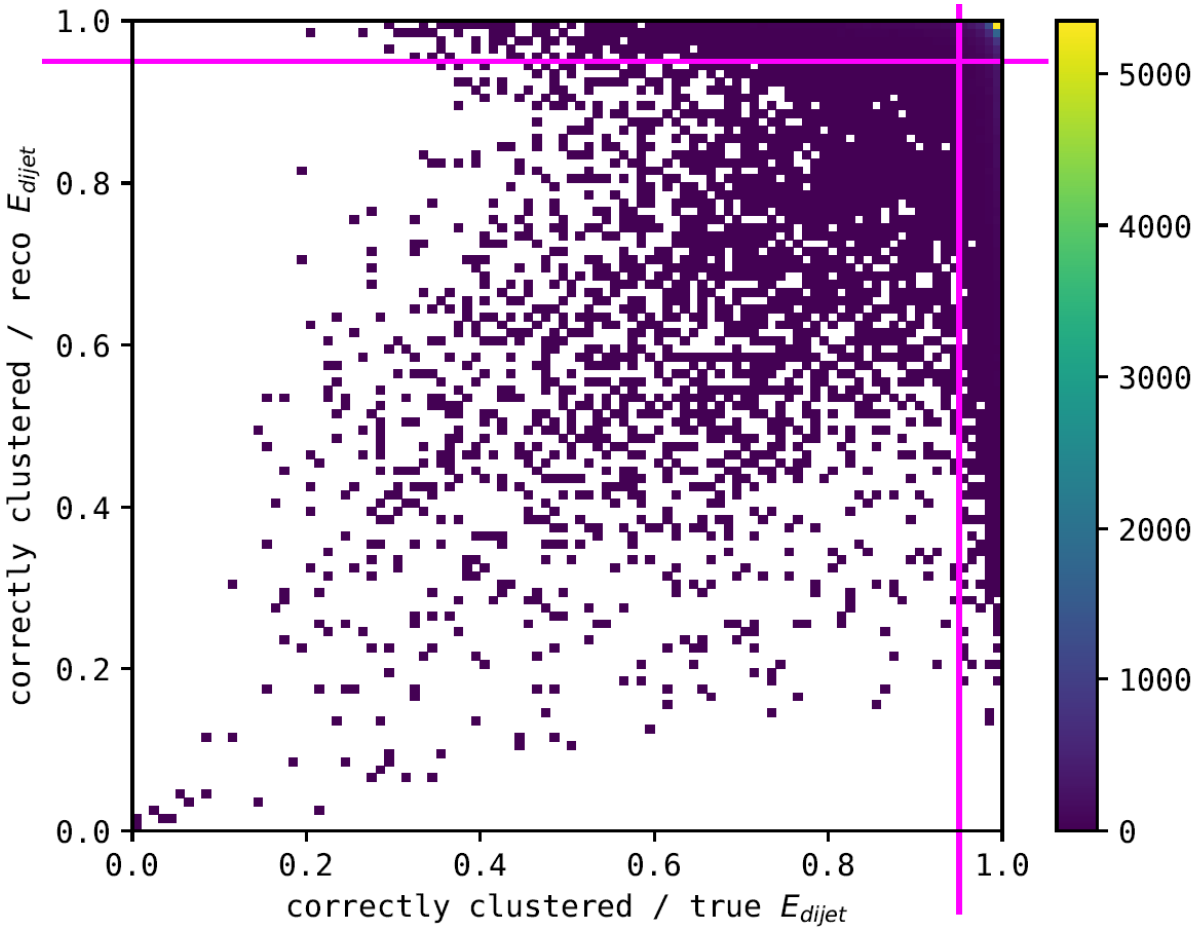


GNN-SC

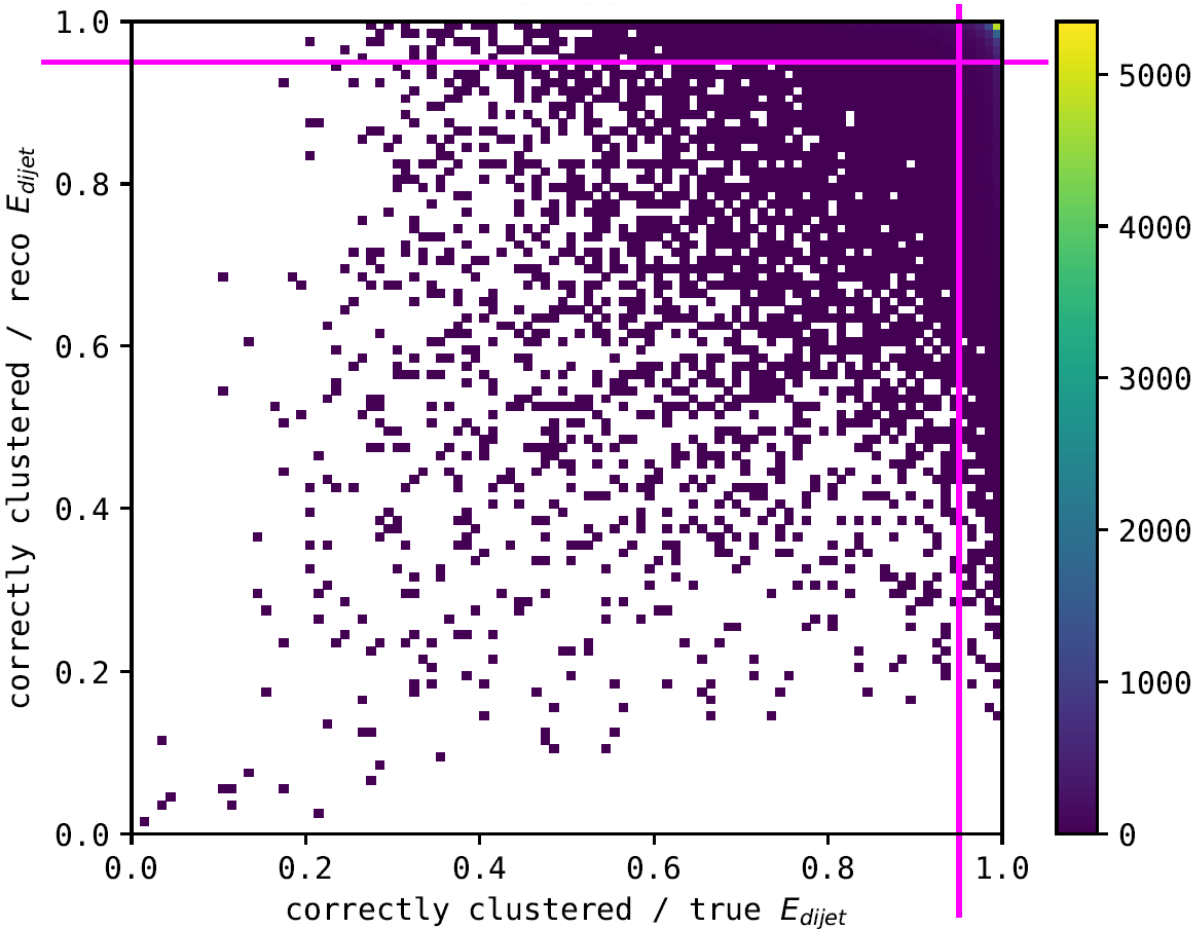


Results: Misclustering

Durham

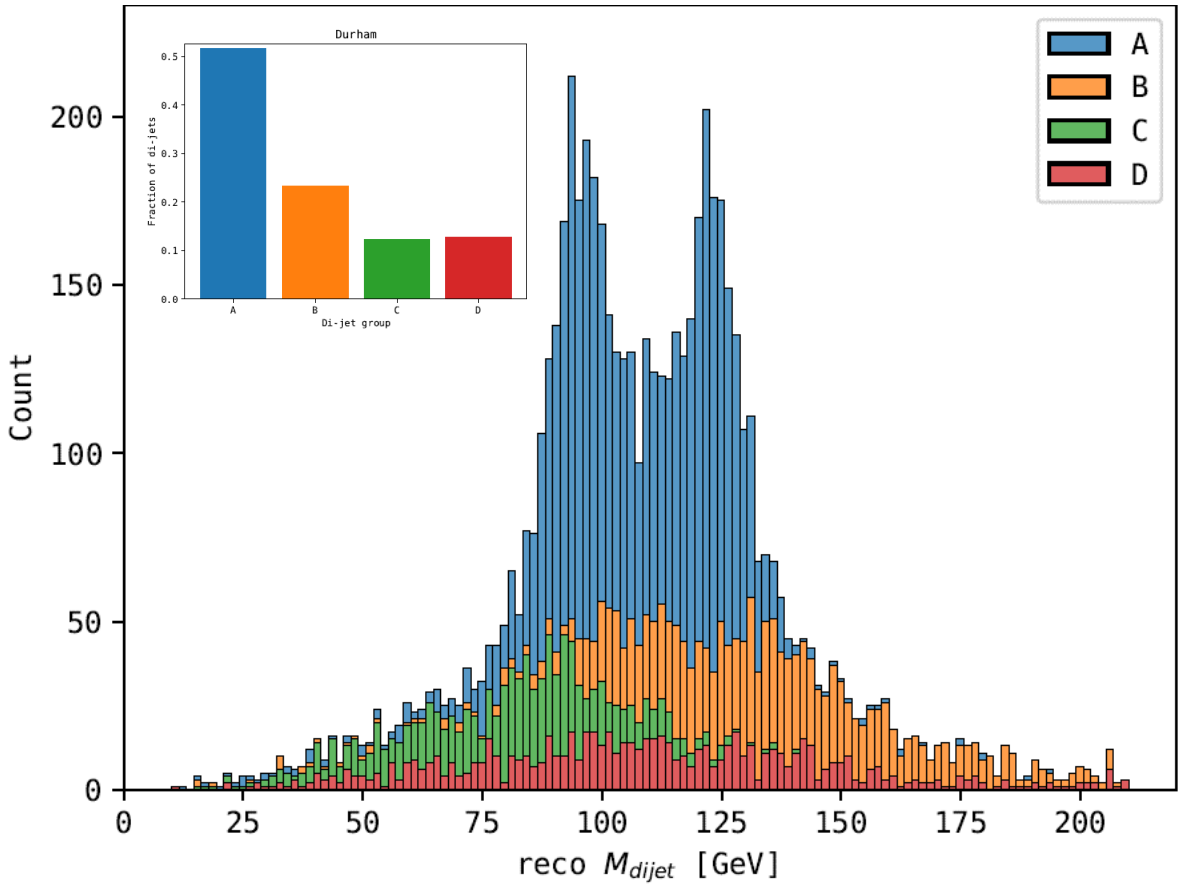


GNN-SC

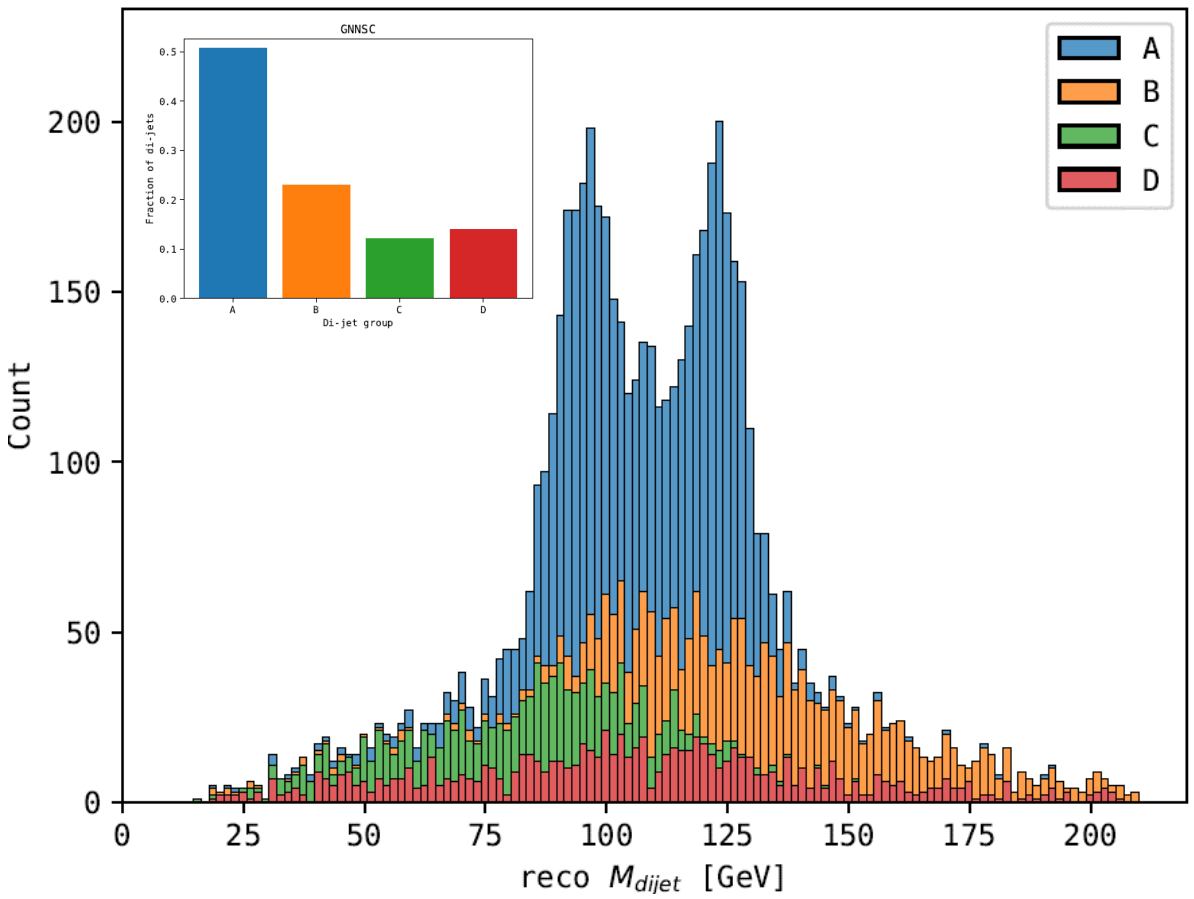


Results on ZZH events

Durham

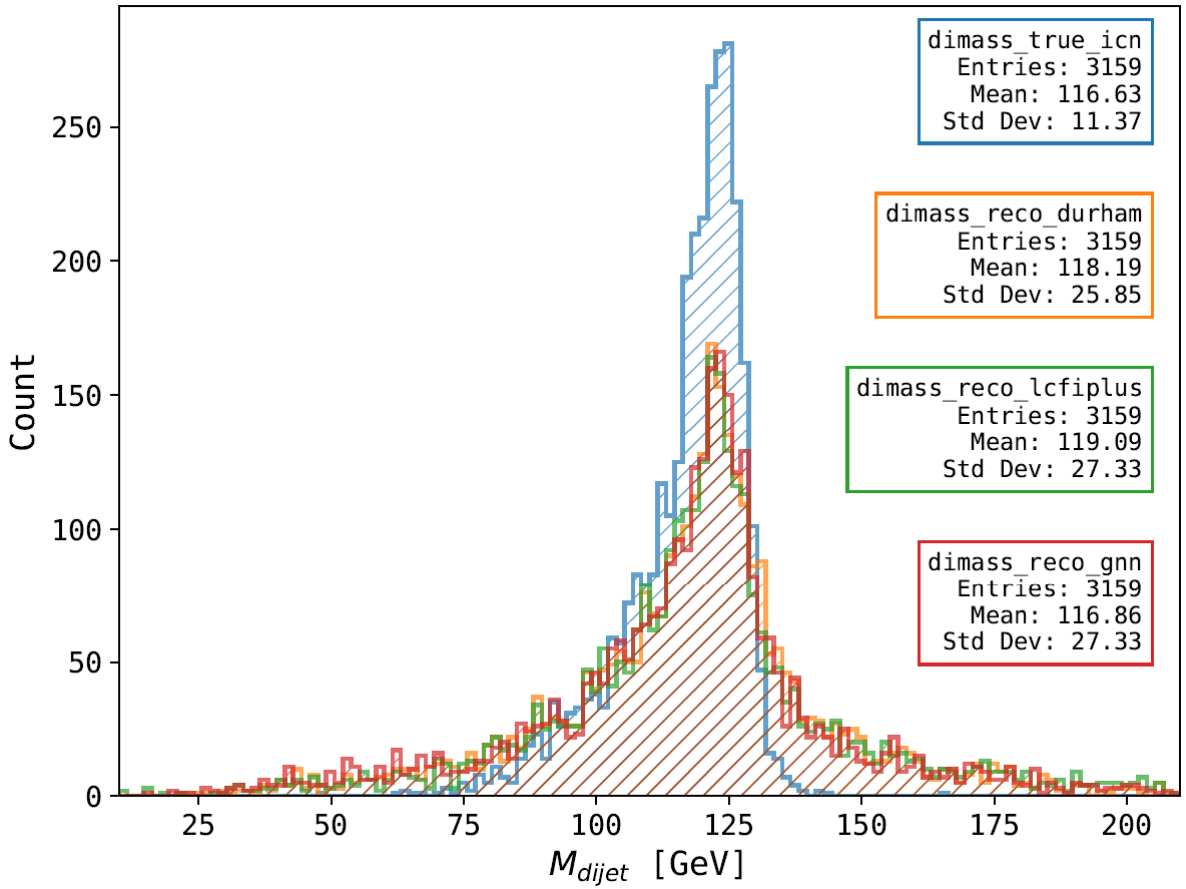


GNN-SC

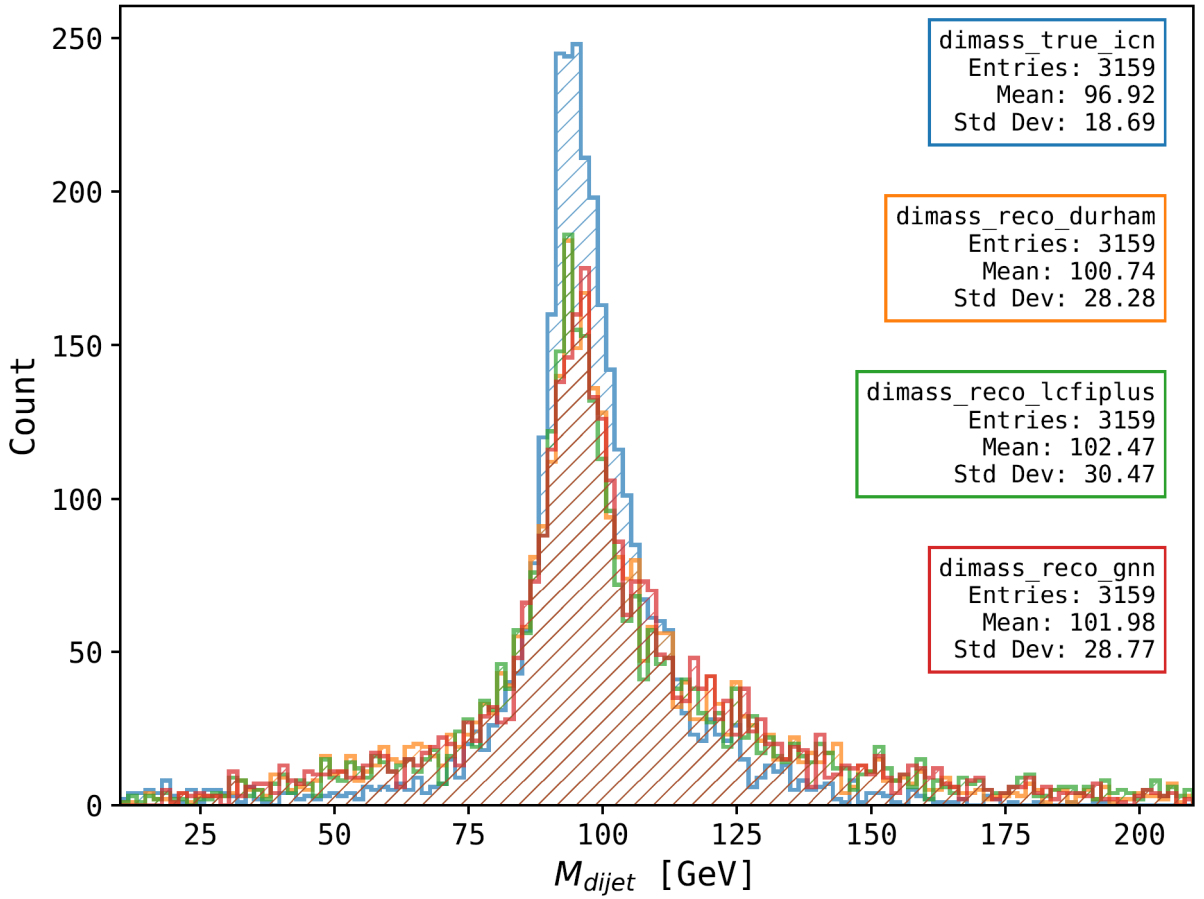


Dijet reconstruction separately in ZZH

H reconstruction



Z reconstruction

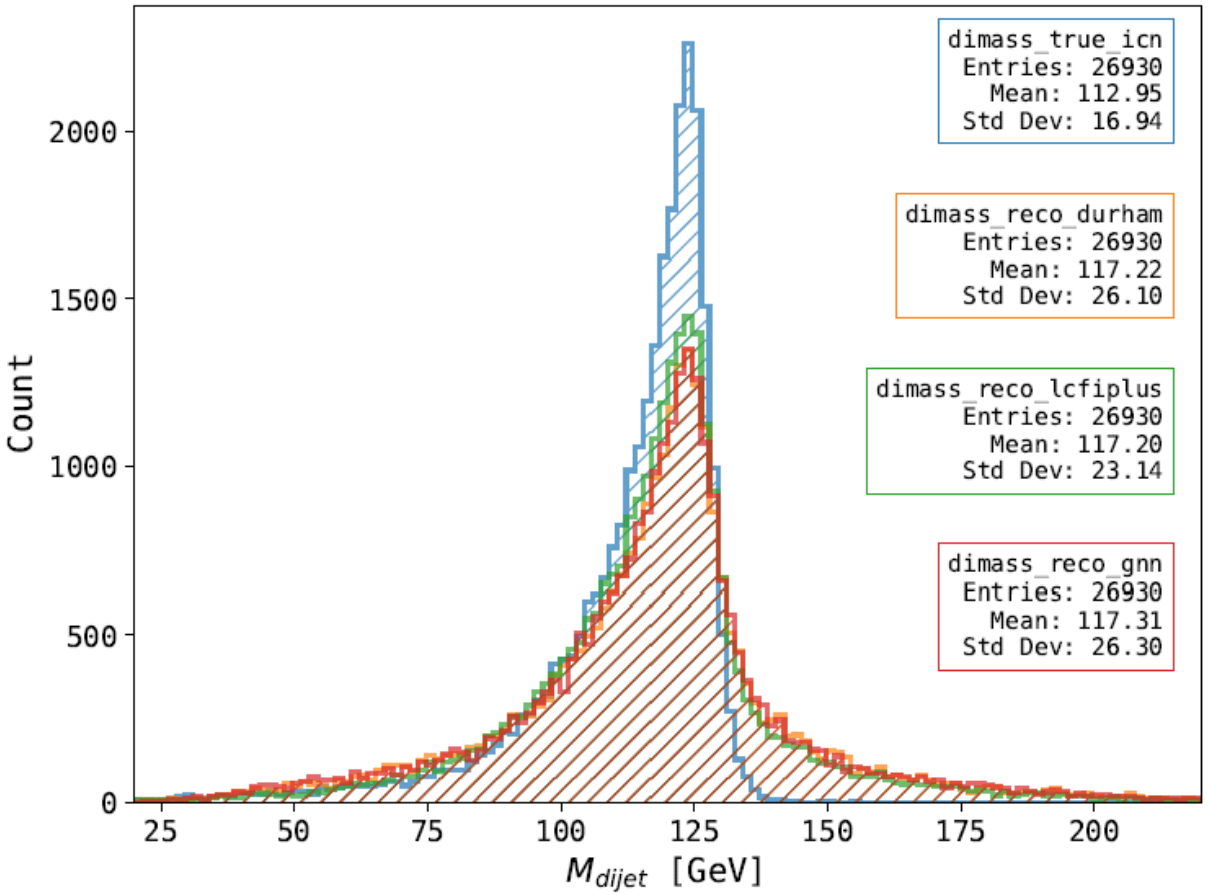


Weighting edges during training by energy

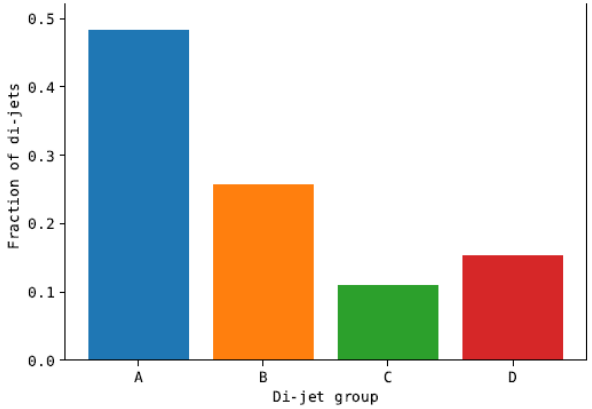
- Idea: weight w_{ij} for each edge between PFOs i, j
- Many possibilities for energy-weighting (GNNSC-EW model):
  - **Here**: normalized geometric mean: $w_{ij} = \frac{\sqrt{E_i^2 + E_j^2}}{A}$
  - Durham-motivated: $2\min\{E_i, E_j\} \cdot (1 - \cos\theta_{ij})$

Energy weighted training

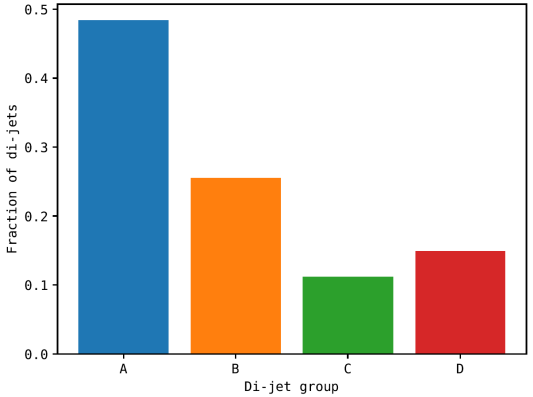
GNNSC-EW



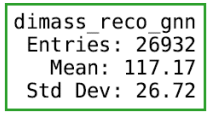
GNNSC-EW



GNNSC



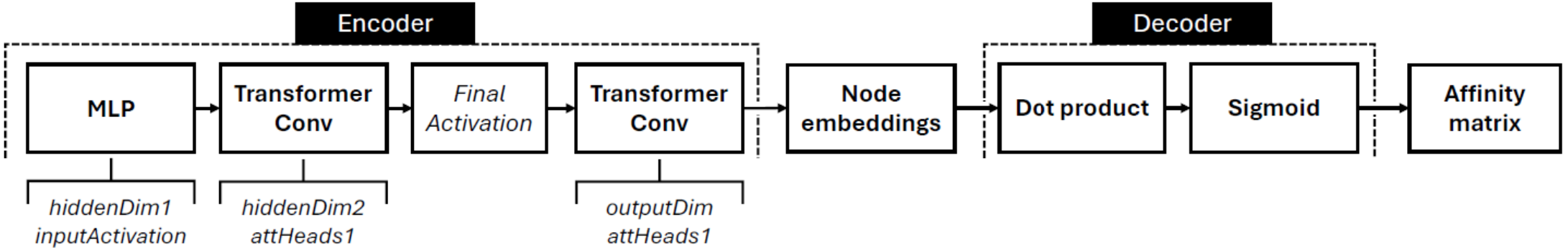
Slightly sharper mass peak compared to before



Thank you!

Backup

Architecture: Edge Scorer



Misclustering on ZZH events