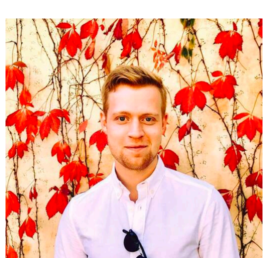
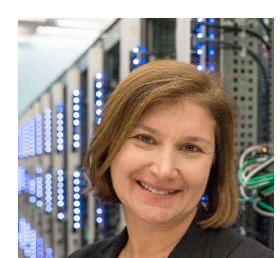
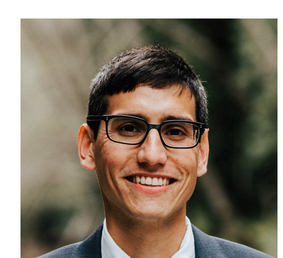
Estonian Research Council

Scalable neural network models and terascale datasets for particle-flow reconstruction

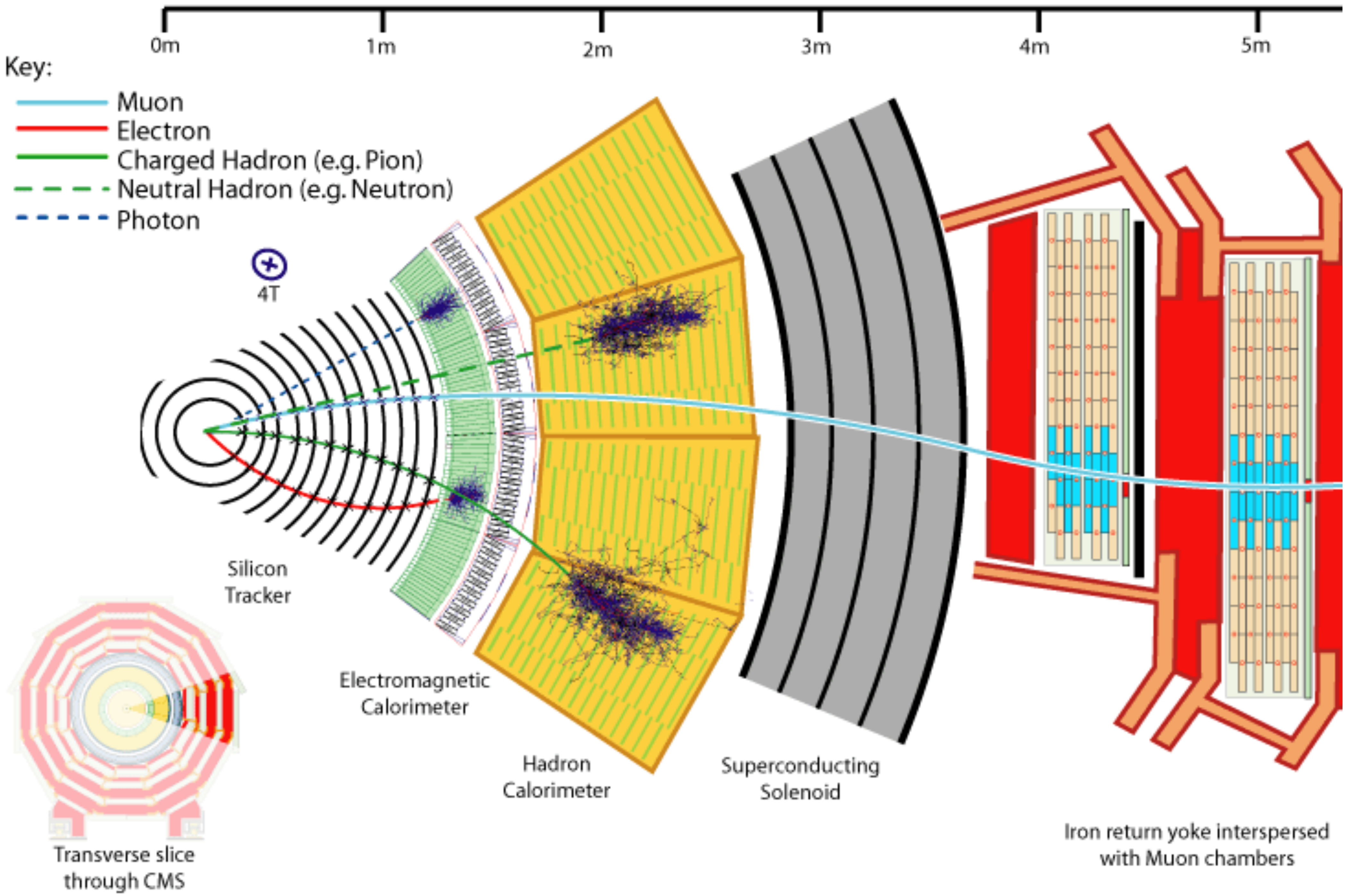
Eric Wulff (CERN), Farouk Mokhtar (UCSD), David Southwick (CERN), Mengke Zhang (UCSD), Maria Girone (CERN), Javier Duarte (UCSD), Joosep Pata (KBFI)

ML4Jets @ Hamburg

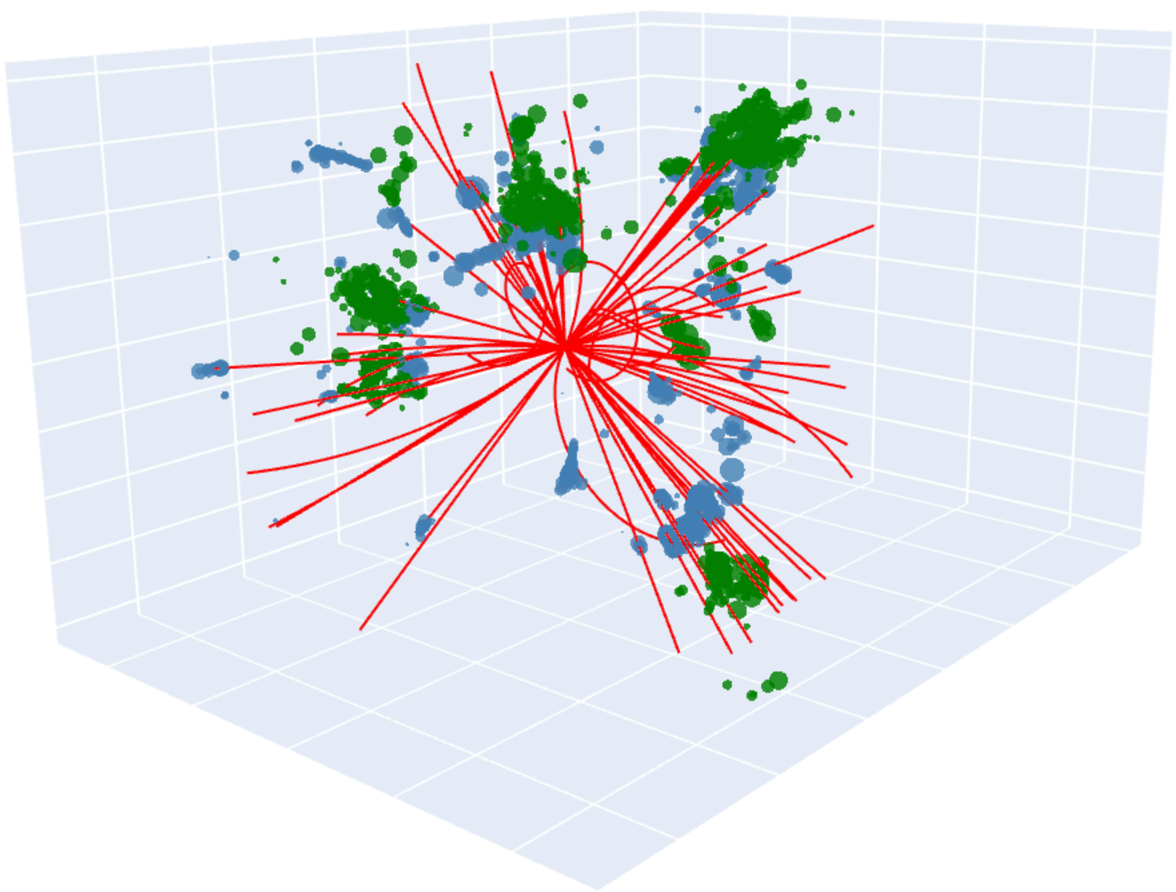
November 6, 2023

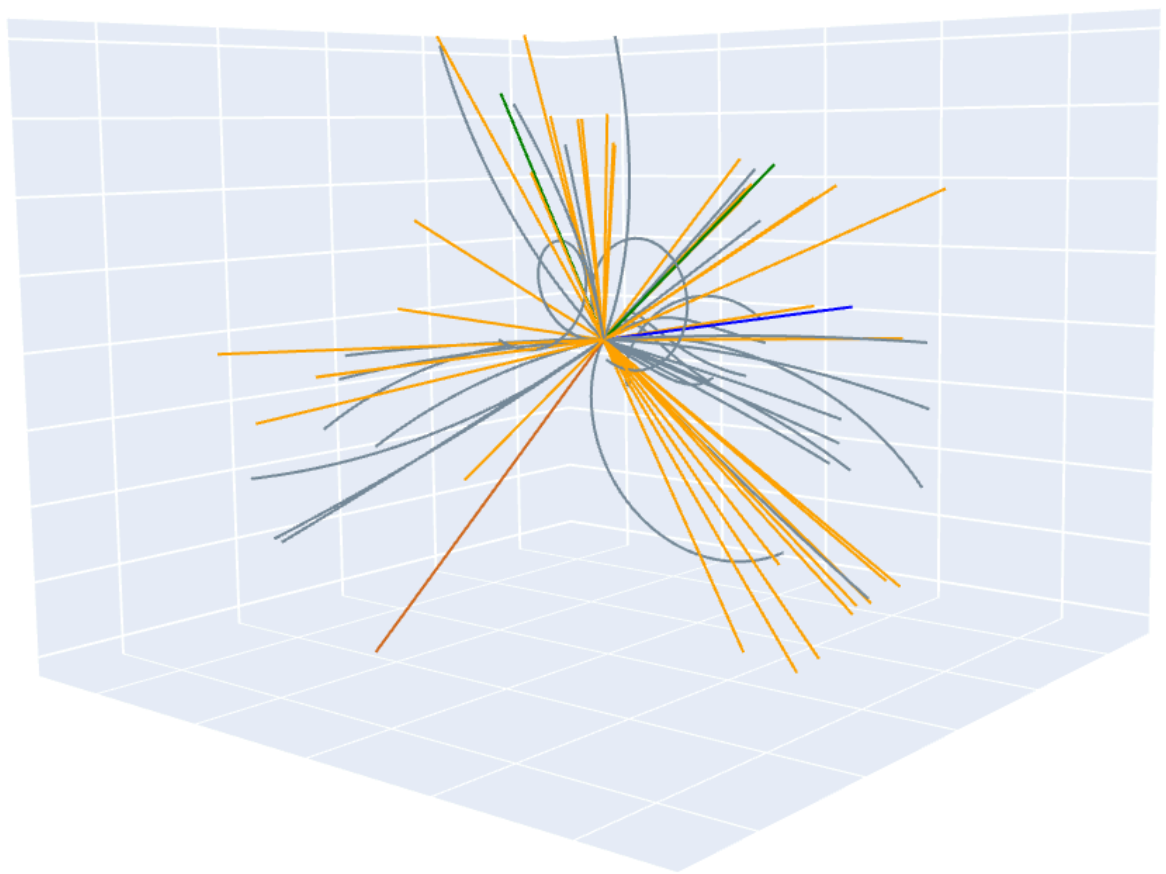
Multilayered detectors



need complex data reconstruction.

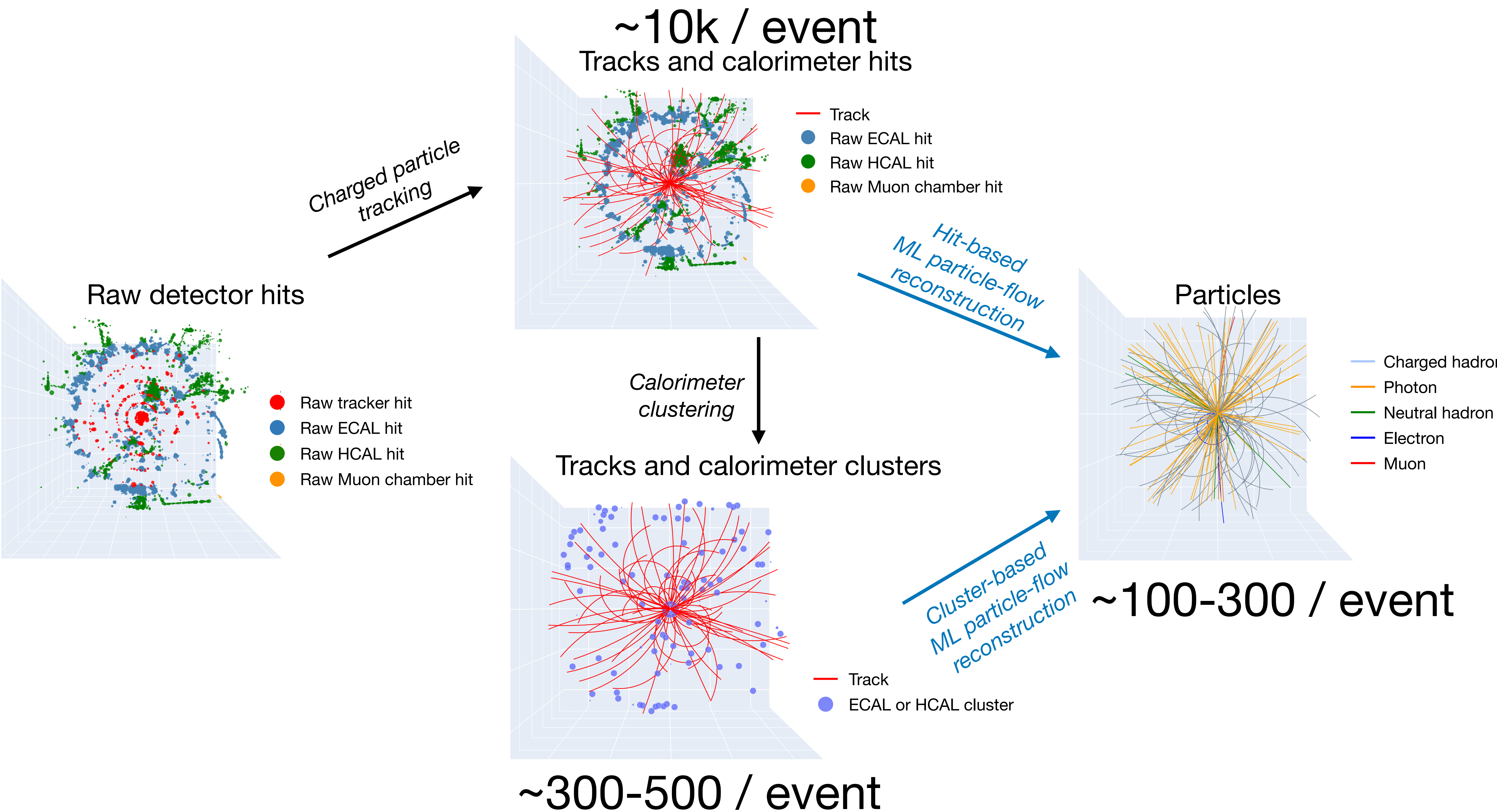


tracks and hits



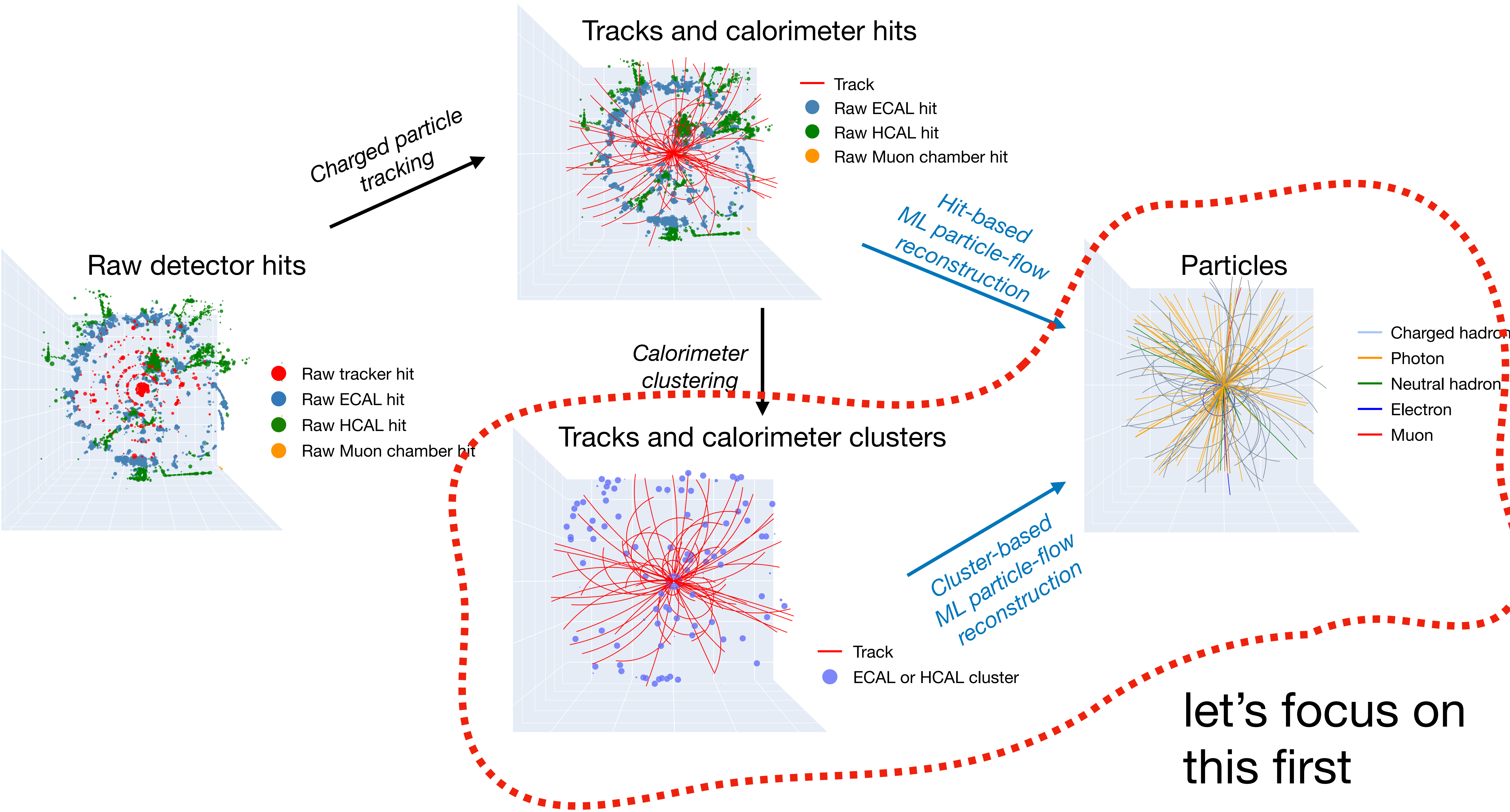
particles

We have created a new **open dataset** with Key4HEP and Geant4

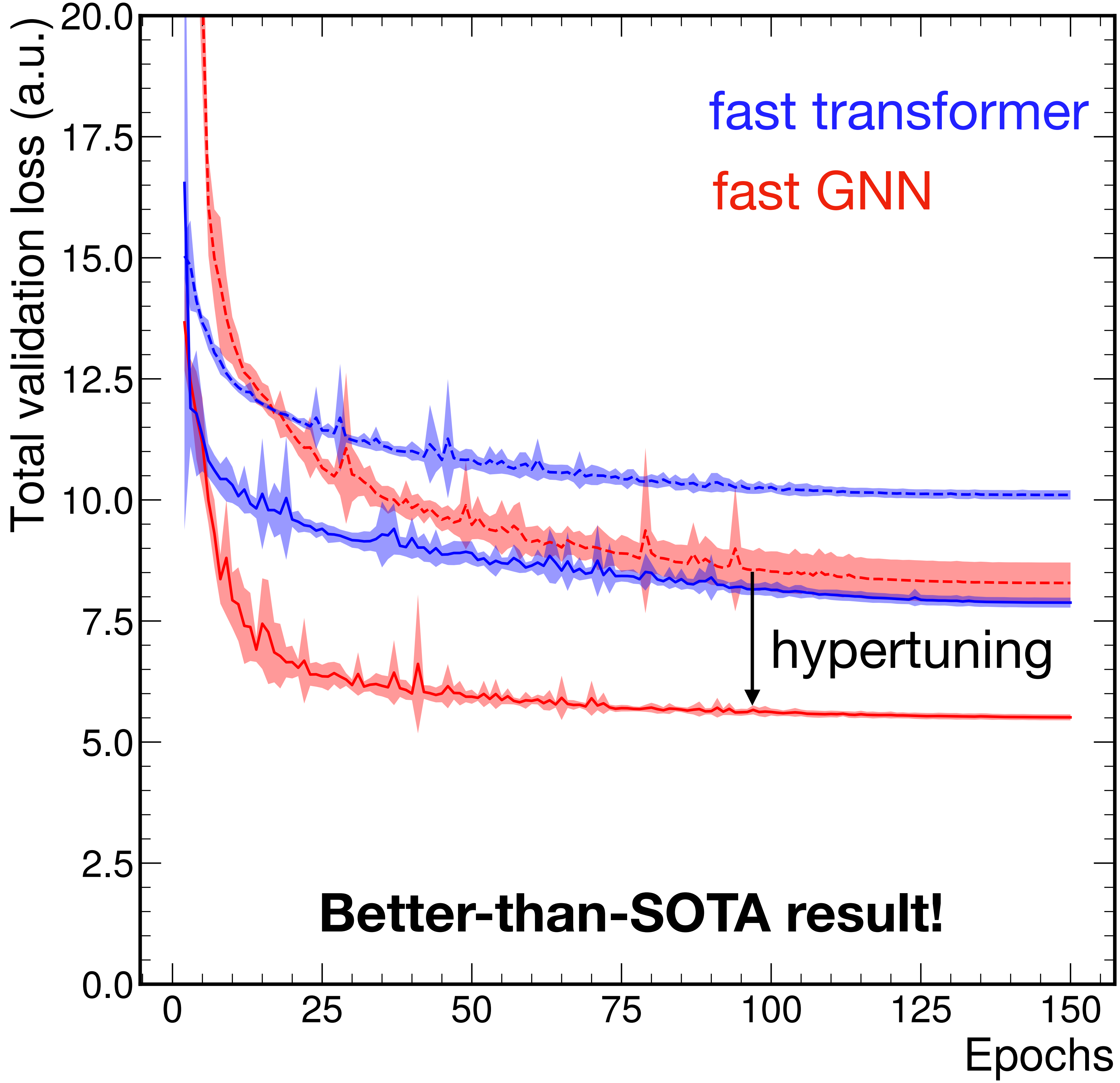


calo/tracker hits, tracks and calo clusters, baseline & target particles

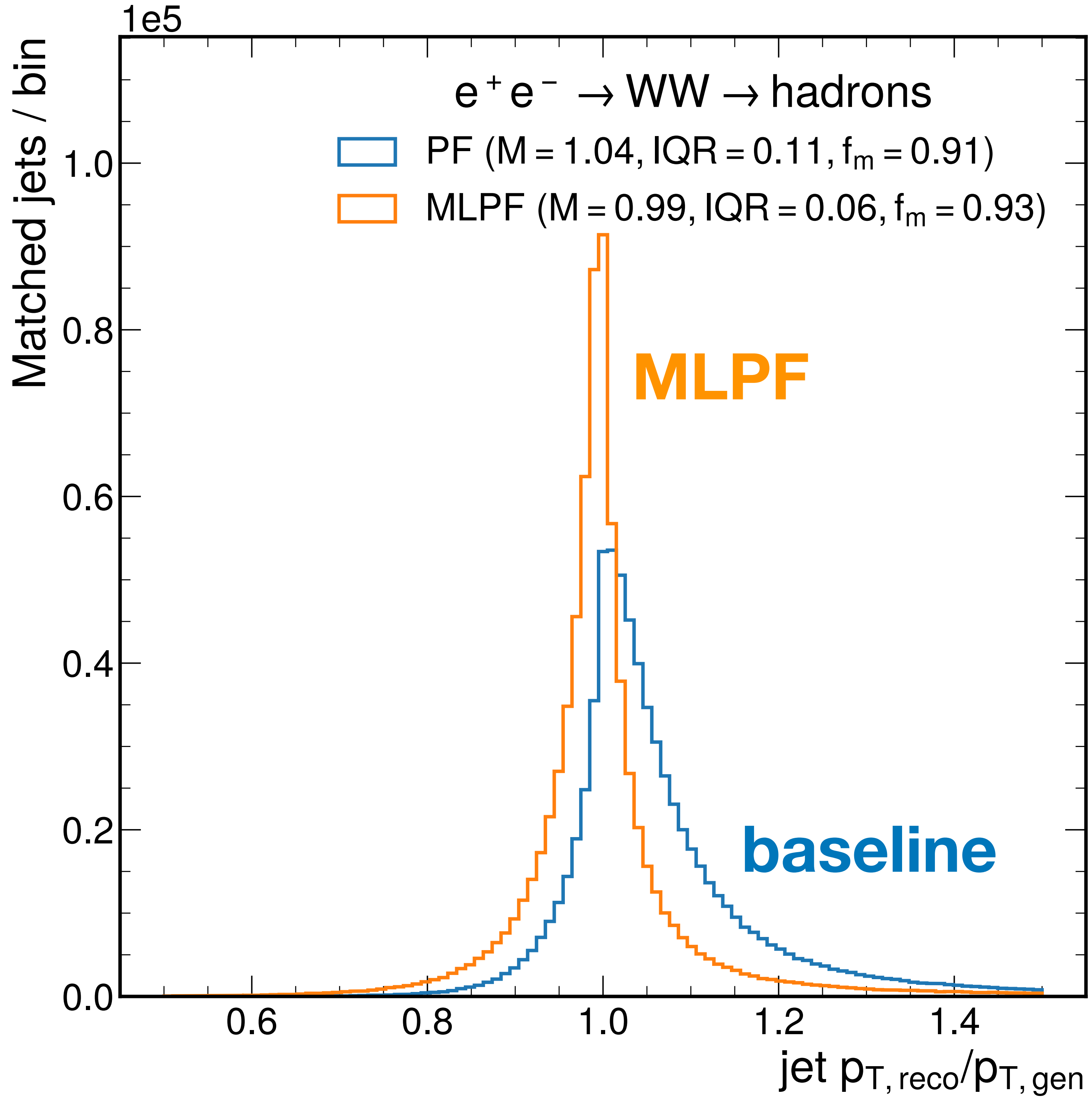
MLPF(tracks and clusters) → particles



Extensive hyperparameter tuning and model comparison…

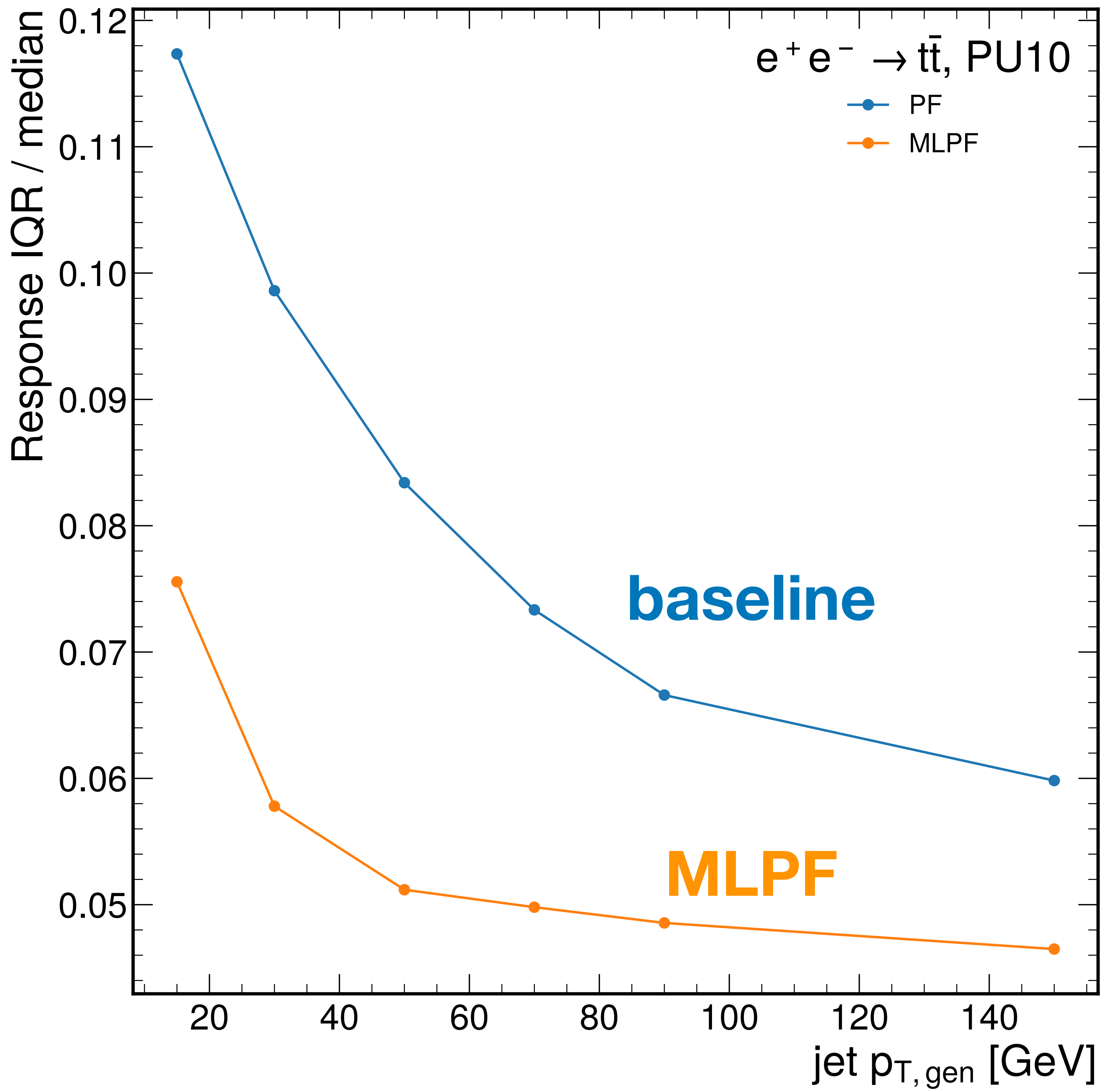


In samples never used in training…



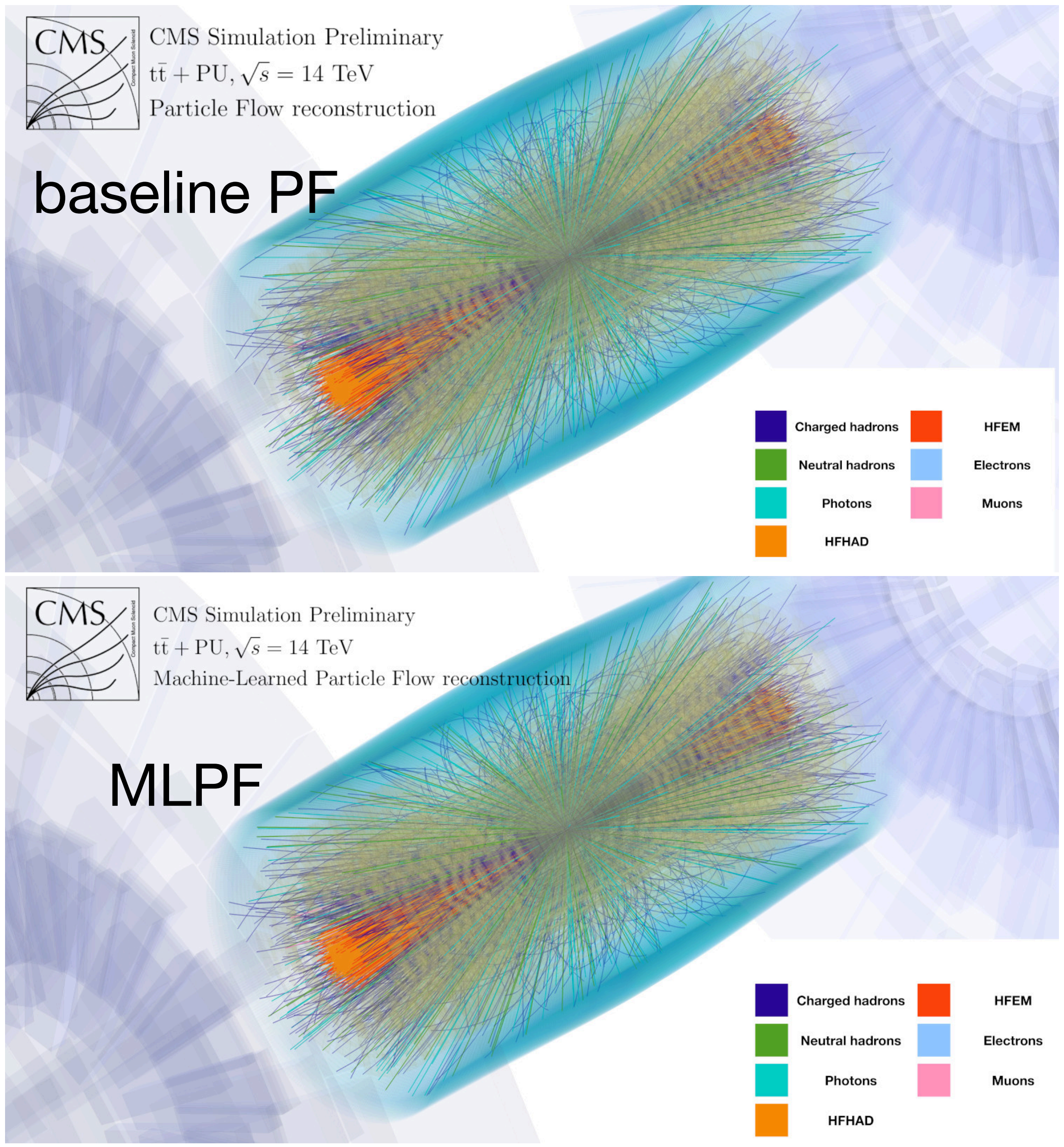
almost 50% improvement in jet response width over the baseline

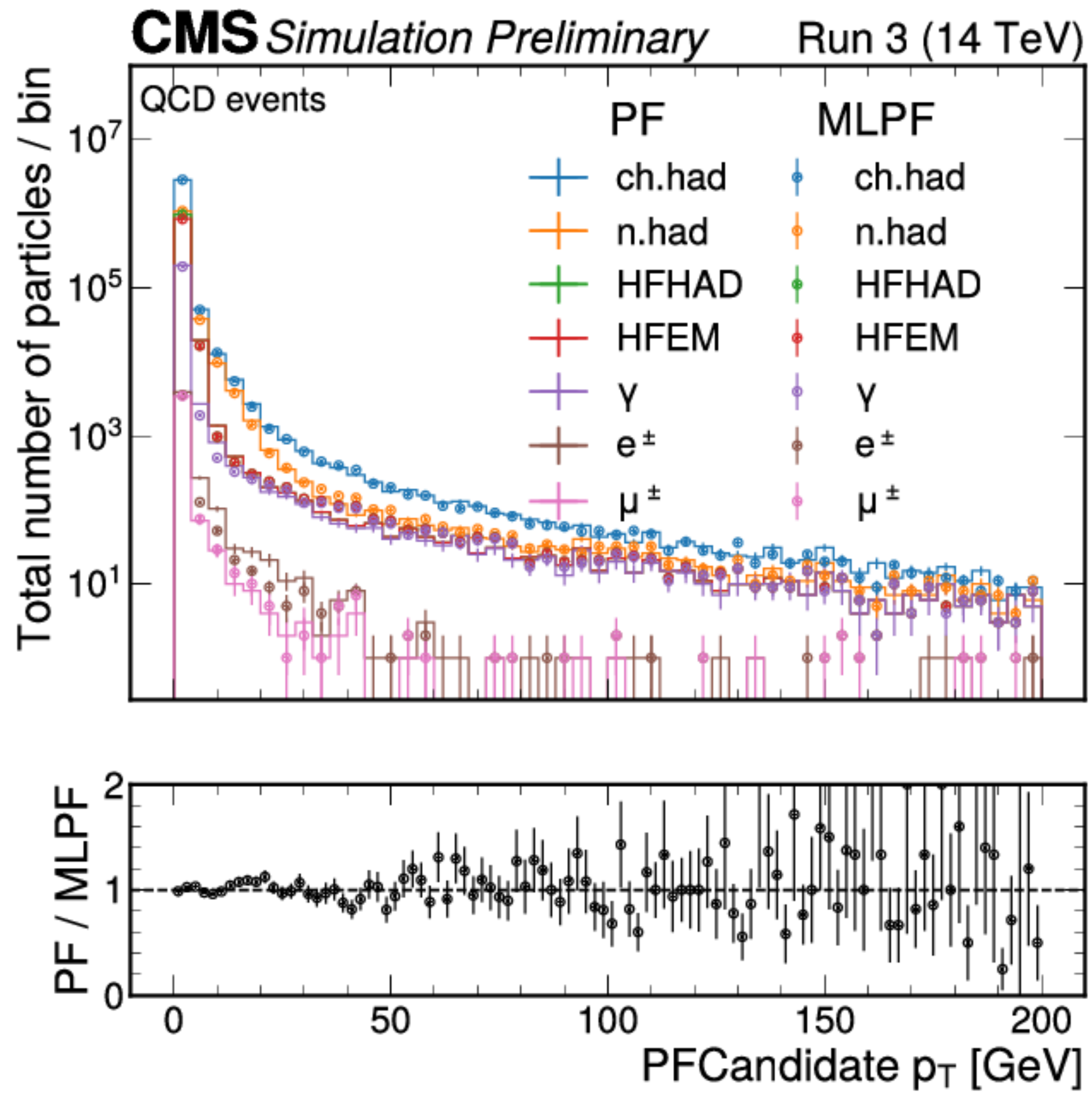
In samples never used in training…



a consistent improvement over the full pT spectrum

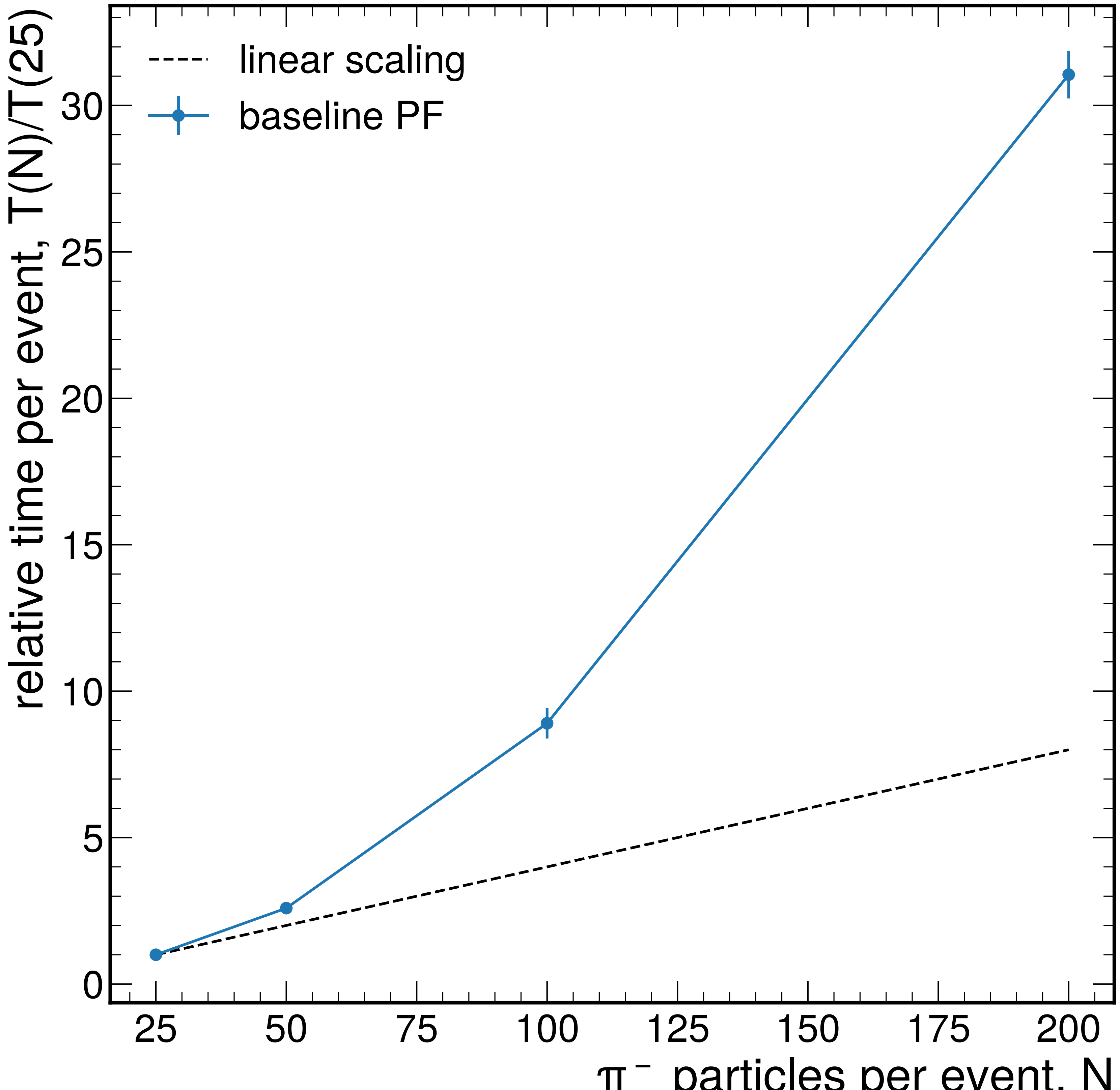
Also tested in a real detector (2022), now in the process of updating



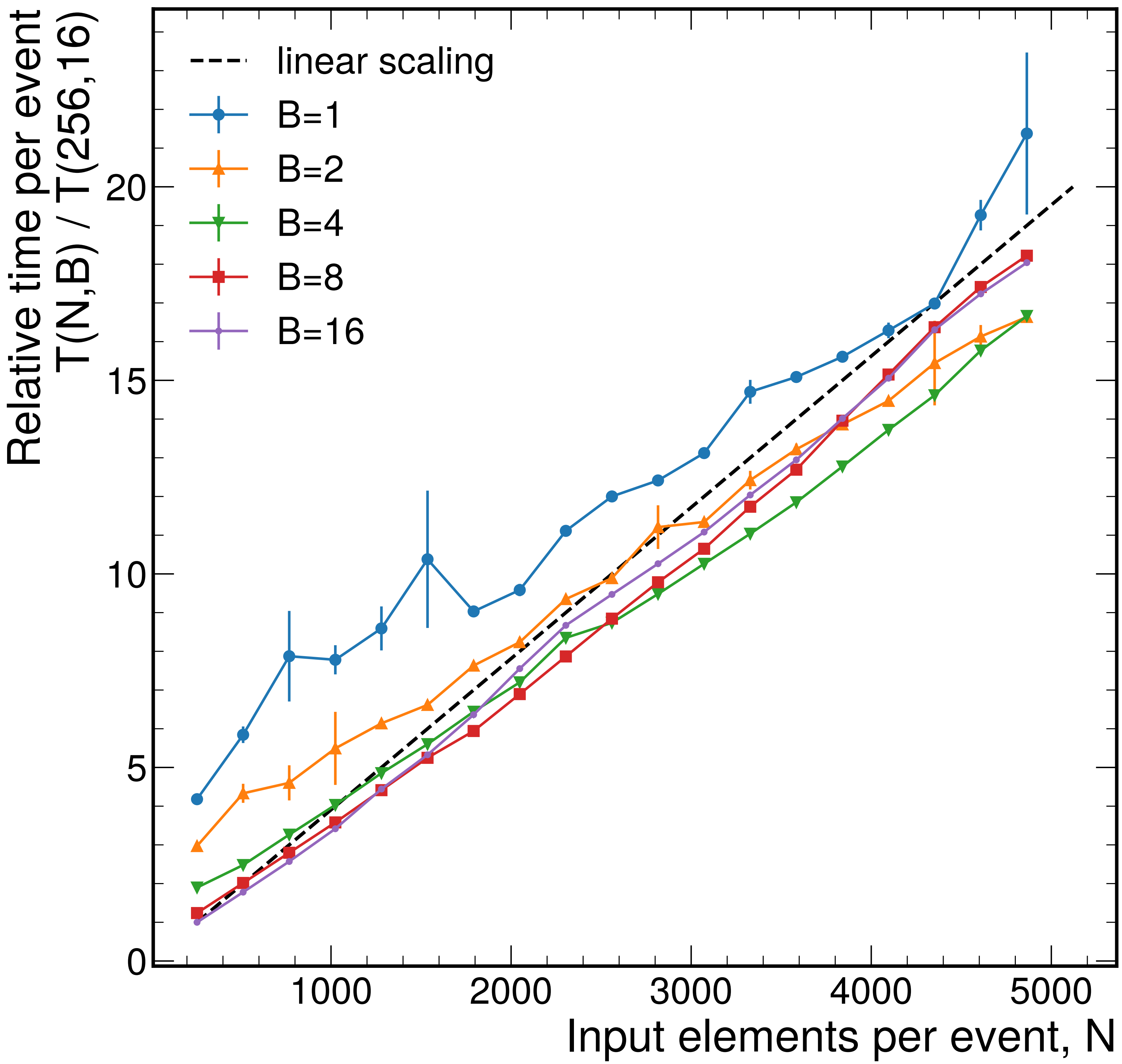


JP, Javier Duarte, Farouk Mokhtar, Eric Wulff, Jieun Yoo, Jean-Roch Vlimant, Maurizio Pierini, Maria Girone. Machine Learning for Particle Flow Reconstruction at CMS. ACAT 2021. https://doi.org/10.48550/arXiv.2203.00330, http://cds.cern.ch/record/2792320

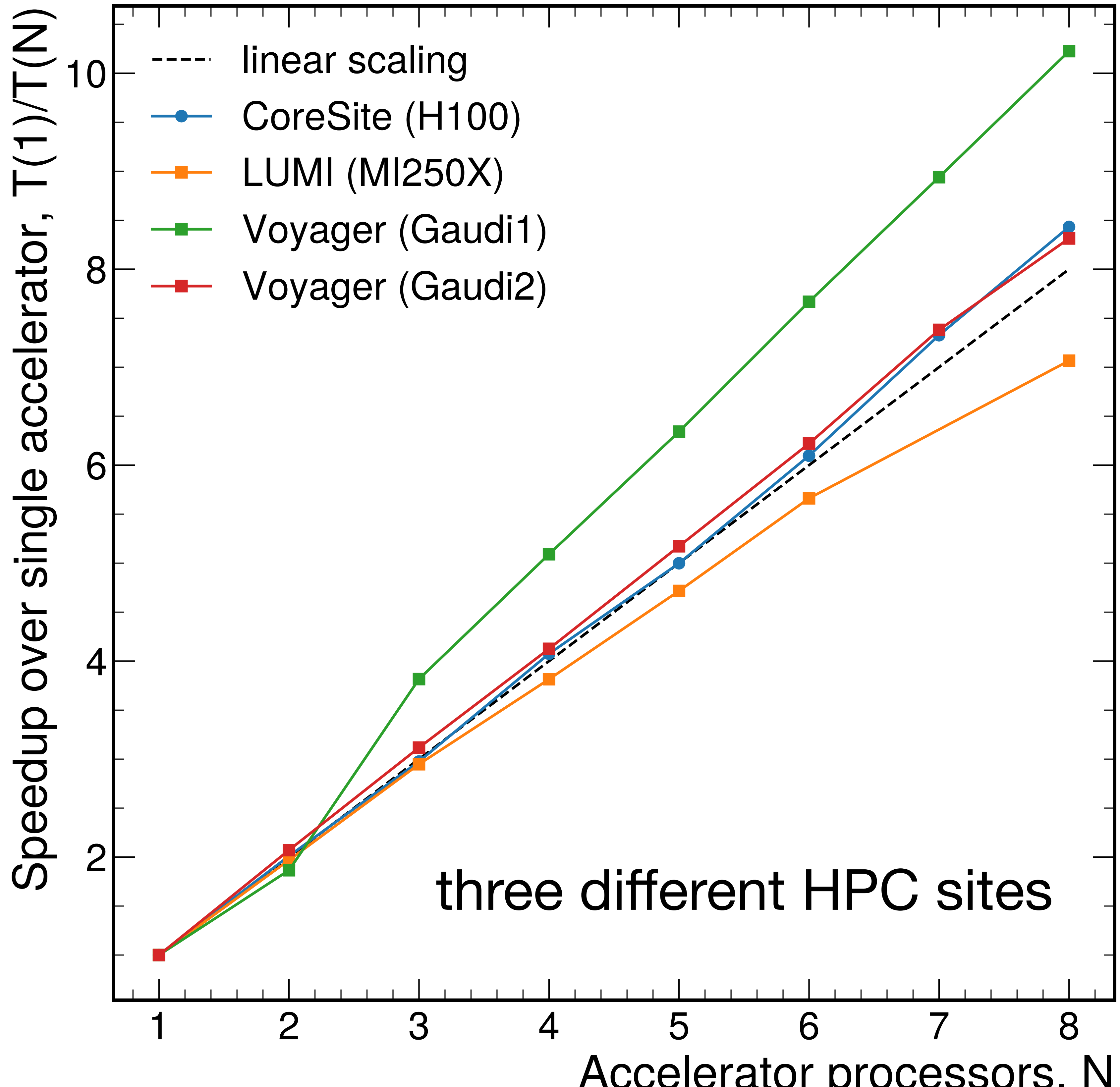
Baseline (untuned) algo runs only on CPU, scales ~quadratically, runtime per event is in seconds.

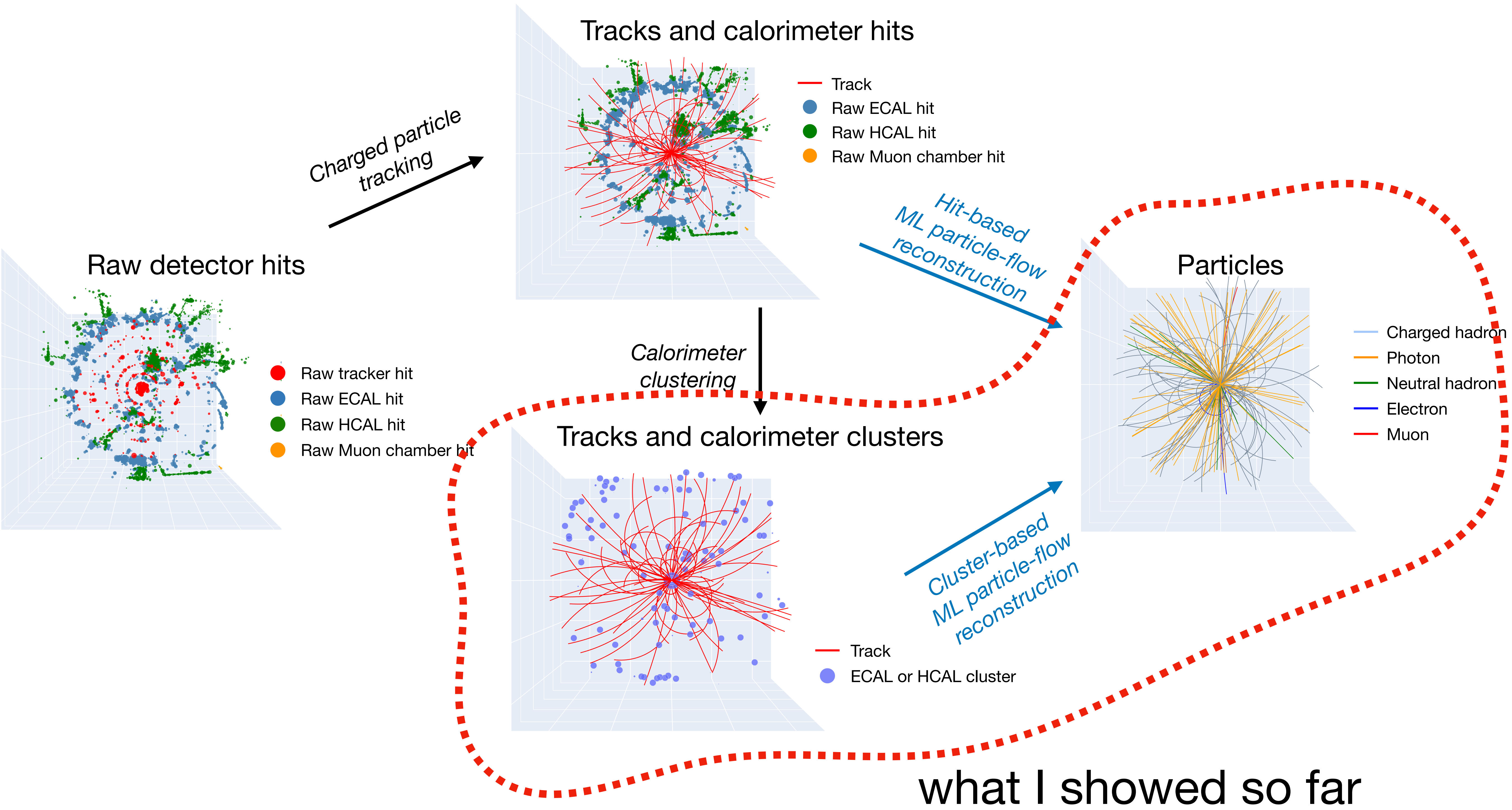


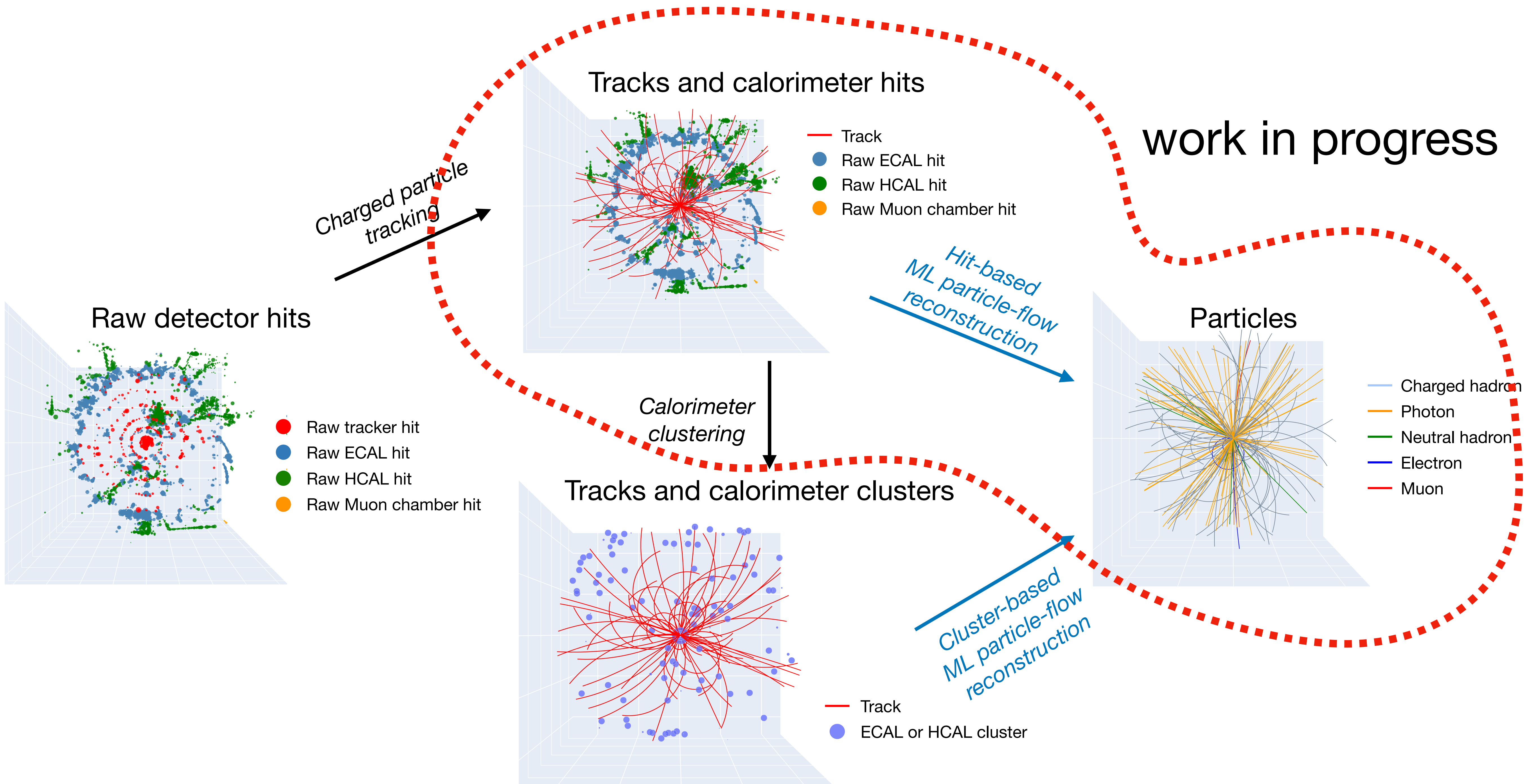
ML model scales linearly, runs in milliseconds



Portable on CPU, nVidia & AMD GPU, Intel Habana Gaudi chips

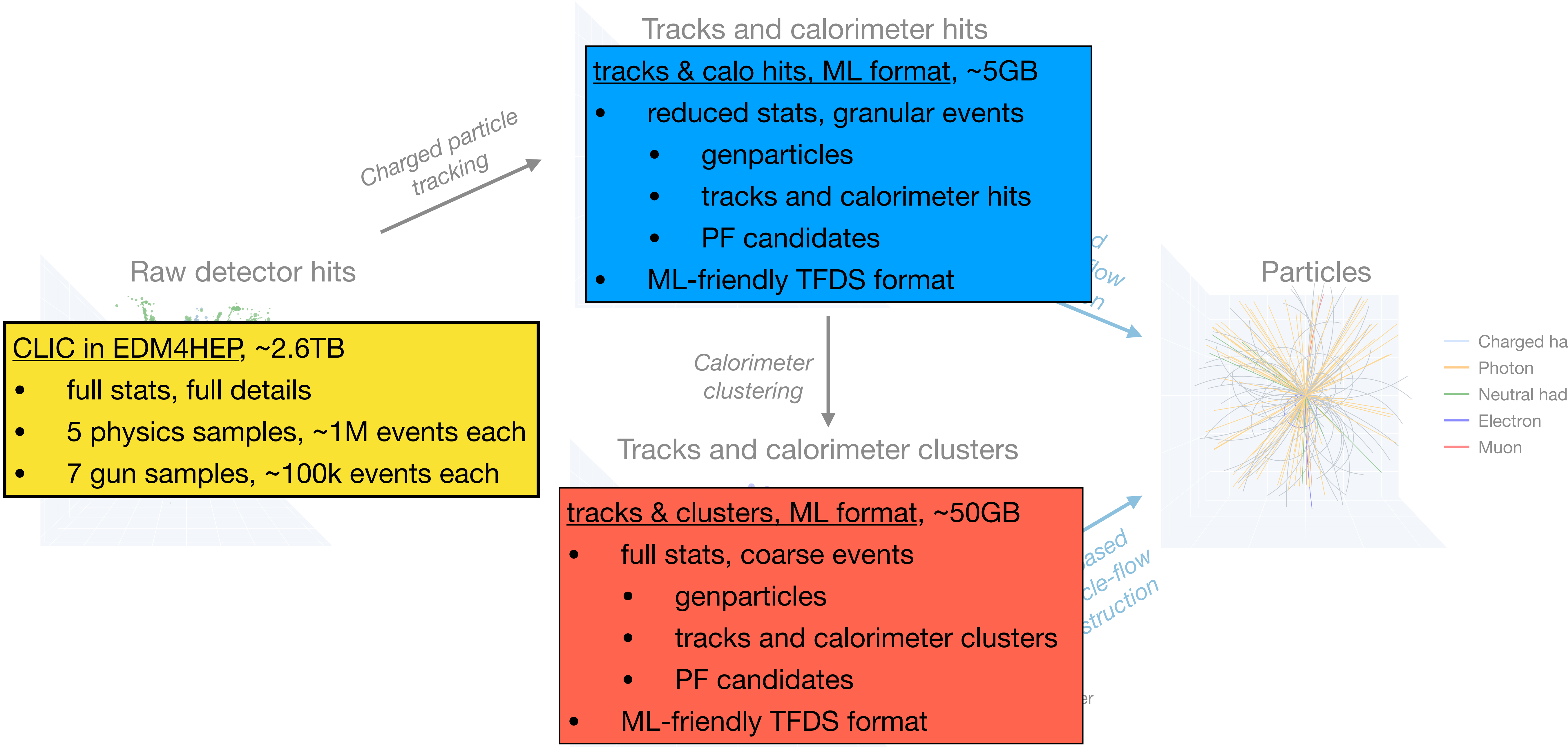



Raw detector hits
Raw tracker hit
Raw ECAL hit
Raw HCAL hit
Raw Muon chamber hit
Charged particle tracking
Tracks and calorimeter hits
Track
Raw ECAL hit
Raw HCAL hit
Raw Muon chamber hit
Hit-based ML particle-flow reconstruction
Calorimeter clustering
Tracks and calorimeter clusters
Track
ECAL or HCAL cluster
Cluster-based ML particle-flow reconstruction
Particles
Charged hadron
Photon
Neutral hadron
Electron
Muon
what I showed so far



work in progress
Raw detector hits
Raw tracker hit
Raw ECAL hit
Raw HCAL hit
Raw Muon chamber hit
Charged particle tracking
Tracks and calorimeter hits
Track
Raw ECAL hit
Raw HCAL hit
Raw Muon chamber hit
Calorimeter clustering
Tracks and calorimeter clusters
Track
ECAL or HCAL cluster
Hit-based ML particle-flow reconstruction
Cluster-based ML particle-flow reconstruction
Particles
Charged hadron
Photon
Neutral hadron
Electron
Muon

Open datasets!



[tracks & calo hits, ML format](https://doi.org/10.5281/zenodo.8414225), ~5GB

- reduced stats, granular events
    - genparticles
    - tracks and calorimeter hits
    - PF candidates
- ML-friendly TFDS format

[CLIC in EDM4HEP](https://doi.org/10.5281/zenodo.8260741), ~2.6TB

- full stats, full details
- 5 physics samples, ~1M events each
- 7 gun samples, ~100k events each

[tracks & clusters, ML format](https://doi.org/10.5281/zenodo.8409592), ~50GB

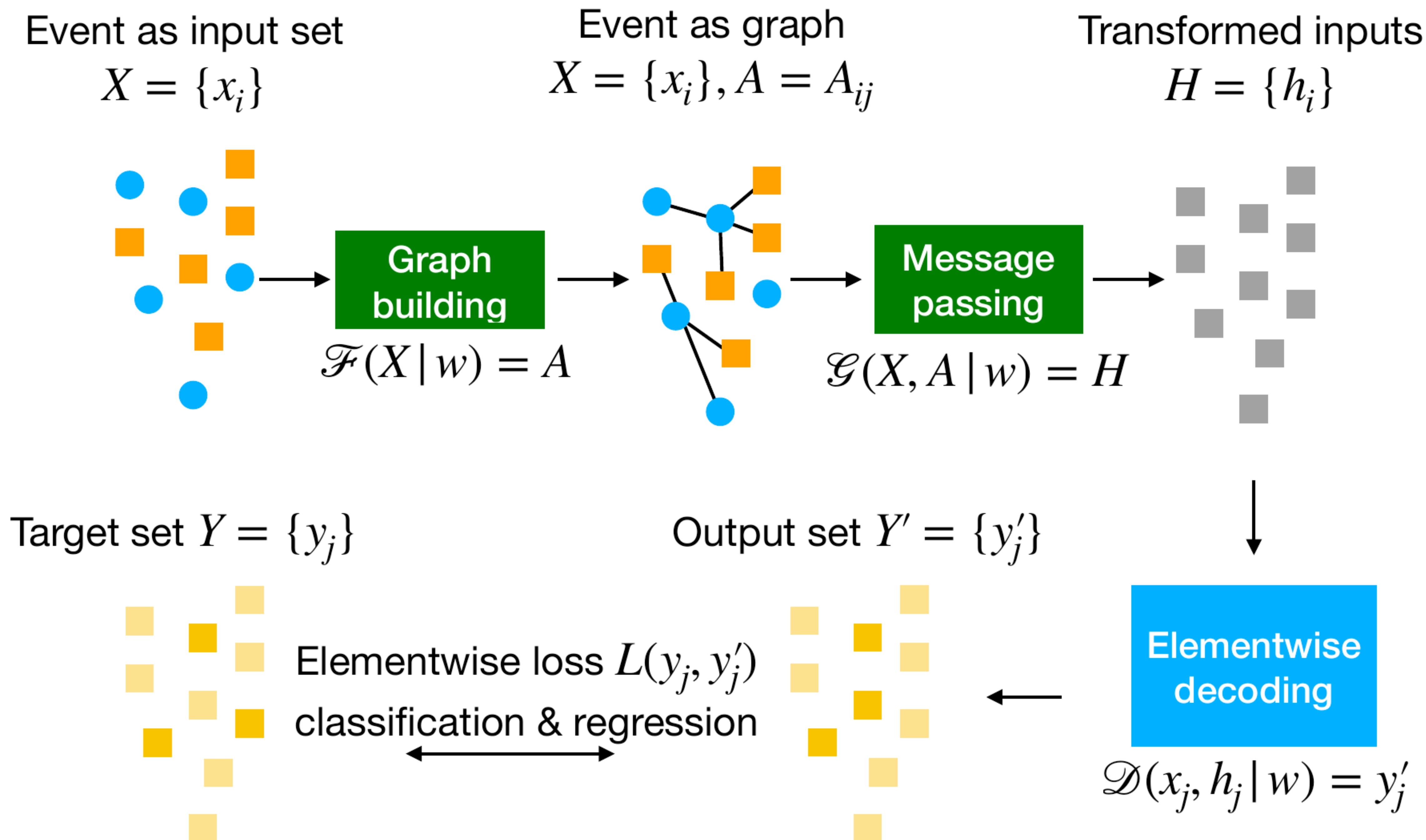
- full stats, coarse events
    - genparticles
    - tracks and calorimeter clusters
    - PF candidates
- ML-friendly TFDS format

- https://doi.org/10.5281/zenodo.8260741
- https://doi.org/10.5281/zenodo.8414225
- https://doi.org/10.5281/zenodo.8409592

Summary

- Particle flow reconstruction is a complex and interesting problem to address with end-to-end ML
- ML can improve jet/MET response significantly over a naive baseline
- Scalable ML models allow processing of full events with high throughput and portability
- Open datasets & code can accelerate research
- More granular events, updates & integration tests with a real detector on the way

Backup



$$x_i = [\mathrm{elem.type}, p_\mathrm{T}, E_\mathrm{ECAL}, E_\mathrm{HCAL}, \eta, \phi, \eta_\mathrm{outer}, \phi_\mathrm{outer}, q, \ldots]$$

$$y_j = [\mathrm{PID}, p_\mathrm{T}, E, \eta, \phi, q], \quad \mathrm{PID} \in \{\text{none, charged hadron, neutral hadron}, \gamma, e^\pm, \mu^\pm, \ldots\}$$

$$h_i \in \mathbb{R}^{N_\mathrm{hidden}}$$

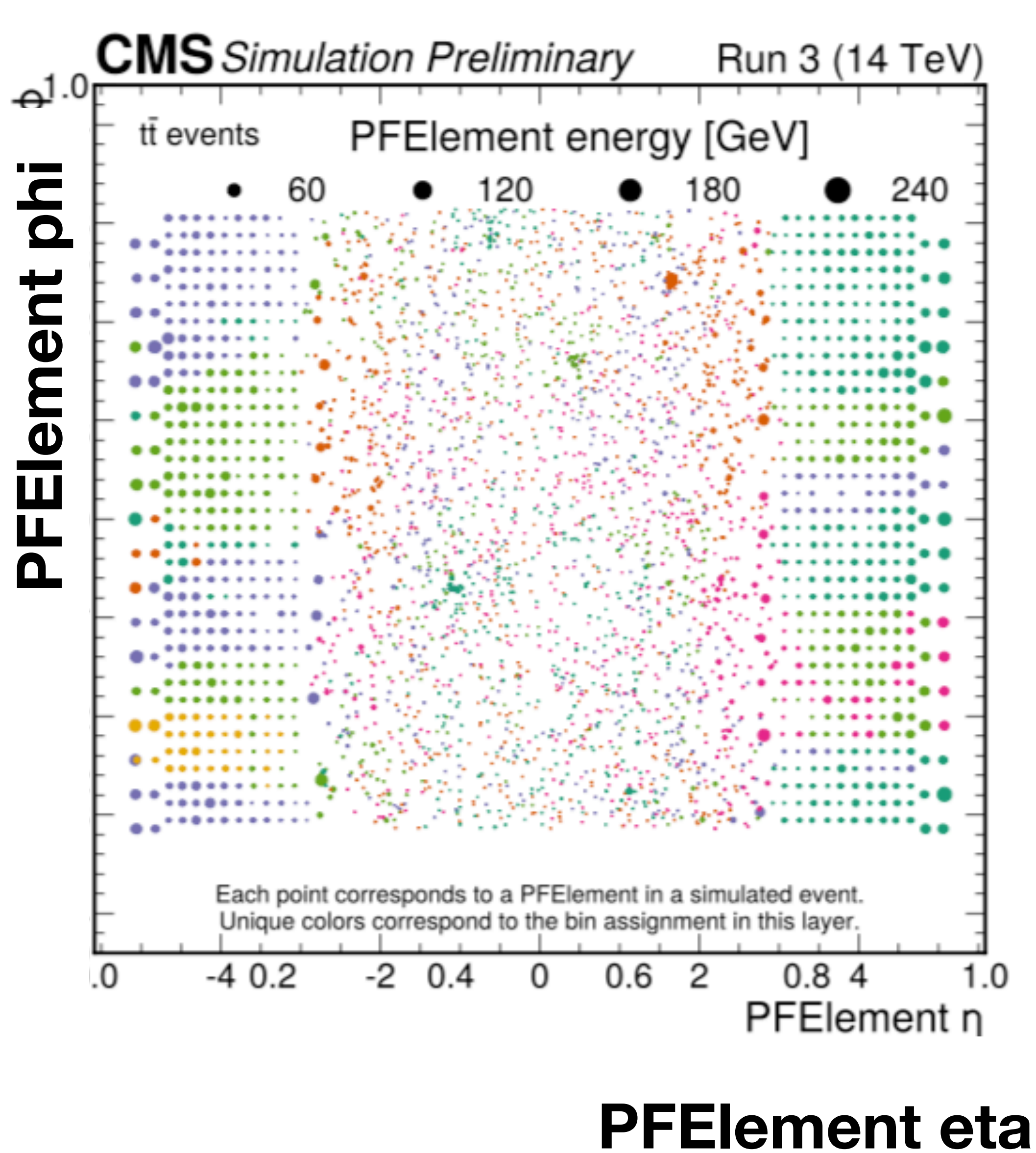
Trainable neural networks: $\mathscr{F}, \mathscr{G}, \mathscr{D}$

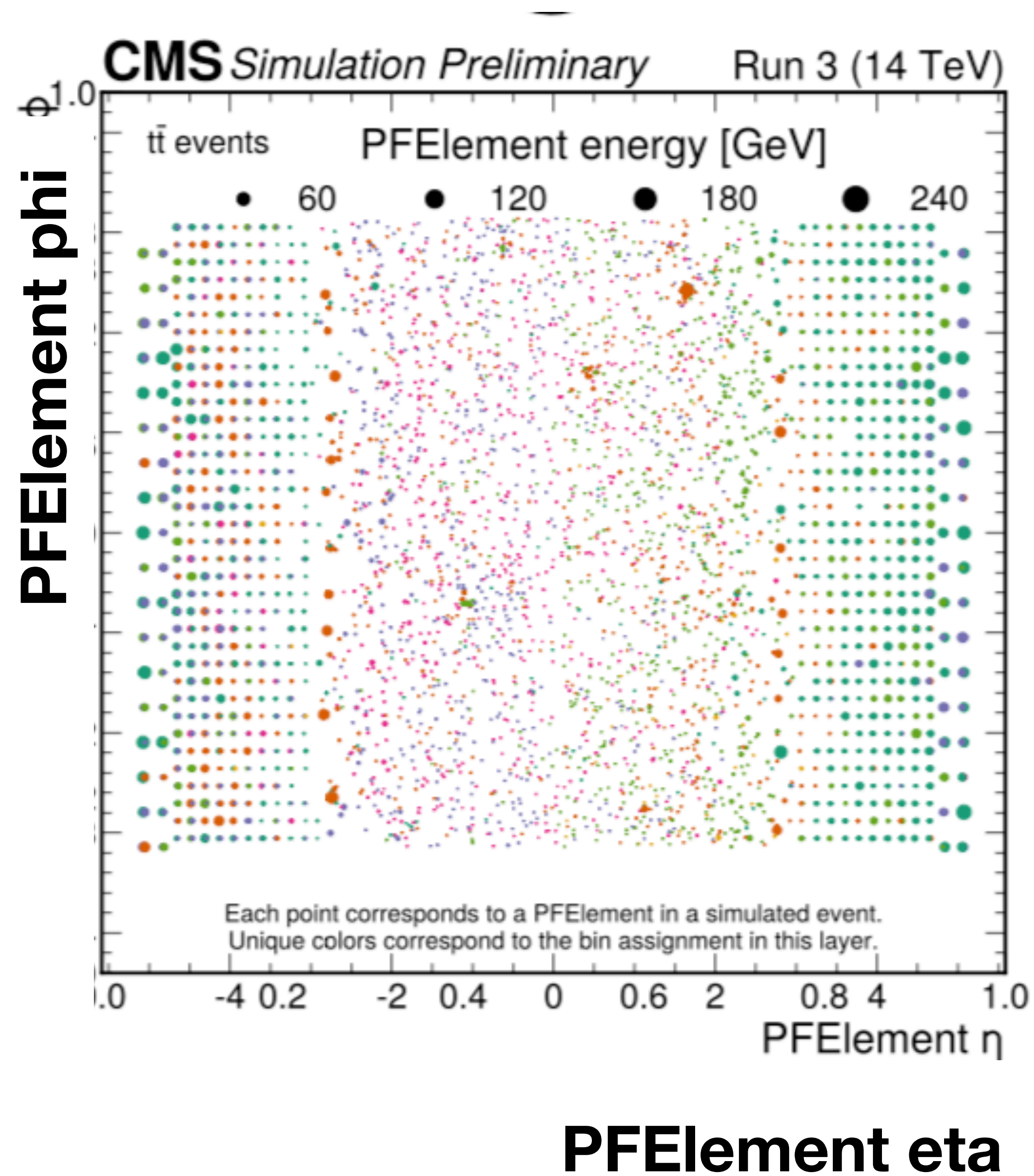
- track, - calorimeter cluster, - encoded element

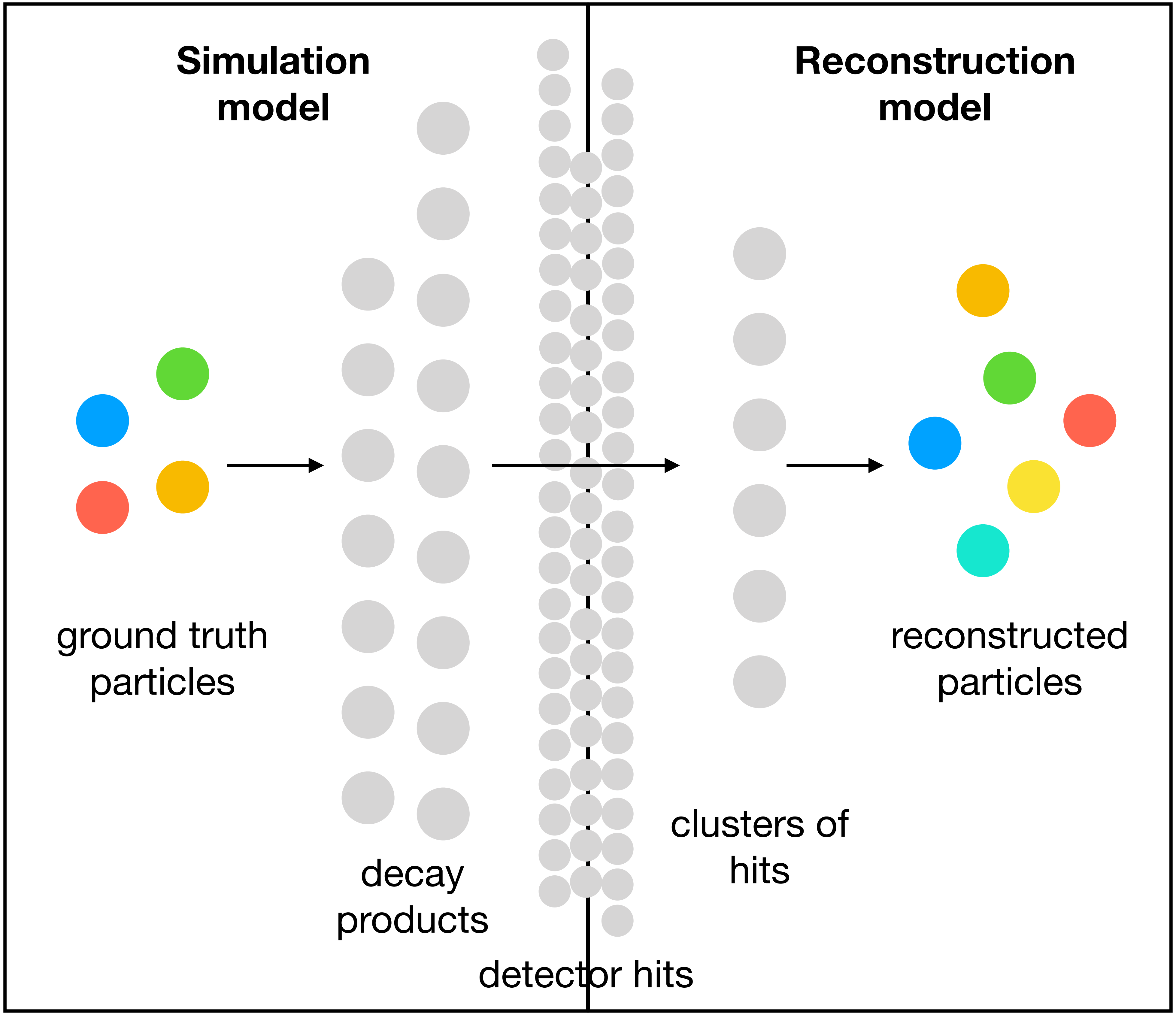
- target (predicted) particle, - no target (predicted) particle

Clustering to reconstruction

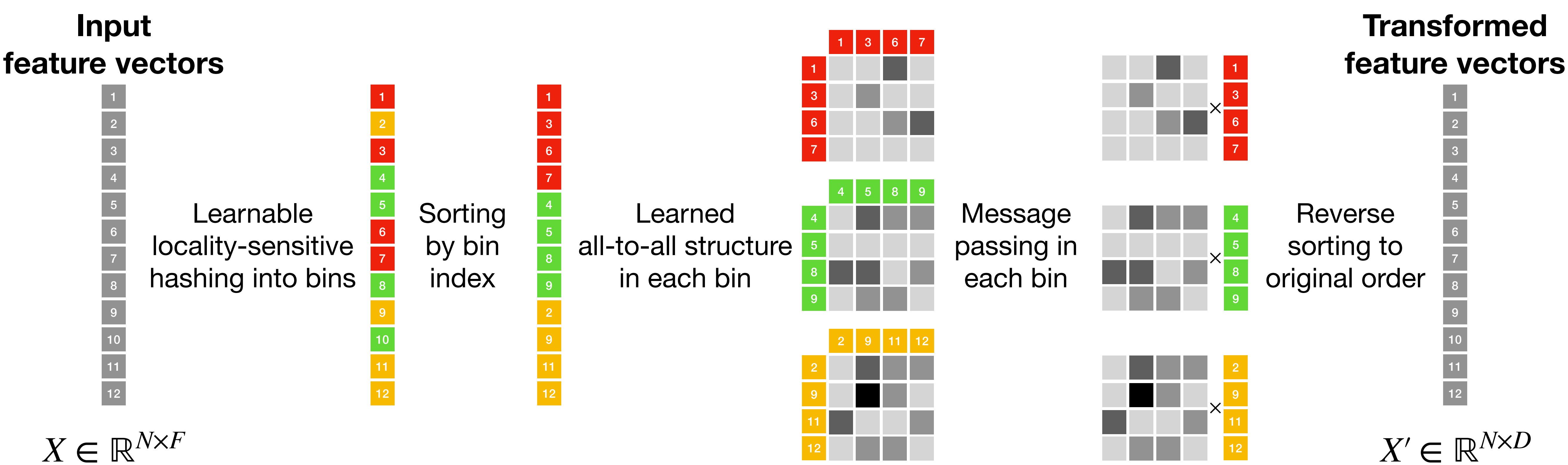
Clustering (graph building) is an internal detail, not a model target. **Particle reconstruction is the real goal!**



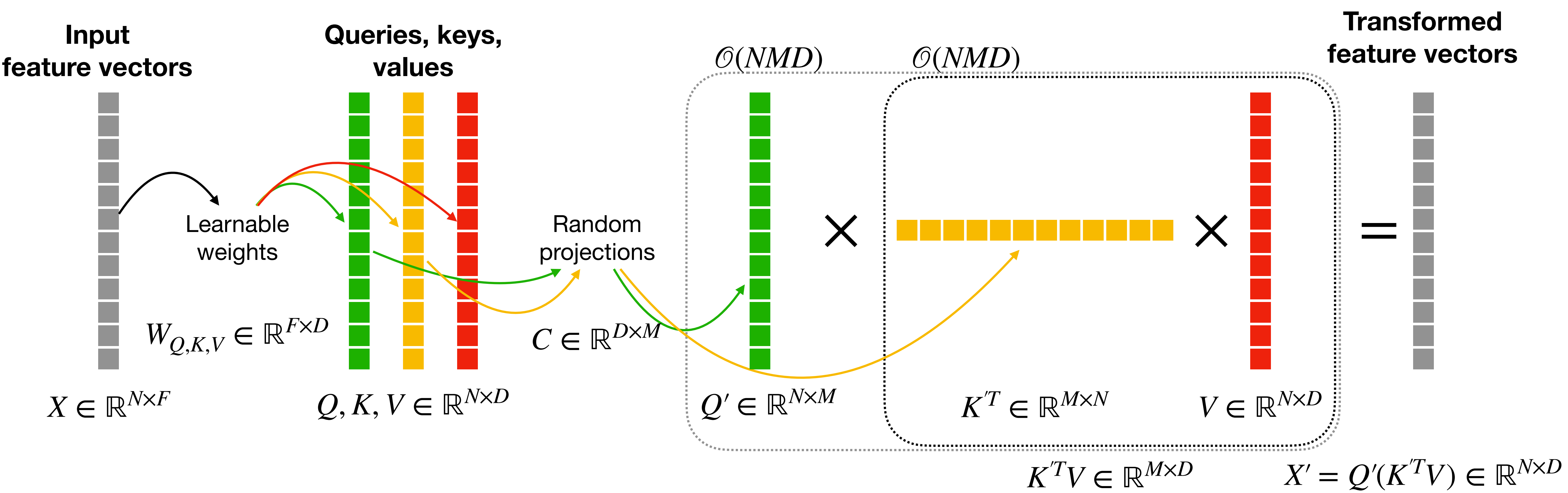


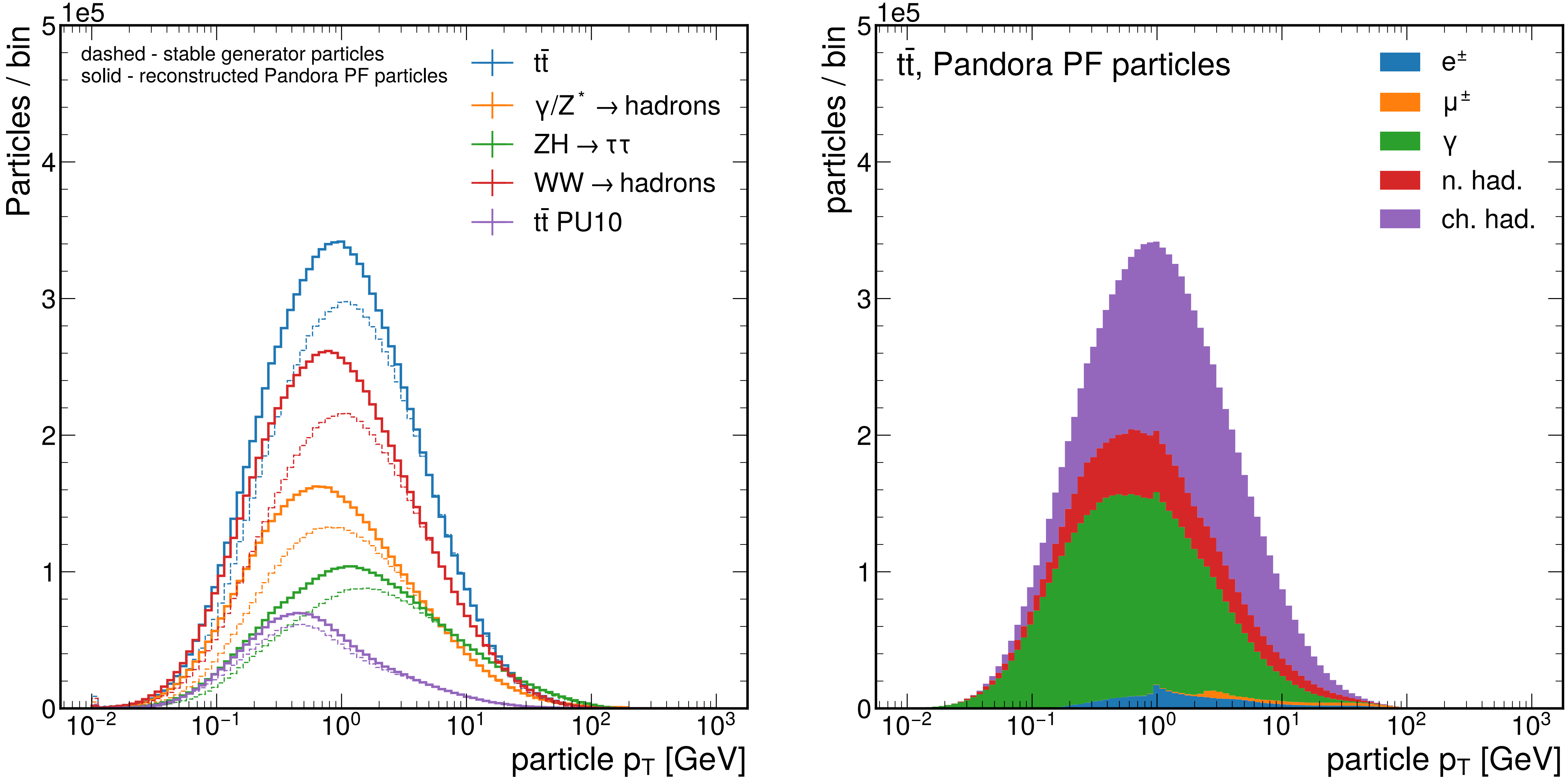

Simulation model
Reconstruction model
ground truth particles
decay products
detector hits
clusters of hits
reconstructed particles

One layer of learnable graph building with locality sensitive hashing and message passing

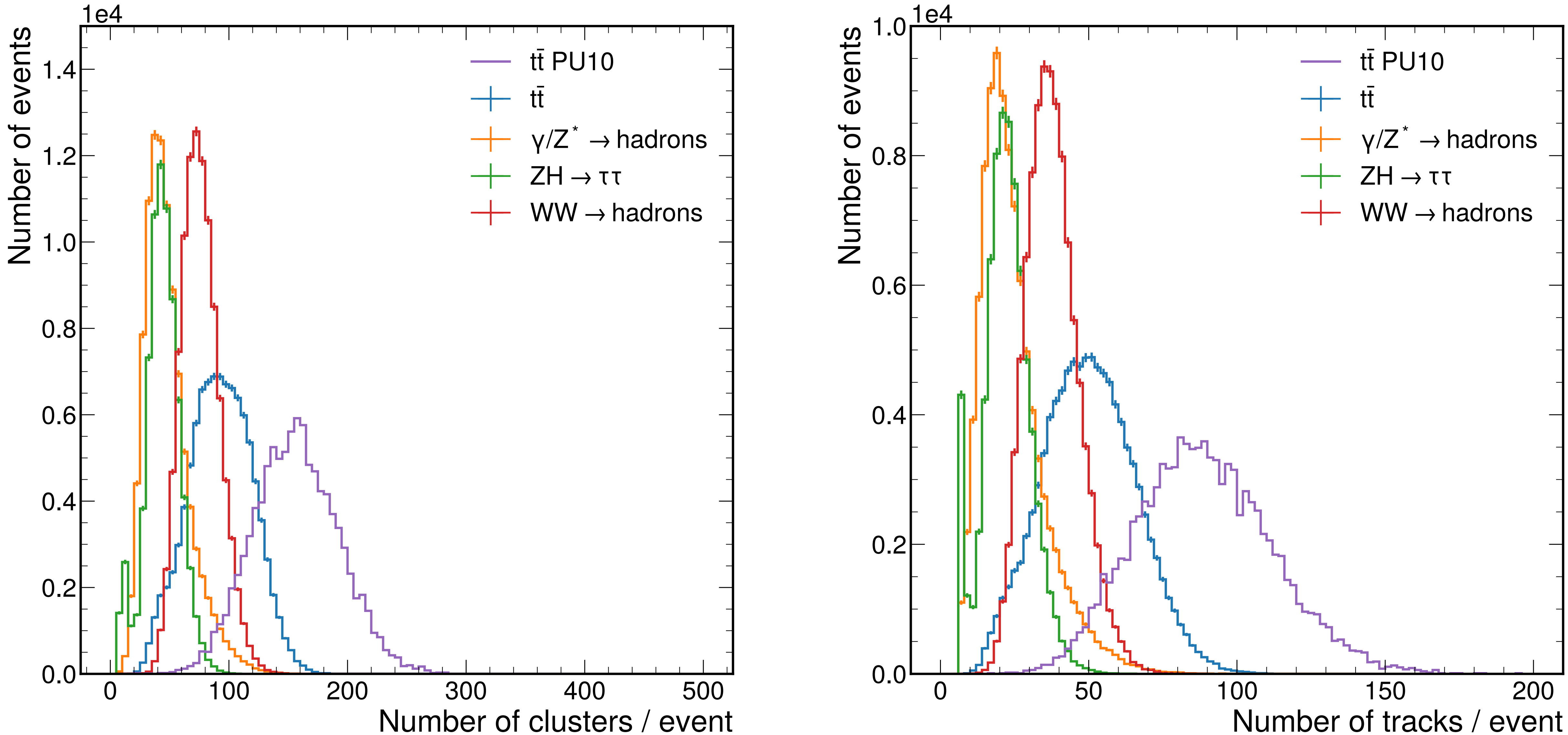


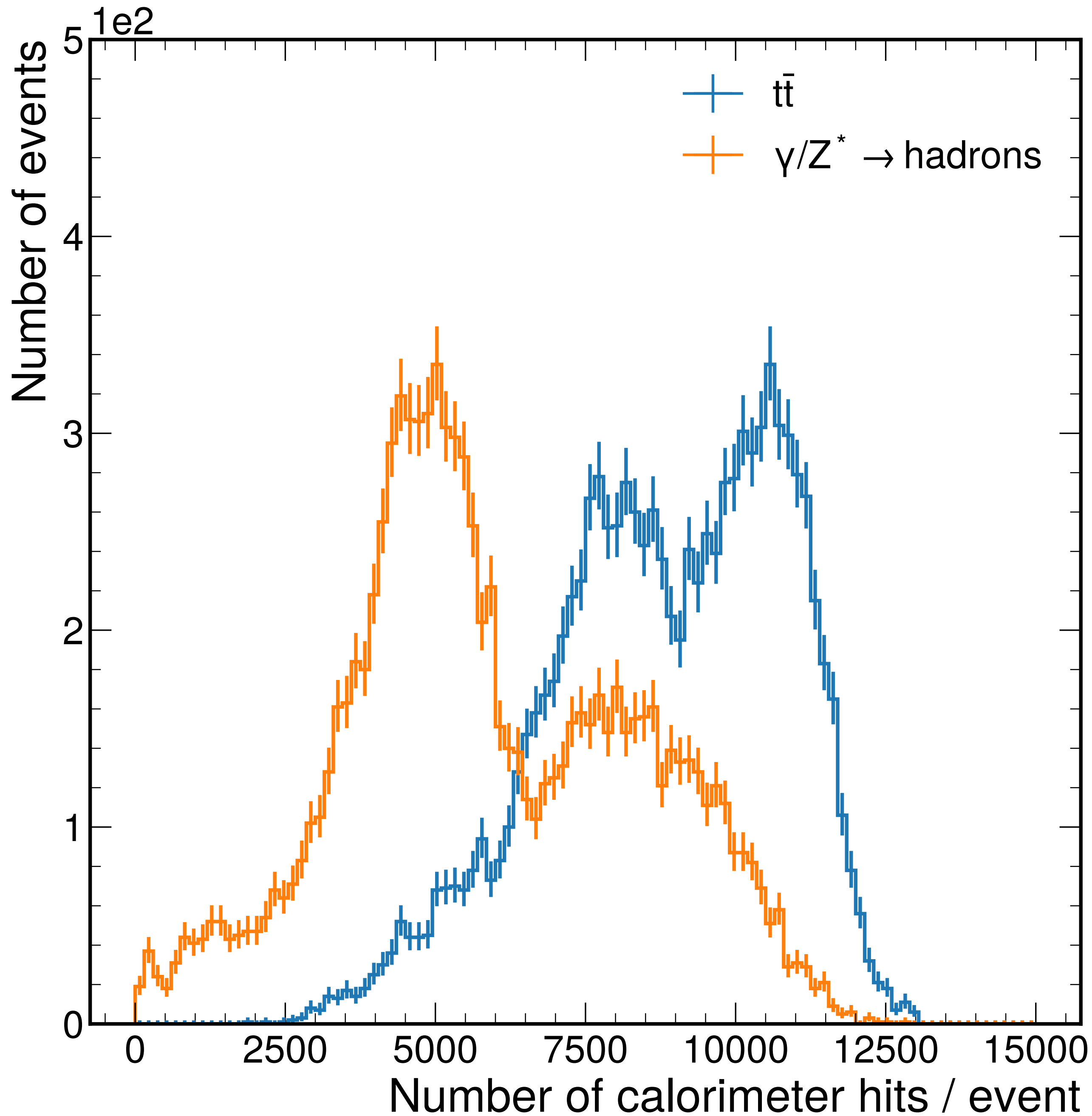
One layer of kernel-based self attention with the FAVOR mechanism.

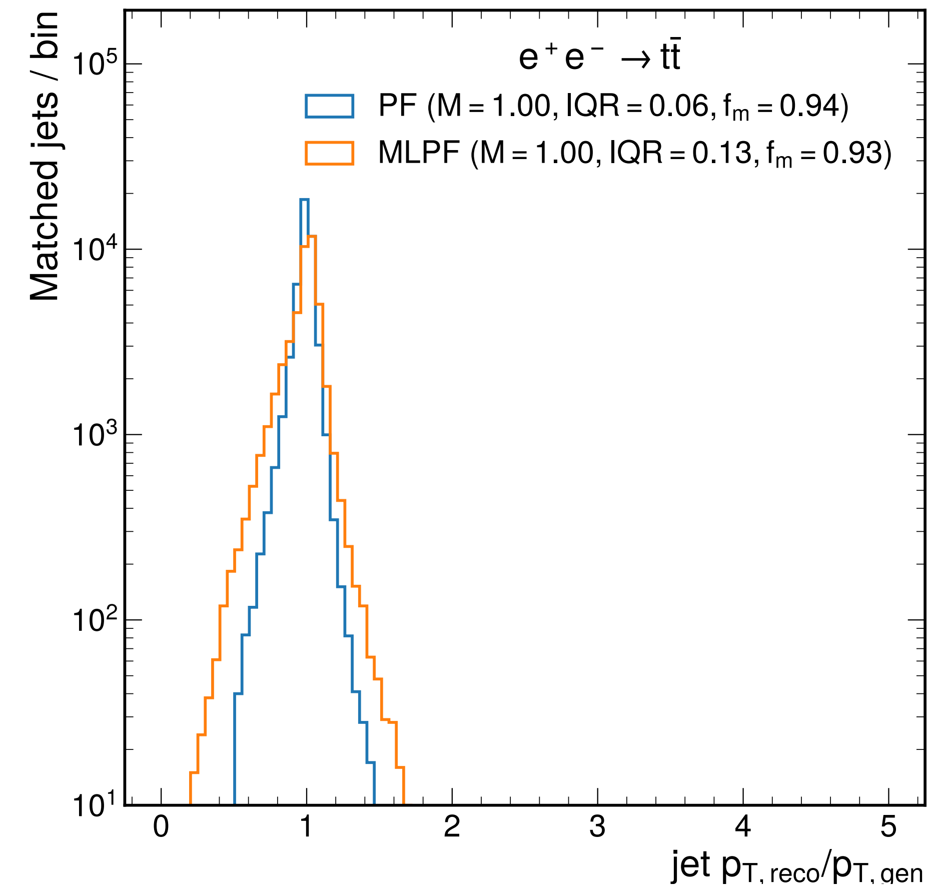


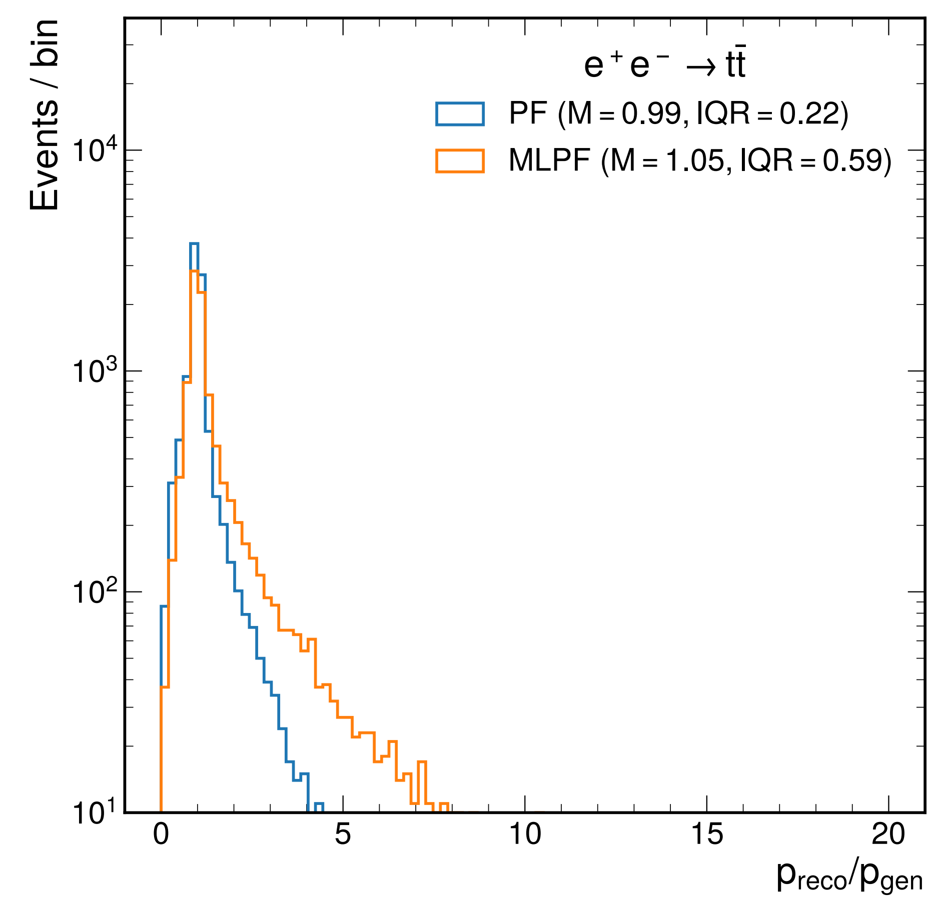


1e5
Particles / bin
dashed - stable generator particles
solid - reconstructed Pandora PF particles
tt̄
γ/Z* → hadrons
ZH → ττ
WW → hadrons
tt̄ PU10
particle pT [GeV]
1e5
particles / bin
tt̄, Pandora PF particles
e±
μ±
γ
n. had.
ch. had.
particle pT [GeV]


1e4
Number of events
Number of clusters / event
1e4
Number of events
Number of tracks / event
tt̄ PU10
tt̄
γ/Z* → hadrons
ZH → ττ
WW → hadrons


1e2
Number of events
tt̄
γ/Z* → hadrons
Number of calorimeter hits / event
0
2500
5000
7500
10000
12500
15000
0
1
2
3
4
5


Matched jets / bin
e⁺e⁻ → tt̄
PF (M = 1.00, IQR = 0.06, f_m = 0.94)
MLPF (M = 1.00, IQR = 0.13, f_m = 0.93)
jet p_T, reco/p_T, gen


Events / bin
e⁺e⁻ → tt̄
PF (M = 0.99, IQR = 0.22)
MLPF (M = 1.05, IQR = 0.59)
p_reco/p_gen