MLPF:
Machine Learning for Particle Flow

Joosep Pata (NICPB, Estonia)
Farouk Mokhtar, Javier Duarte, Michael Zhang, Jinya Jiang (UCSD)
Eric Wulff, David Southwick, Maria Girone (CERN)

April 18, 2023
IML meeting

- Belayneh, D., Carminati, F., Farbin, A. *et al.* Calorimetry with deep learning: particle simulation and reconstruction for collider physics. *Eur. Phys. J. C* **80,** 688 (2020). https://doi.org/10.1140/epjc/s10052-020-8251-9

- Jan Kieseler, Object condensation: one-stage grid-free multi-object reconstruction in physics detectors, graph, and image data. Eur. Phys. J. C 80, 886 (2020). https://doi.org/10.1140/epjc/s10052-020-08461-2

- Saptaparna Bhattacharya, Nadezda Chernyavskaya, Saranya Ghosh, Lindsey Gray, Jan Kieseler et al. GNN-based end-to-end reconstruction in the CMS Phase 2 High-Granularity Calorimeter. ACAT 2021. https://doi.org/10.48550/arXiv.2203.01189

- Shah Rukh Qasim, Nadezda Chernyavskaya, Jan Kieseler, Kenneth Long, Oleksandr Viazlo et al. End-to-end multi-particle reconstruction in high occupancy imaging calorimeters with graph neural networks. https://doi.org/10.48550/arXiv.2204.01681

- Di Bello, F.A., Ganguly, S., Gross, E. et al. Towards a computer vision particle flow. Eur. Phys. J. C 81, 107 (2021). https://doi.org/10.1140/epjc/s10052-021-08897-0

- **JP, Javier Duarte, Jean-Roch Vlimant, Maurizio Pierini & Maria Spiropulu. MLPF: efficient machine-learned particle-flow reconstruction using graph neural networks. Eur. Phys. J. C 81, 381 (2021). https://doi.org/10.1140/epjc/s10052-021-09158-w**

- **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**

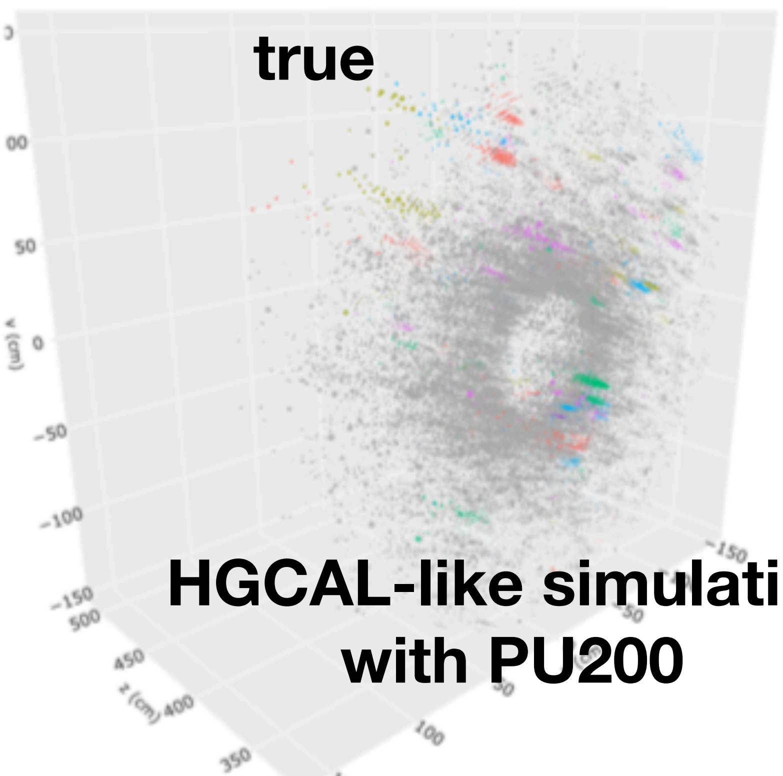
- Francesco Armando Di Bello, Etienne Dreyer, Sanmay Ganguly, Eilam Gross, Lukas Heinrich, Anna Ivina, Marumi Kado, Nilotpal Kakati, Lorenzo Santi, Jonathan Shlomi, Matteo Tusoni, Reconstructing particles in jets using set transformer and hypergraph prediction networks. https://arxiv.org/abs/2212.01328

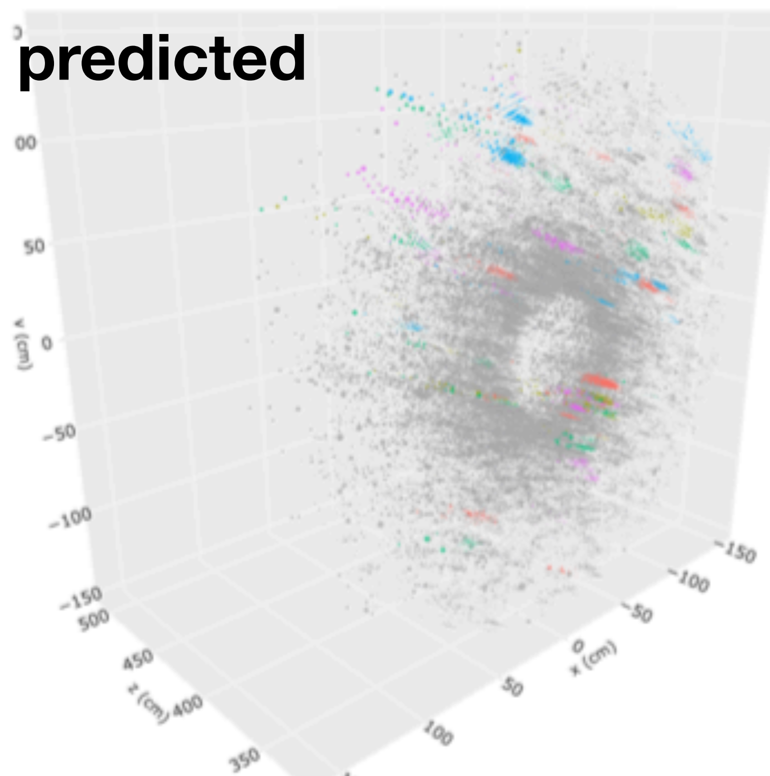
- **Farouk Mokhtar, JP, Javier Duarte, Eric Wulff, Dylan Rankin, Maurizio Pierini, Jean-Roch Vlimant. Progress towards an improved particle flow algorithm at CMS with machine learning. ACAT 2022 and ML4Jets 2022. https://arxiv.org/abs/2303.17657, http://cds.cern.ch/record/2842375**

ML based reco is an active area of research. In the interest of time, I will focus on the MLPF-related publications I'm more familiar with.

ML for HGCal

The physics task involves in using the clustering to predict the energy of the particle initiating the clustered shower.





Loss function over cluster, regressing the energy

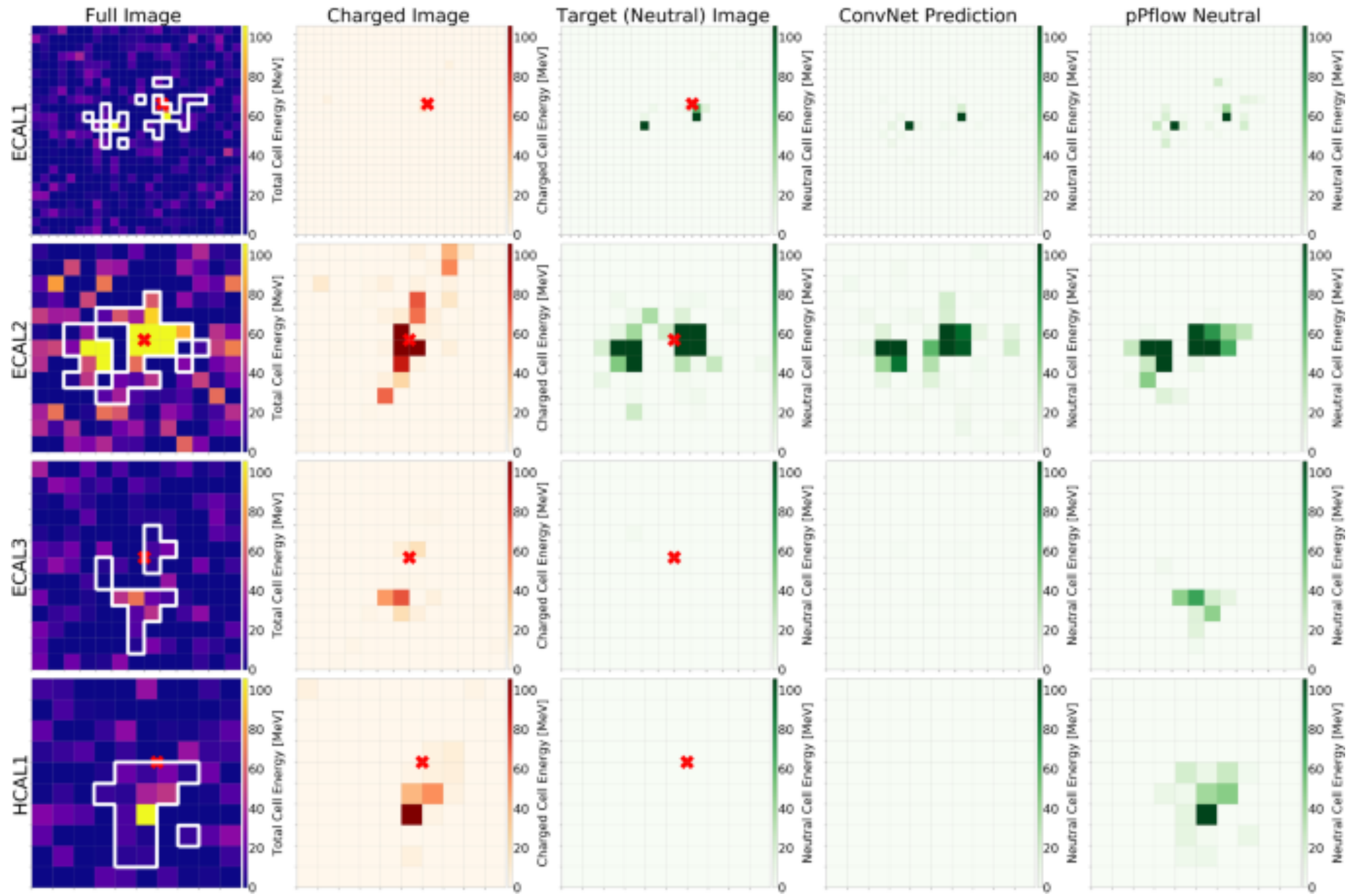
$$L_P = \sum_{t \in T} \frac{w_t}{\sum_{h \in H_t} \xi(h)} \sum_{h \in H_t} \xi(h) L_E,$$

$$L_E = \log\left(\left(\frac{E_{\mathrm{true},t} - \psi_h E_{\mathrm{dep},t}}{\sqrt{E_{\mathrm{true},t}} + 0.003}\right)^2 + 1\right),$$

Shah Rukh Qasim, Nadezda Chernyavskaya, Jan Kieseler, Kenneth Long, Oleksandr Viazlo et al. End-to-end multi-particle reconstruction in high occupancy imaging calorimeters with graph neural networks. https://doi.org/10.48550/arXiv.2204.01681

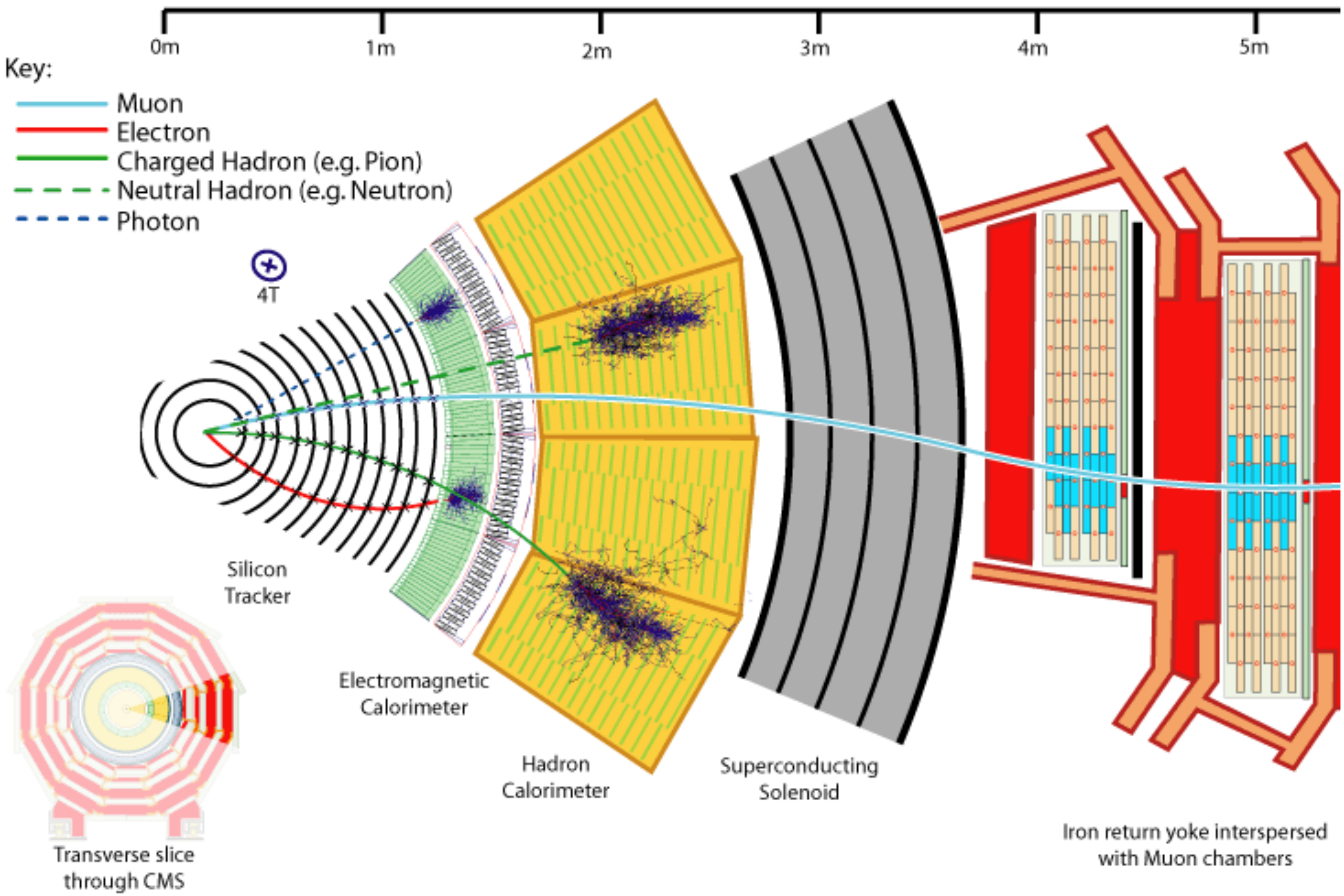
ML for neutral energy

The full event: a multilayered calorimetric image + tracks.
Predict the neutral energy deposits.

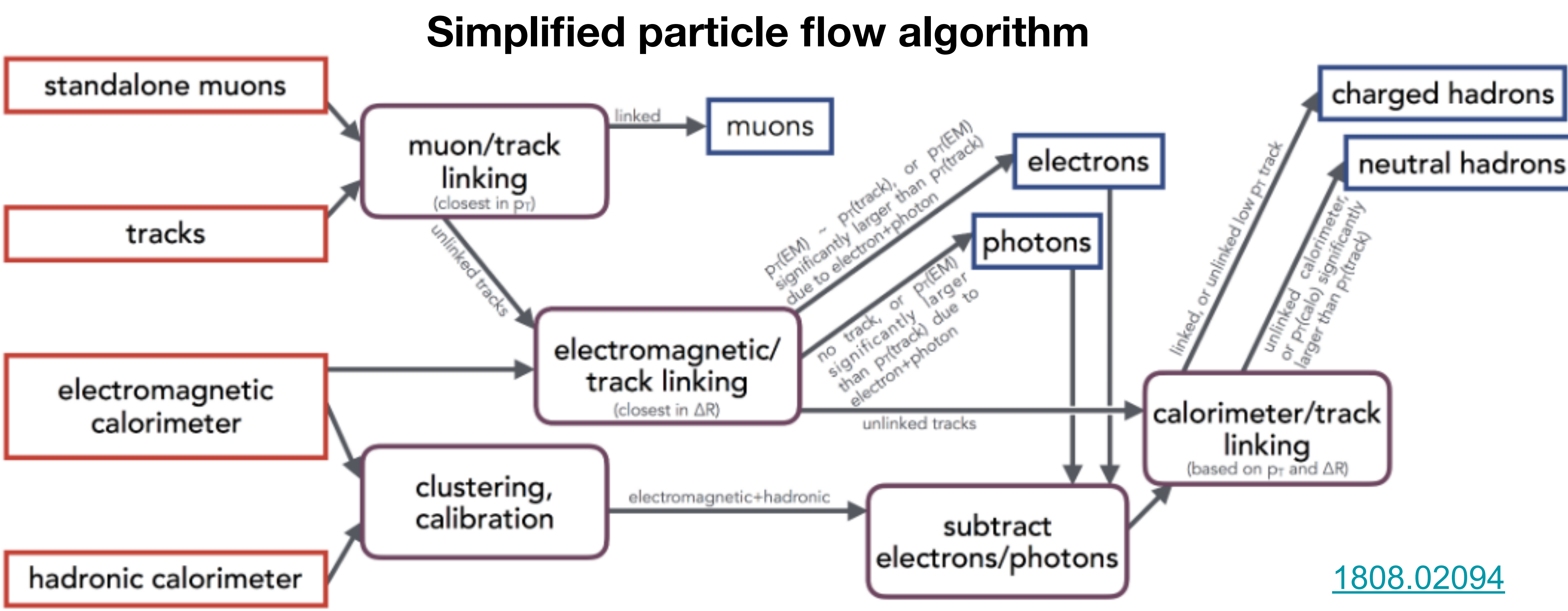


Di Bello, F.A., Ganguly, S., Gross, E. et al. Towards a computer vision particle flow. Eur. Phys. J. C 81, 107 (2021). https://doi.org/10.1140/epjc/s10052-021-08897-0

Multilayered detectors



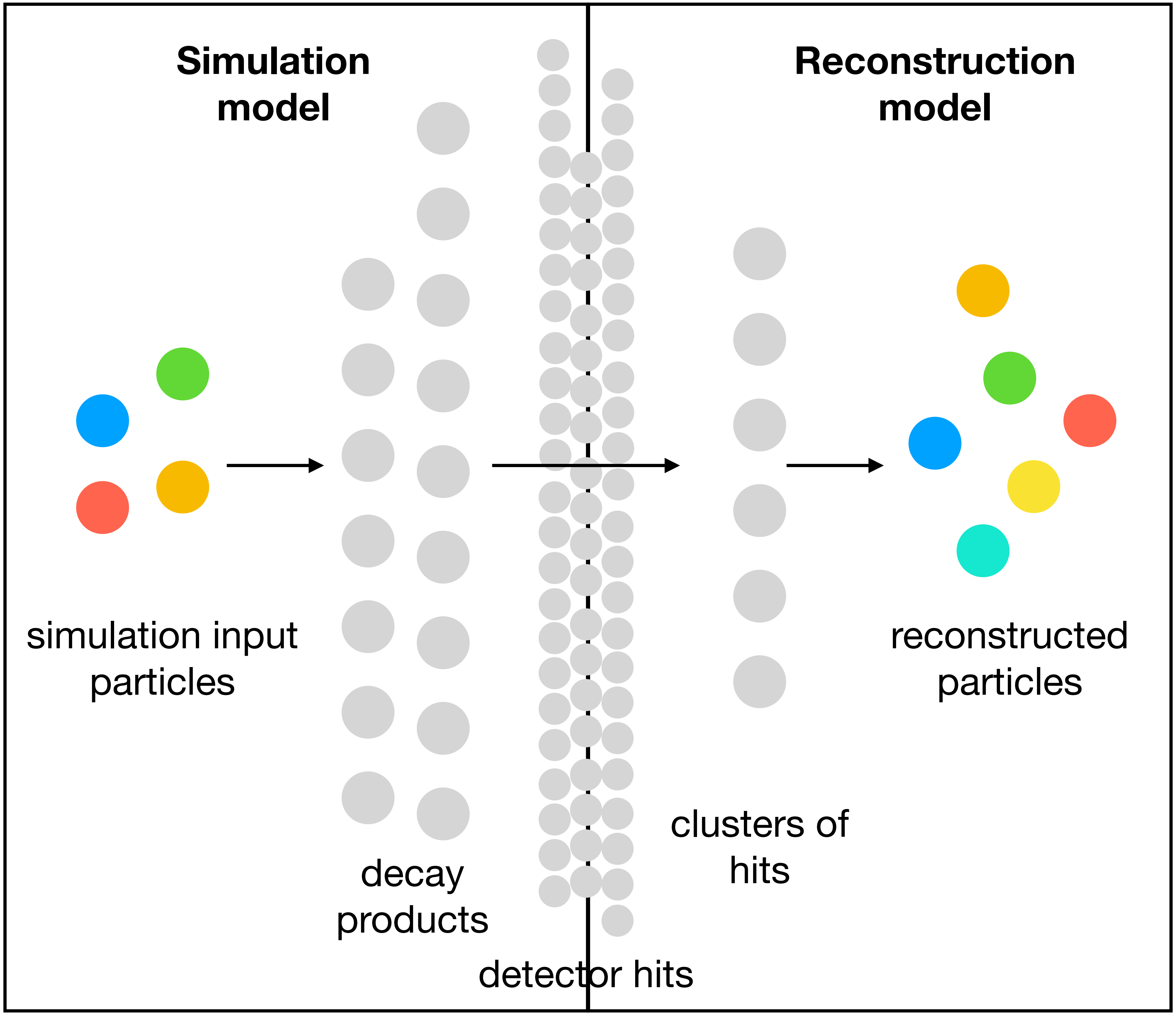
Algorithmic reconstruction



1808.02094

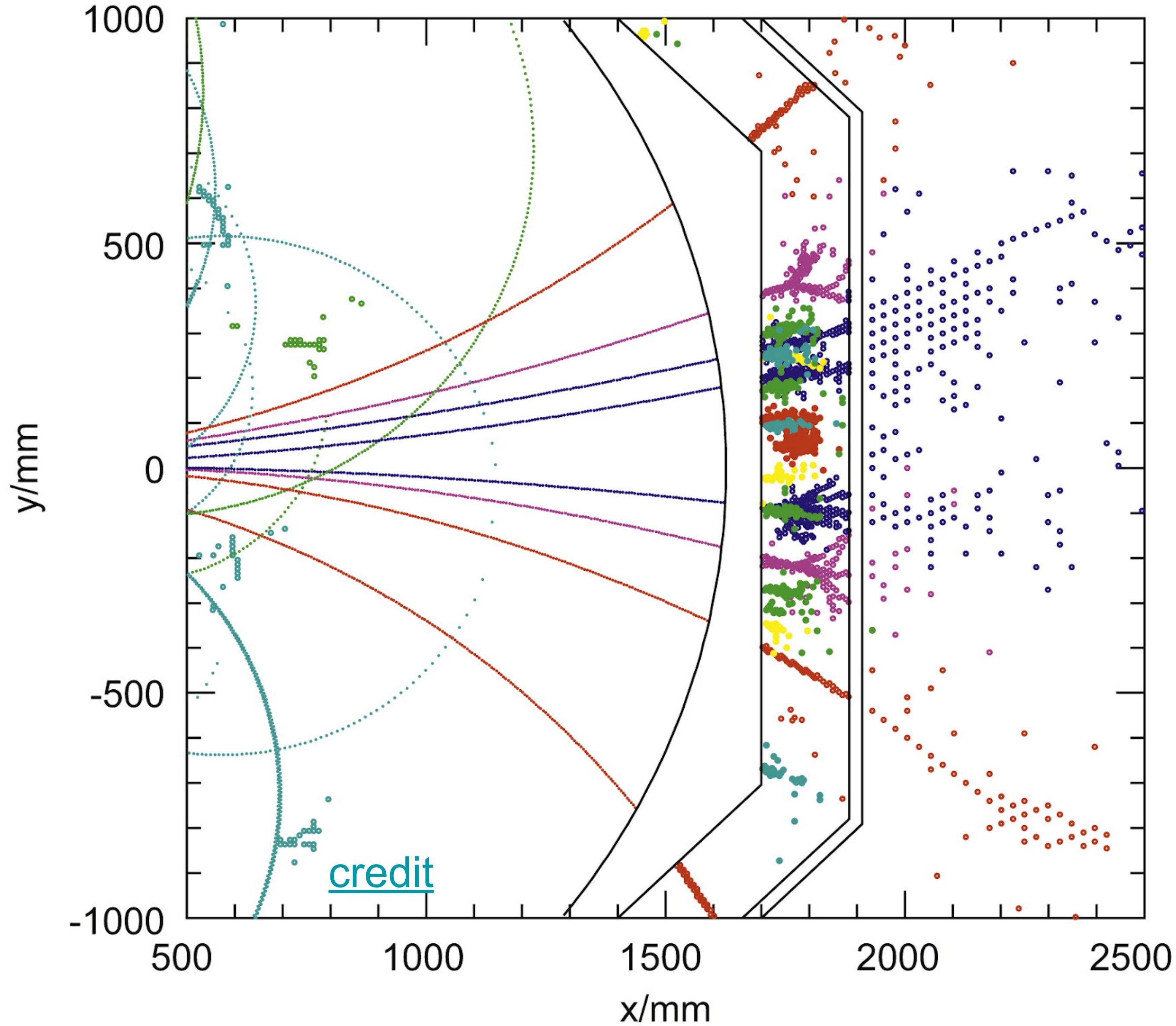
Figure 2. Schematic of particle flow algorithm for CMS Level-1 trigger correlator.

Simulation to reconstruction



Calorimeter clustering

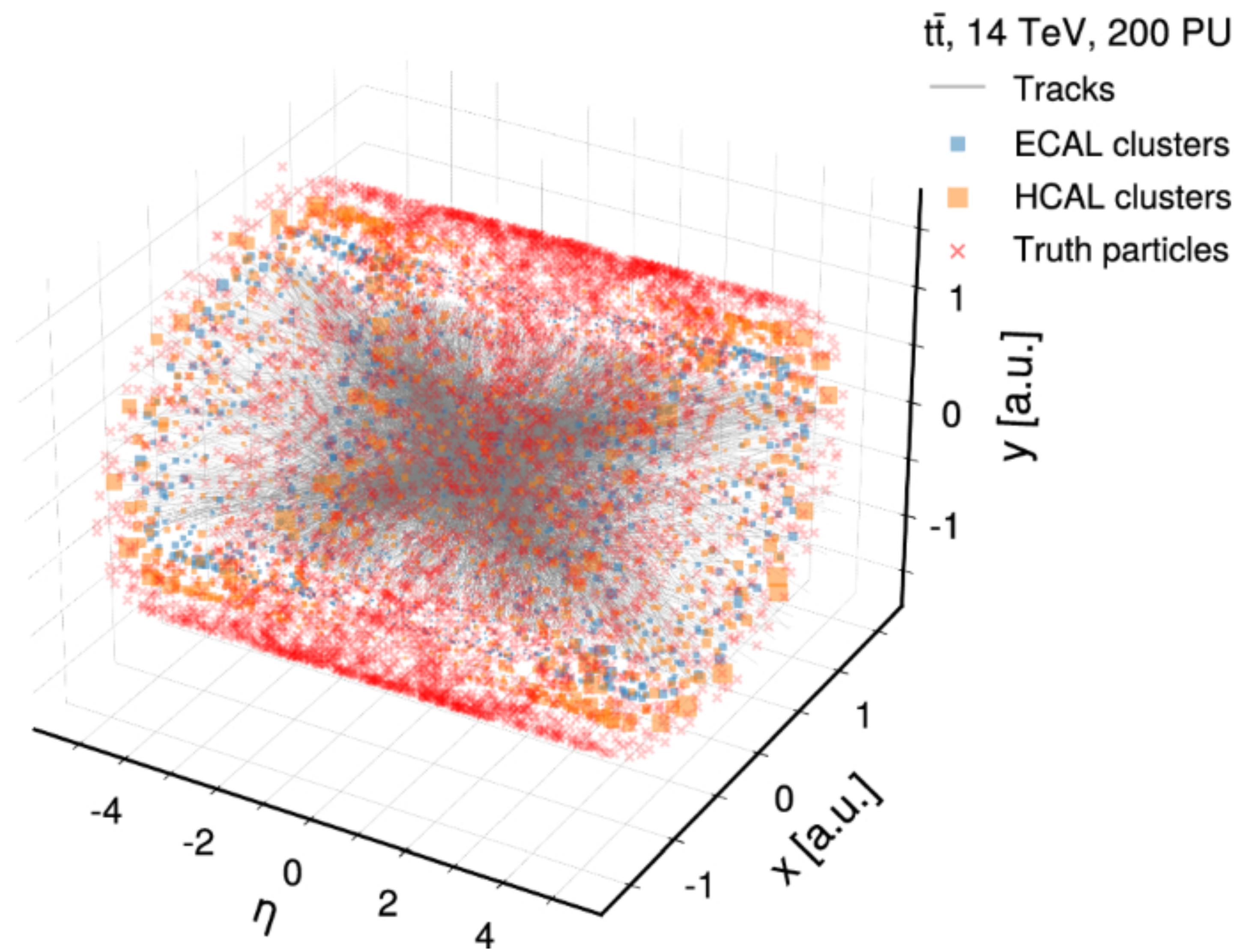
- Segment the energy deposits (hits) according to the originator particles
- The hits are embedded in a complicated feature space (Cartesian position, energy, signal significance, timing, layer information, ...)
- Showers from different particles may overlap spatially
- **Standard heuristic approaches** based on seeding & collecting neighbors, typically iterative



Sparse representations

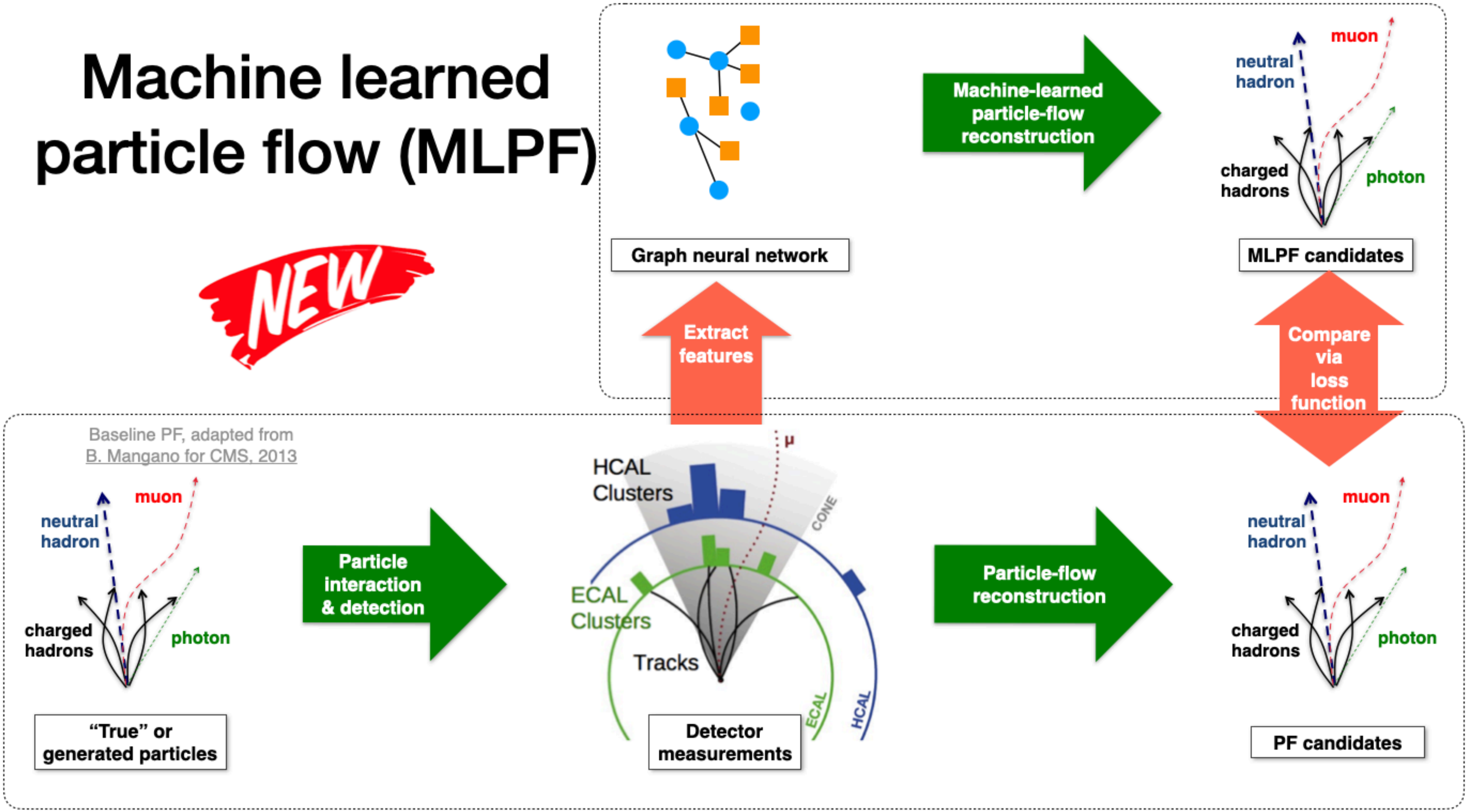
Starting from tracks and calorimeter clusters, aim to reconstruct the full set of input particles.

Inputs are heterogeneous, no natural underlying topology or associations.

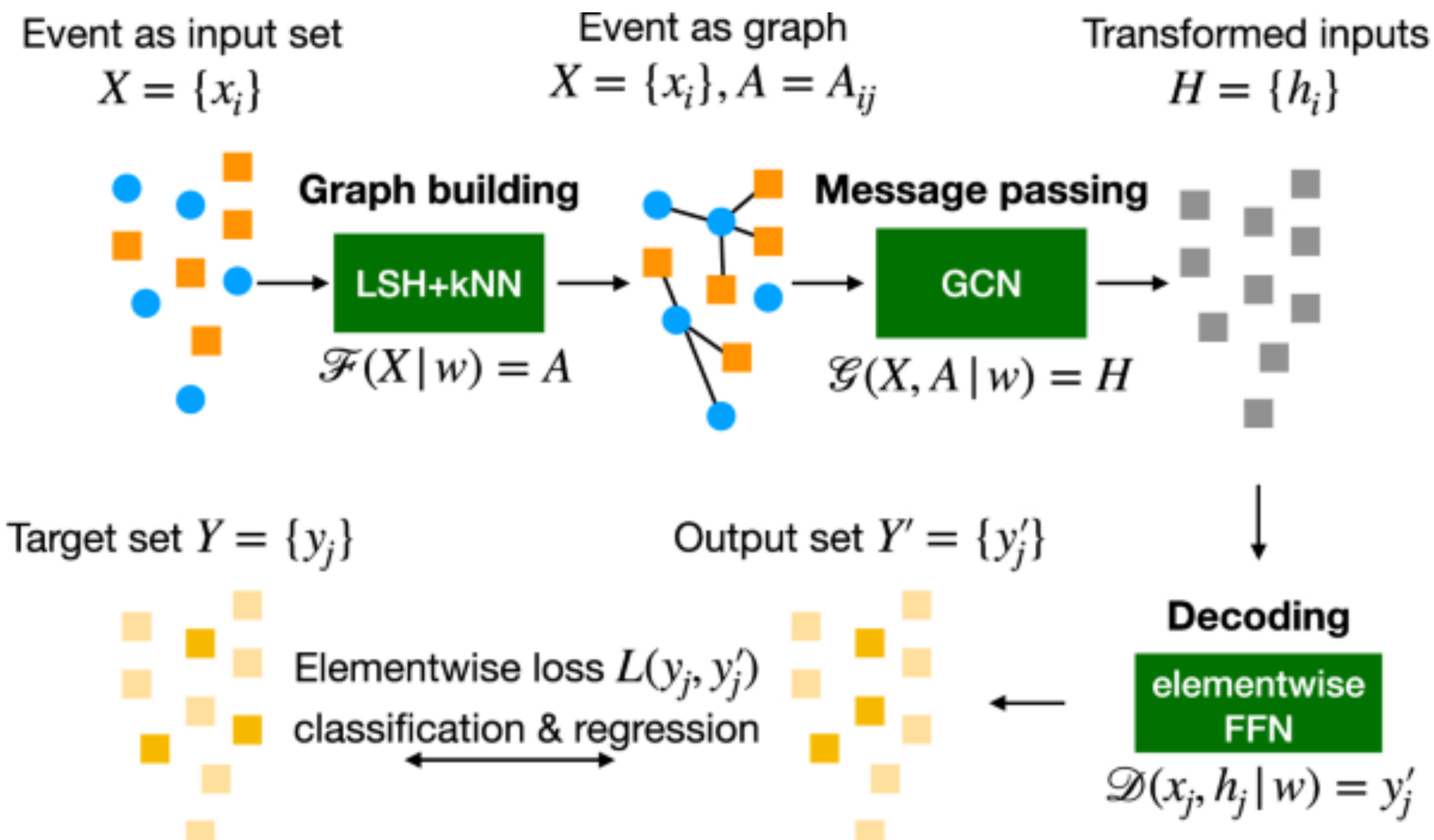


Pata, J., Duarte, J., Vlimant, JR. et al. MLPF: efficient machine-learned particle-flow reconstruction using graph neural networks. Eur. Phys. J. C 81, 381 (2021). https://doi.org/10.1140/epjc/s10052-021-09158-w

Machine learned particle flow (MLPF)



A simplification: treat the inputs as a homogenous set.



$$x_i = [\text{type}, p_{\mathrm{T}}, E_{\mathrm{ECAL}}, E_{\mathrm{HCAL}}, \eta, \phi, \eta_{\mathrm{outer}}, \phi_{\mathrm{outer}}, q, \ldots], \quad \text{type} \in \{\text{track}, \text{cluster}\}$$

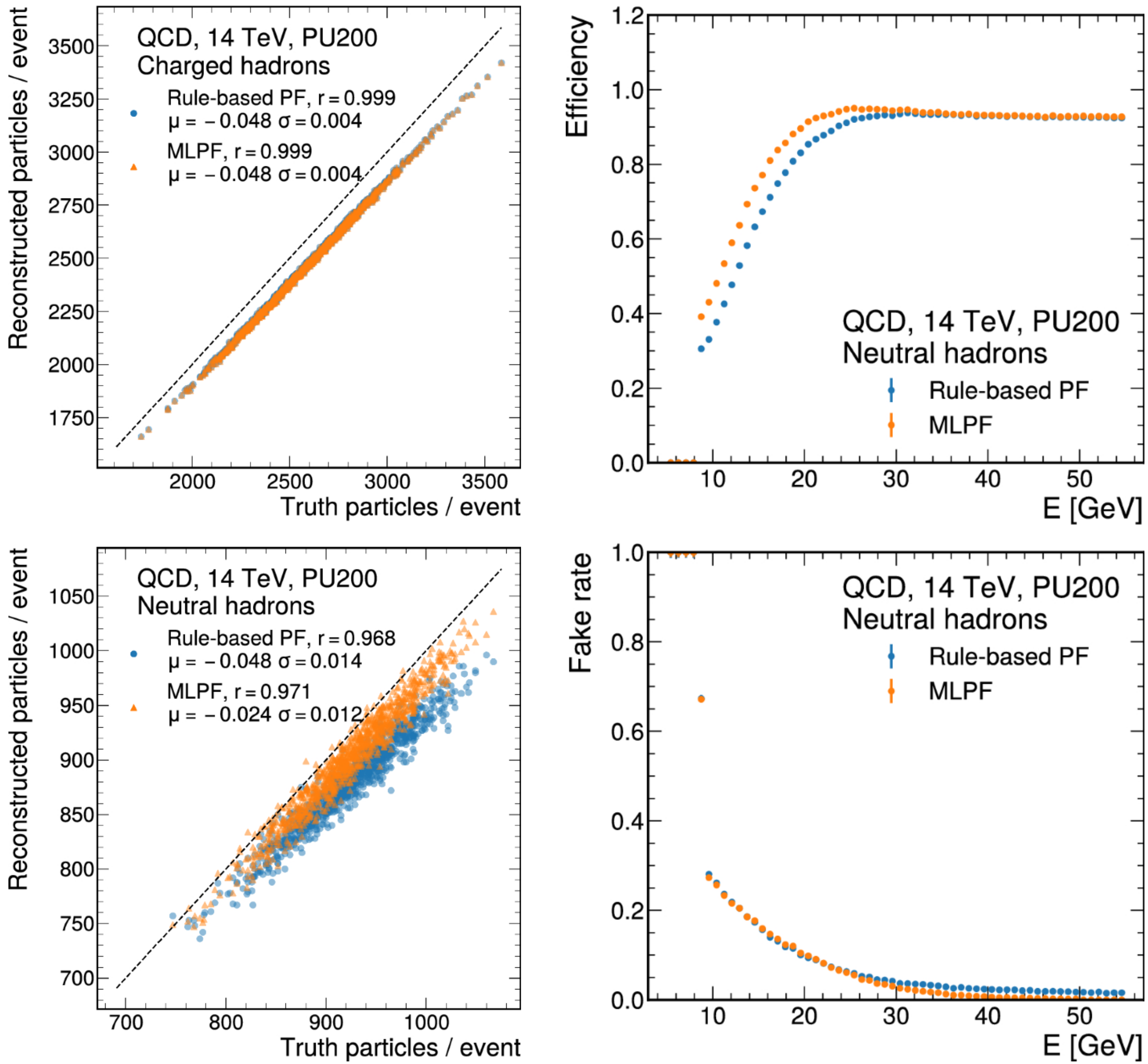
$$y_j = [\mathrm{PID}, p_{\mathrm{T}}, E, \eta, \phi, q, \ldots], \quad \mathrm{PID} \in \{\text{none}, \text{charged hadron}, \text{neutral hadron}, \gamma, e^{\pm}, \mu^{\pm}\}$$

$$h_i \in \mathbb{R}^{256}$$

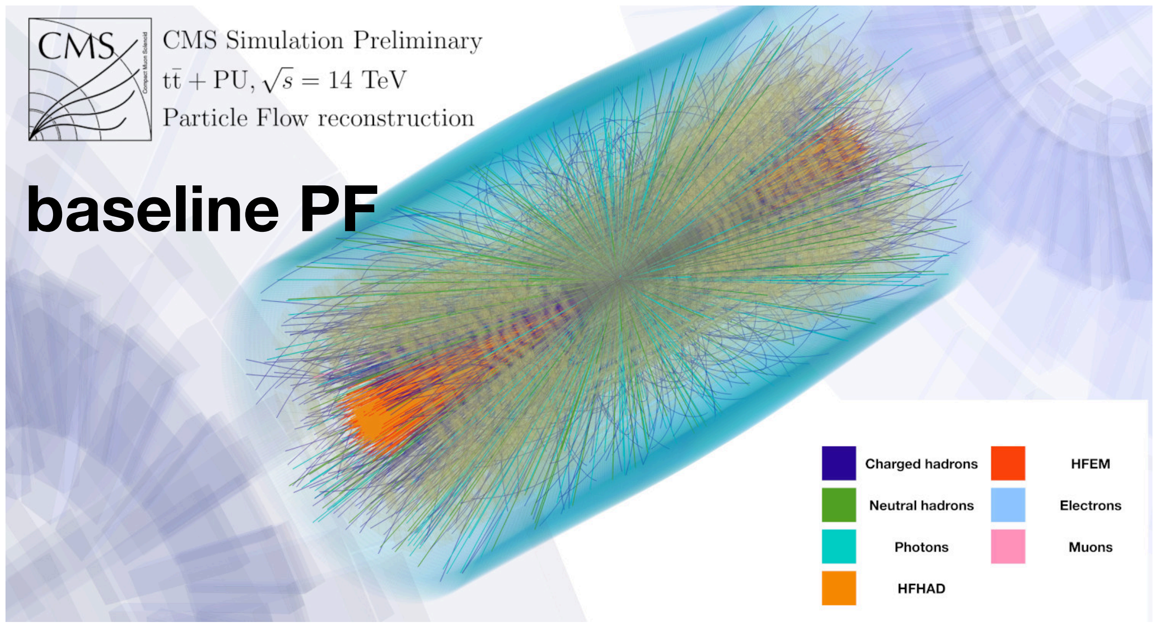
Trainable neural networks: $\mathcal{F}, \mathcal{G}, \mathcal{D}$

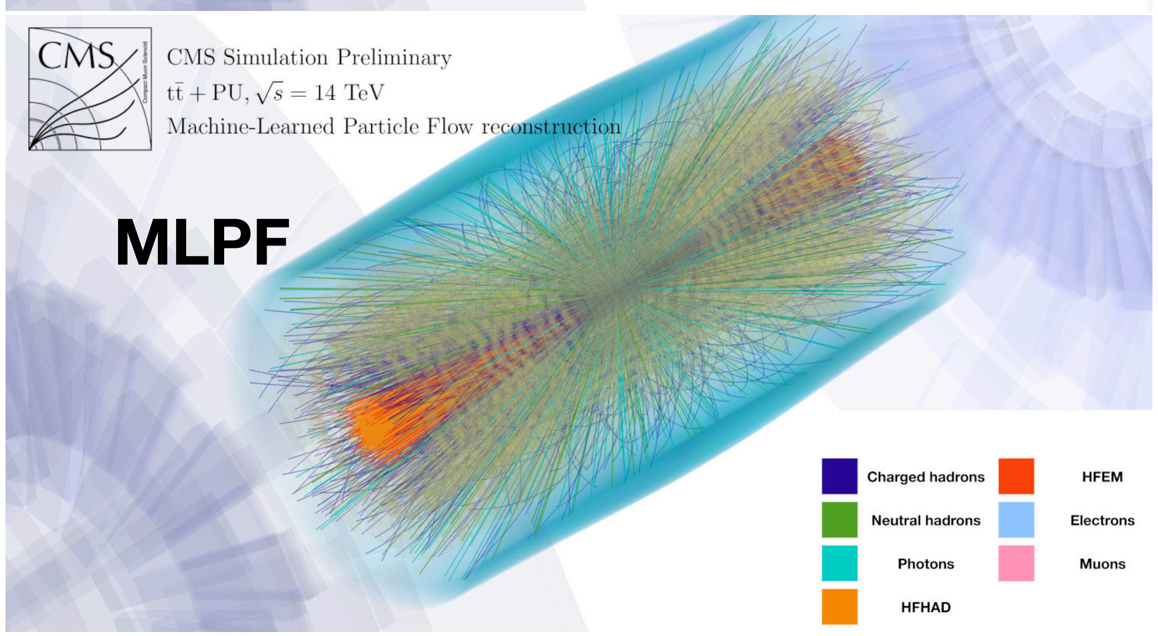
- track, - calorimeter cluster, - encoded element

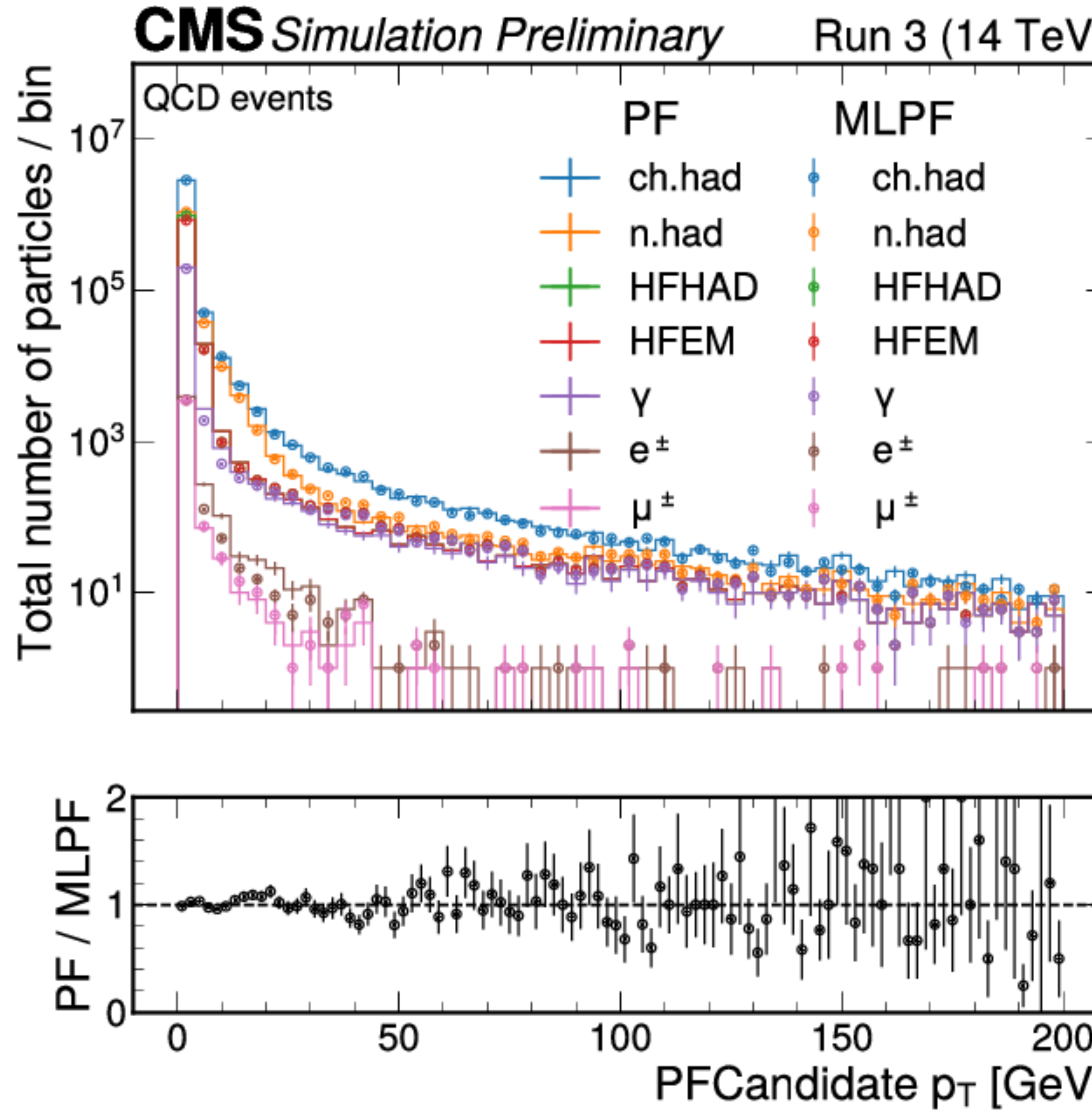
- target (predicted) particle, - no target (predicted) particle


QCD, 14 TeV, PU200
Charged hadrons
Rule-based PF, r = 0.999
μ = −0.048 σ = 0.004
MLPF, r = 0.999
μ = −0.048 σ = 0.004
Reconstructed particles / event
Truth particles / event
Efficiency
E [GeV]
QCD, 14 TeV, PU200
Neutral hadrons
Rule-based PF
MLPF
QCD, 14 TeV, PU200
Neutral hadrons
Rule-based PF, r = 0.968
μ = −0.048 σ = 0.014
MLPF, r = 0.971
μ = −0.024 σ = 0.012
Reconstructed particles / event
Truth particles / event
Fake rate
E [GeV]
QCD, 14 TeV, PU200
Neutral hadrons
Rule-based PF
MLPF

Test in a real detector



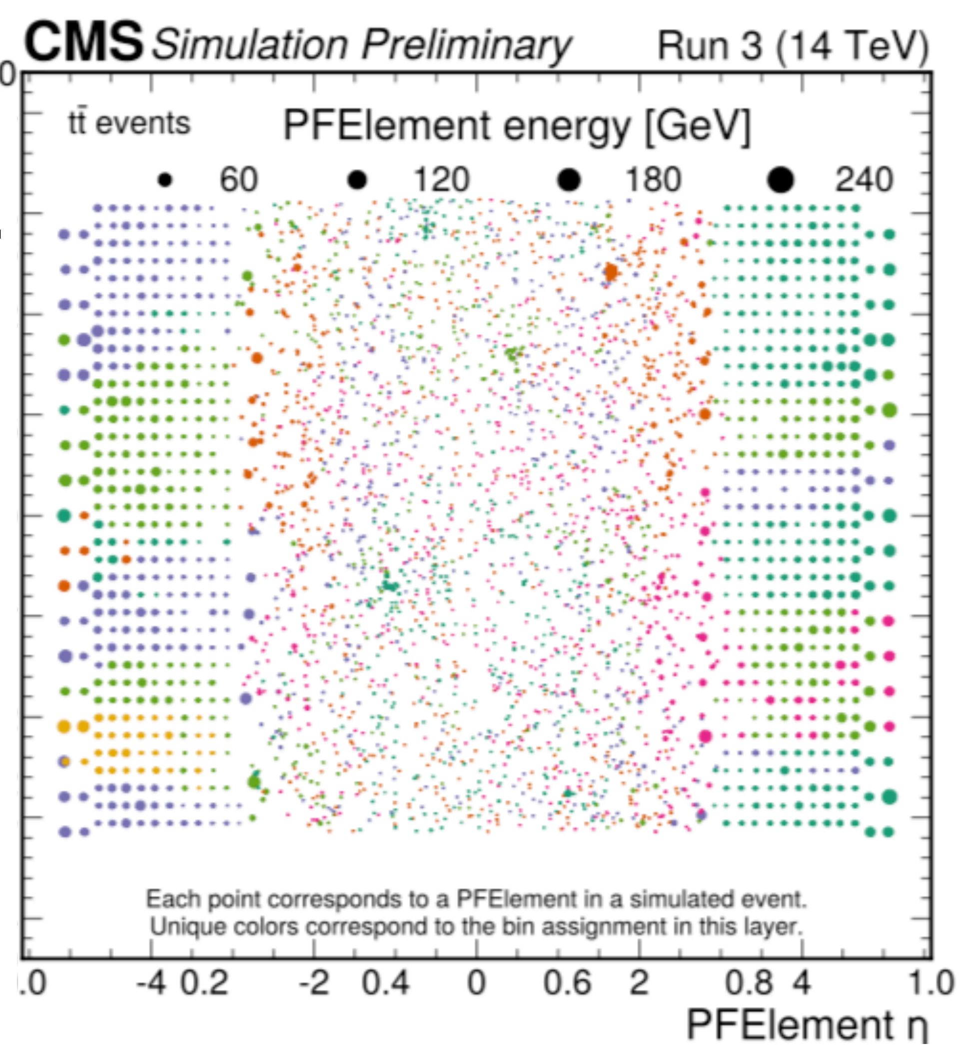


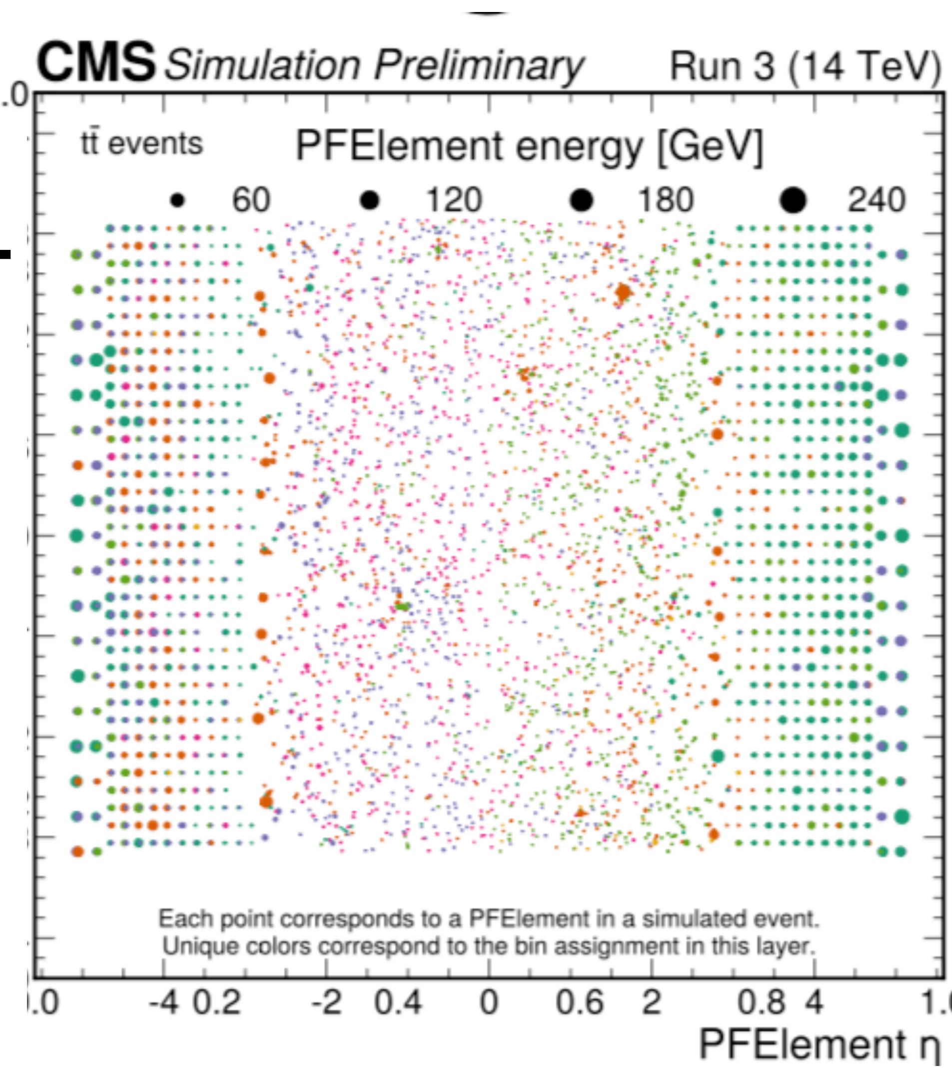


JP, Javier Duarte, Farouk Mokhtar, Eric Wulff, Jieun Yoo, Jean-Roch Vlimant, Maurizio Pierini, Maria Girone. Machine Learning for Particle Flow Reconstruction at CMS. ACAT 2021. https://doi.org/10.48550/arXiv.2203.00330, http://cds.cern.ch/record/2792320

Clustering to reconstruction

In this case, clustering (graph building) is an internal detail, not a model target. Reconstructing particles is the physical optimization target.





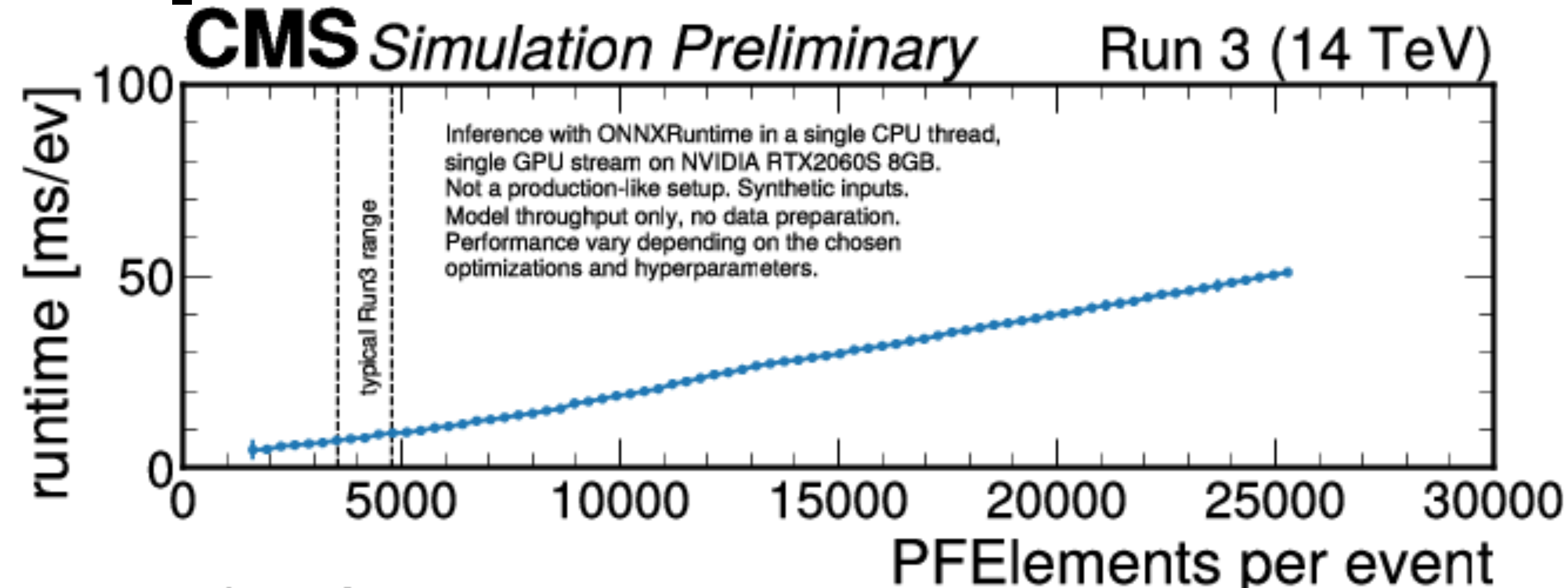
PFElement phi

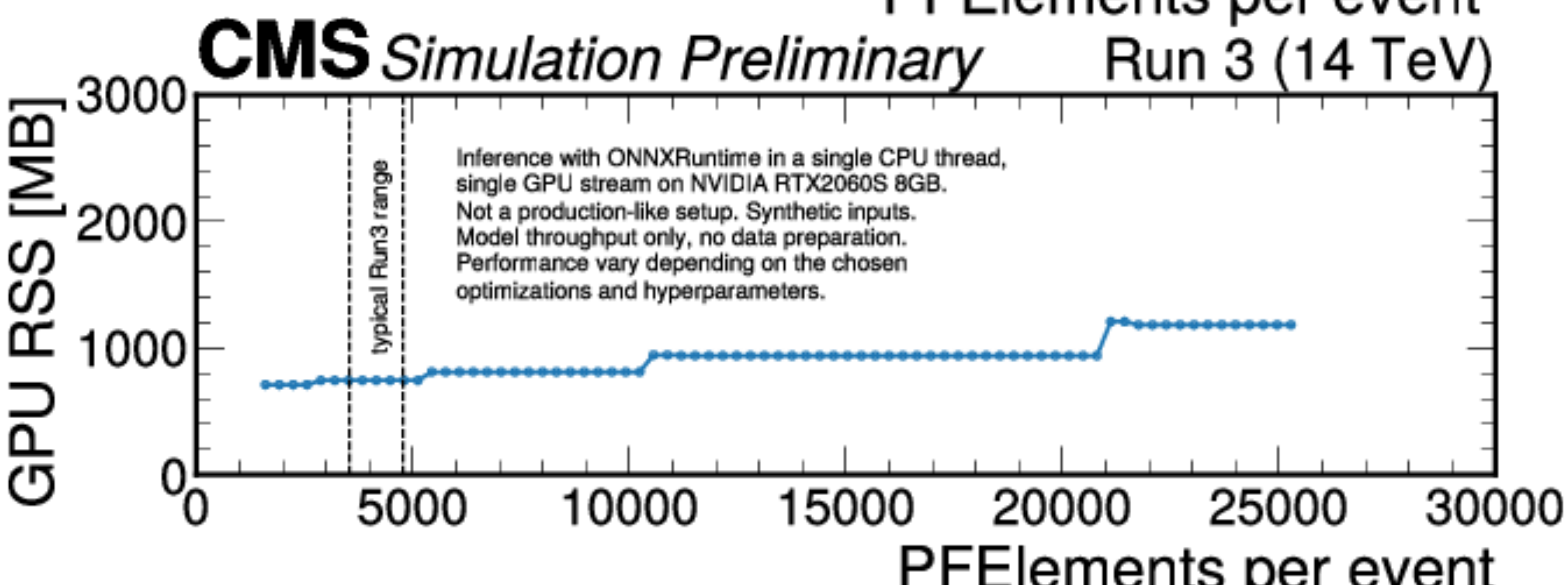
PFElement eta

PFElement phi

PFElement eta

Computational scalability

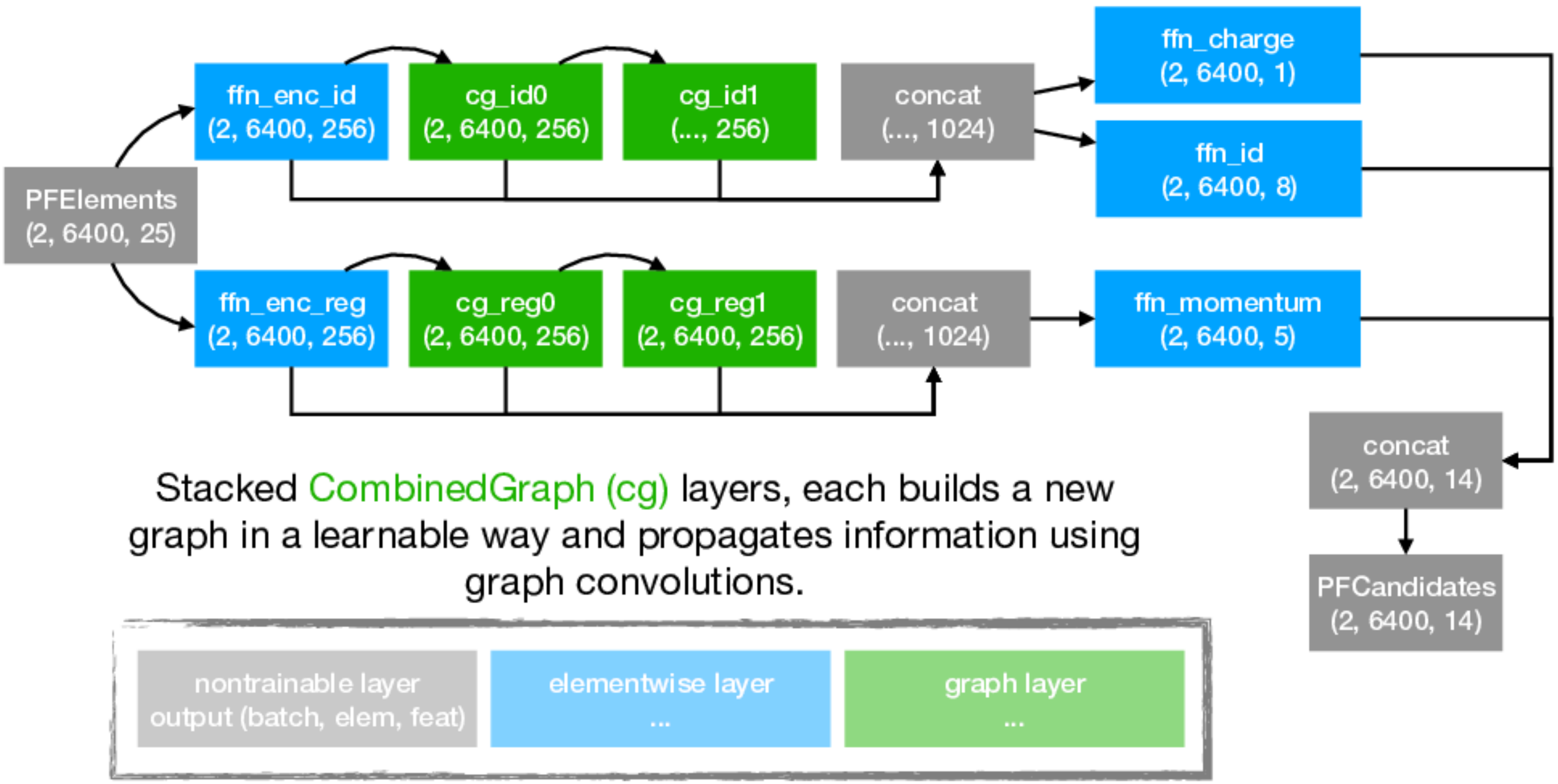




Export and deploy via ONNX. Avoid nonportable code, currently testing on AMD, Habana, Nvidia, CPU…

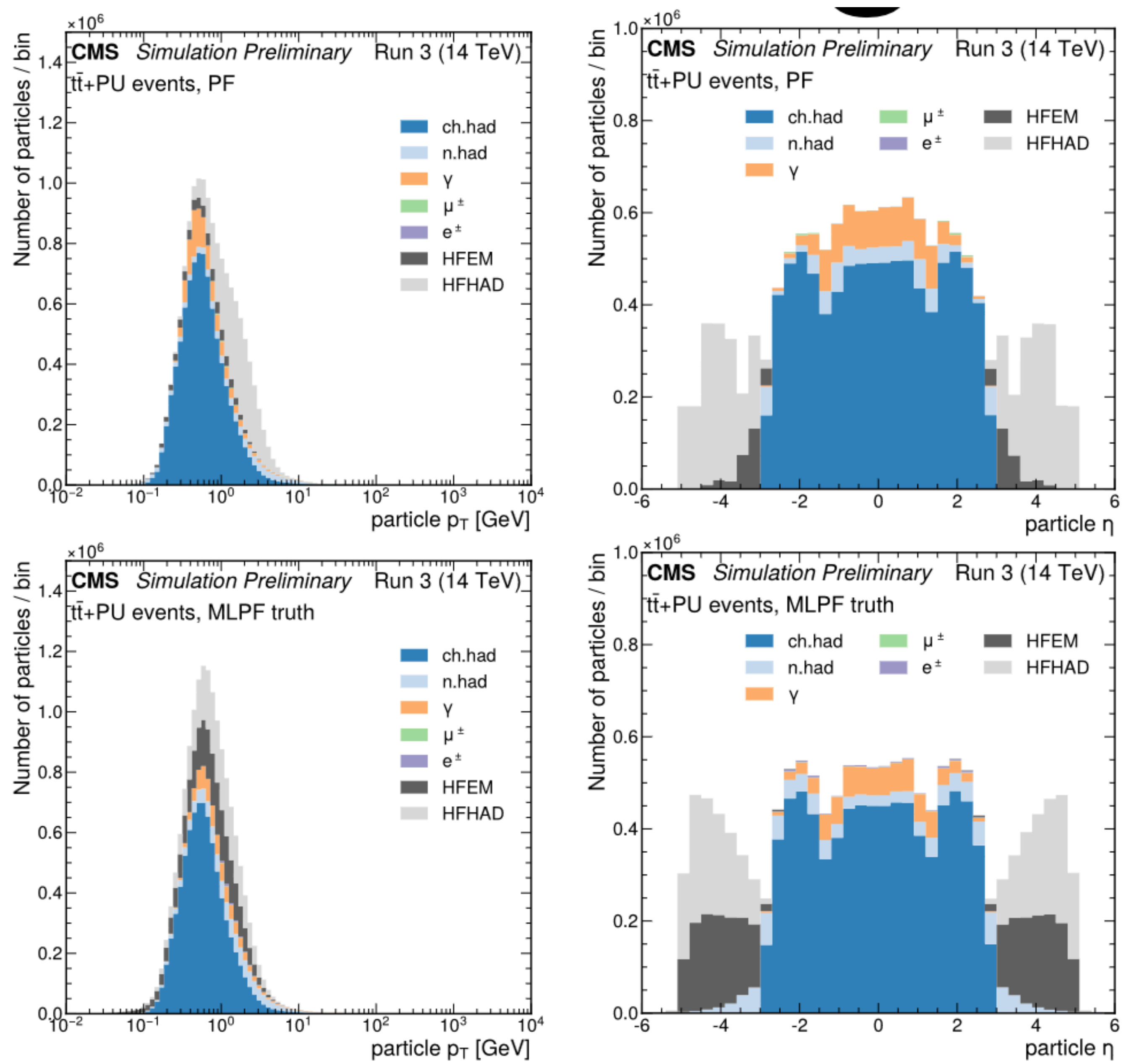
Stacked models

As an example (batch, elem, feat) = (2, 6400, 25)



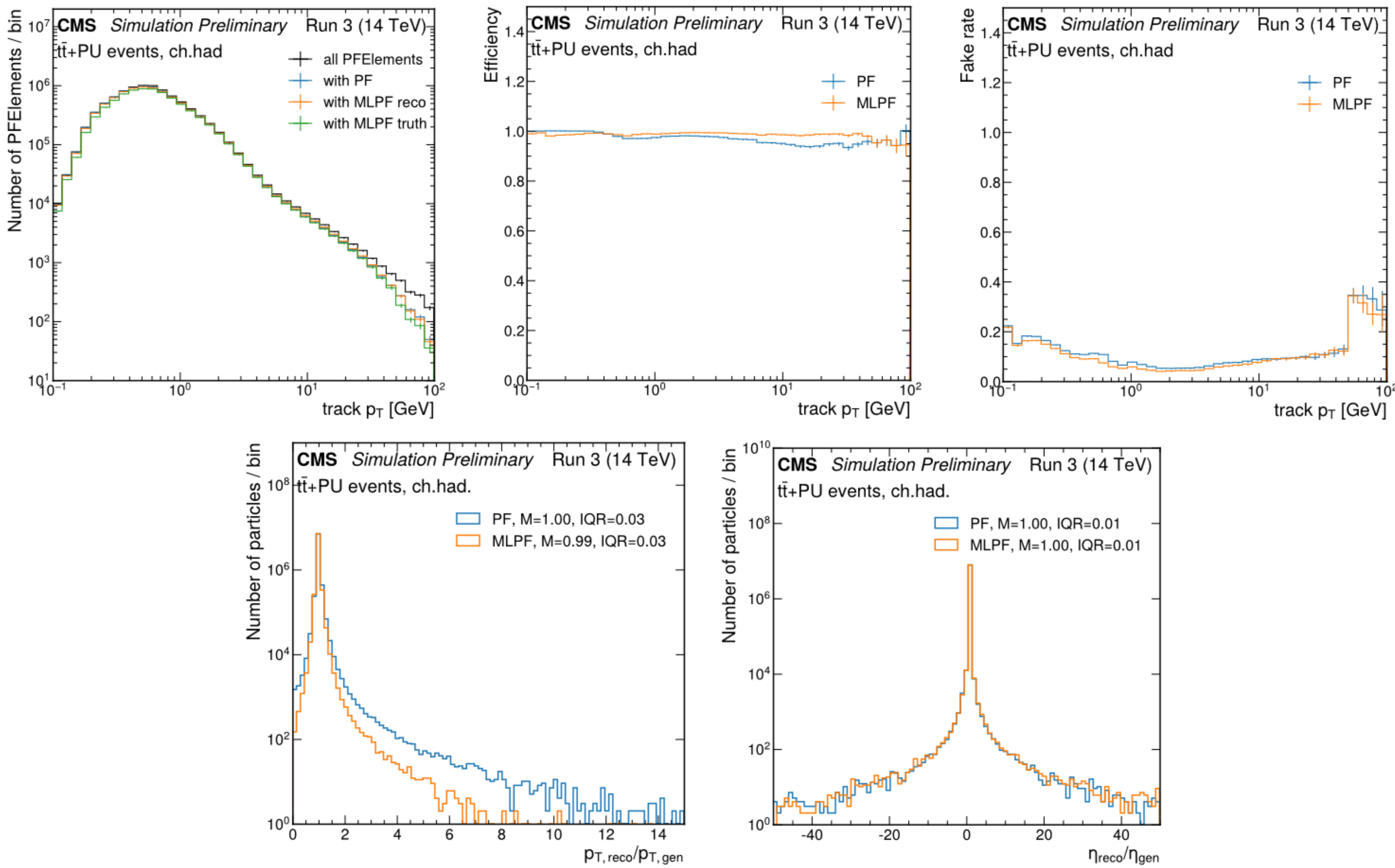
Graph building can happen at multiple layers in the model.

Truth-level training in CMS



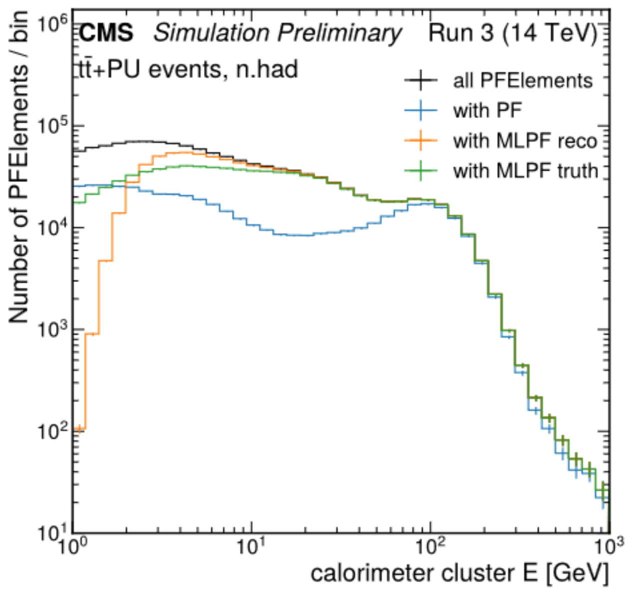
JP, Javier Duarte, Farouk Mokhtar, Eric Wulff, Jieun Yoo, Jean-Roch Vlimant, Maurizio Pierini, Maria Girone. Machine Learning for Particle Flow Reconstruction at CMS. ACAT 2021. https://doi.org/10.48550/arXiv.2203.00330, http://cds.cern.ch/record/2792320

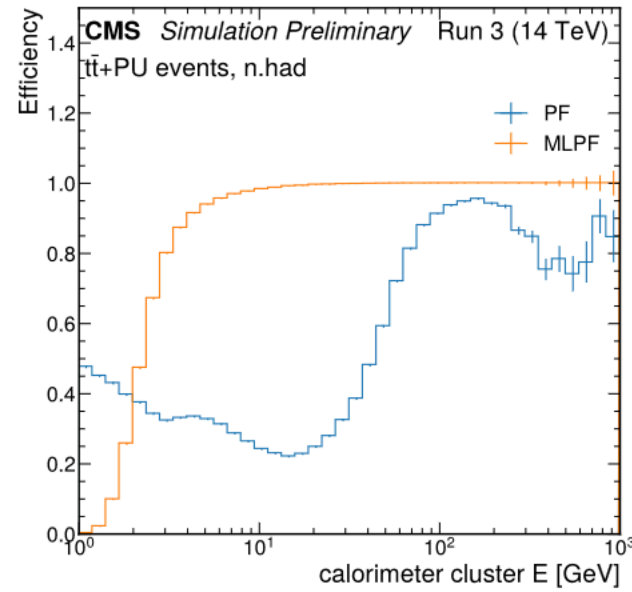
Charged hadrons

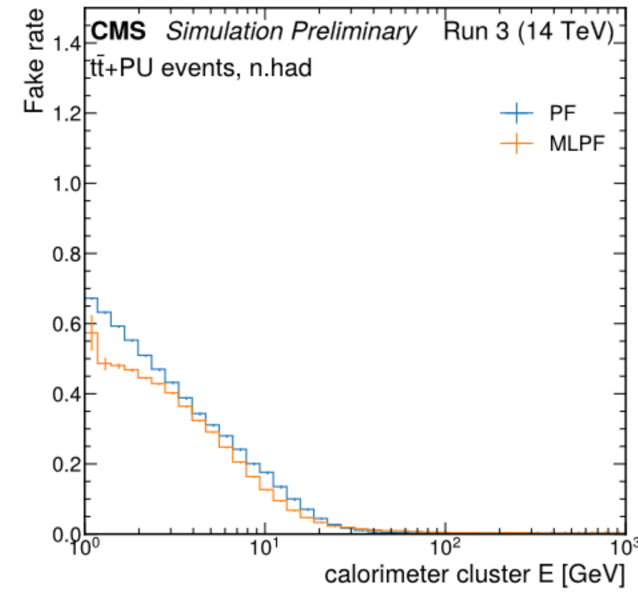


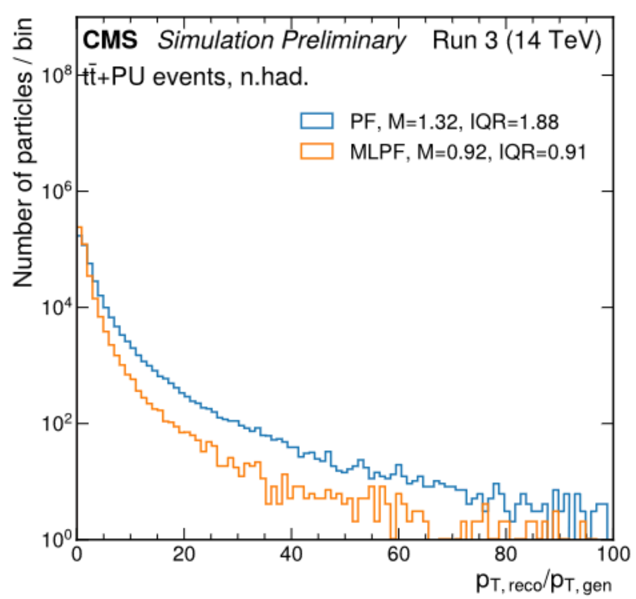
We can improve charged hadron eff/fake-rate and resolution.

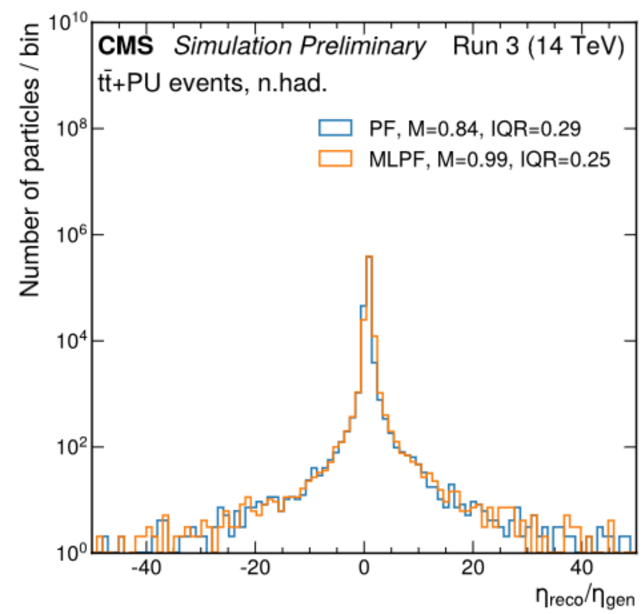
Neutral hadrons





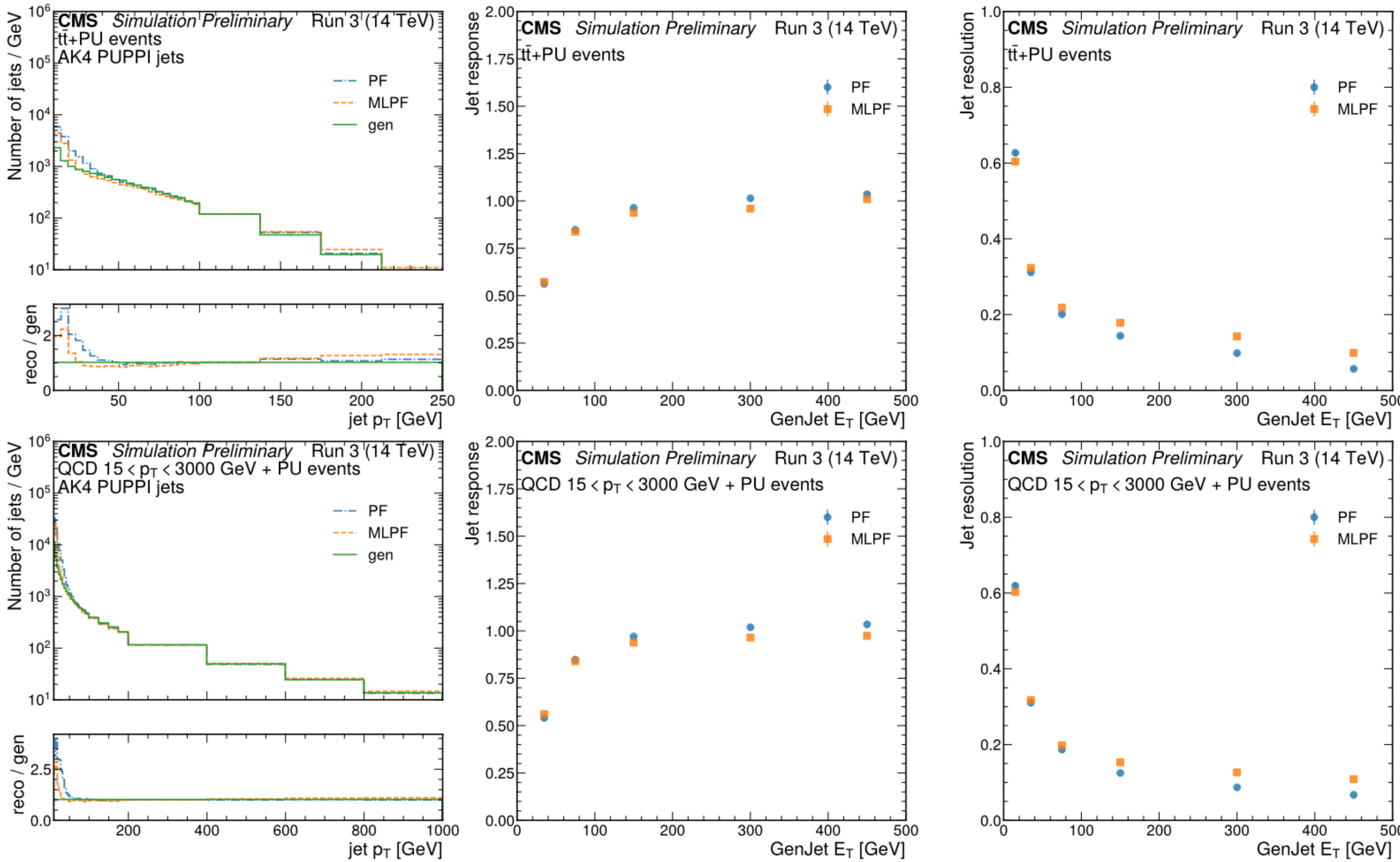






Neutral hadrons now have a clear turn-on and improved resolution.

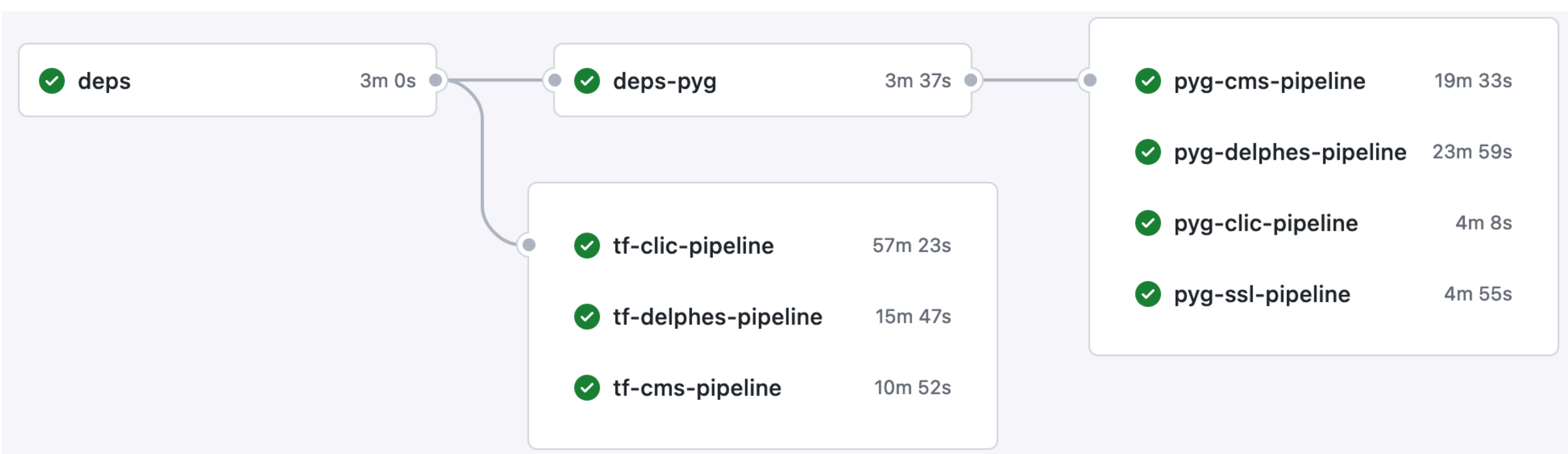
Jets



Jet response is compatible between PF and MLPF trained on a gen/sim level target.

Common repo

- MLPF in https://github.com/jpata/particleflow aims at providing
  - a set of PF-related datasets for the community in a single repository
  - a platform for testing and comparing ML-based reco models for PF under comparable circumstances
  - computationally scalable baseline models that can be run on various platforms for ML engineering studies



Discussion

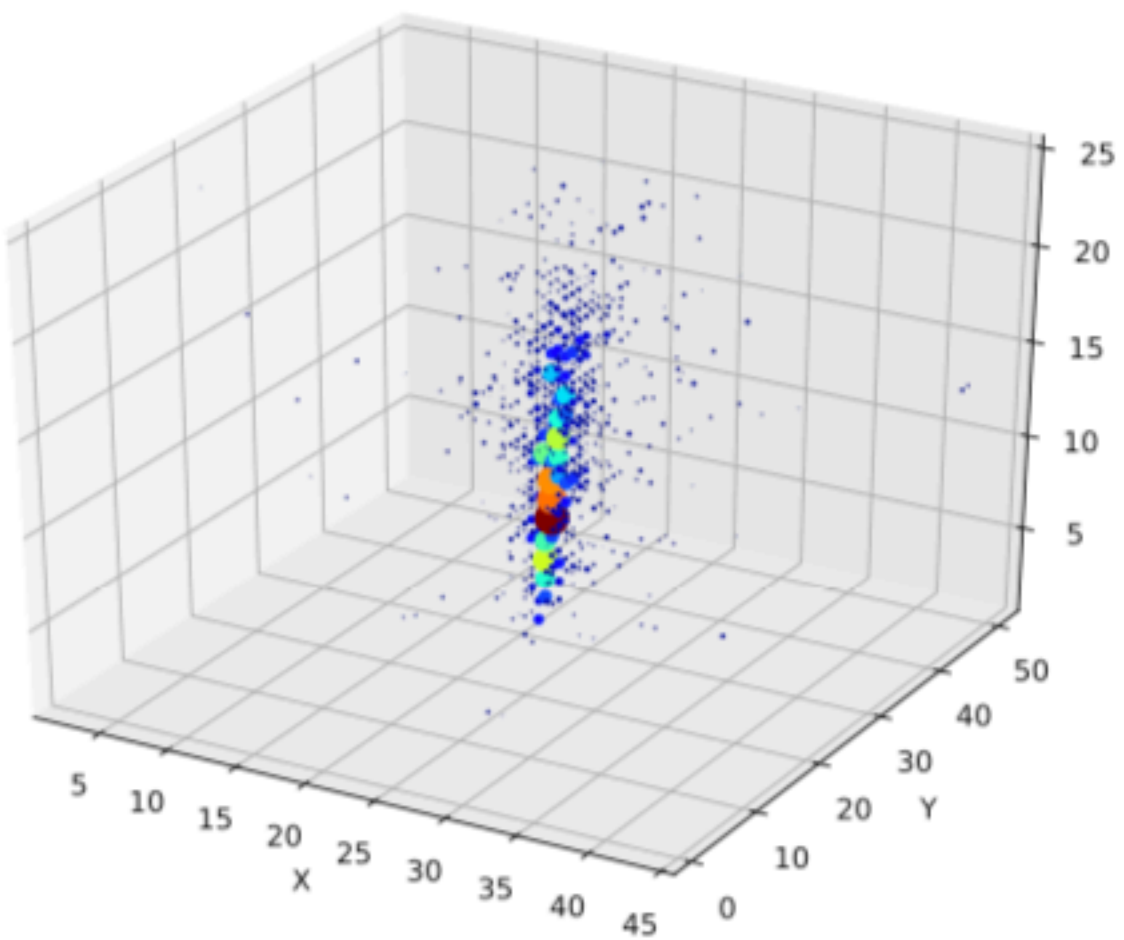
- **Aim towards realistic, open benchmark datasets and baseline algorithms!**
- **What are the goals of reconstruction?** Unique clustering/ segmentation, particle-level physics reconstruction, event-level (jets, MET) physics reconstruction?
- **To what extent can ML for simulation and ML for reco approaches inform each other?** Are they the inverse of each other? Can one construct a model that does both?
- So far, synthetic data (simulation) is driving the efforts. What is the role of data-driven approaches, e.g. **learning representations from data**, fine-tuning on specific tasks?

Backup

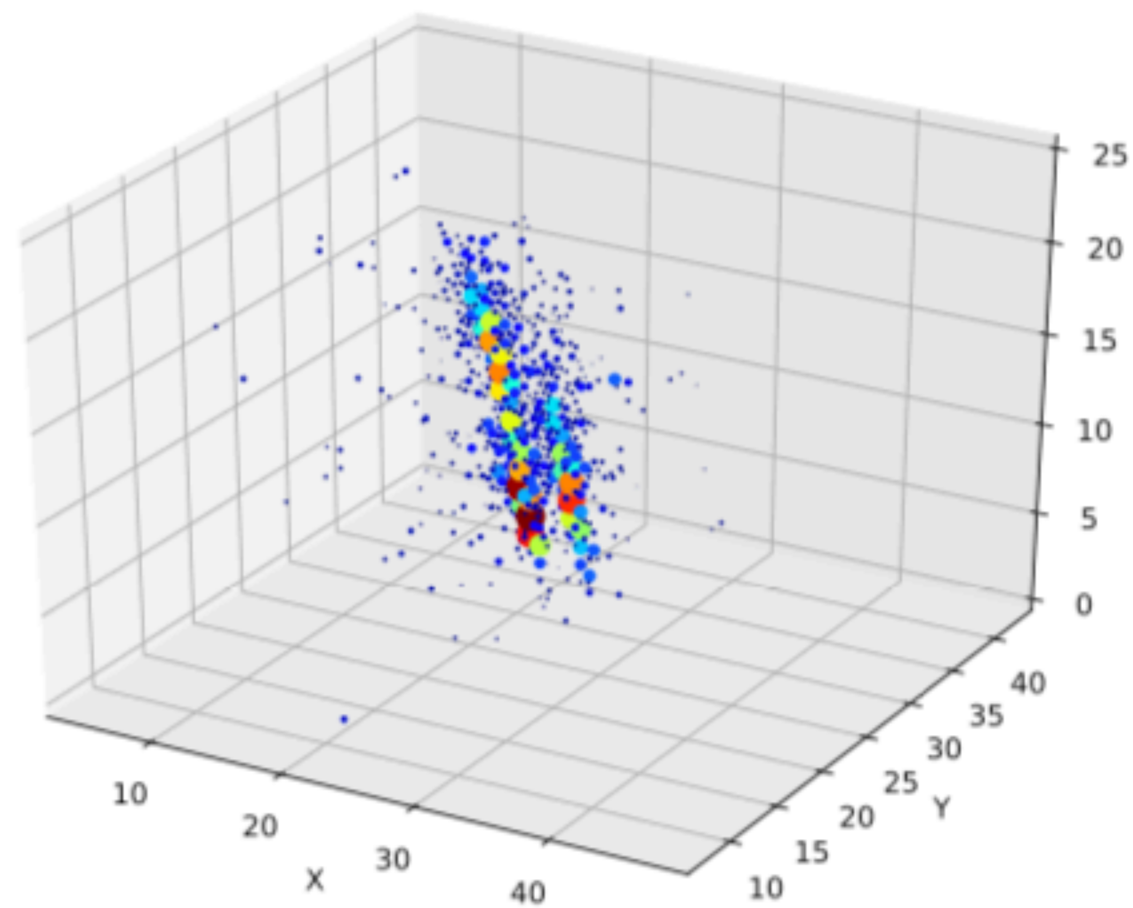
Single particle showers

Data are very sparse!

photon

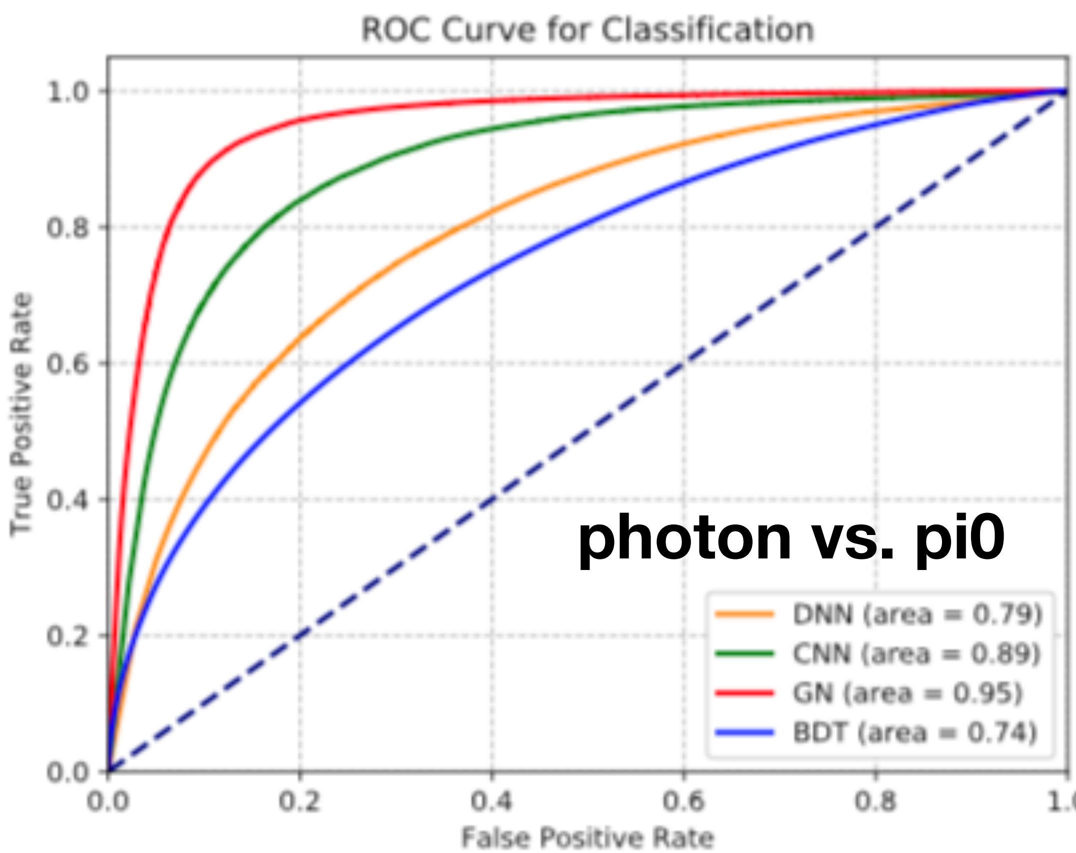


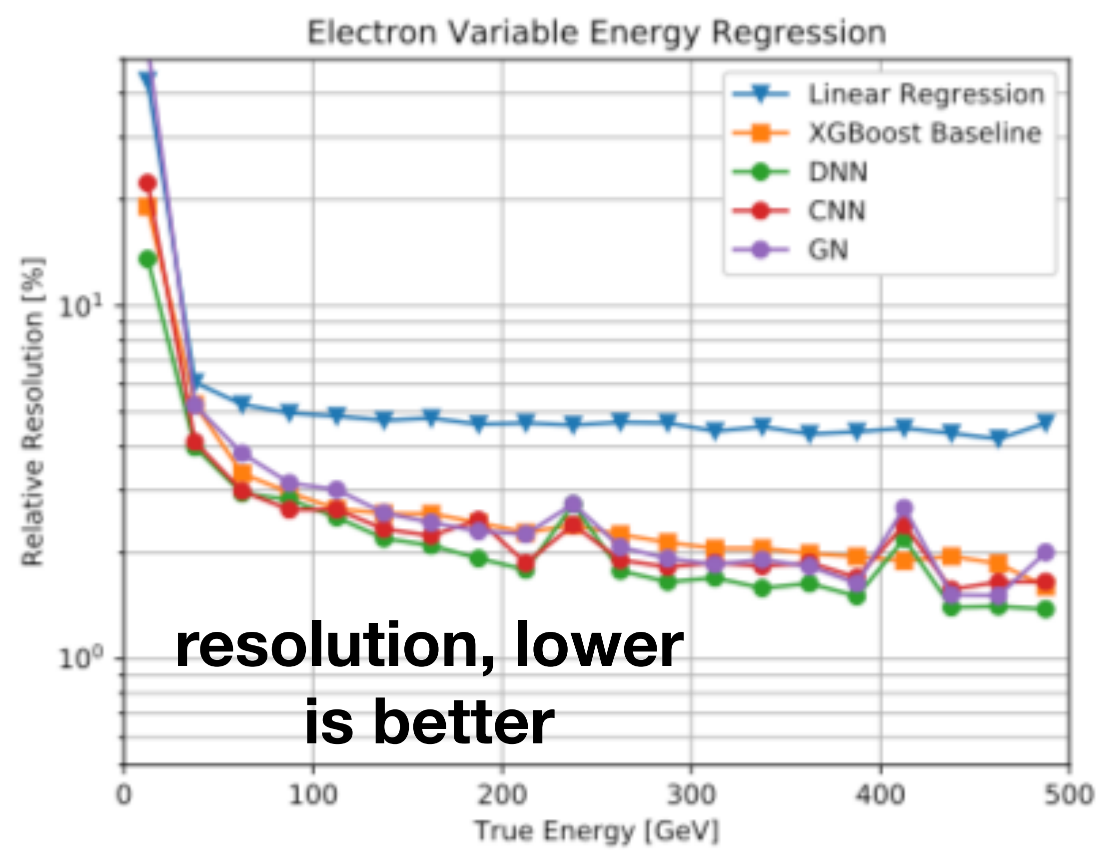
pi0 -> two photons



Belayneh, D., Carminati, F., Farbin, A. *et al.* Calorimetry with deep learning: particle simulation and reconstruction for collider physics. *Eur. Phys. J. C* **80,** 688 (2020). https://doi.org/10.1140/epjc/s10052-020-8251-9

Multi-task learning





DNNs/CNNs on granular detetors are performant for shower identification and energy regression.

Belayneh, D., Carminati, F., Farbin, A. *et al.* Calorimetry with deep learning: particle simulation and reconstruction for collider physics. *Eur. Phys. J. C* **80,** 688 (2020). https://doi.org/10.1140/epjc/s10052-020-8251-9

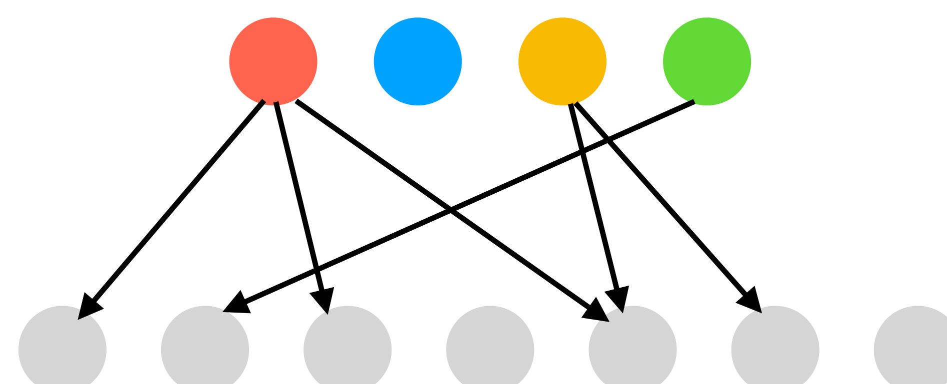
Particle representation

- The ground truth is a set of simulation particles (p4, ID)
- The input is the set of all calorimeter hits (energy, location)

truth particles

simulation information

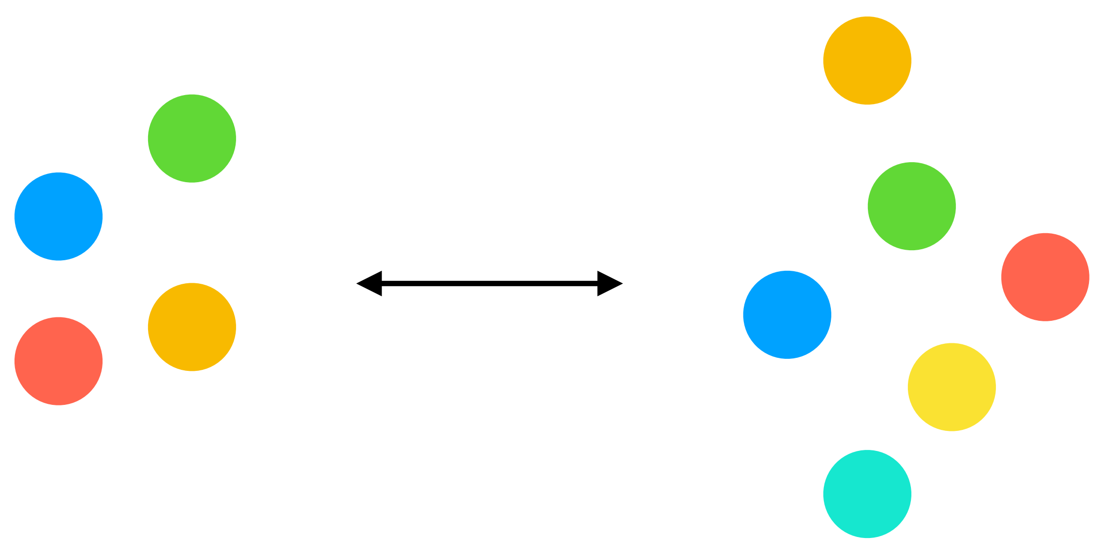
input hits



An unknown number of different truth particles (segmentation labels).

Set-to-set problem

Each particle is described by a multi-class label, and is embedded in a complex, problem-dependent feature space.



ground truth
simulation input particles

prediction
reconstructed particles

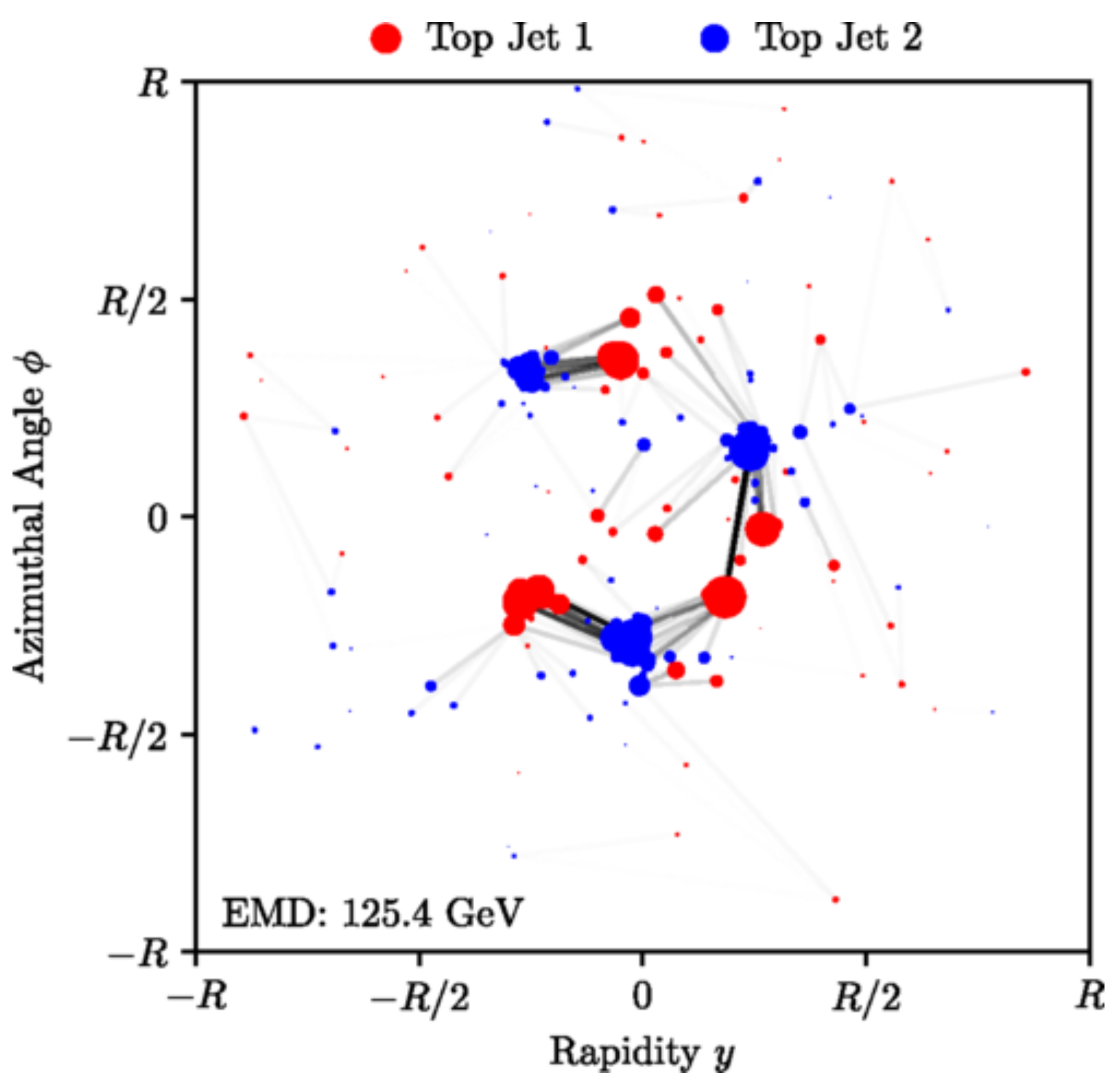
How to compare two sets of arbitrary size with complex features?
How to do it differentiably, in a performant way?

Energy flow

$$\mathrm{EMD}(\mathcal{E},\mathcal{E}') = \min_{\{f_{ij}\ge 0\}} \sum_{ij} f_{ij}\frac{\theta_{ij}}{R} + \left|\sum_i E_i - \sum_j E'_j\right|,$$

$$\sum_j f_{ij} \le E_i, \qquad \sum_i f_{ij} \le E'_j, \qquad \sum_{ij} f_{ij} = E_{\min},$$

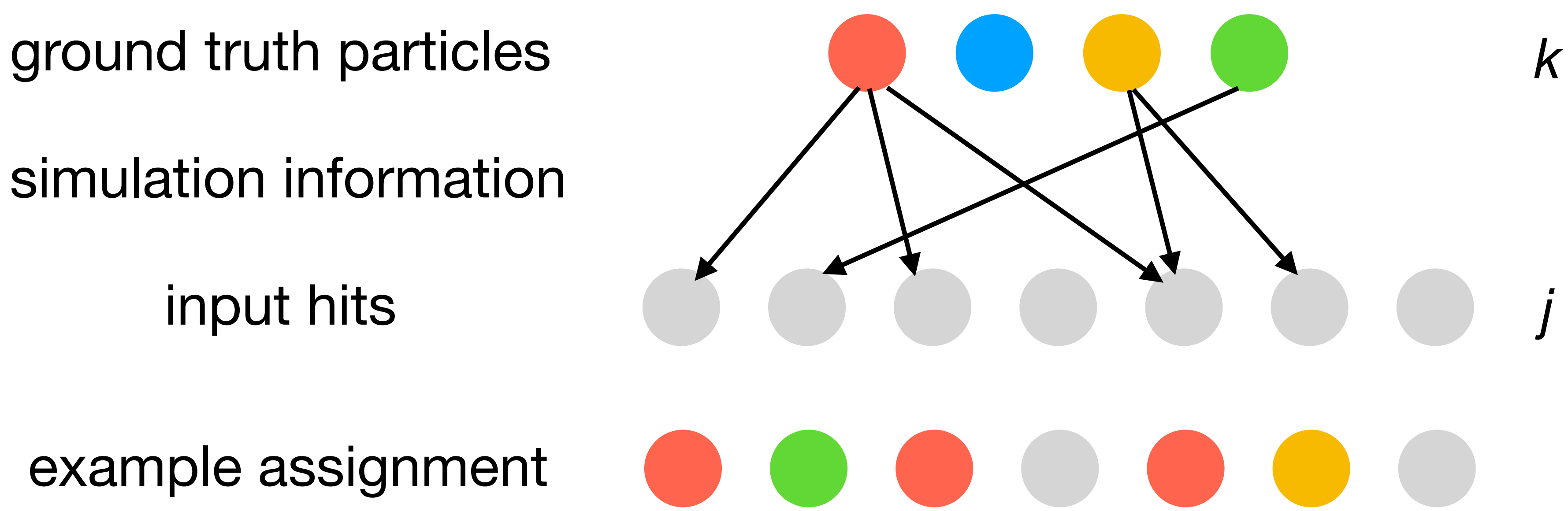
- Use Earth Mover's Distance to define a differentiable loss between two sets of particles described by (E, eta, phi)
- Good theoretical properties, not sensitive to soft particles / collinear radiation
- Optimal Transport is challenging to practically compute on large sets



Patrick T. Komiske, Eric M. Metodiev, and Jesse Thaler; Phys. Rev. Lett. **123**, 041801; https://doi.org/10.1103/PhysRevLett.123.041801

Object condensation

Boundedness: the number of truth particles usually cannot be larger than the number of inputs (typically it's much smaller).



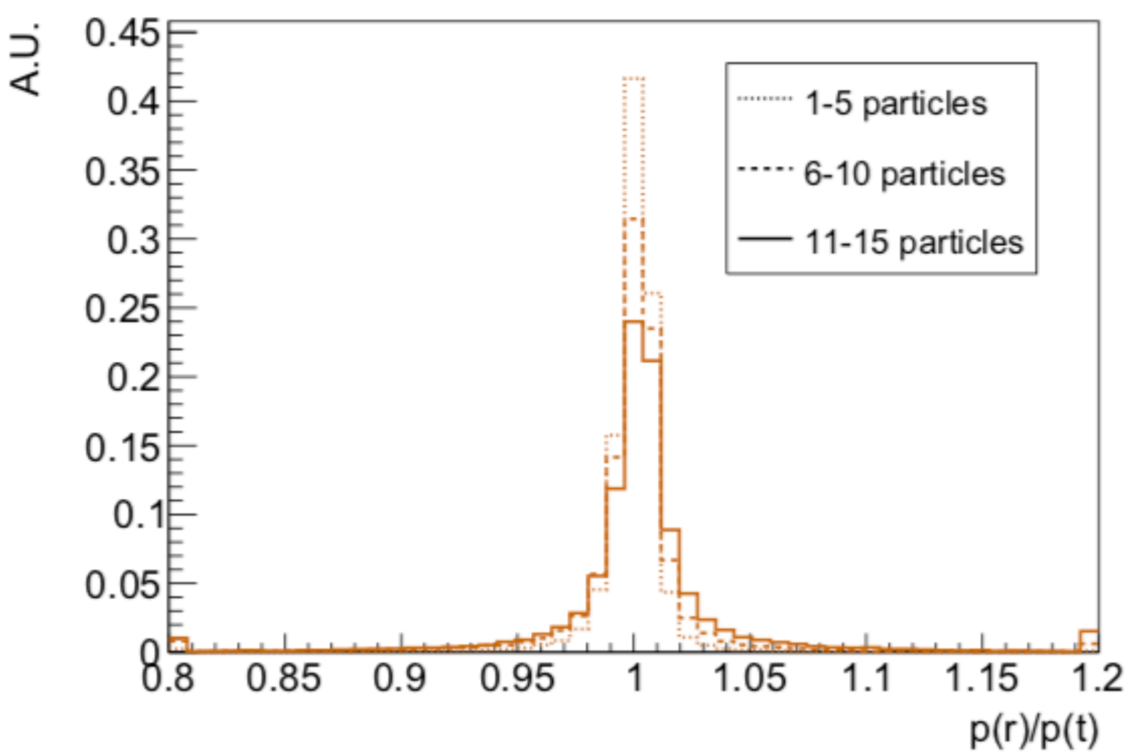
Each input represents exactly one truth particle, with attractive/repulsive potentials in a learned space x_j between correct/incorrect assignments.

$$L_V = \frac{1}{N} \sum_{j=1}^{N} q_j \sum_{k=1}^{K} \Big(\underbrace{M_{jk} \check{V}_k(x_j)}_{\text{attractive}} + \underbrace{(1 - M_{jk}) \hat{V}_k(x_j)}_{\text{repulsive}} \Big) .$$

Kieseler, J. Object condensation: one-stage grid-free multi-object reconstruction in physics detectors, graph, and image data. *Eur. Phys. J. C* **80,** 886 (2020). https://doi.org/10.1140/epjc/s10052-020-08461-2

A simplified set-to-set loss

This approximation is fairly model-independent (e.g. not tied to GNNs). The exact form of the potentials is a hyperparameter.



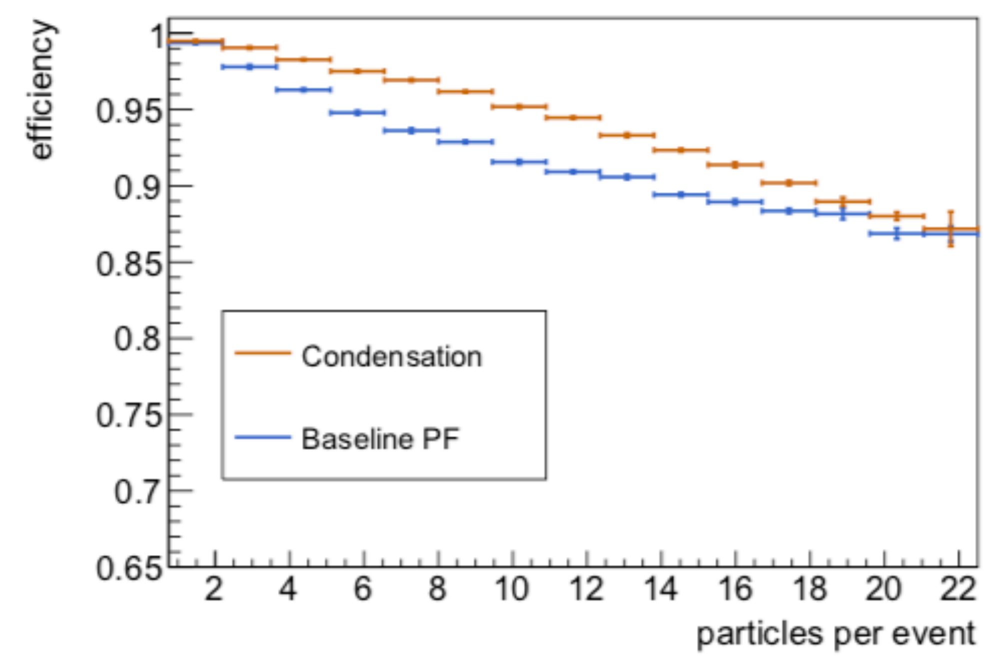


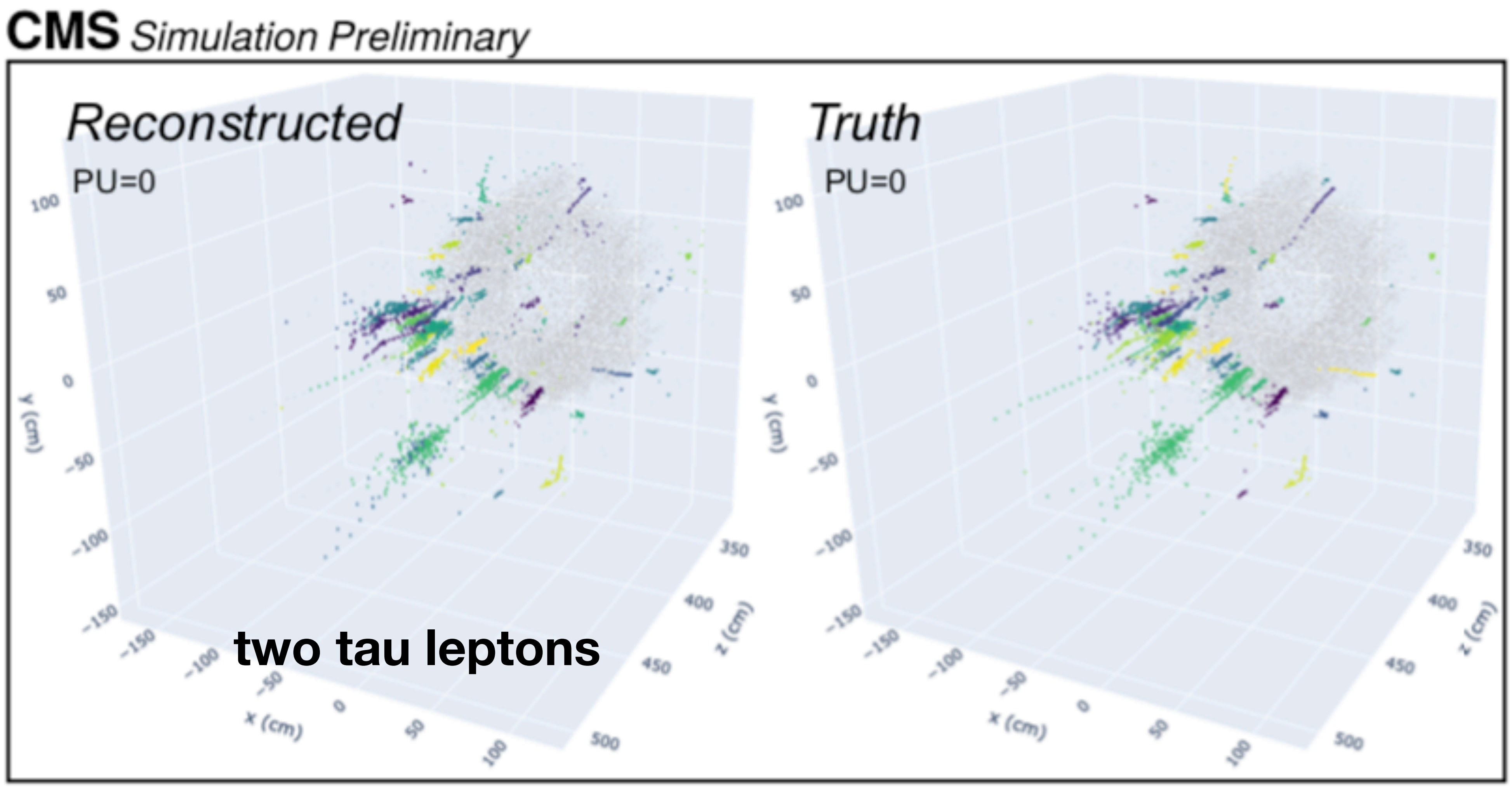
Fig. 6 Reconstruction efficiency as a function of the particle multiplicity in the event

Can be used for constructing particle reconstruction models across a varied number of inputs.

Kieseler, J. Object condensation: one-stage grid-free multi-object reconstruction in physics detectors, graph, and image data. *Eur. Phys. J. C* **80,** 886 (2020). https://doi.org/10.1140/epjc/s10052-020-08461-2

Realistic clustering with ML

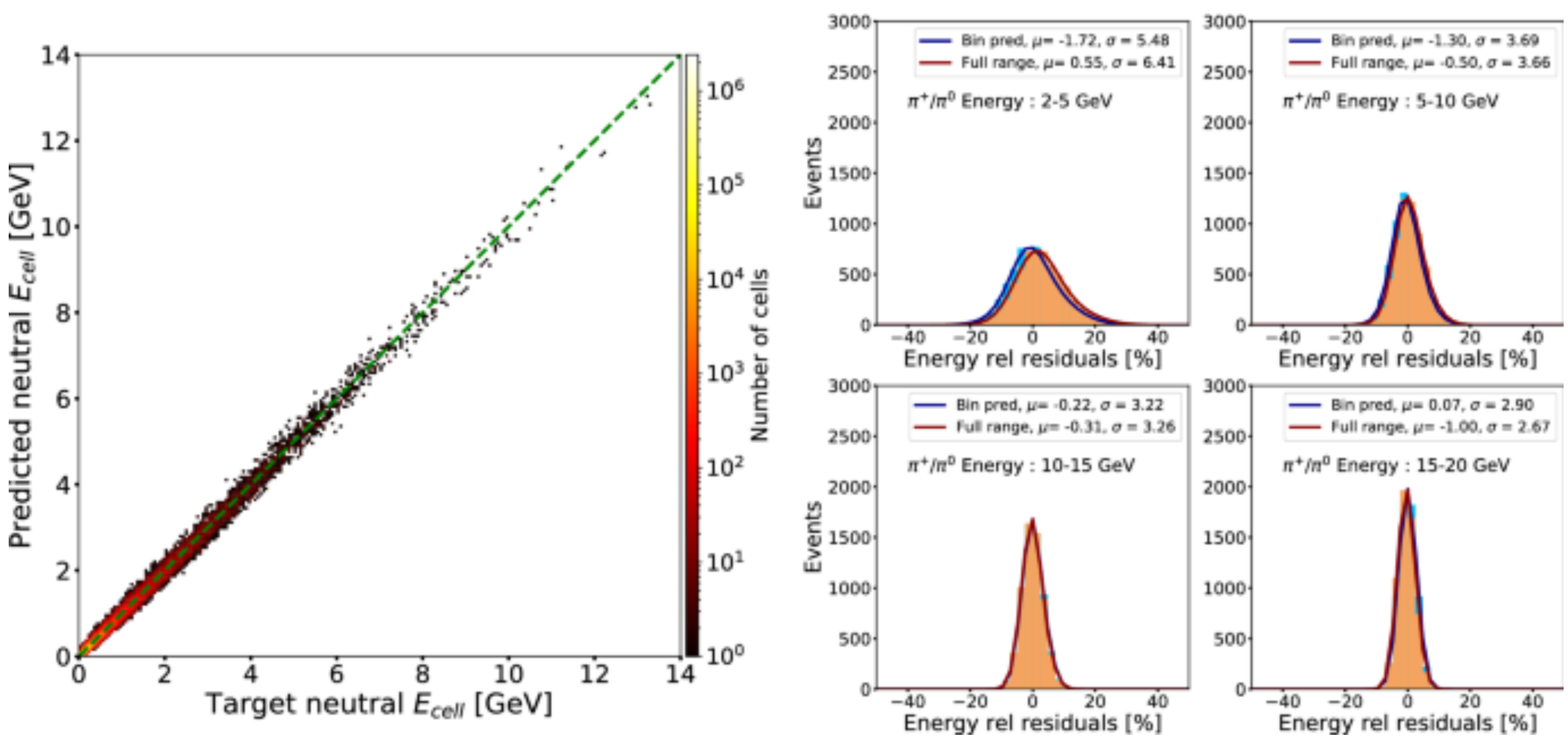
Simulation-level particles → simulation energy deposits → reconstructed energy deposits → predict the cluster label (or noise) for each hit.



Saptaparna Bhattacharya, Nadezda Chernyavskaya, Saranya Ghosh, Lindsey Gray, Jan Kieseler et al. GNN-based end-to-end reconstruction in the CMS Phase 2 High-Granularity Calorimeter. ACAT 2021. https://doi.org/10.48550/arXiv.2203.01189

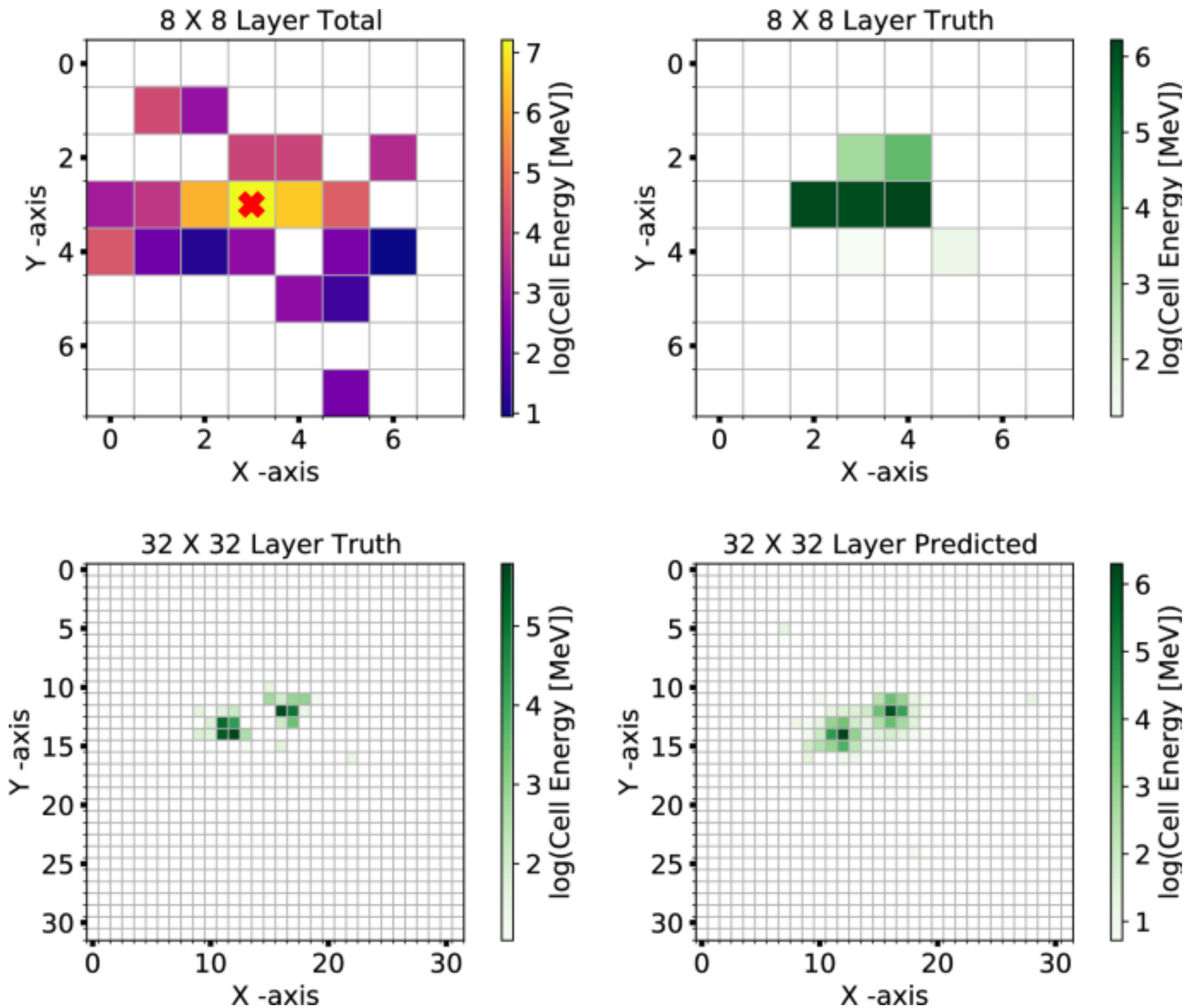
Neutral energy regression

The image-based approach is competitive for the cell neutral energy prediction compared to the algorithmic baseline.



Di Bello, F.A., Ganguly, S., Gross, E. et al. Towards a computer vision particle flow. Eur. Phys. J. C 81, 107 (2021). https://doi.org/10.1140/epjc/s10052-021-08897-0

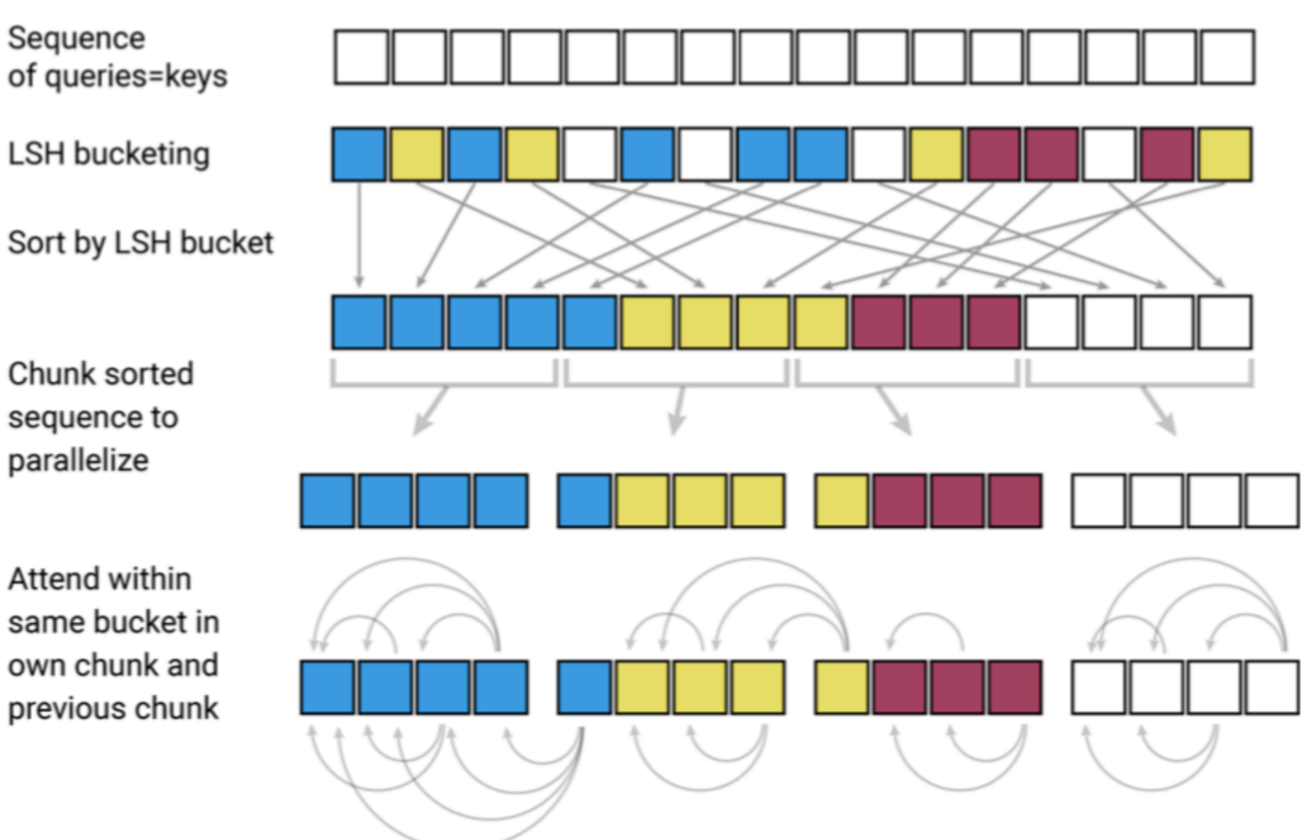
Super-resolution

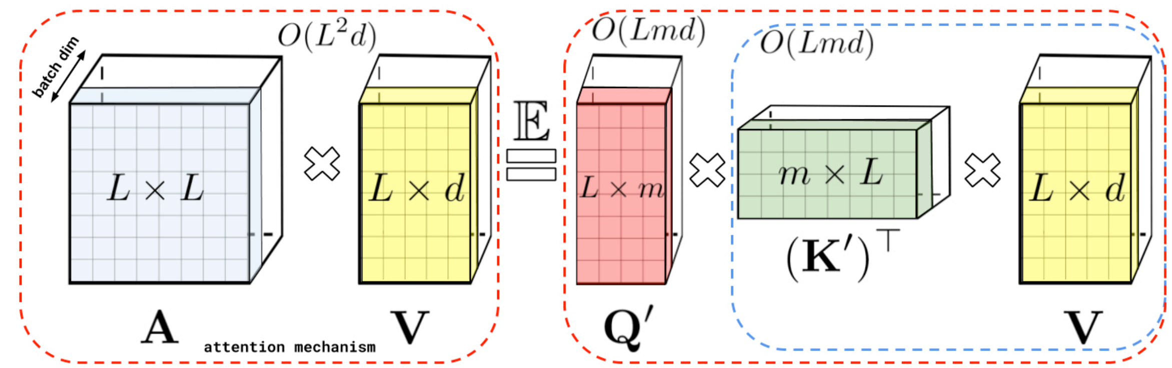


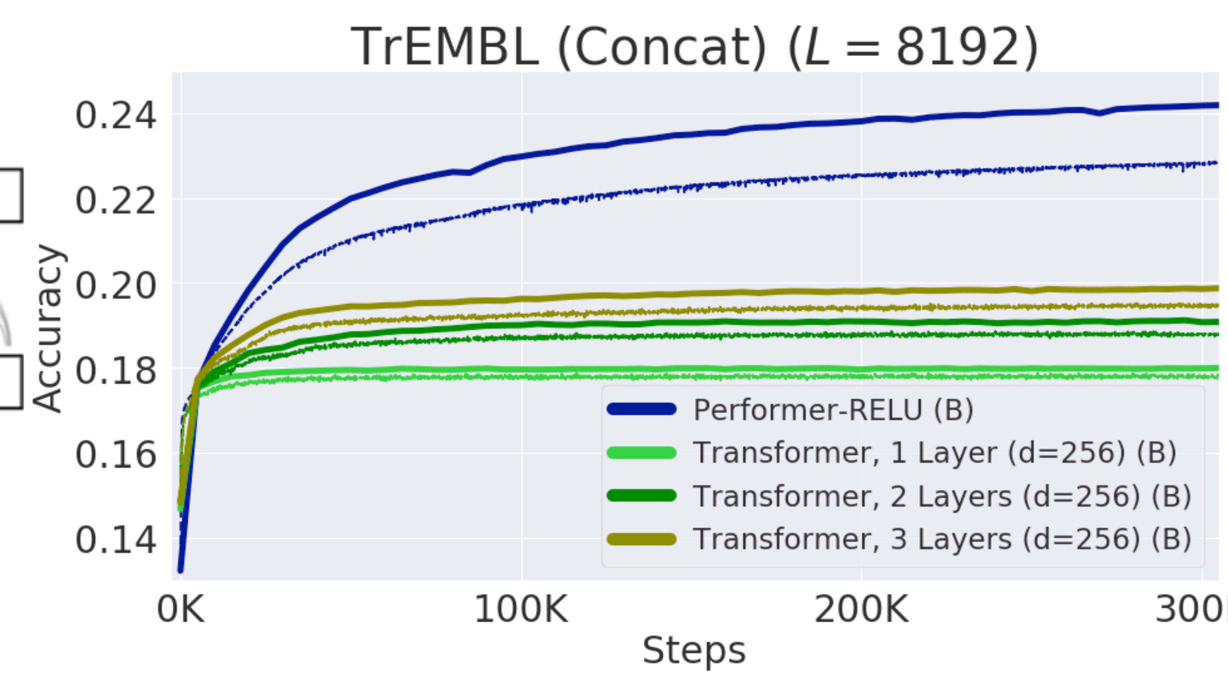
Di Bello, F.A., Ganguly, S., Gross, E. et al. Towards a computer vision particle flow. Eur. Phys. J. C 81, 107 (2021). https://doi.org/10.1140/epjc/s10052-021-08897-0

Scalable models

The computational scaling of models on large sets/ sequences is an active topic.



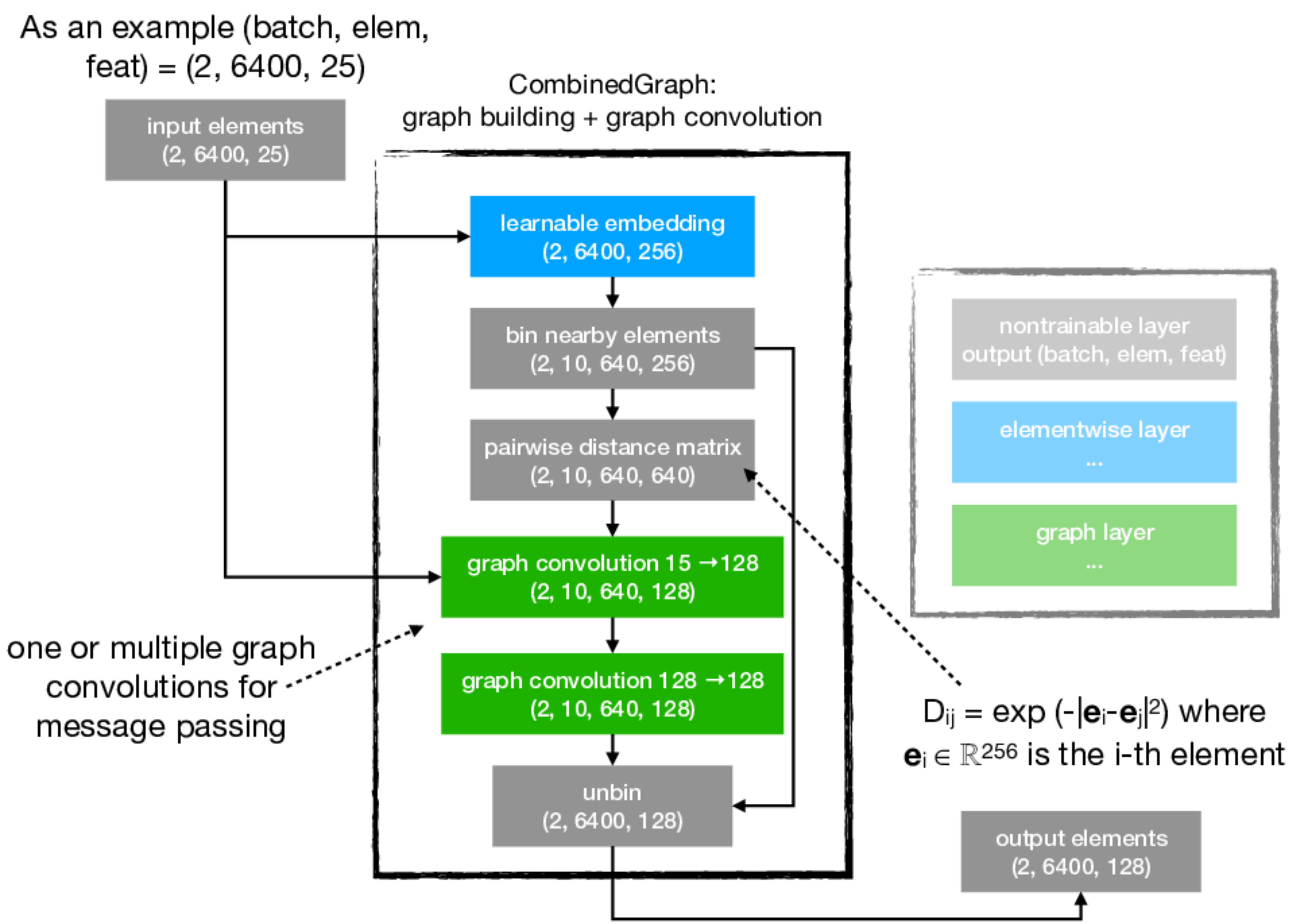




https://arxiv.org/pdf/2001.04451.pdf

https://arxiv.org/abs/2001.04451

Implementation for sets

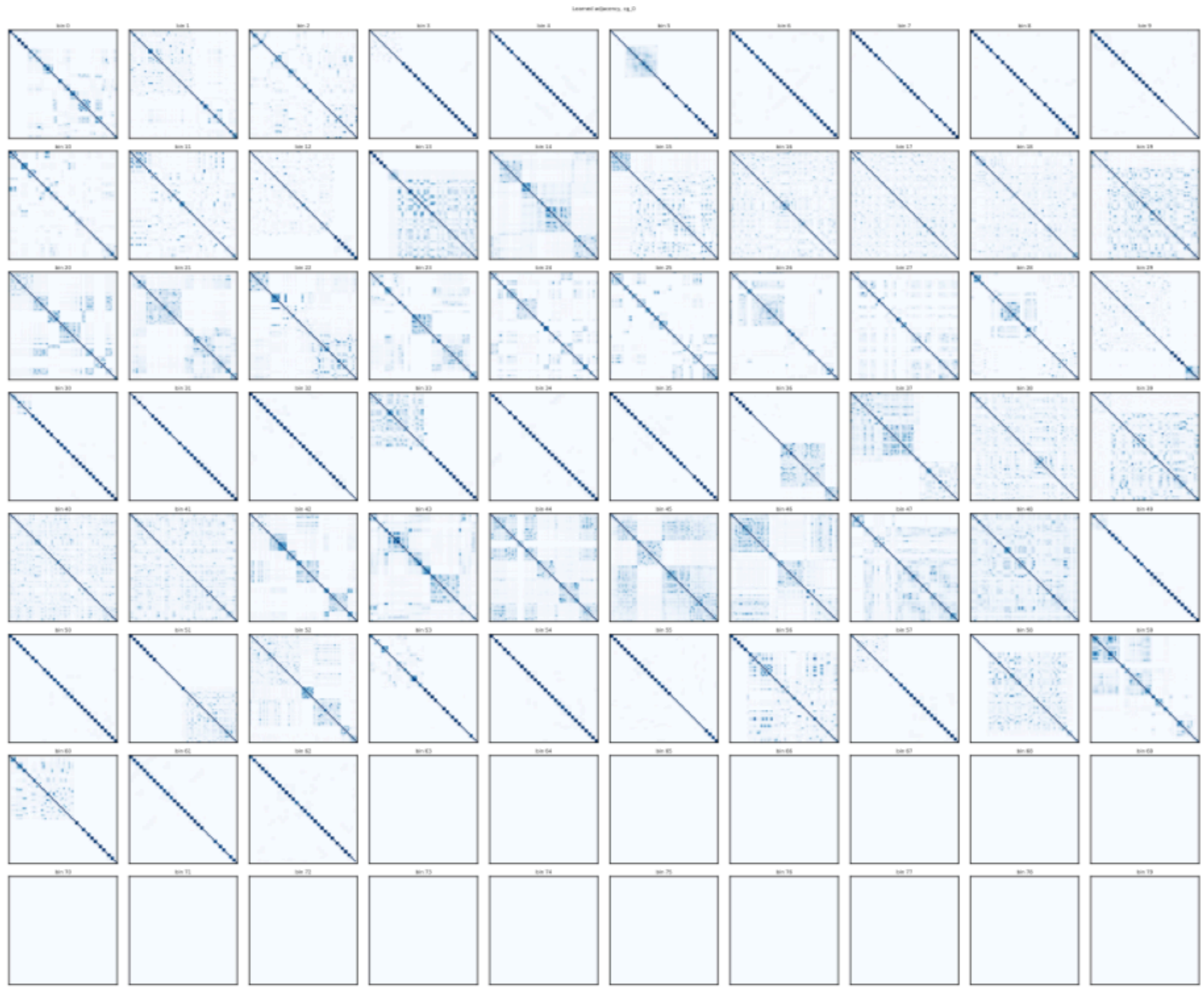


Uses built-in dense matrix, reshape and scatter/gather operations in TF.

Requires batch-mode graphs. No N^2 allocation or computation needed.

Pata, J. et al. Machine Learning for Particle Flow Reconstruction at CMS. ACAT 2021. https://doi.org/10.48550/arXiv.2203.00330

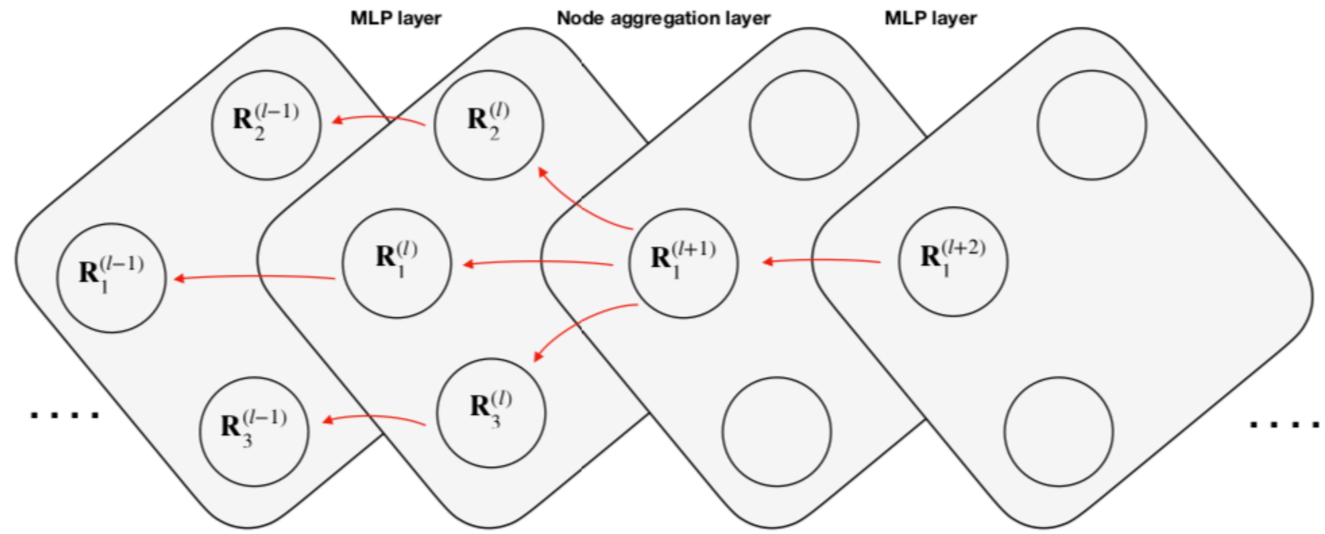
Disjoint event graphs



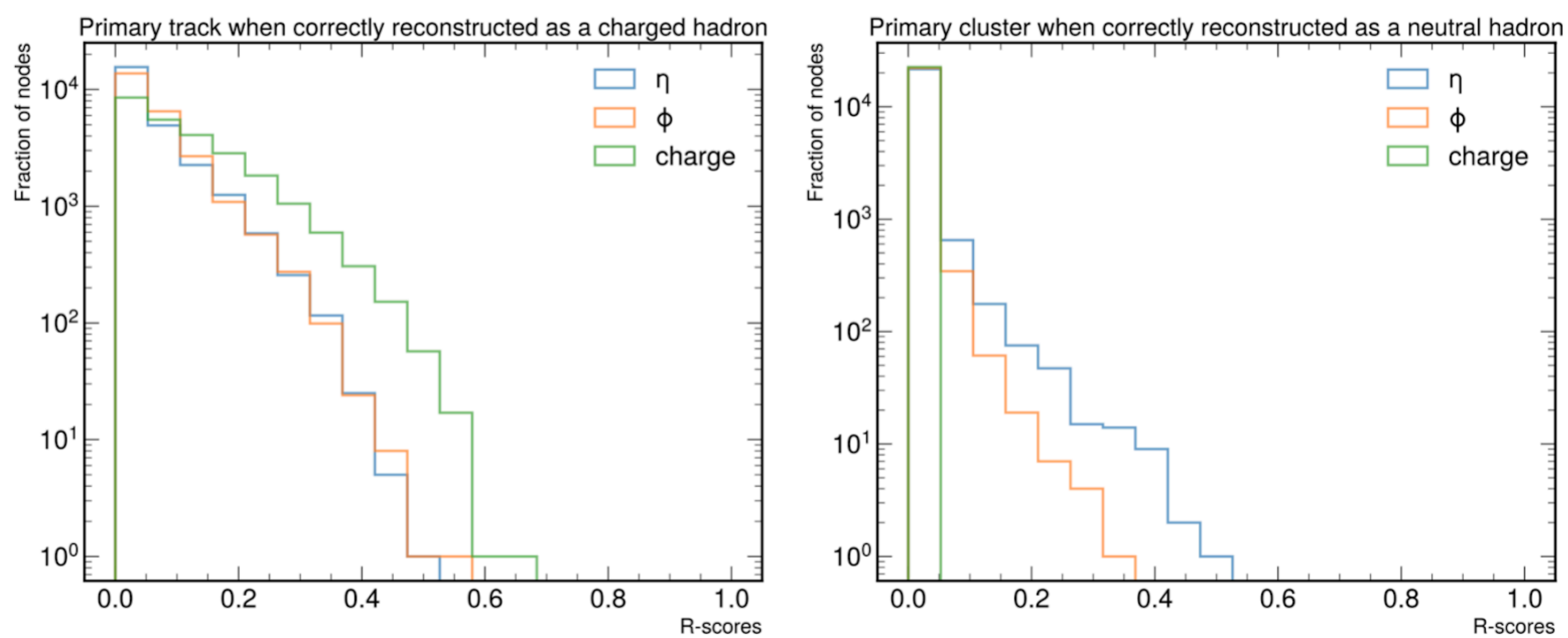
Pata, J. et al. Machine Learning for Particle Flow Reconstruction at CMS. ACAT 2021. https://doi.org/10.48550/arXiv.2203.00330

Interpretability

- What inputs are relevant for a particular model output?
- Compute layerwise relevance scores **R**
- Aggregate along the graph structure



$$\mathbf{R}_j^{(l)} = \sum_k \frac{x_j A_{jk}}{\sum_m x_m A_{mk}} \mathbf{R}_k^{(l+1)}$$



"Explaining machine-learned particle-flow reconstruction"; Farouk Mokhtar, Raghav Kansal, Daniel C Diaz Javier Duarte, **JP**, Maurizio Pierini, Jean-Roch Vlimant. NeurIPS 2021, Machine Learning and the Physical Sciences, https://doi.org/10.48550/arXiv.2111.12840