Machine Learning in reconstruction and calibration at the LHC

LHCP conference 2024

Christian Sonnabend - Boston, 07.06.2024

Introduction

ML = {Statistics; Linear algebra; Computer science; …}, but nowadays ML ≈ {Neural networks}

Cluster-level ⟶ Track-level ⟶ Tagging / triggering / calibrations / PID

CERN

ATLAS EXPERIMENT

ALICE

CMS

LHCb

Challenge

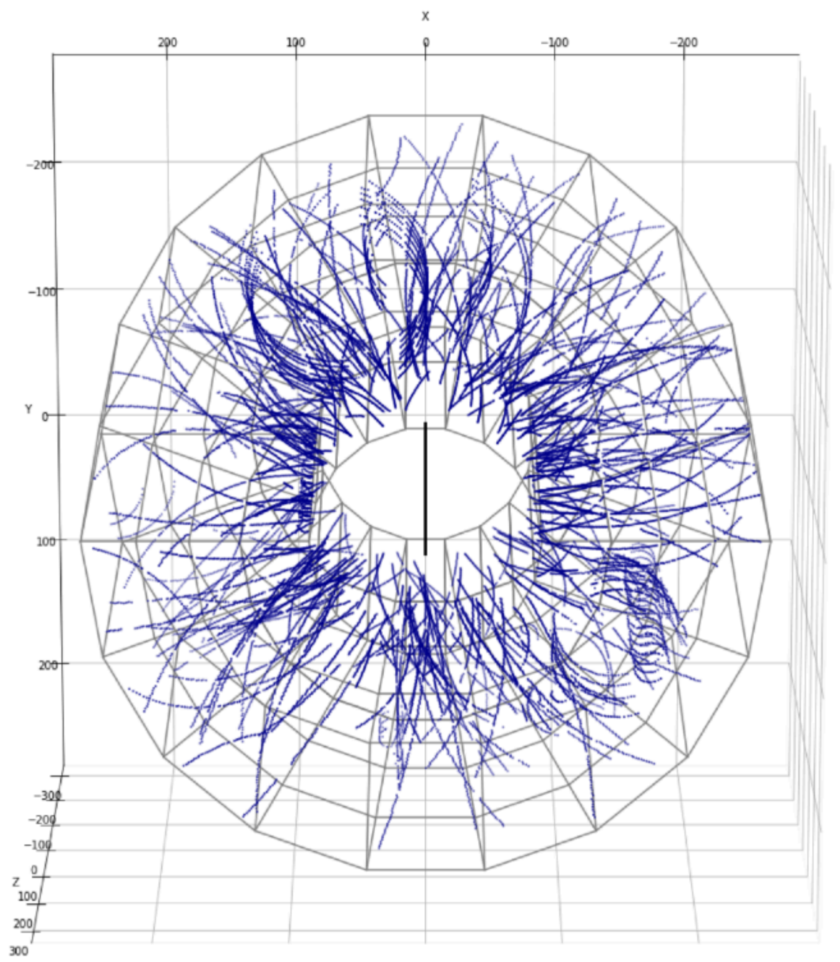
Informed / logical guesses to optimize all of the steps above for best performance

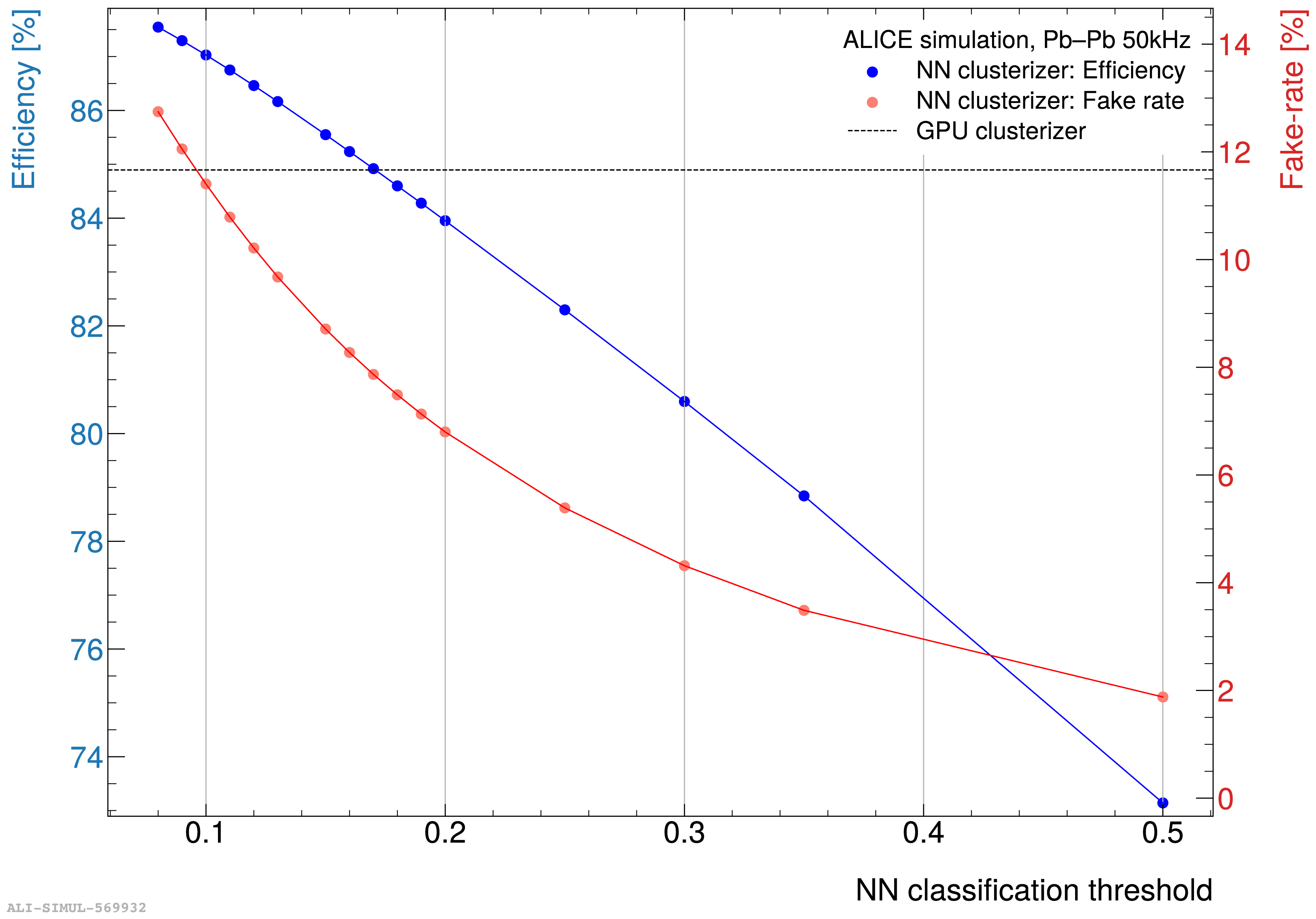
Cluster-level ⟶ Track-level ⟶ Tagging / triggering / calibrations / PID

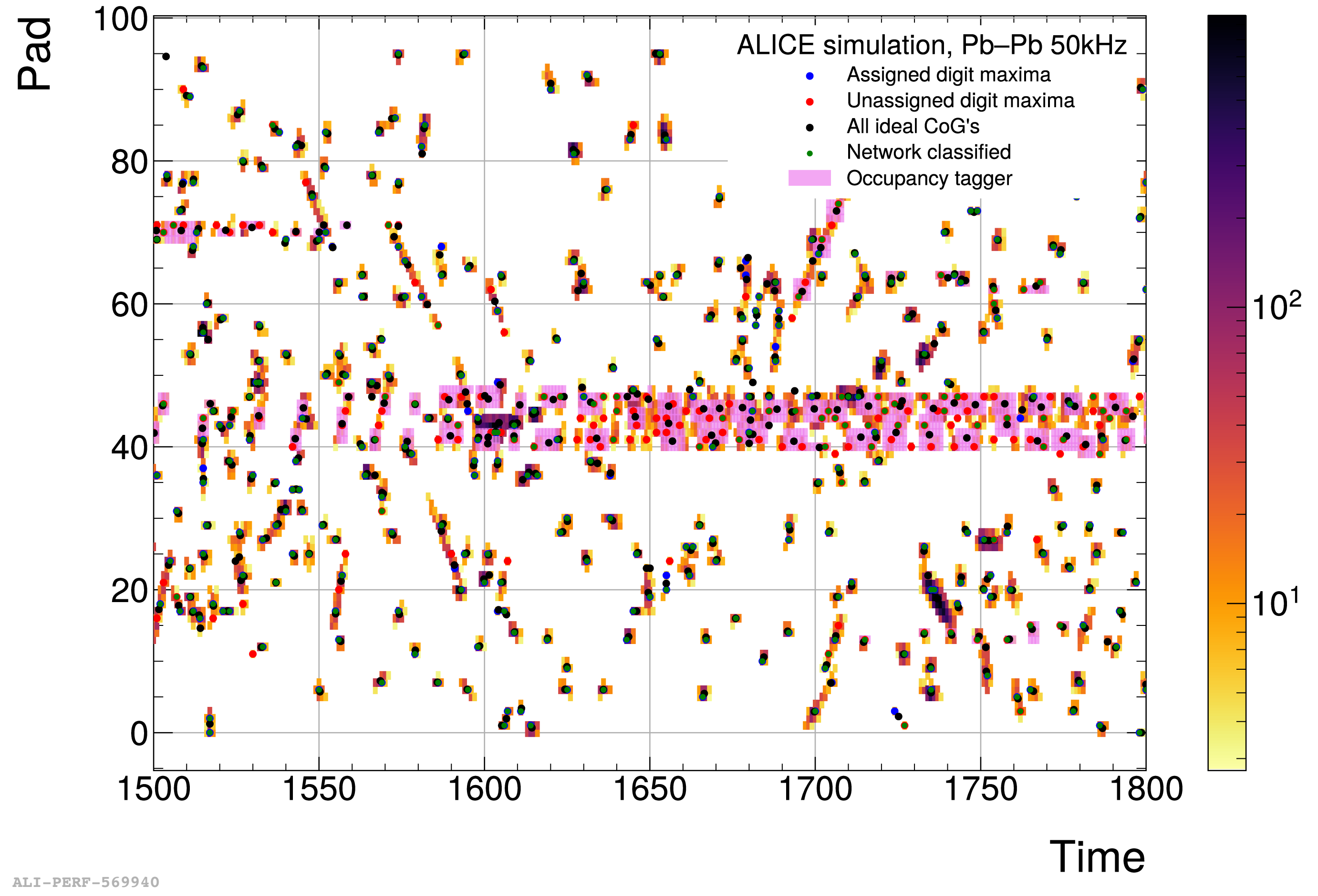
ALICE – DNN for online clusterization

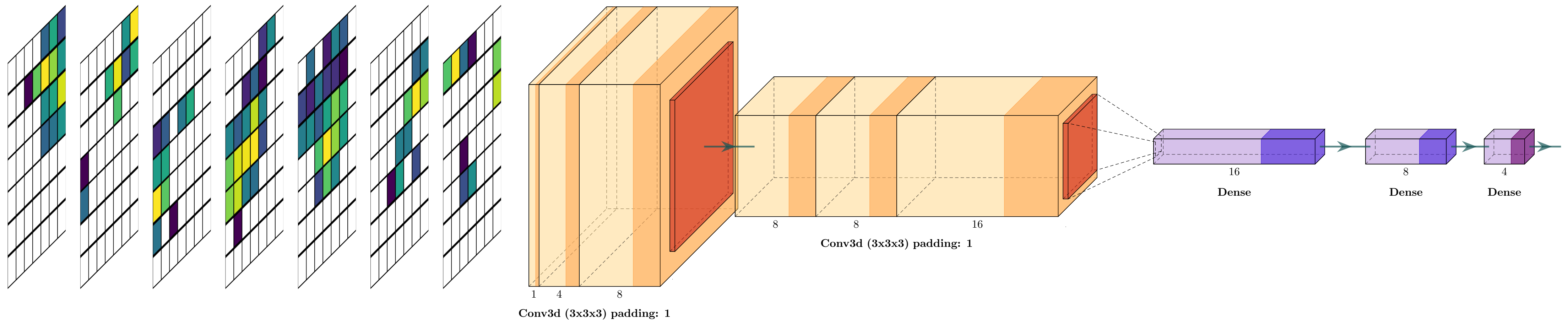
Number of clusters ≈ stored data size

- First step in online reconstruction: ~50 mio. clusters /s /gpu
- Reject clusters while maintaining tracking performance
- Add momentum vector estimate to improve track seeding
- Deploy in online GPU reconstruction of ALICE
- Reduce total number of clusters by ~10% while maintaining tracking performance





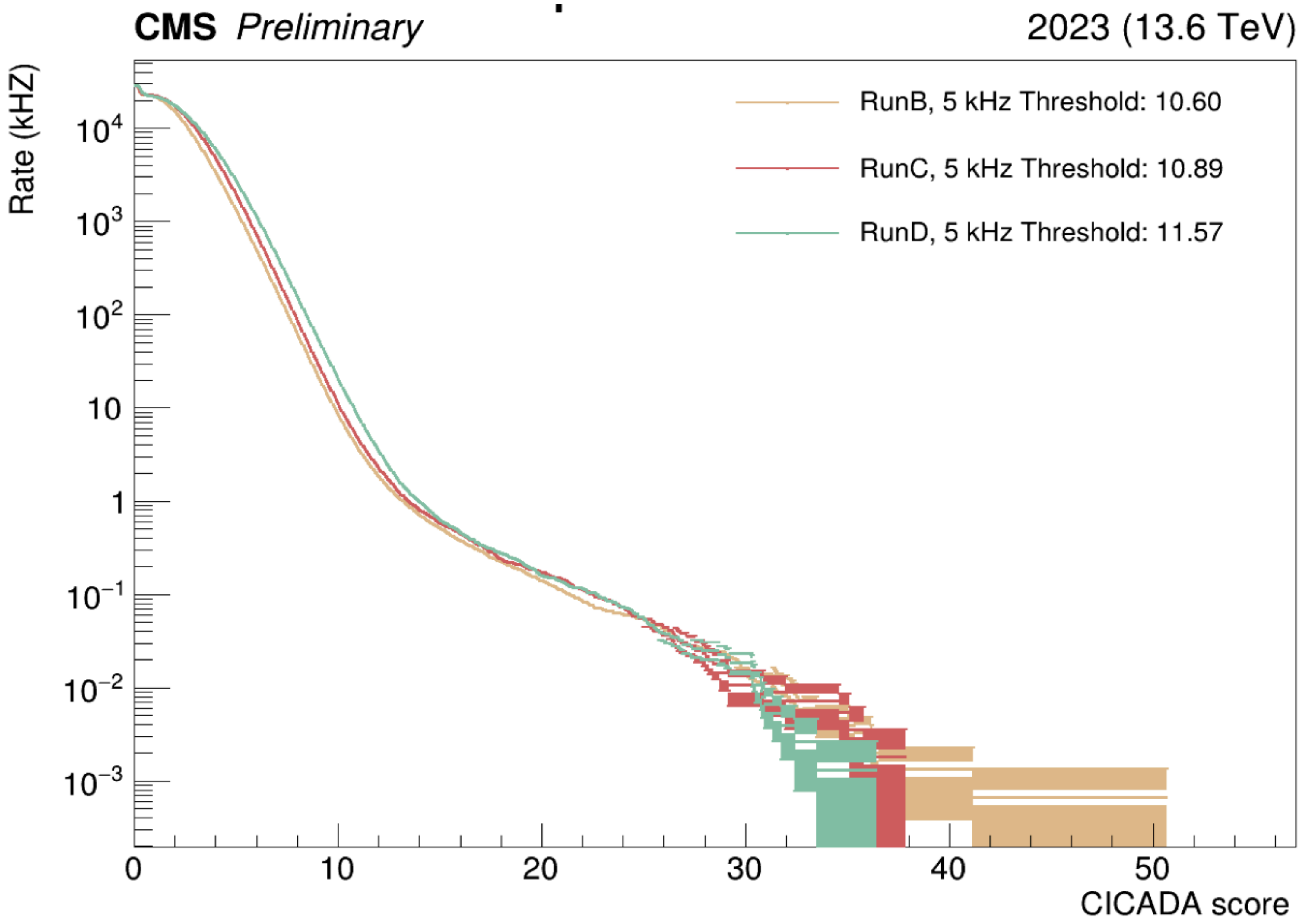


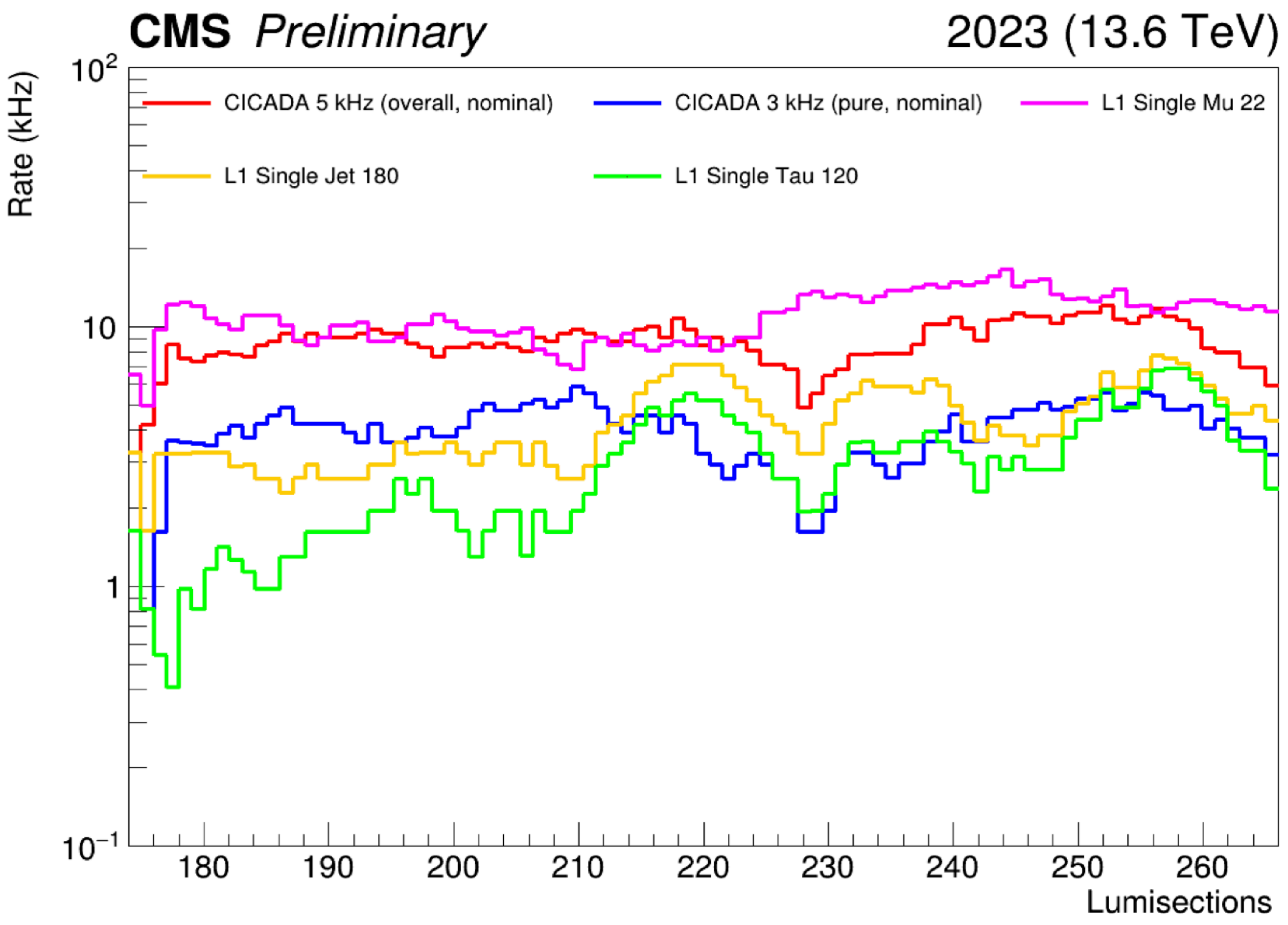


CMS – CICADA

CICADA - Calorimeter image convolutional anomaly detection algorithm

- Operates on a 14x18 (iEta x iPhi) calorimeter image of energy depositions
- Encoder learns to produce low mean square error score for normal (zero-bias) event and spikes for anomalies





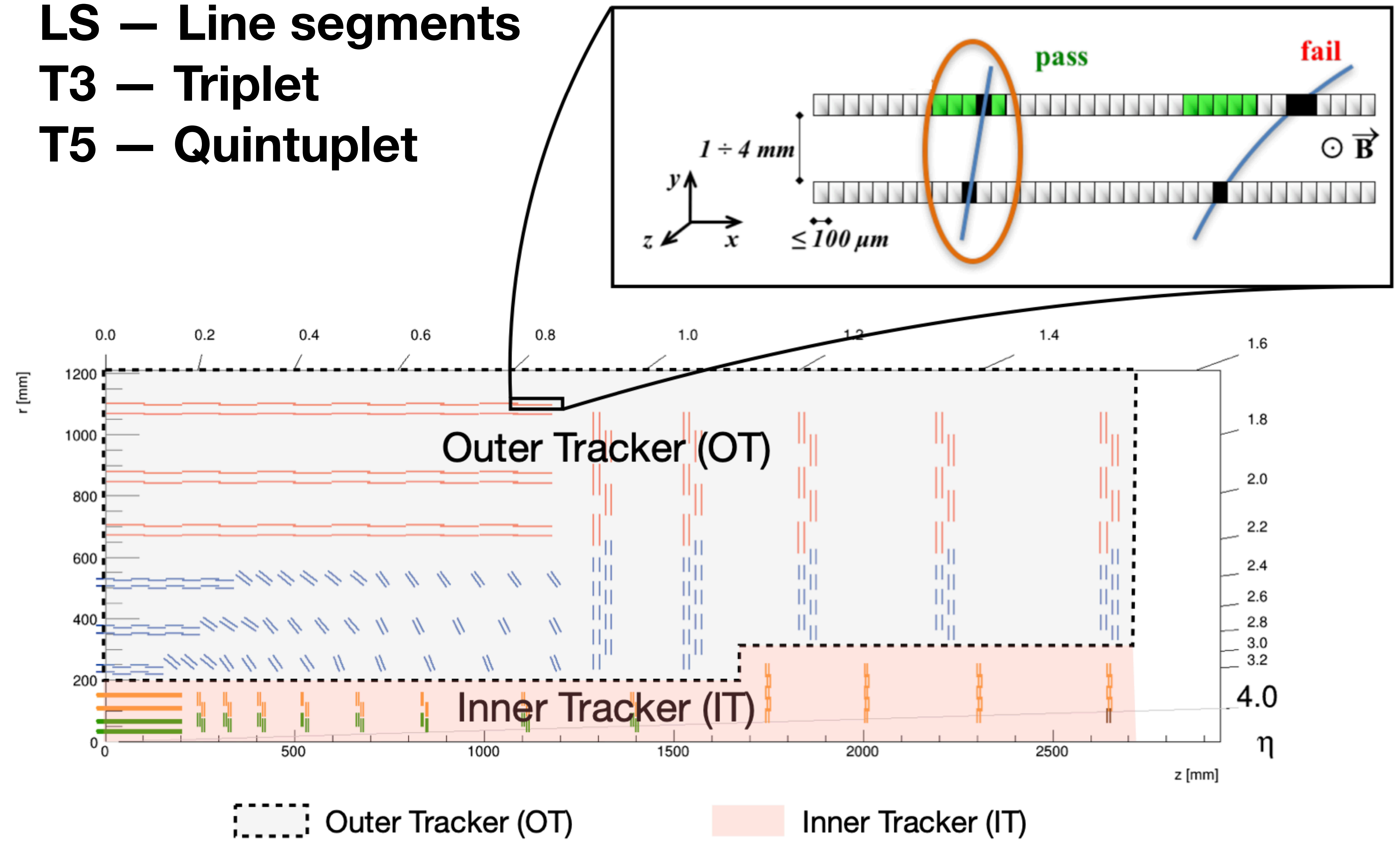
Source: https://cds.cern.ch/record/2879816?ln=en

Cluster-level ⟶ **Track-level** ⟶ Tagging / triggering / calibrations / PID

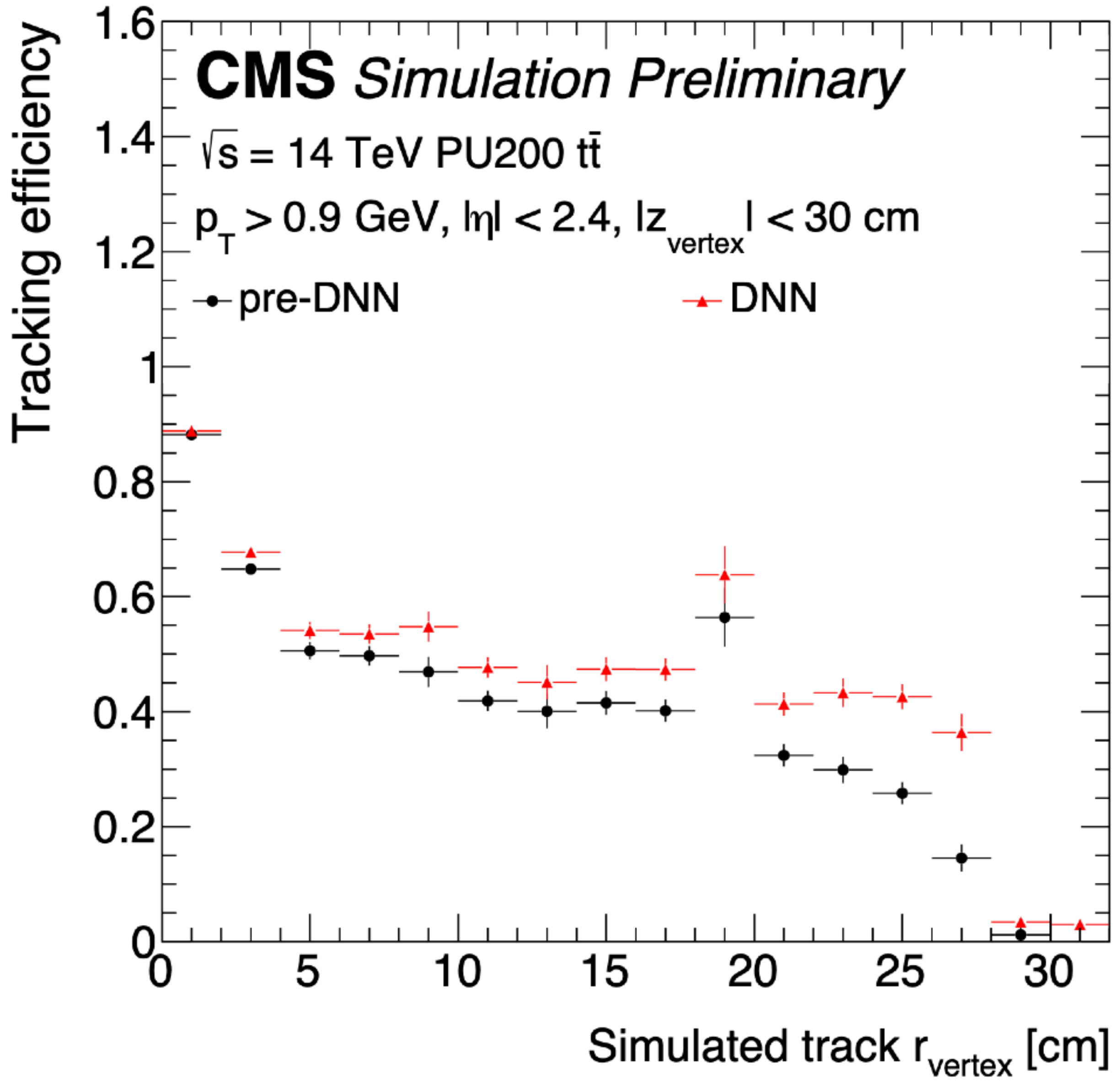
CMS - DNN for line-segment tracking

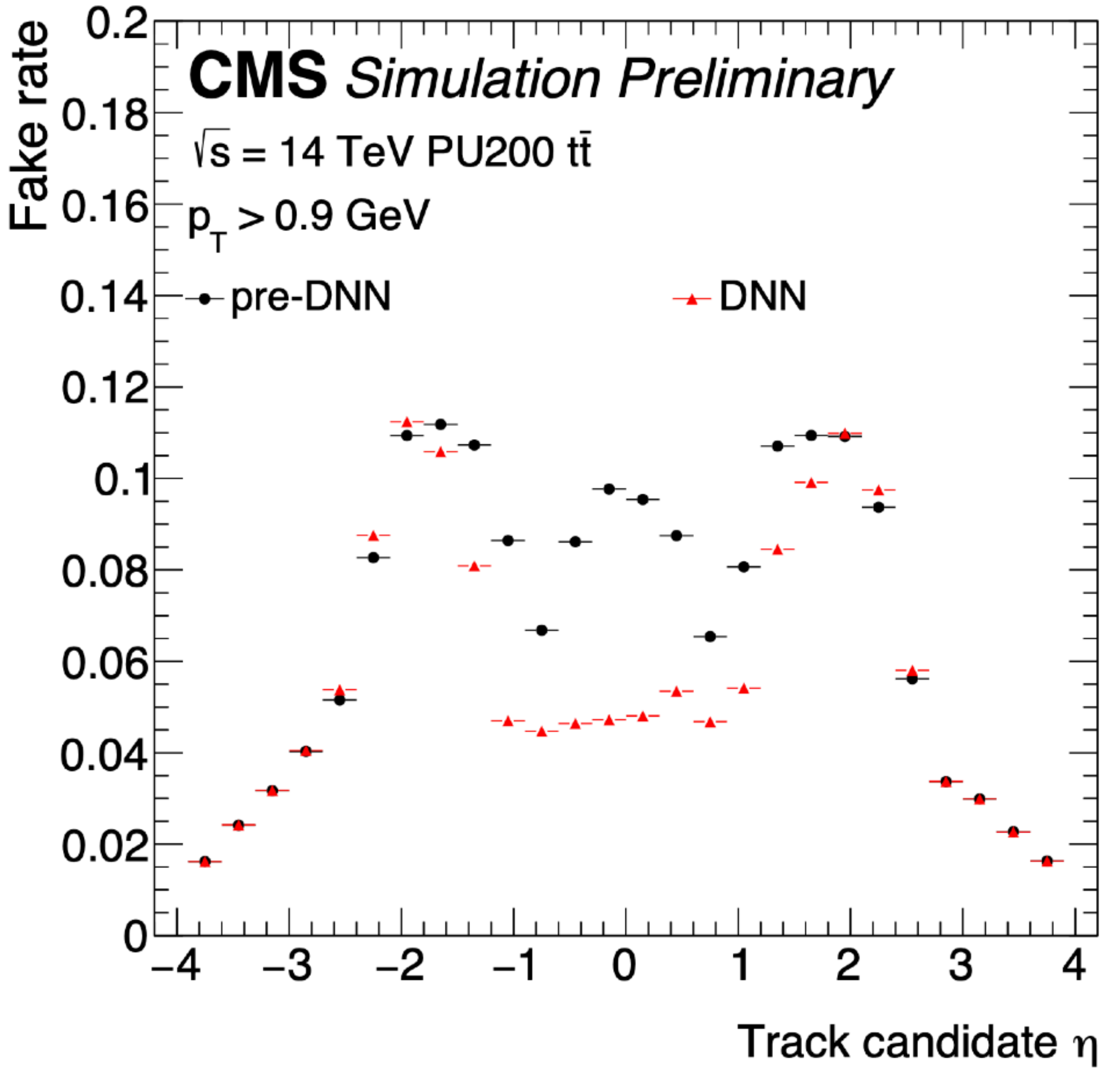
Combination of detector hits into tracklets and connection of tracklets to track candidates with DNN

Tracking builds:
MD — MiniDoublets
LS — Line segments
T3 — Triplet
T5 — Quintuplet

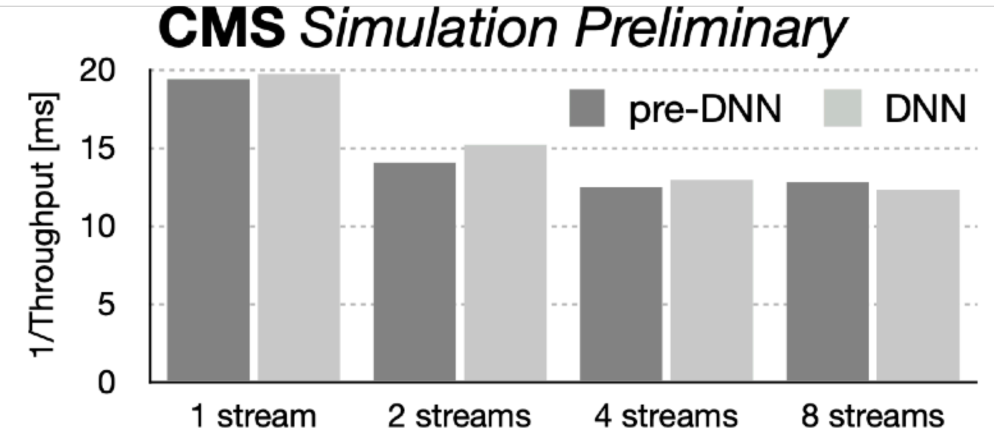


1000 simulated $t\bar{t}$ events with a pileup of 200



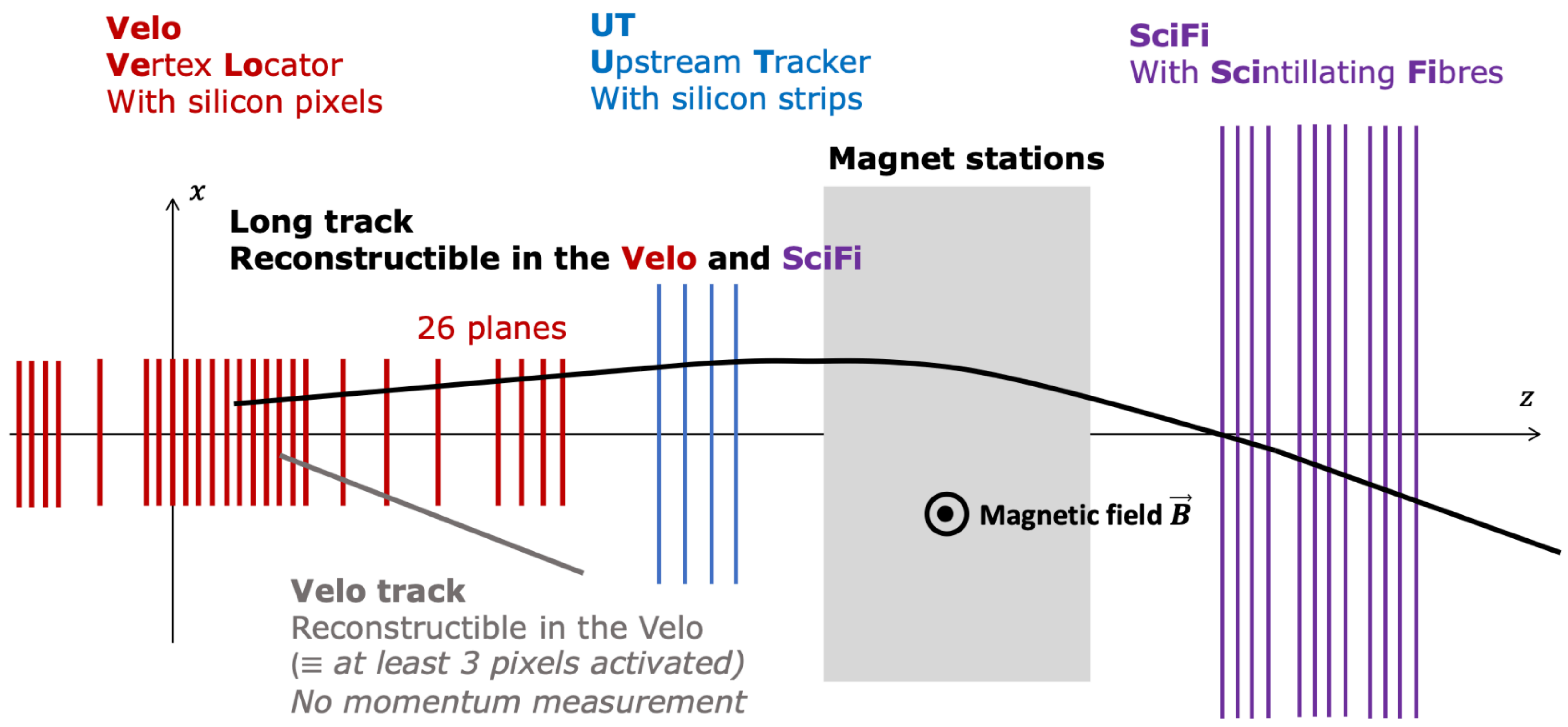


	T5	1/Throughput	N streams
pre-DNN	3.37 ± 0.13	28.4 ± 1.5	1
DNN	3.39 ± 0.07	28.7 ± 1.1	1



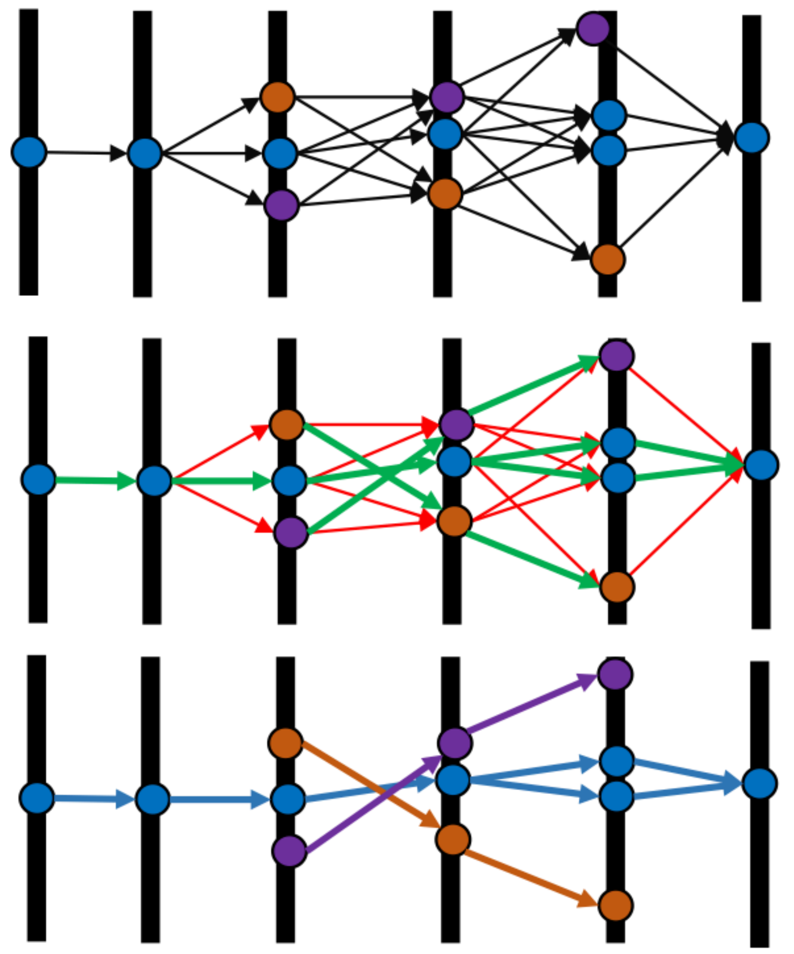
Source: https://cds.cern.ch/record/2872904?ln=en

LHCb - Graph NN for tracking



Strategy

1. Build a **"rough" graph**
   → Embedding Network + Nearest-Neighbour Network
2. **Classify the edges** as genuine or fake
   → Graph Neural Network
3. Keep only the genuine edges
   Identify **connected hits** as tracks
   → Weakly connected component algorithm



Source: https://indico.cern.ch/event/1252748/contributions/5521484/attachments/2731094/4748485/etx4velo_ctd2023.pdf

Long tracks

Category	Metric	Allen	Etx4velo $s_{\text{triplet}} > 0.32$, $d^2_{\max} = 0.010$	Etx4velo $s_{\text{triplet}} > 0.36$, $d^2_{\max} = 0.020$
Long, no electrons ✓ In acceptance ✓ Reconstructible in the velo ✓ Reconstructible in the SciFi ✓ Not an electron	Efficiency	99.26%	99.28%	99.51%
	Clone rate	2.54%	0.96%	0.89%
	Hit efficiency	96.46%	98.73%	98.90%
	Hit Purity	99.78%	99.94%	99.94%
Long electrons ✓ In acceptance ✓ Reconstructible in the velo ✓ Reconstructible in the SciFi ✓ Electron	Efficiency	97.11%	98.80%	99.22%
	Clone rate	4,25%	7.42%	7.31%
	Hit efficiency	95.24%	96.54%	96.79%
	Hit purity	97.11%	98.46%	98.46%
Long, from strange ✓ In acceptance ✓ Reconstructible in the velo ✓ Decays from a strange *Good proxy for displaced tracks*	Efficiency	97.69%	97.50%	98.06%
	Clone rate	2.50%	0.92%	0.81%
	Hit efficiency	97.69%	98.22%	98.77%
	Hit purity	99.34%	99.68%	99.68%
X	Ghost rate	2.18%	0.76%	0.81%

Velo-only tracks

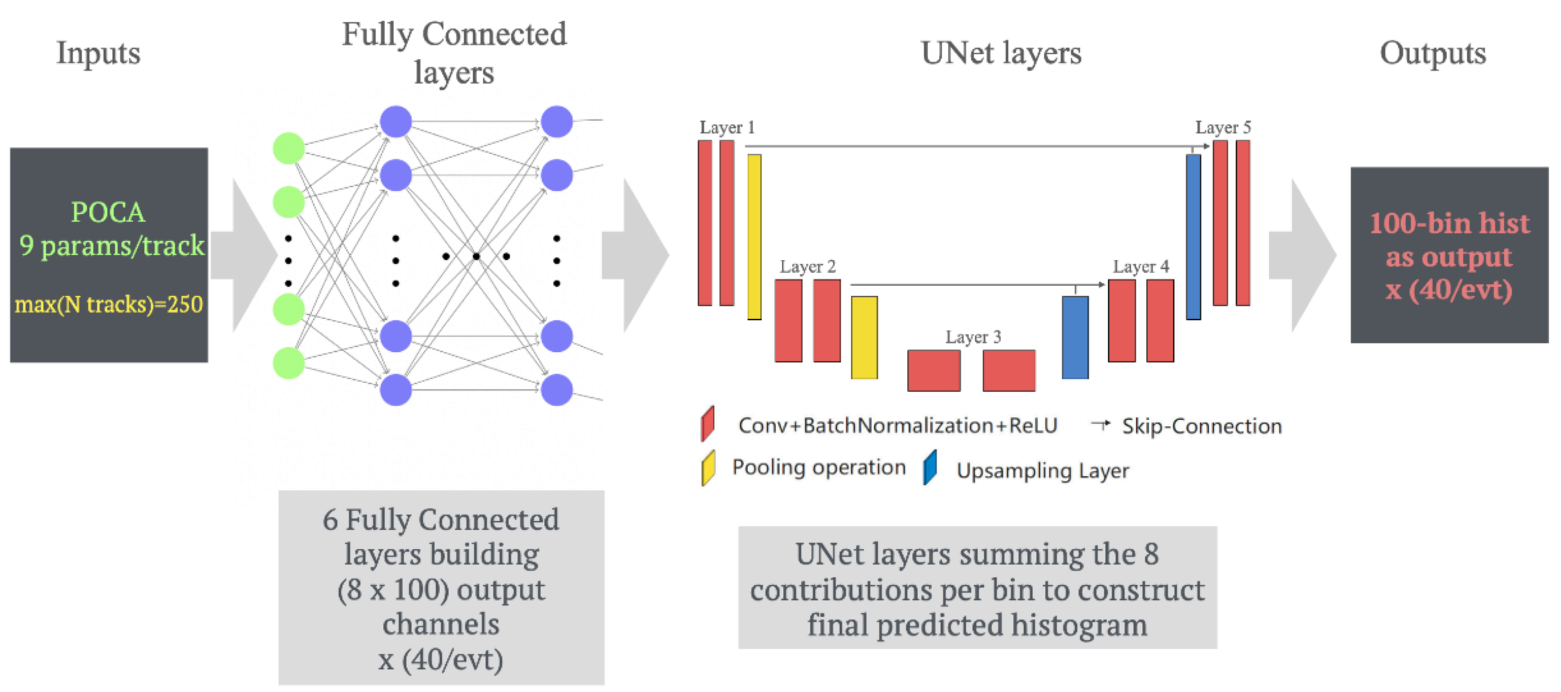
Category	Metric	Allen	Etx4velo $s_{\text{triplet}} > 0.32$, $d^2_{\max} = 0.010$	Etx4velo $s_{\text{triplet}} > 0.36$, $d^2_{\max} = 0.020$
Velo-only, no electrons ✓ In acceptance ✓ Reconstructible in the velo ✓ Not reconstructible in the SciFi ✓ Not an electron	Efficiency	96.84%	97.03%	97.86%
	Clone rate	3.84%	1.08%	1.02%
	Hit efficiency	93.89%	97.93%	98.32%
	Hit Purity	99.50%	99.84%	99.82%
Velo-only electrons ✓ In acceptance ✓ Reconstructible in the velo ✓ Not reconstructible in the SciFi ✓ Electron	Efficiency	67.81%	85.10%	86.69%
	Clone rate	10.27%	5.02%	4.97%
	Hit efficiency	79.21%	93.33%	93.88%
	Hit purity	97.35%	99.07%	98.99%
Velo-only, from strange ✓ In acceptance ✓ Not reconstructible in the velo ✓ Decays from a strange *Good proxy for displaced tracks*	Efficiency	93.53%	93.07%	96.05%
	Clone rate	5.60%	1.97%	1.77%
	Hit efficiency	90.05%	93.92%	96.05%
	Hit purity	99.36%	99.67%	99.64%

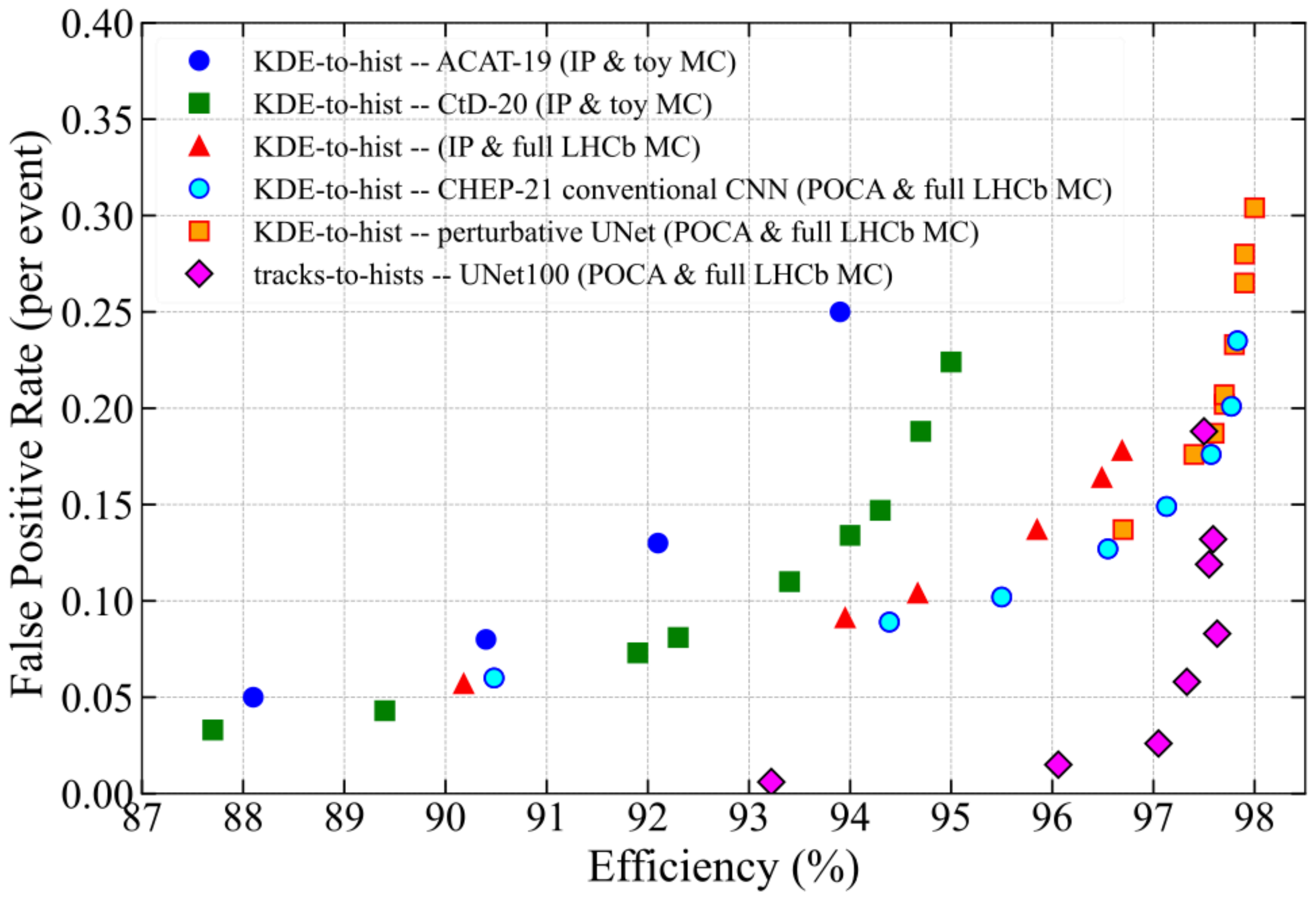
Comparable or better performance for efficiencies, clone-rates and purities

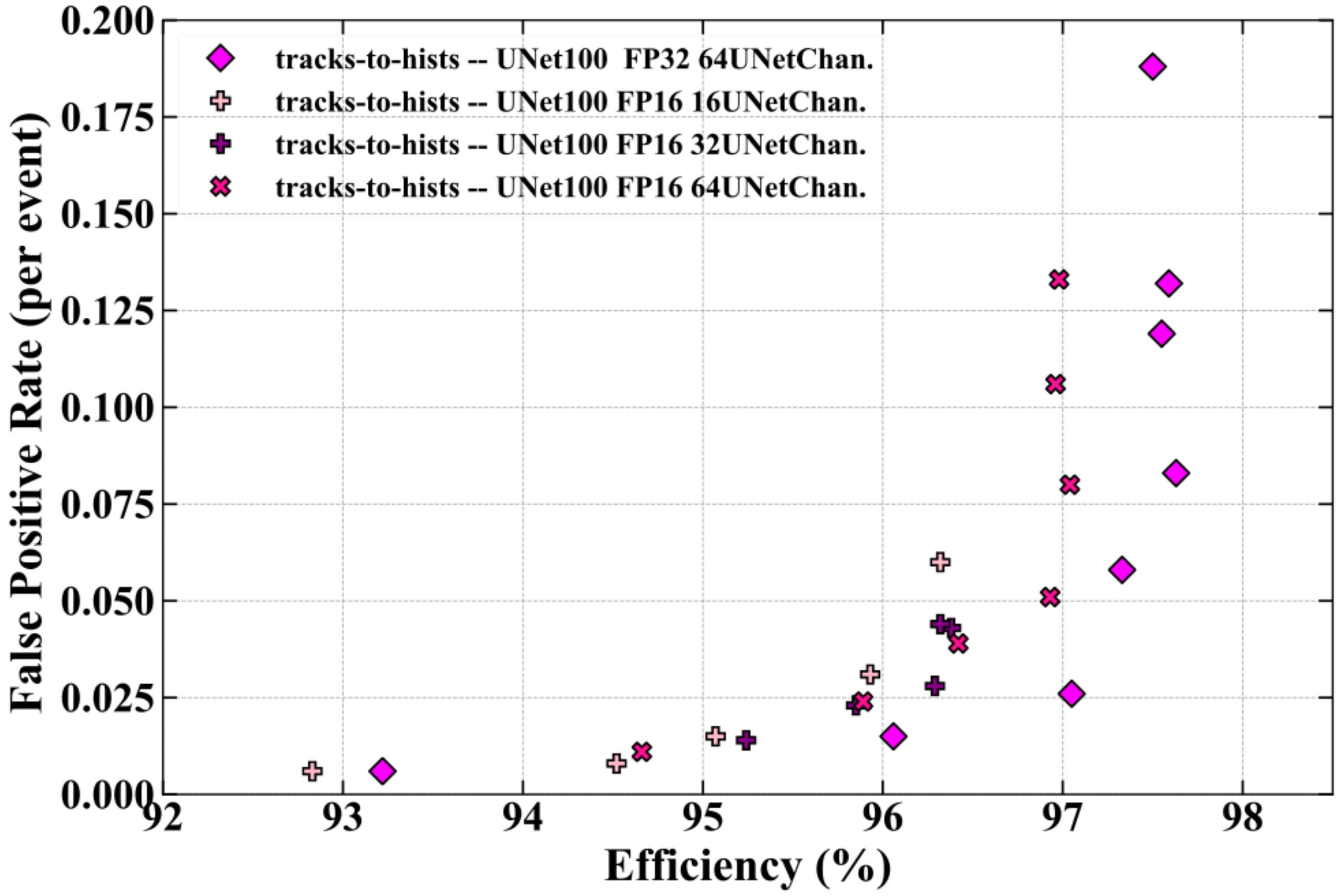
LHCb – DNN for primary vertex finding

Kernel density estimate (KDE) is replaced by DNN's for PV finding

- KDE estimates PV locations from tracks crossing the beamline (tracks-to-KDE) + CNN (KDE-to-hist)
- Now: DNN replaces KDE estimate and uses track features directly (track-to-KDE) + U-NET -> (tracks-to-hist)
- Similar approach is also taken by ATLAS and shows significant improvement to standard AMVF







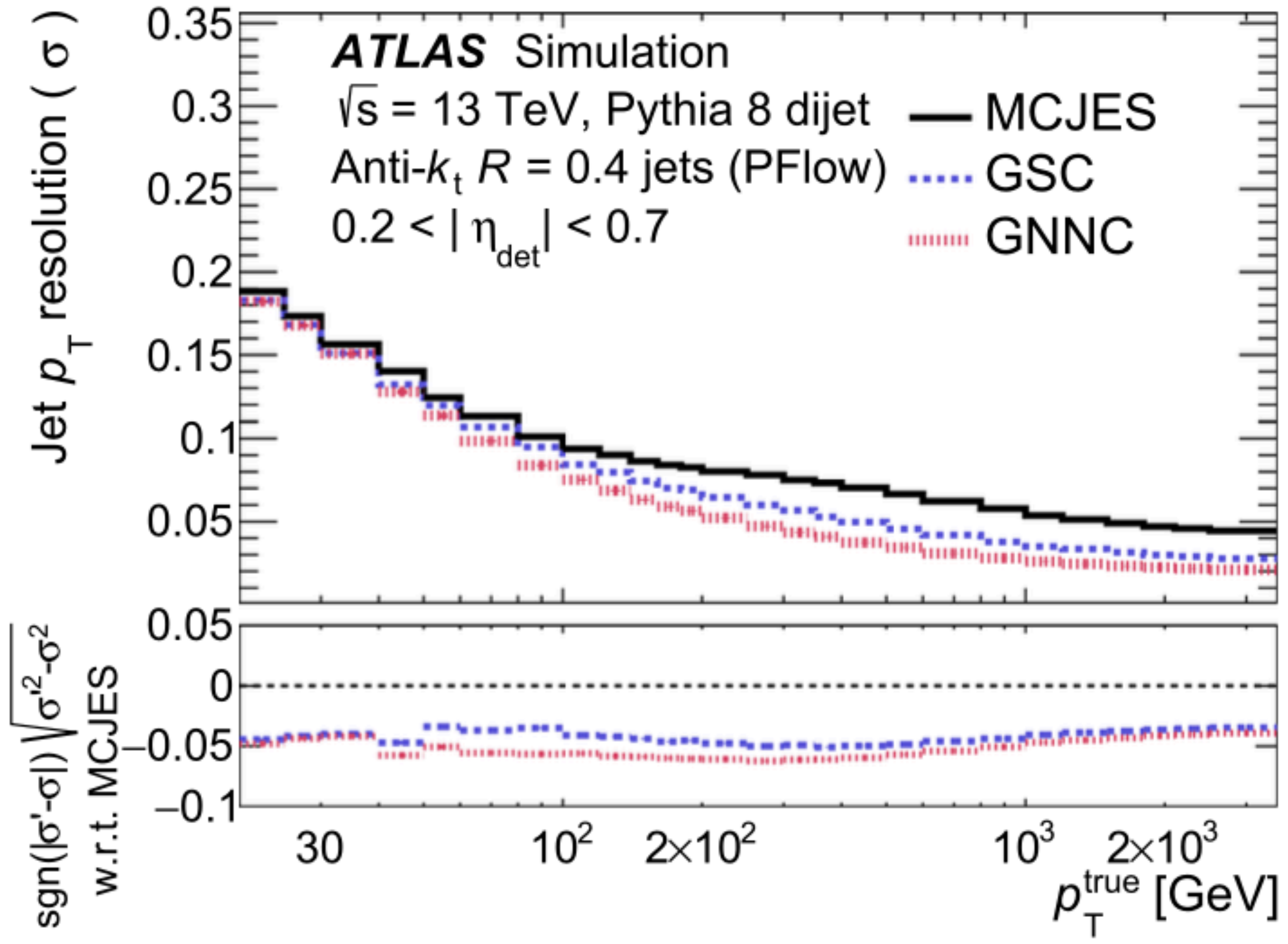
Source: https://arxiv.org/abs/2309.12417

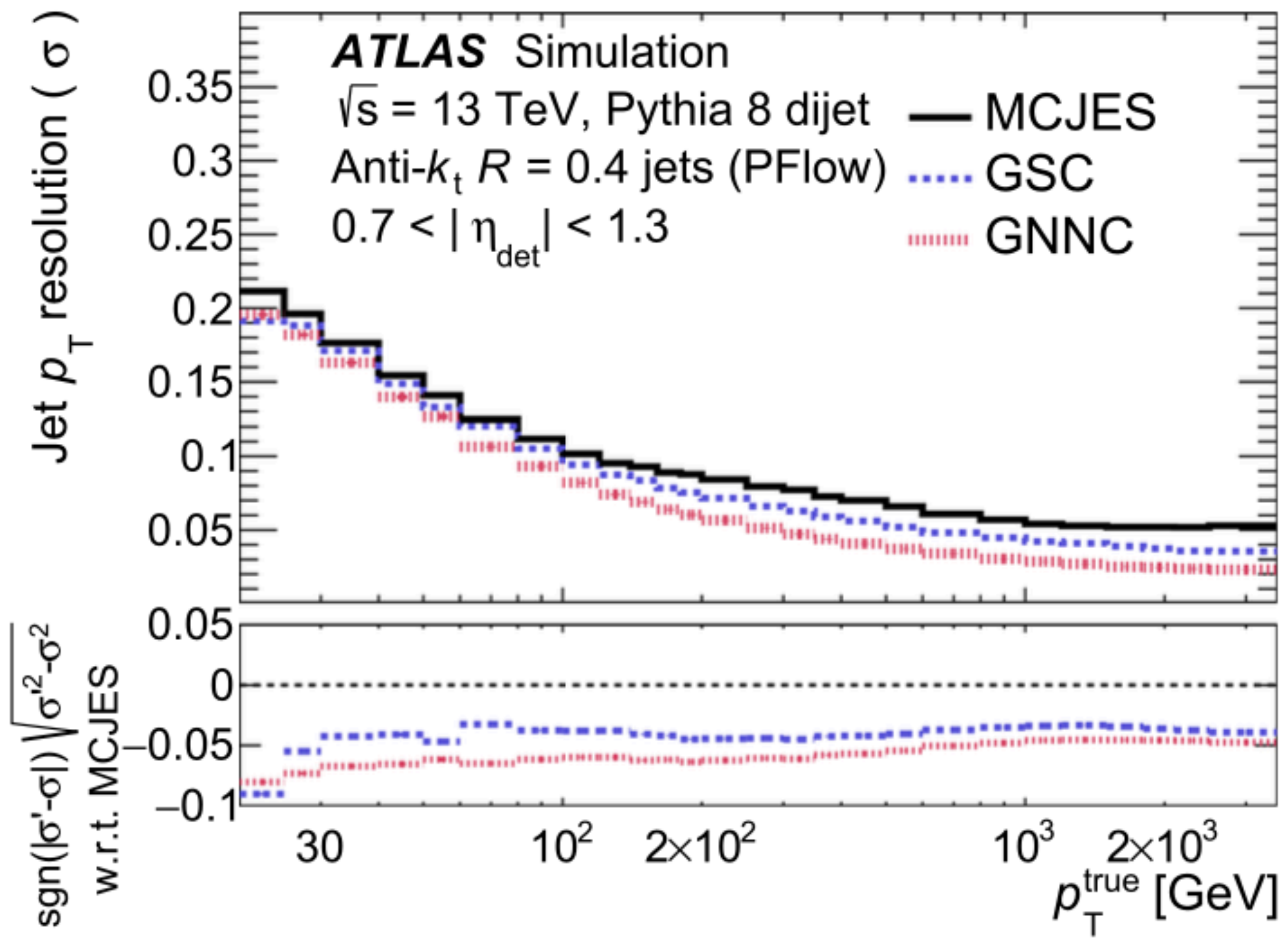
Cluster-level ⟶ Track-level ⟶ **Tagging / triggering / calibrations / PID**

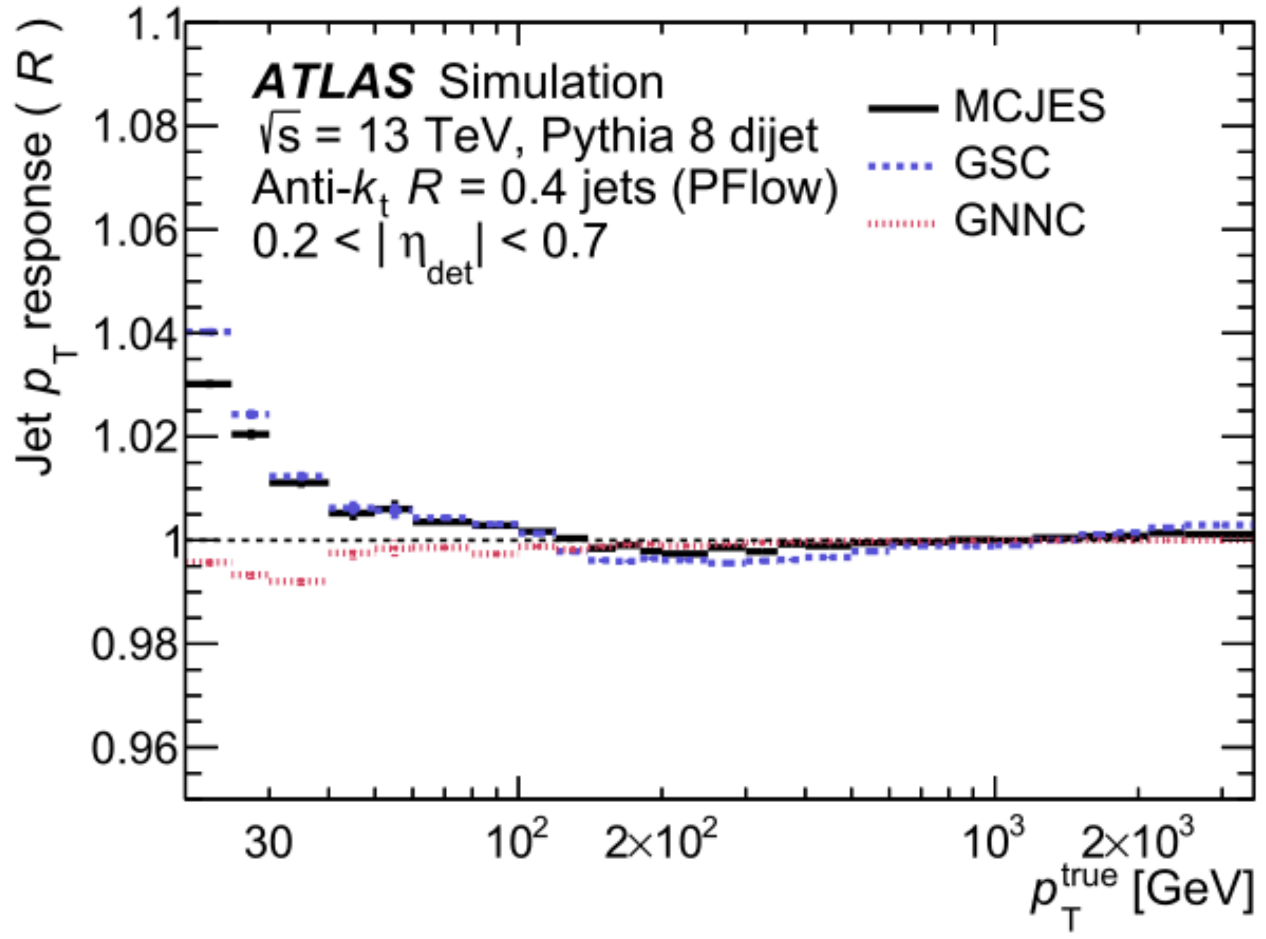
ATLAS - GNNC for small R-jet calibration

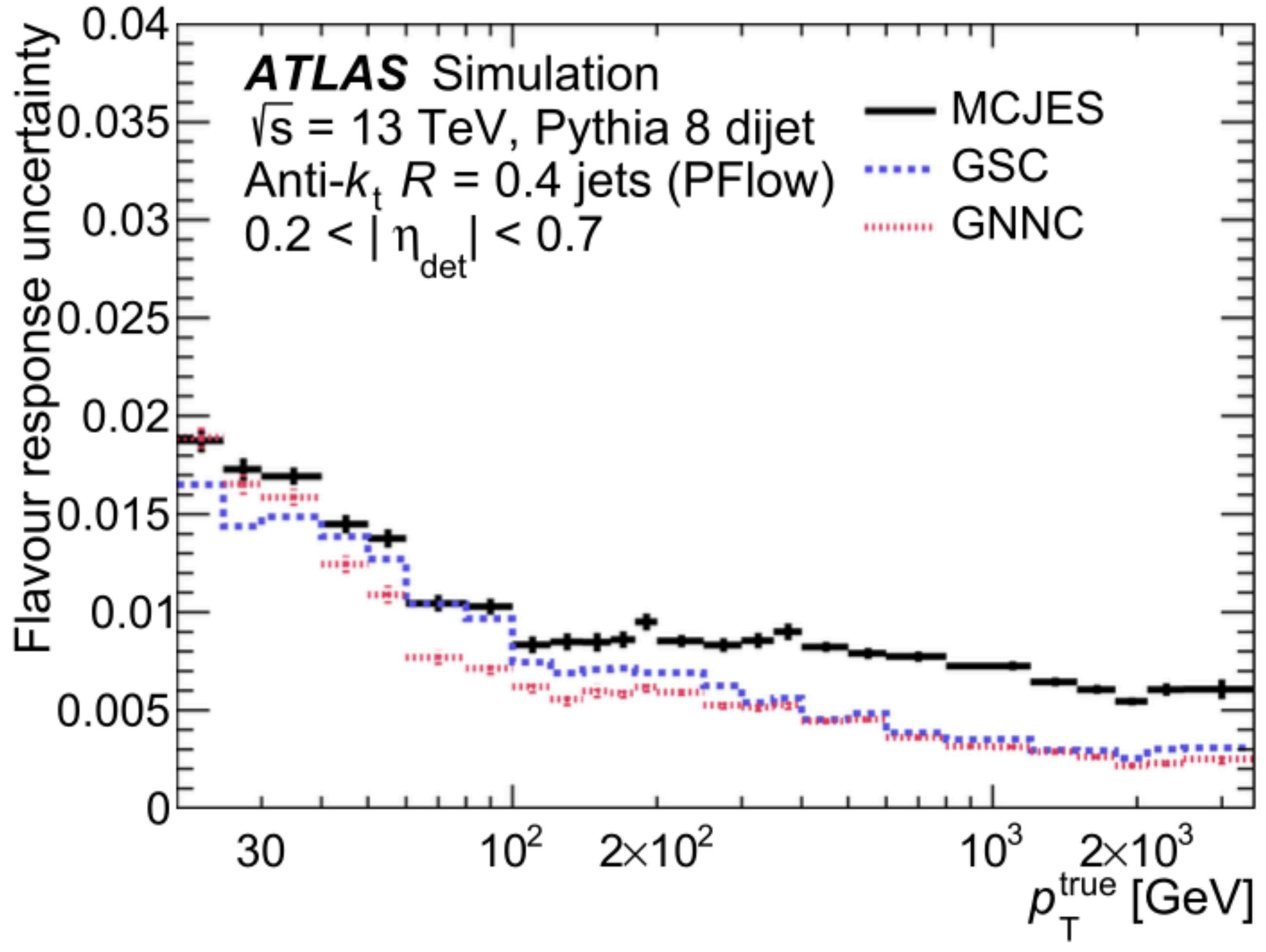
DNN as improvement to sequential calibration

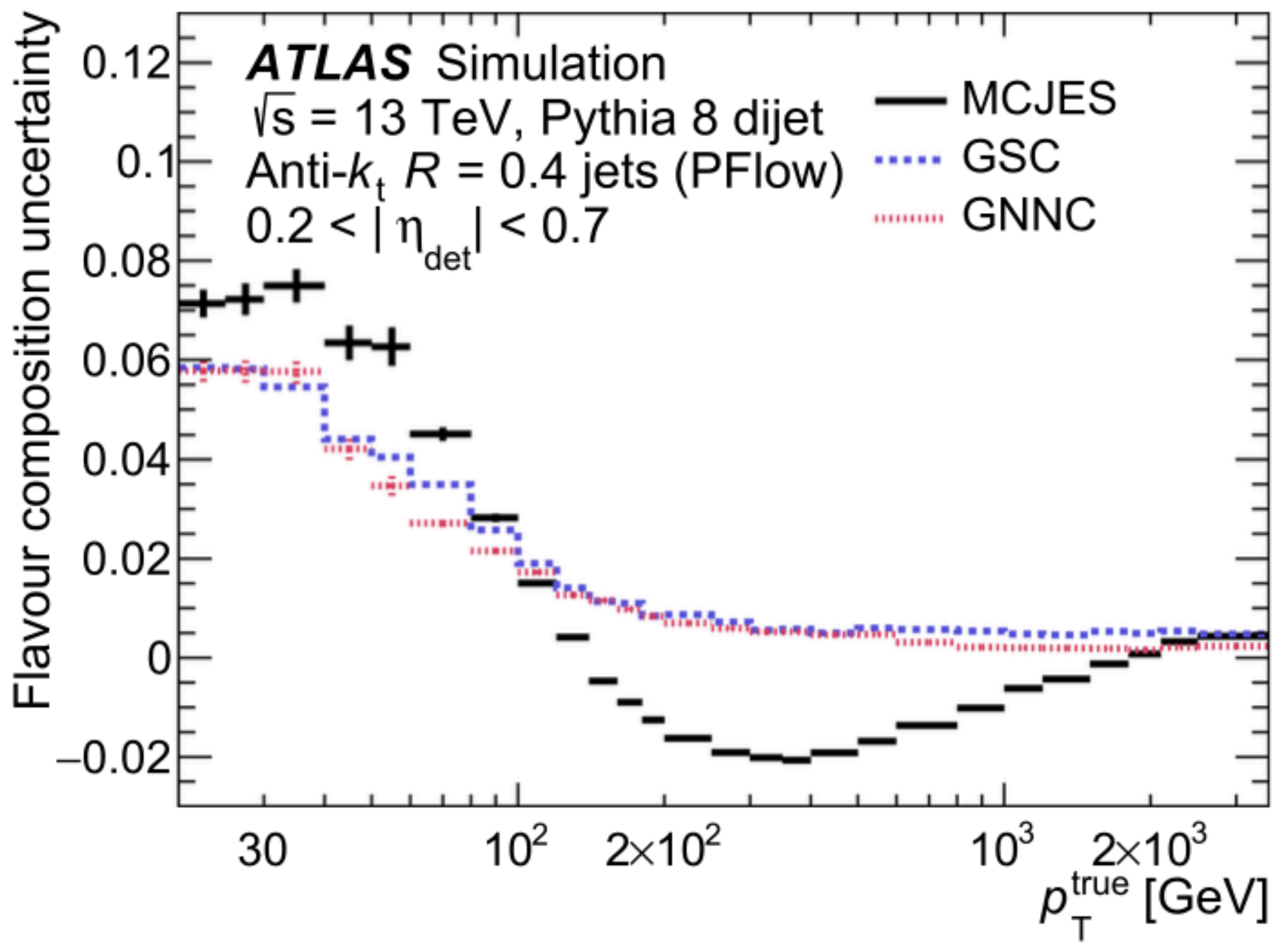
- Global sequential calibration (GSC) receives 6 informative jet variables and performs sequential, multiplicative corrections to jet p_T
- Correlations can be taken into account -> DNN for individual η bins with additional information about kinematics, energy depositions and pile-up









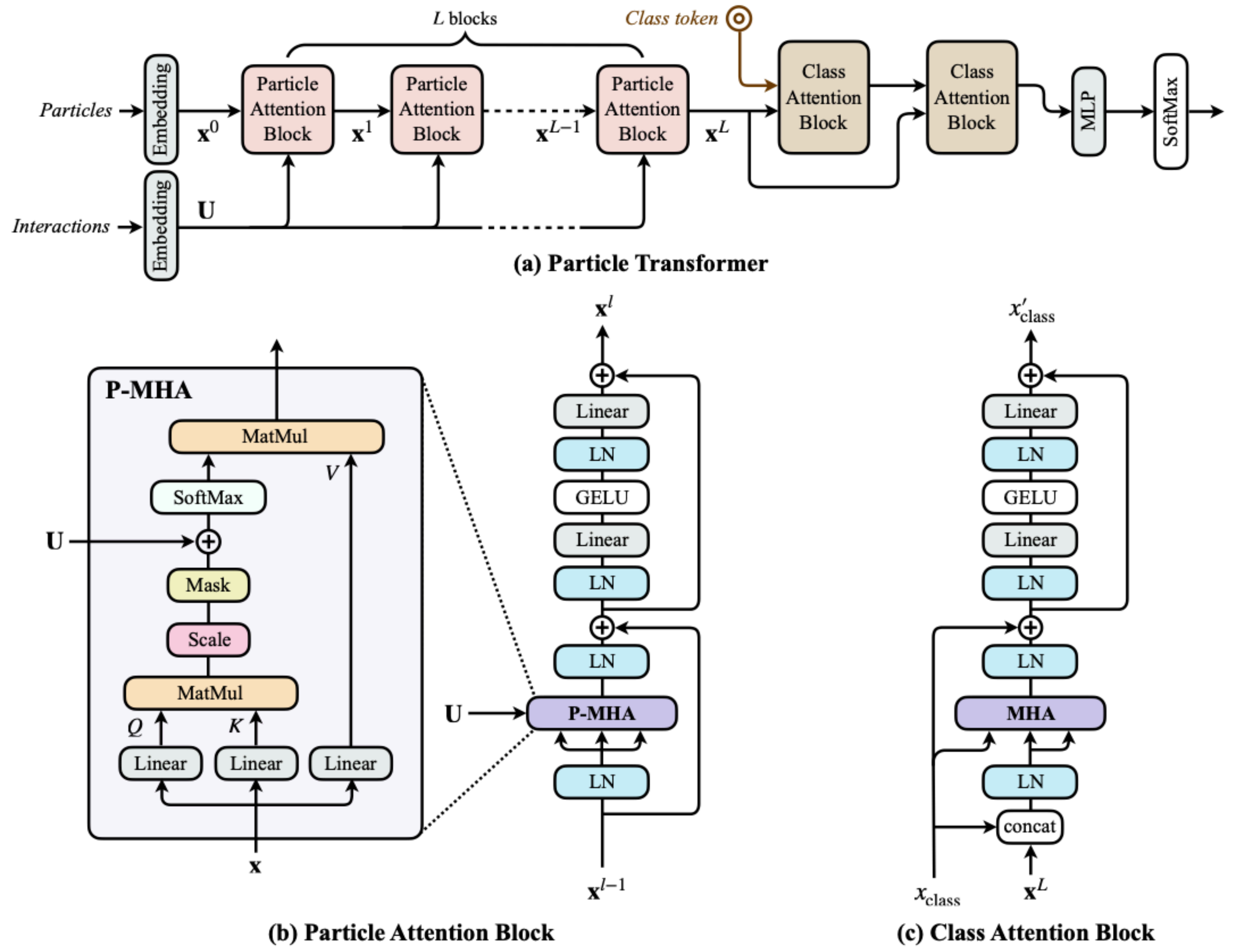


Source: https://link.springer.com/article/10.1140/epjc/s10052-023-11837-9

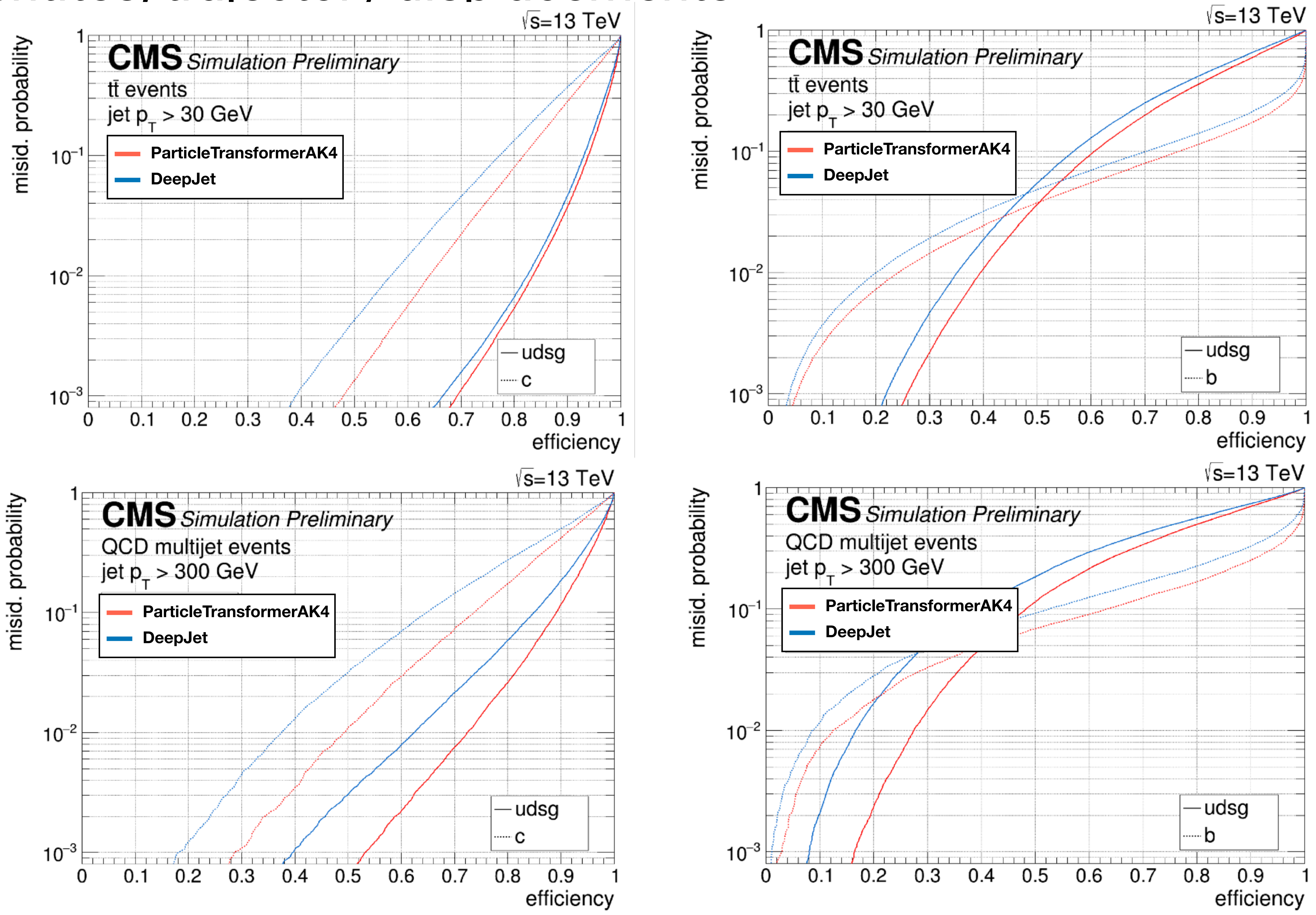
CMS - ParticleTransformerAK4

Heavy flavour jet tagging using transformer model

- Major success of deep learning models for jet tagging: DeepJet, ParticleNet
- Physics augmented attention mechanism: Pair-wise "interactions" of constituents as model input:
  - Inputs: kinematics (4-momentum), particle identites, trajectory displacements



The architecture of (a) Particle Transformer (b) Particle Attention Block (c) Class Attention Block.

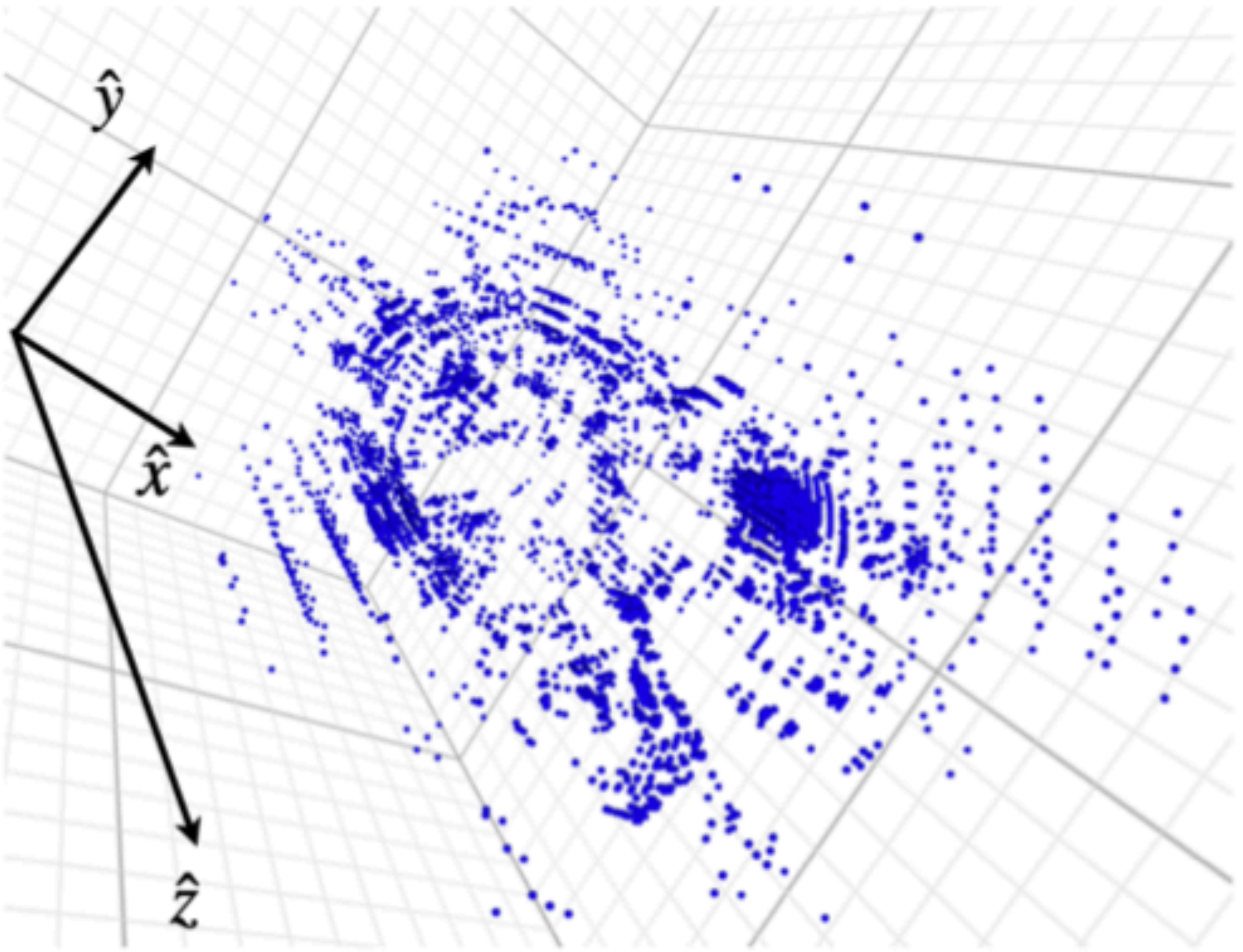


Source: https://cds.cern.ch/record/2839920?ln=en

ATLAS - Point-cloud based pion identification

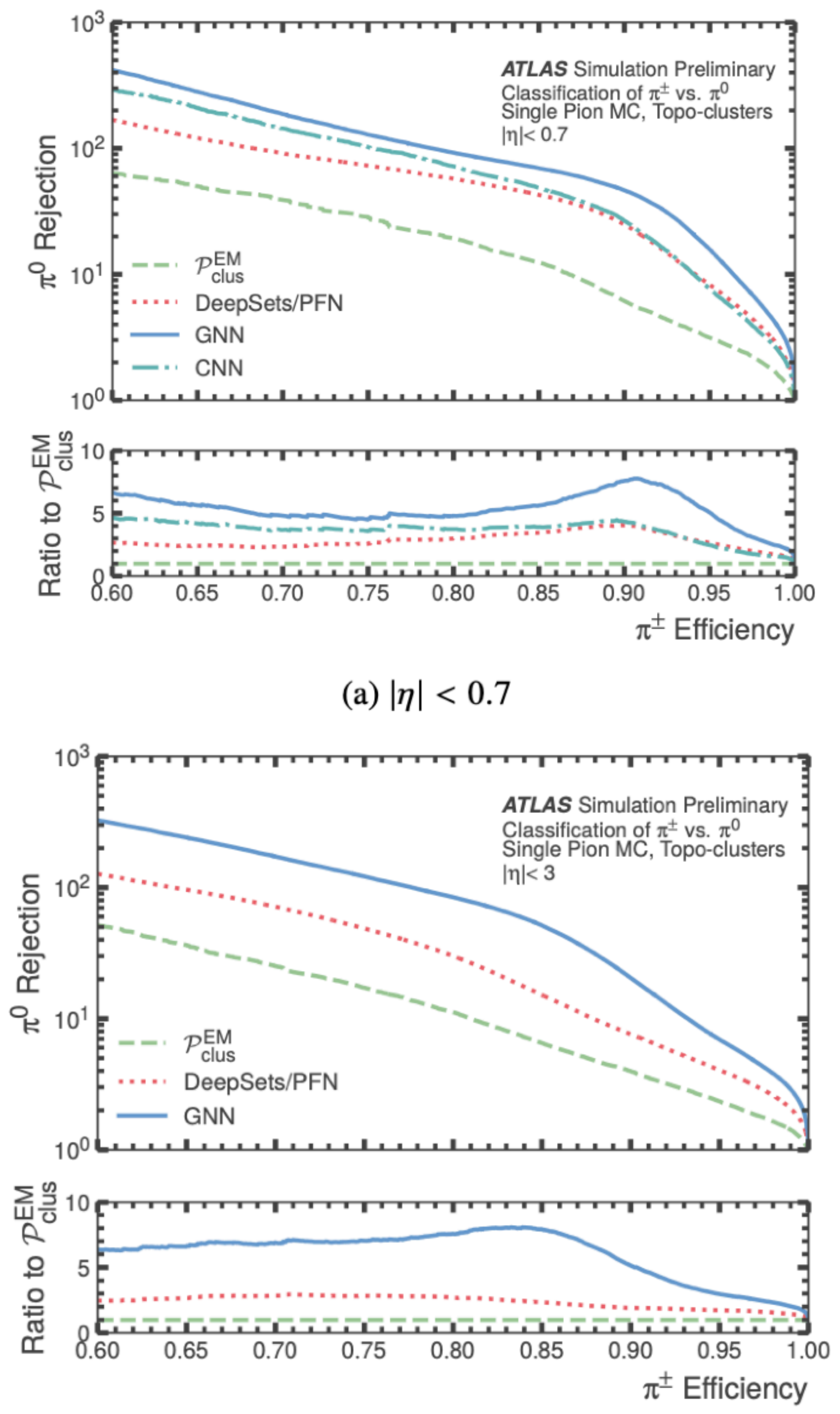
Vast set of ML models tested for calorimeter pion rejection / identification / regression

- Point cloud based: High dimensional input feature space of cluster images + track properties
- Model architectures: DeepSets, GNN's and Transformers
- All ML models outperform the traditional π^0 / π^{+-} rejection algorithm and perform well on energy calibration



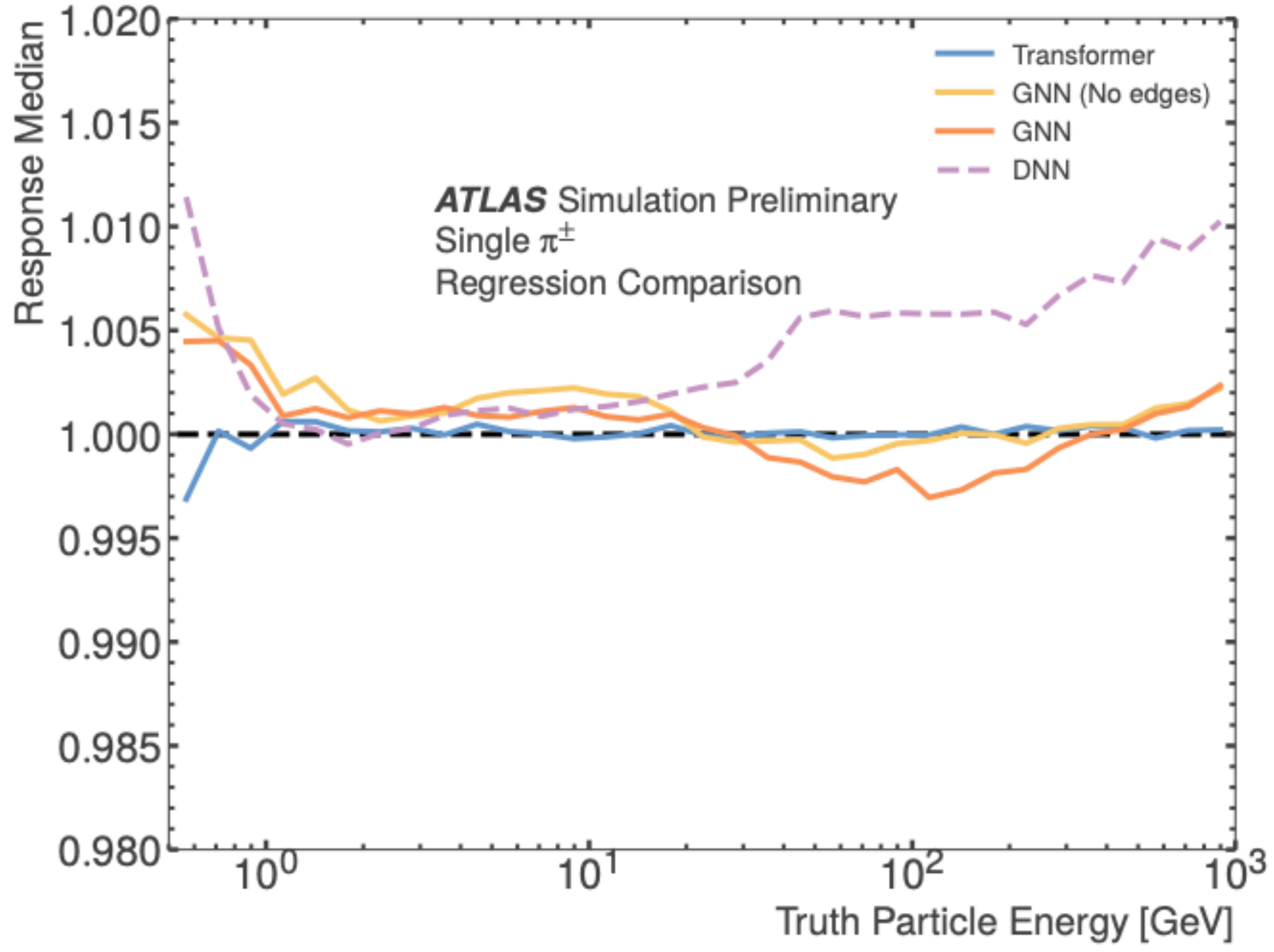
Point-cloud representation of clusters

Source: https://cds.cern.ch/record/2825379?ln=en



(a) $|\eta| < 0.7$

(b) $|\eta| < 3$

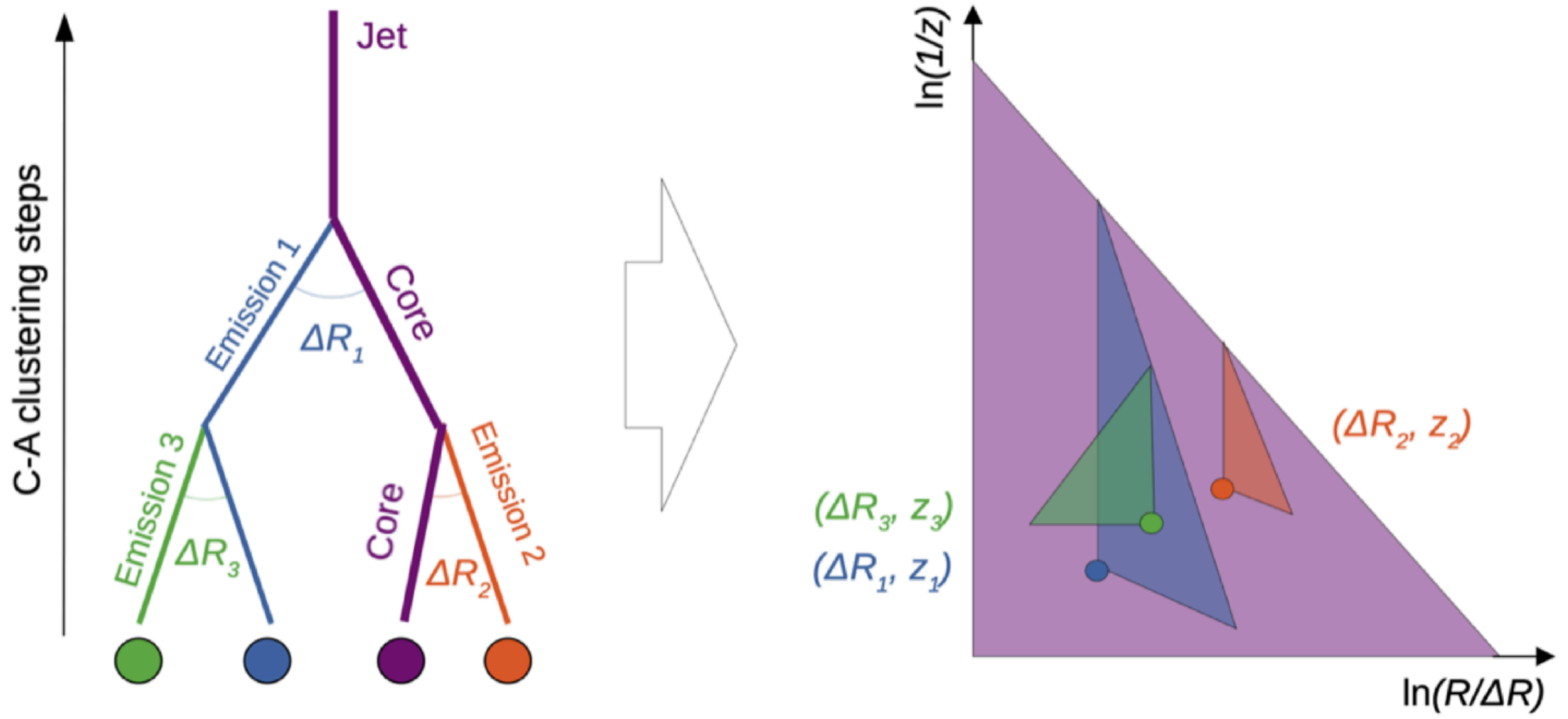


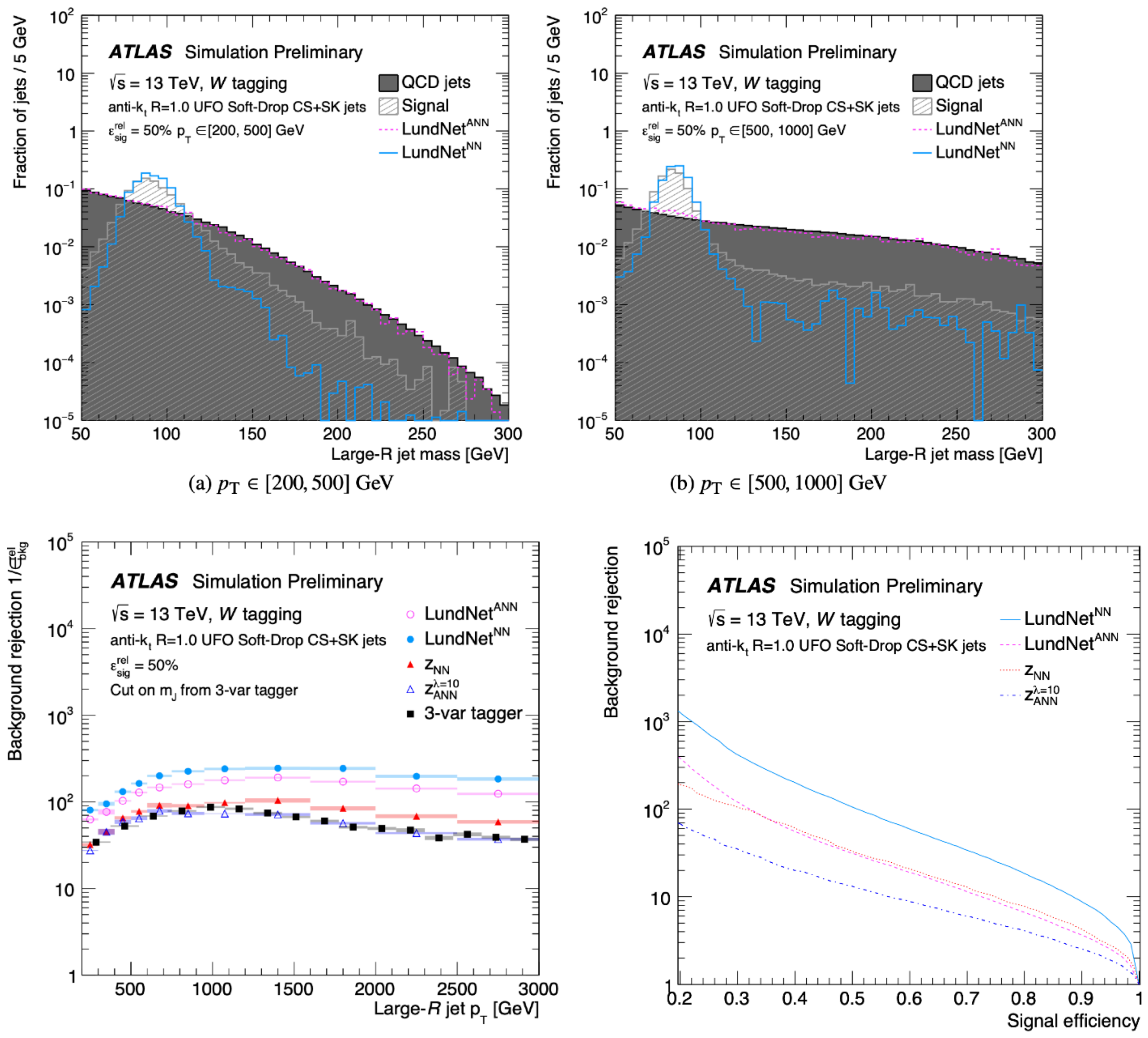
[**New study with Bayesian NN's**]()

ATLAS - Lund-plane W-tagging with GNN

Large-R jets contain both prongs of the W decay -> GNN's can find jet-constituent correlations

- GNN learns declustered Lund plane variables (such as k_T, z, ΔR) as graph representation
- Adversarial network trained on gaussian mixture model to decouple mass correlation





Simulation:
Signal: W' -> WZ -> qqqq
Background: light quark or gluon jets

For sub-jets i,j:

$$\Delta R_{ij} = \sqrt{\Delta y_{ij}^2 + \Delta\phi_{ij}^2}, \qquad z = \frac{p_T^j}{p_T^i + p_T^j}, \qquad k_t = p_T^j \Delta R_{ij}$$

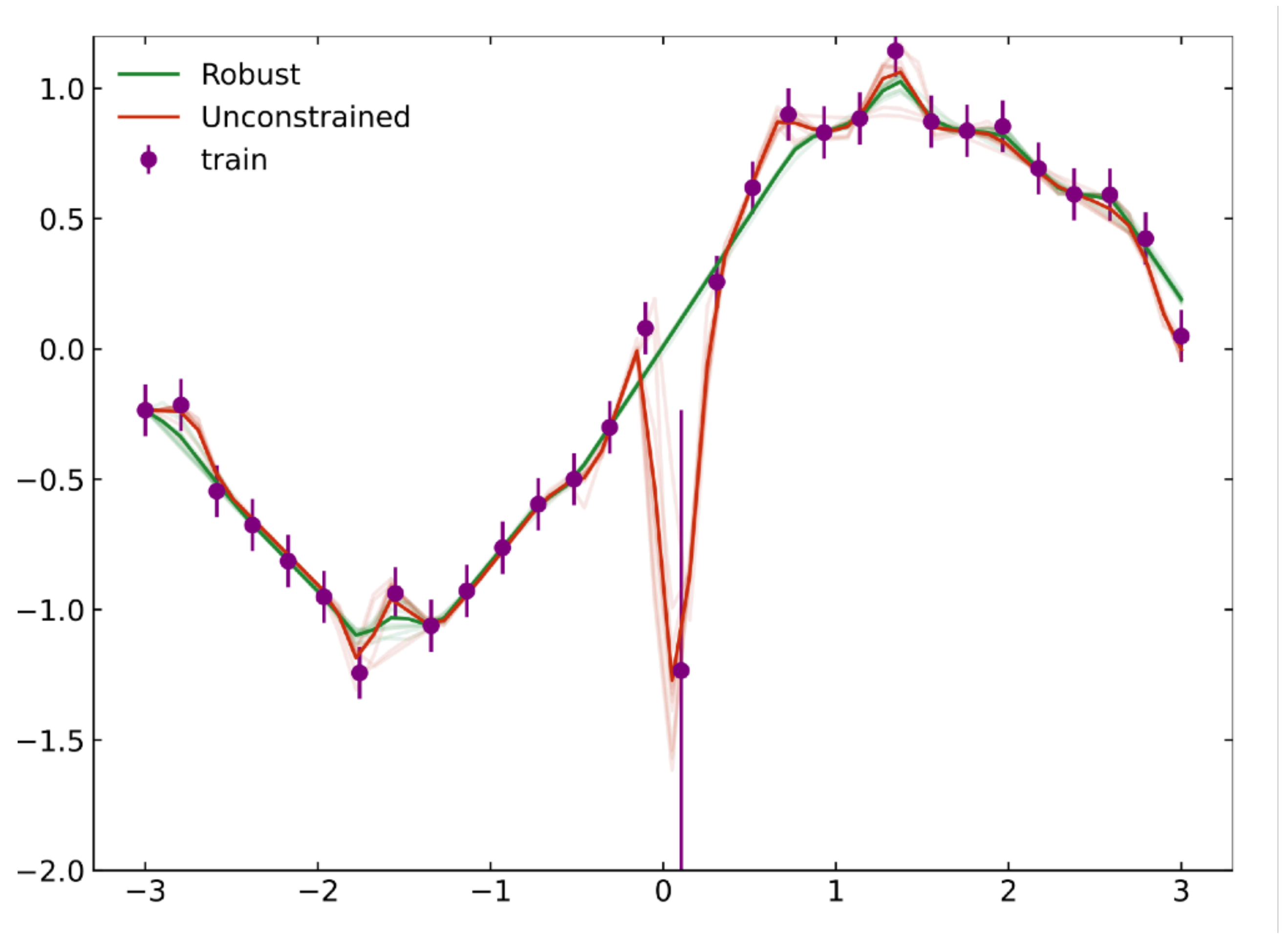
Source: https://cds.cern.ch/record/2864131

LHCb – Lipschitz NN for the online triggering

Monotonic Lipschitz NN - Regularising the weights and makes output robust

- Layer-wise weight-based normalization leads to Lipschitz constraint and robustness
- Adding linear term to activation function leads to monotonicity
- Inclusive heavy-flavour triggers (e.g. b-hadron secondary vertices)

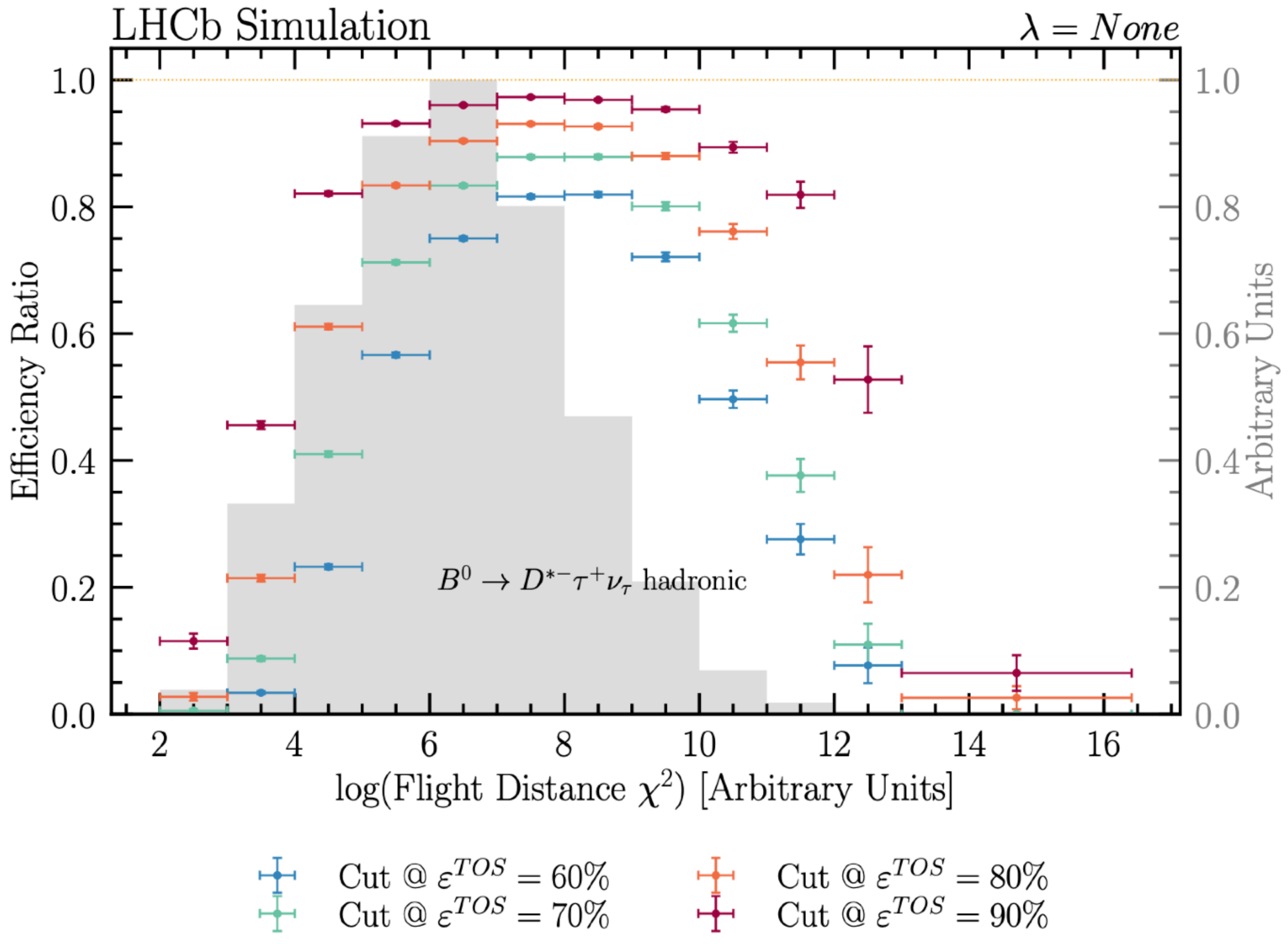
A function, $f : X \to Y$, is Lipschitz continuous if
$$D_Y\left(f(\vec{x}_1), f(\vec{x}_2)\right) \leq k D_X(\vec{x}_1, \vec{x}_2) \quad \forall \vec{x}_1, \vec{x}_2 \in X, k \in \mathbb{R}$$



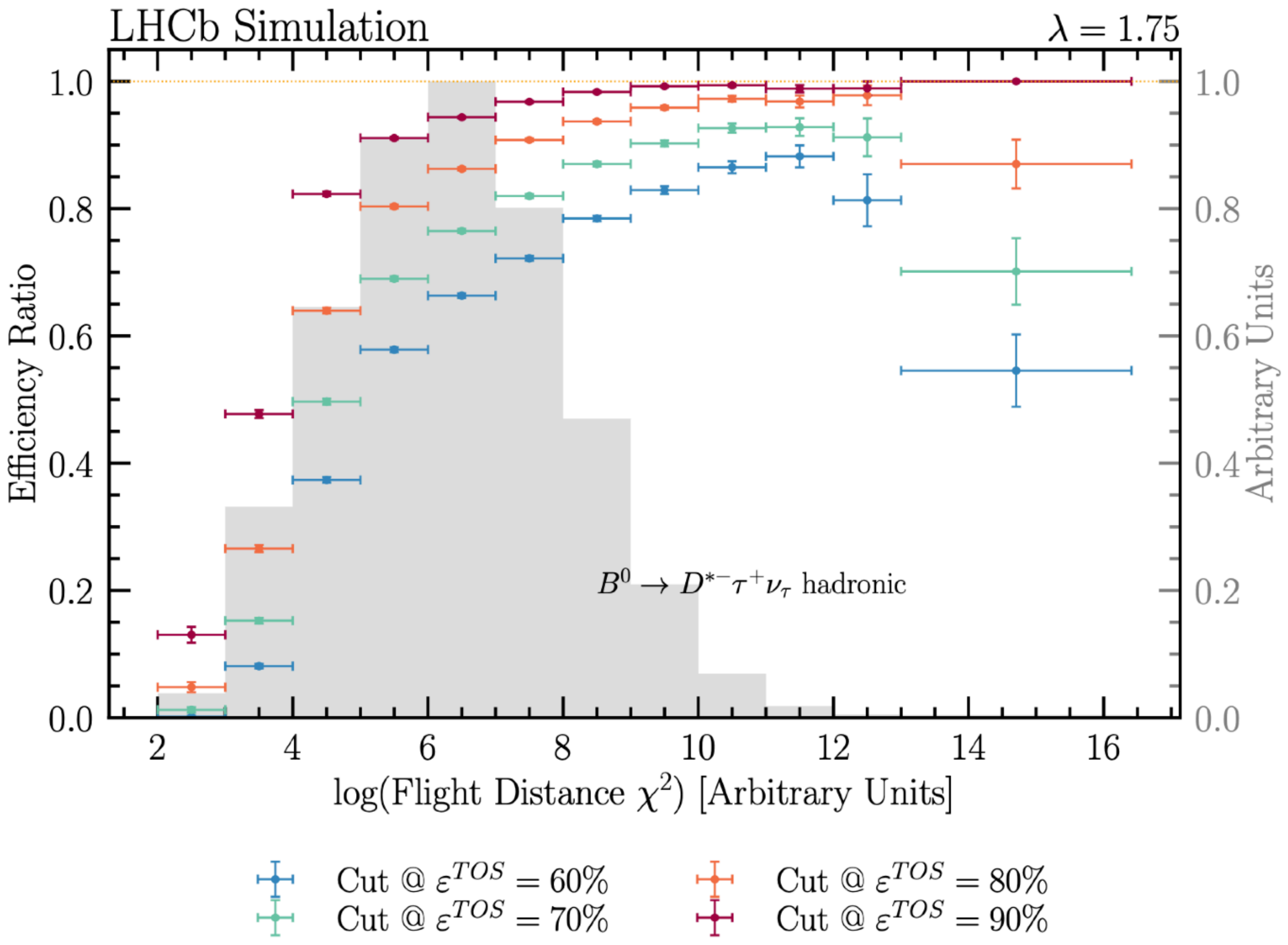
Robustness demonstration of monotonic Lipschitz DNN

Source: https://arxiv.org/abs/2312.14265

Regular, unconstrained DNN



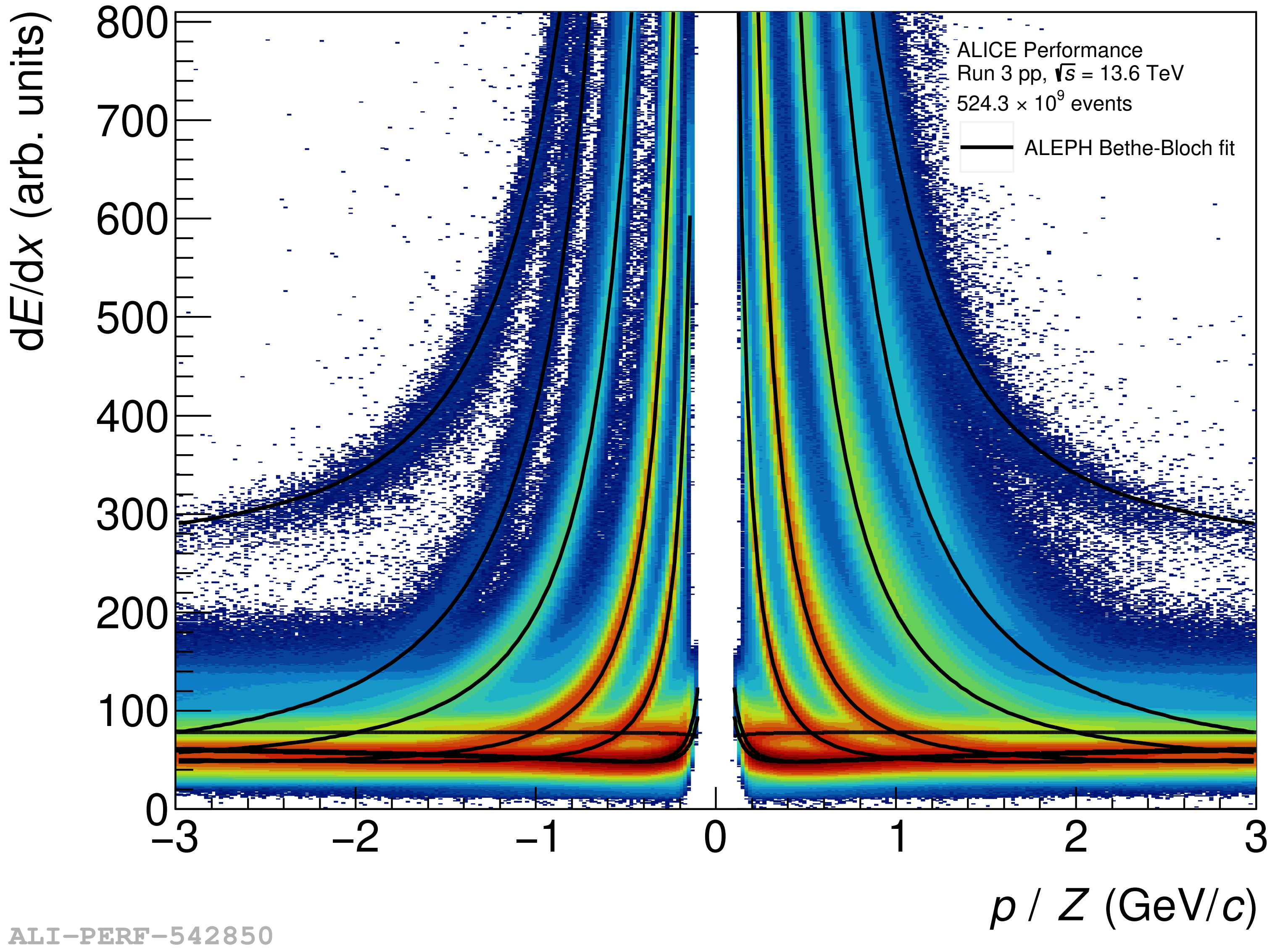
Monotonic Lipschitz DNN



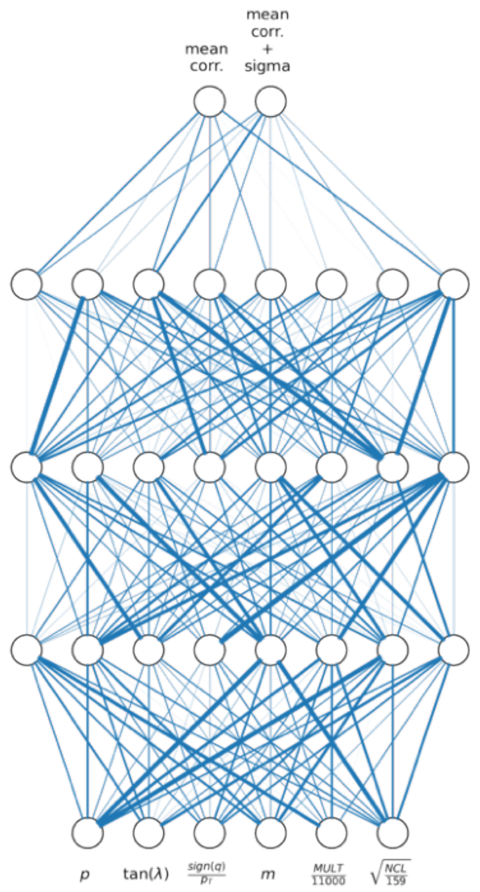
ALICE – TPC particle identification

TPC PID is central component in particle identification

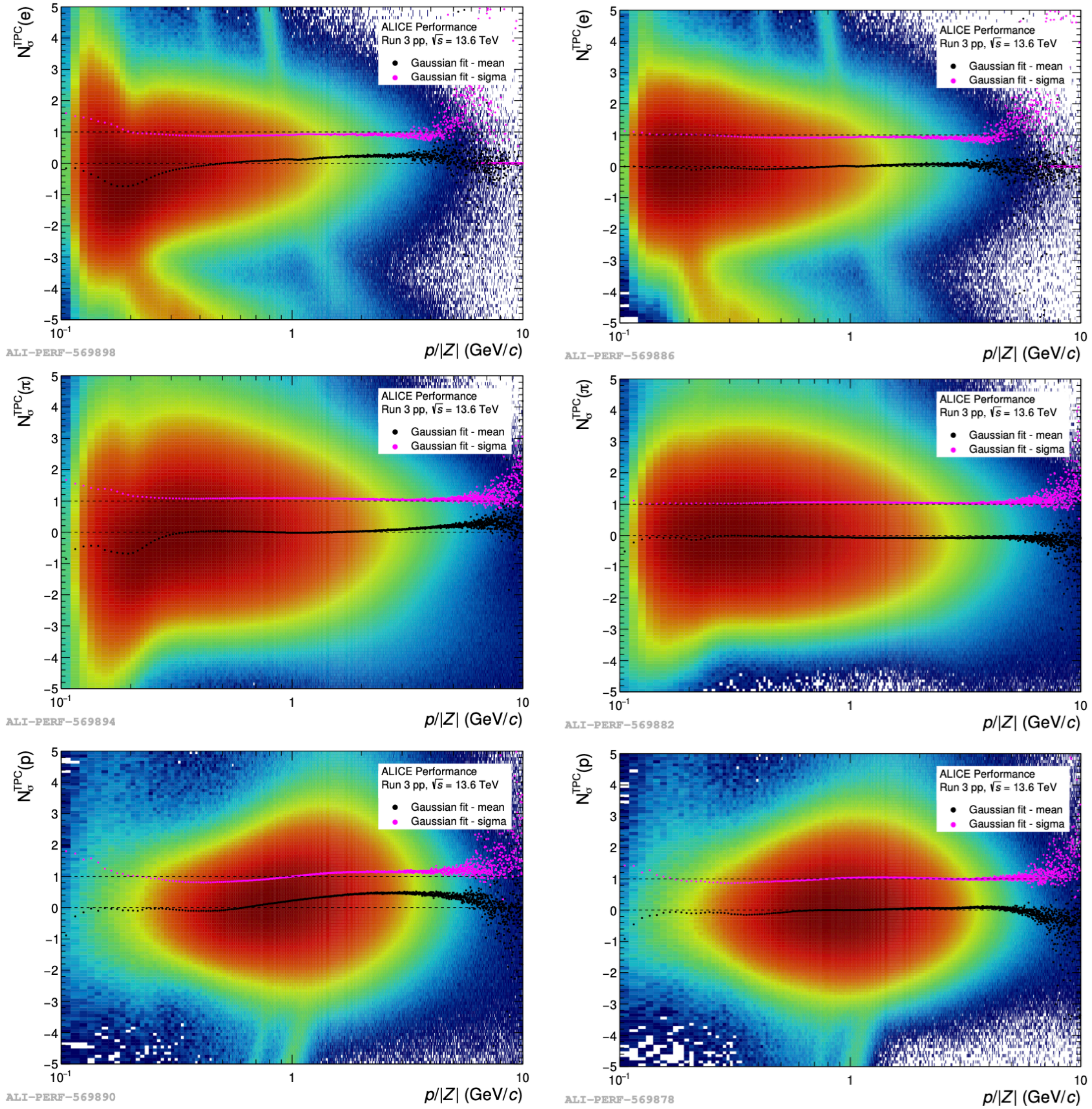
- Full 6D, data-driven corrections and sigma estimation to the Bethe-Bloch
- In full production release and utilised by analysers



→



→



Source: https://cds.cern.ch/record/2856252

News

Software and hardware for training

ML community wishlist for centralised training / testing

- Centralised guidelines and solutions for (C++) code integration
- Support for heterogeneous architectures
- Maintained, monitored and shared hardware availability through central CERN services

—> GPU's are expensive but chances are they will get even more expensive in the future
—> Experience across industry suggests that models will get larger
—> Local (e.g. institute resources) will not suffice to deal with LHC-scale data

---

CERN IT

- Review of action-list in progress and to be made public by end of summer
- Growing infrastructure for training and HPO at ml.cern.ch (from within CERN network)
- Infrastructure for ML for NextGen triggers: ETA is end of the year -> O(100) H100 GPU's

Conclusion

LHC experiments make heavy use of ML!

- Model architectures: Close to cutting-edge developments in industry
- Large scale inferencing on huge (TB/s) data
- Infrastructure is ever growing

Where can we improve?

- Centralised training infrastructure
- Adopt C++ framework for inference on heterogeneous architectures
- With increasing model size, energy consumption grows

Some advertisement

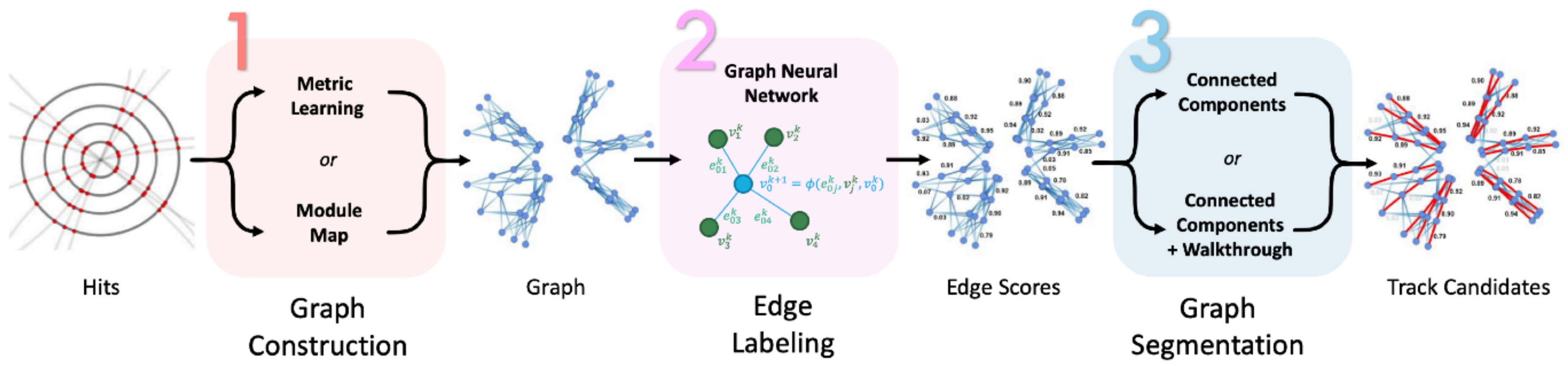
- Inter-experimental machine learning (IML) group: https://iml.web.cern.ch/homepage

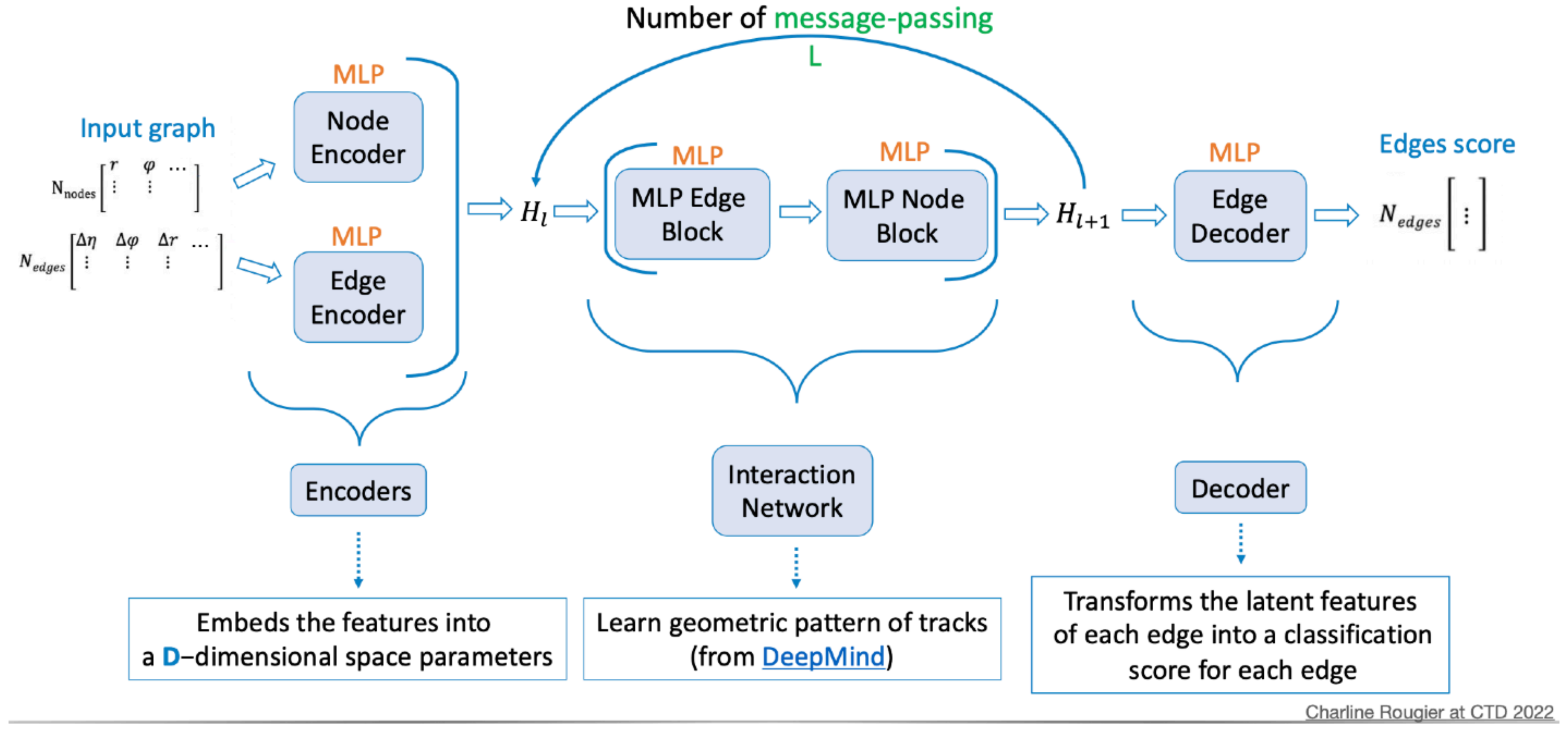
Backup

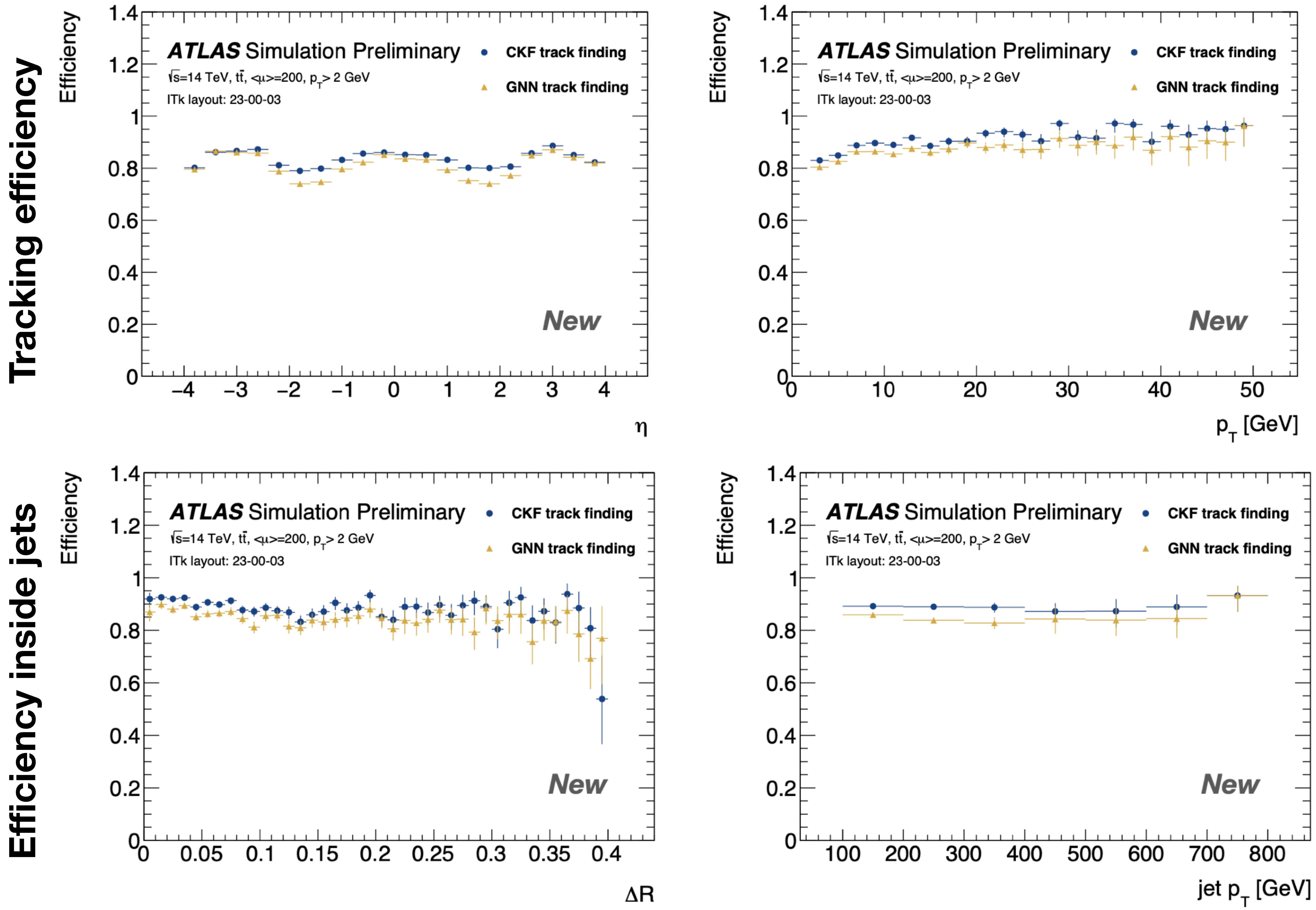
ATLAS - GNN for tracking in Run 4

GNN4ITk - Full GNN tracking chain for Run 4

- Competitive to CKF tracker both in terms of physics and computing performance

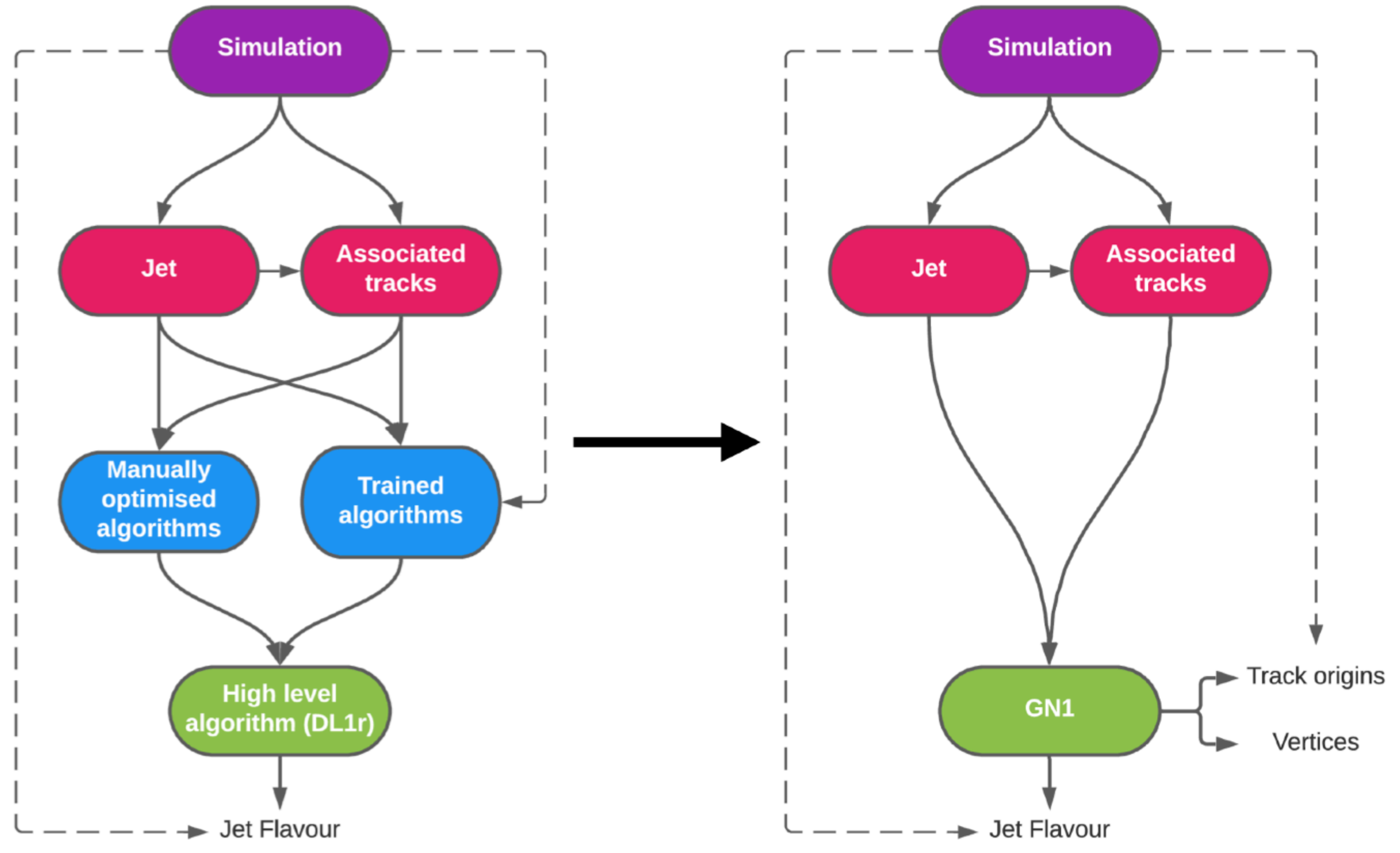


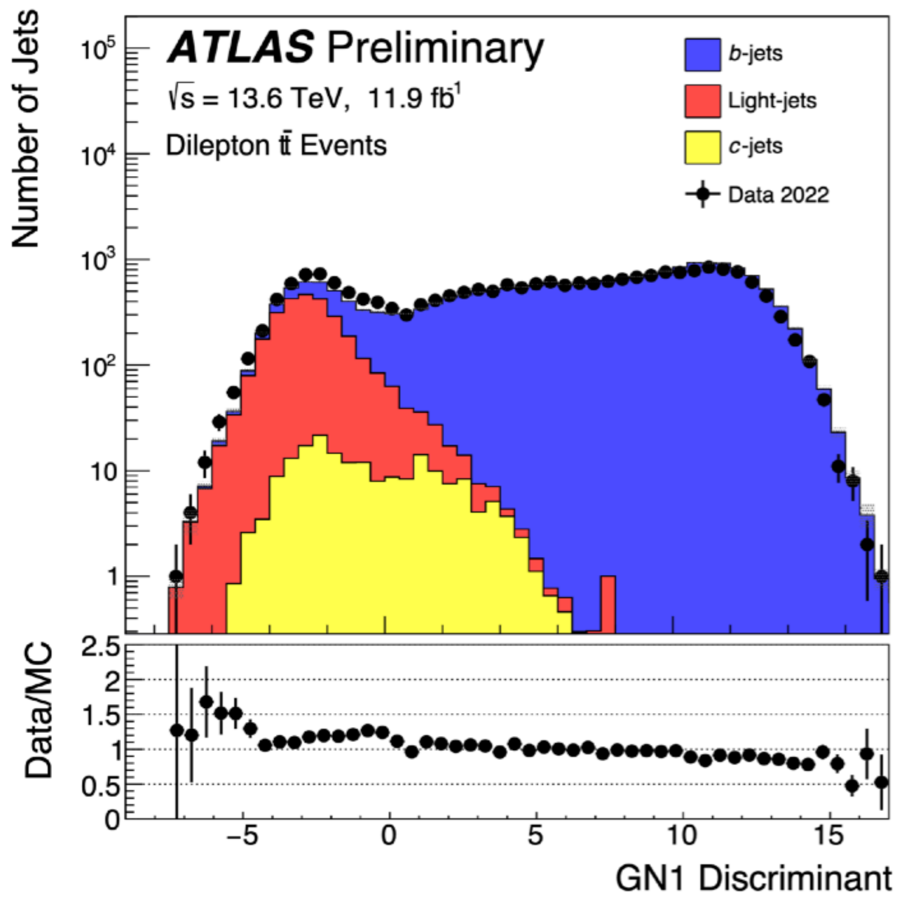


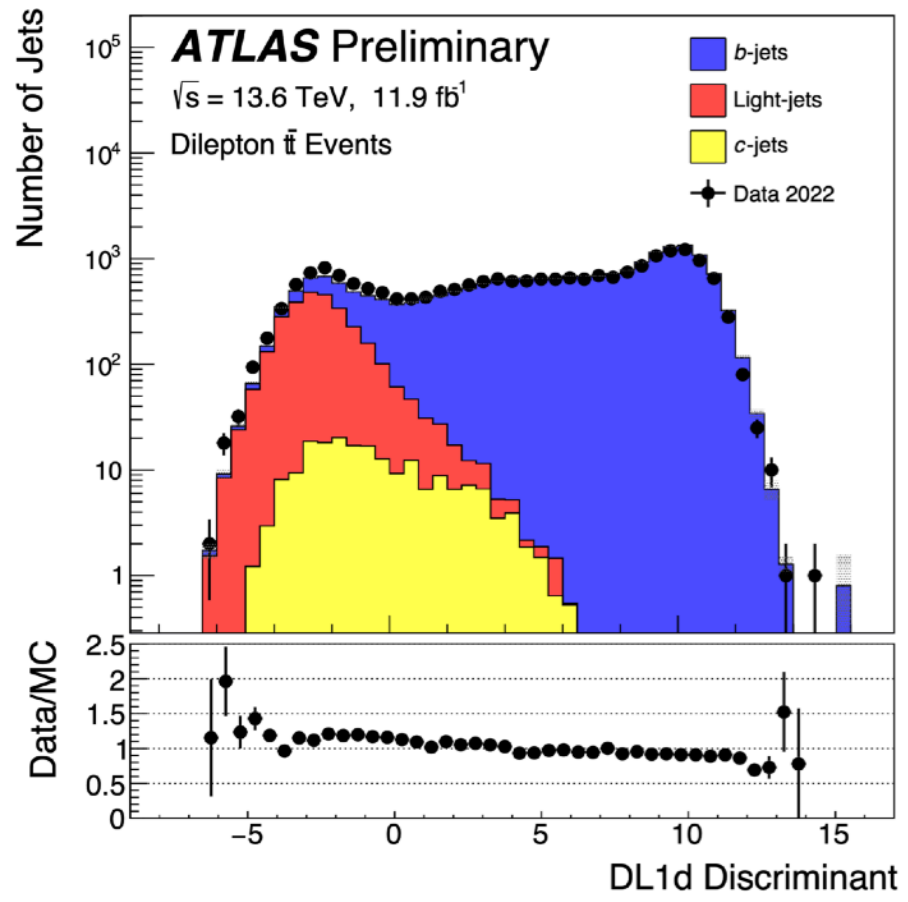


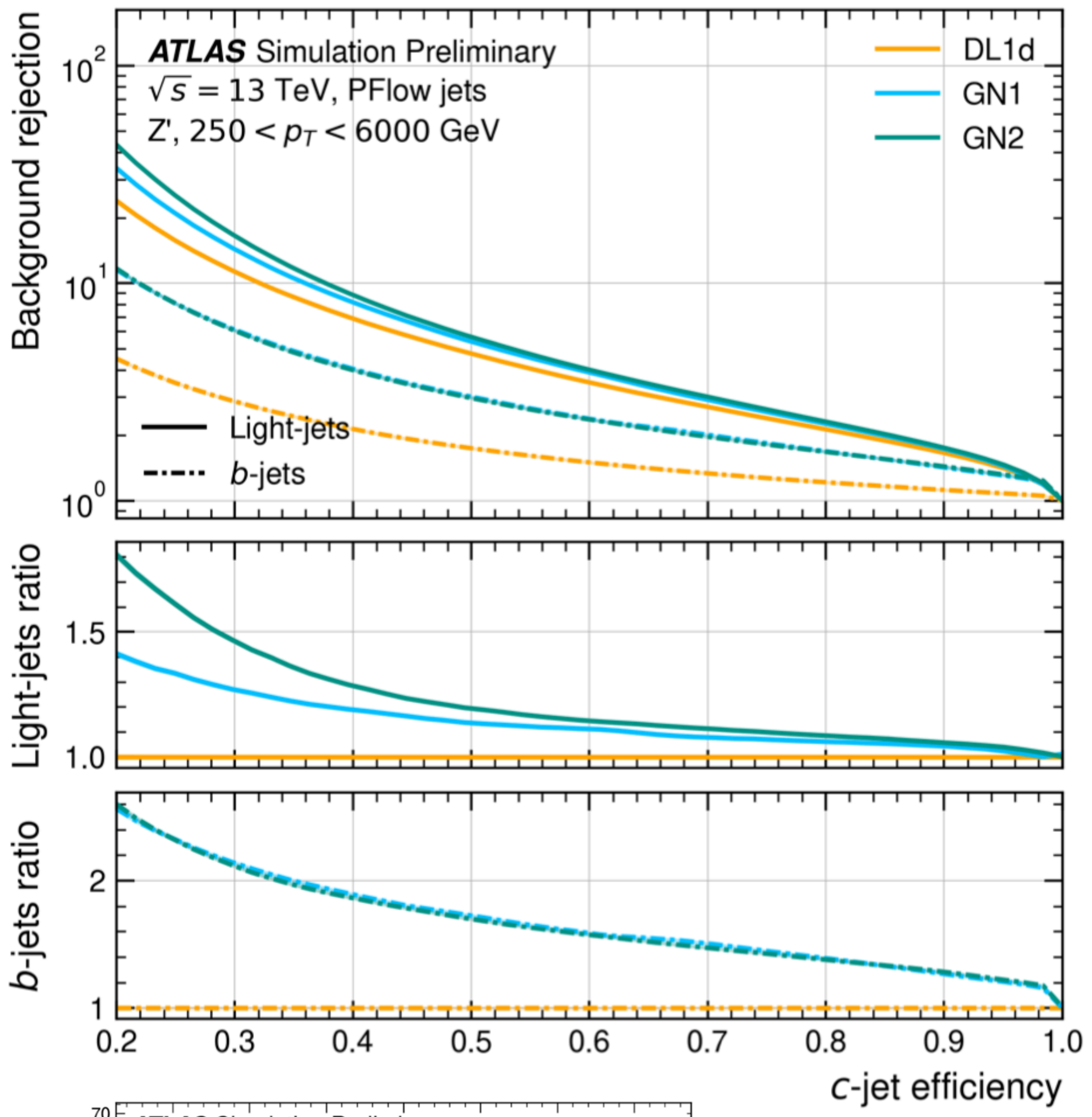
Source: https://indico.cern.ch/event/1252748/contributions/5576737/

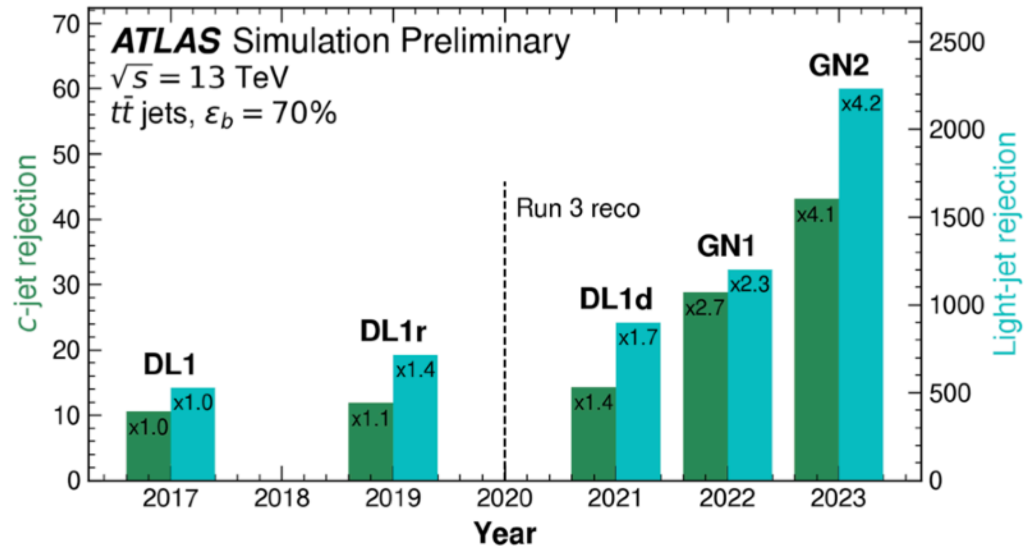
ATLAS - GNN for jet-flavour tagging











At the 70% $t\bar{t}$ working point (WP) for GN1:

- ➢ **2.25x** increase in c-jet rejection
- ➢ **1.8x** increase in light-jet rejection

- 1.5x c-rejection and 2x light-rejection on **ttbar**
- 1.75x c-rejection and 1.2x light-rejection on **Z'**

New HLT performance plots!

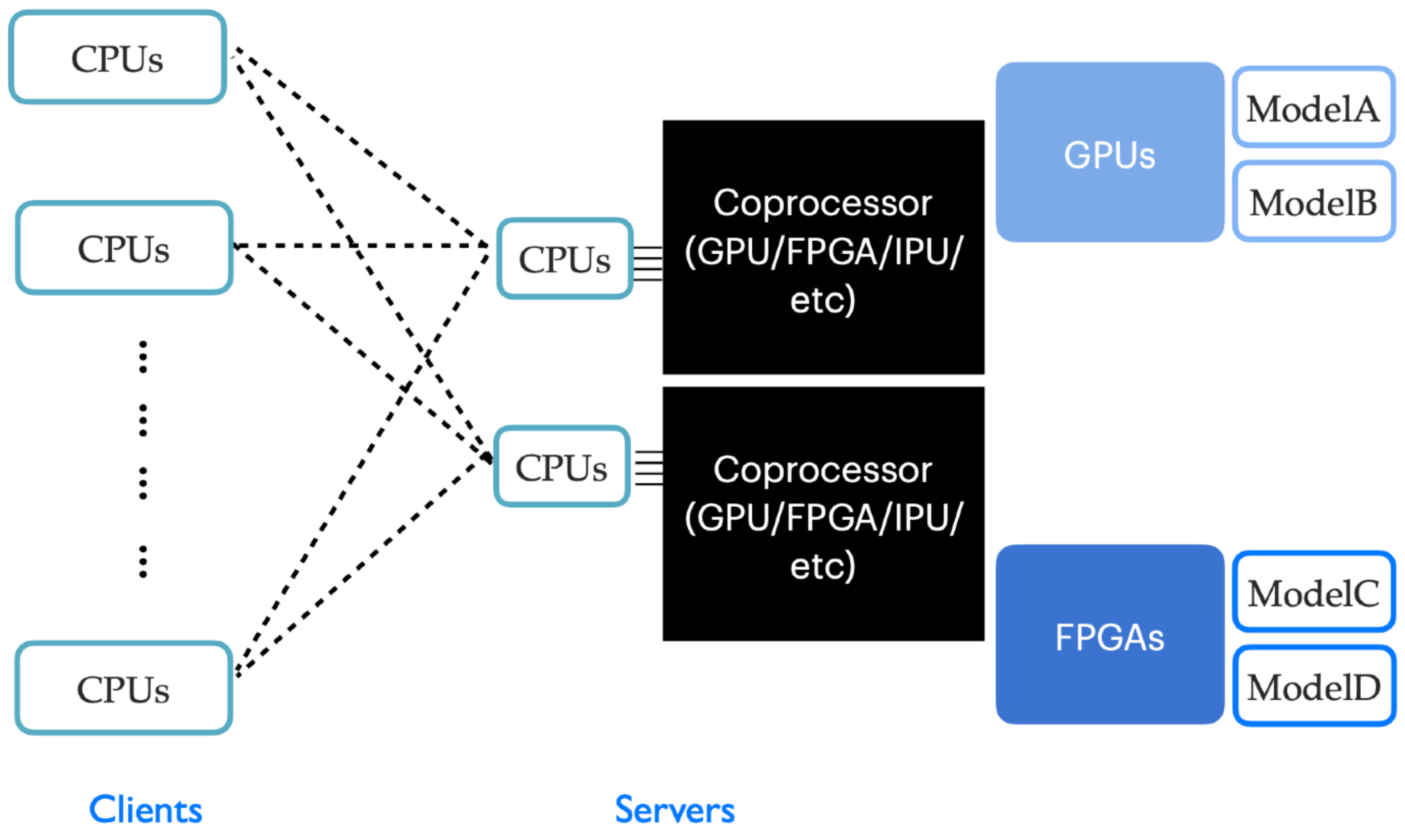
Source: https://cds.cern.ch/record/2855275/files/ATL-PHYS-SLIDE-2023-048.pdf

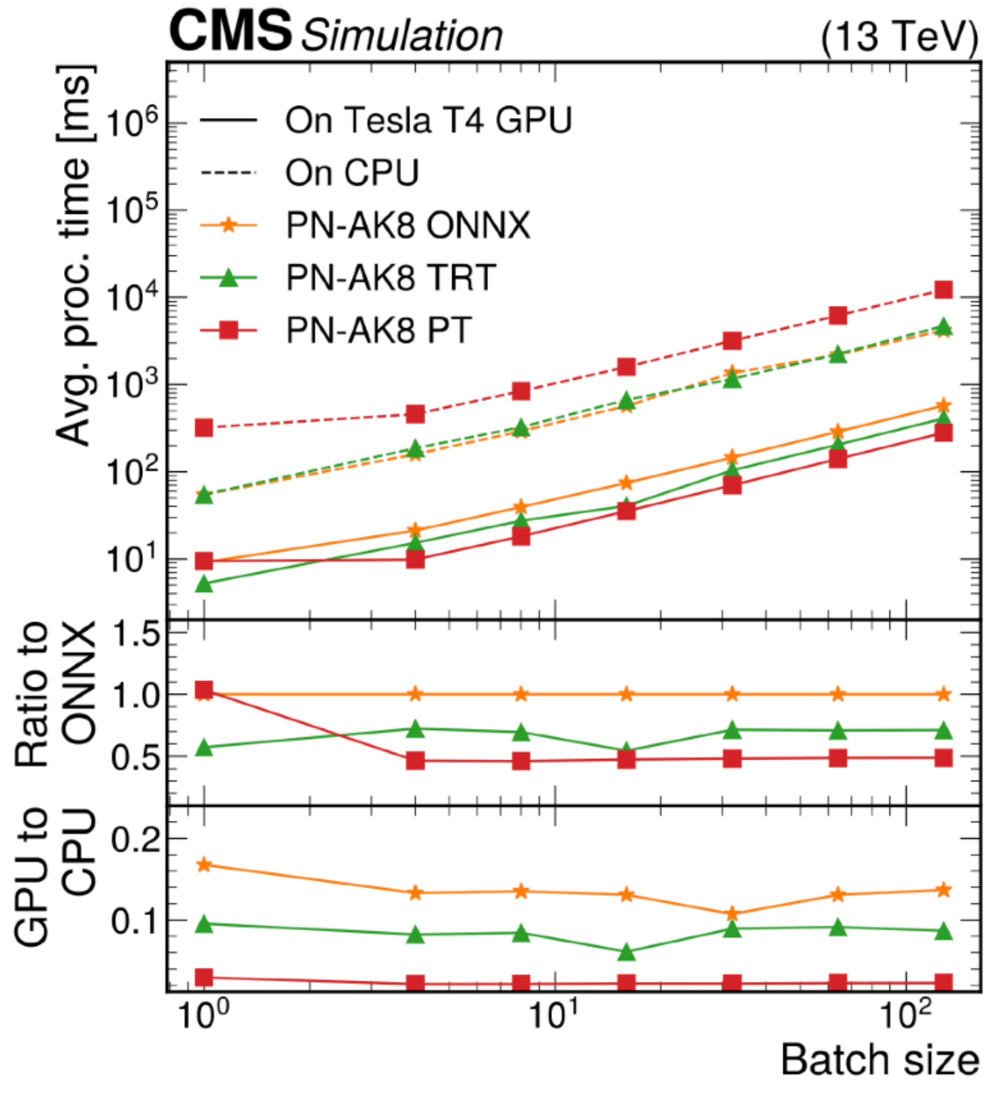
CMS - SONIC framework

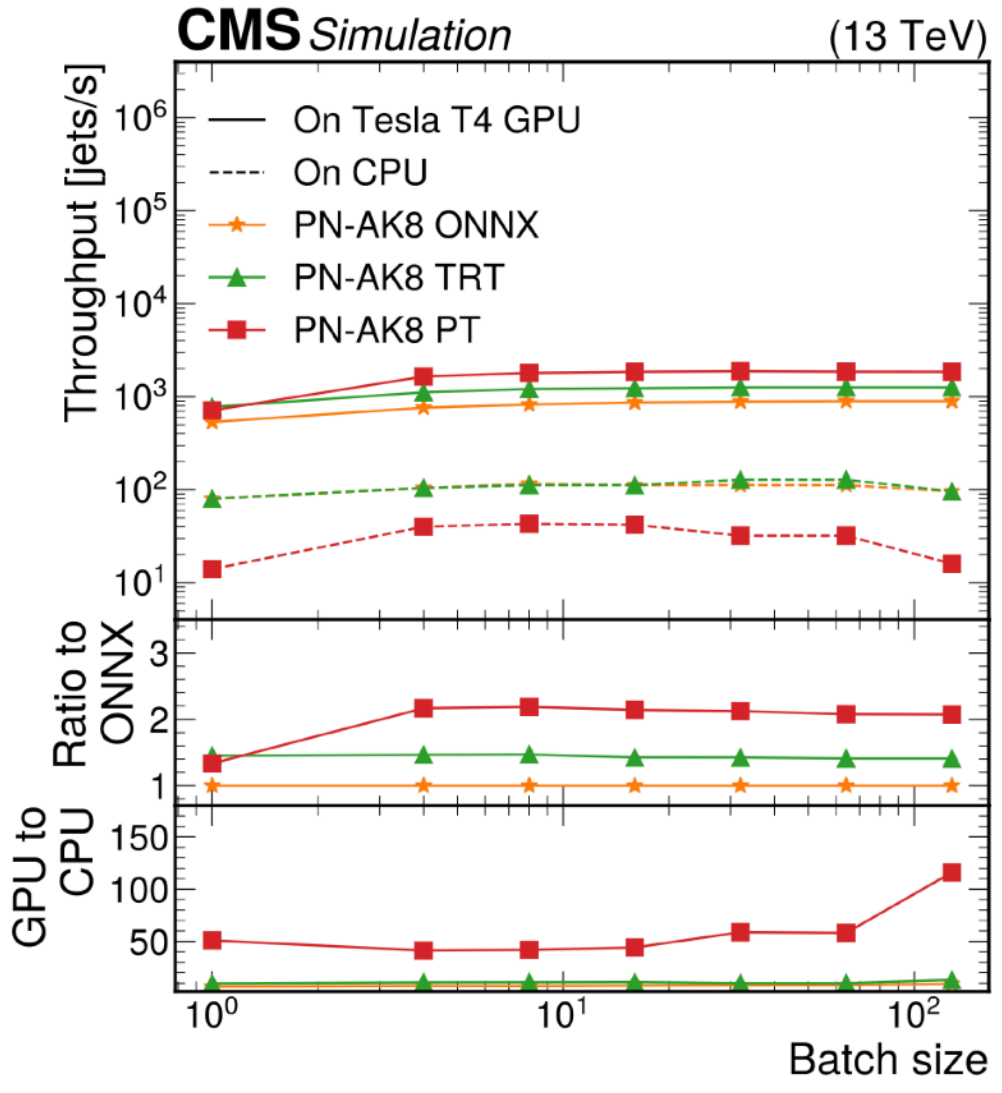
Services for Optimized Network Inference on Coprocessors — SONIC

Coprocessors (GPU's, FPGA's, ASIC's, IPU's) as-a-service application for large-scale data processing

- NVIDIA Triton framework for inference on co-processors (support for PyTorch, TensorRT, ONNX Runtime, TensorFlow and XGBOOST models)
- Advantages on-server: Multiple model instances, dynamic batching, model analyzer, ragged batching
- Advantages of SONIC: Containerisation, simplicity, efficiency, flexibilty







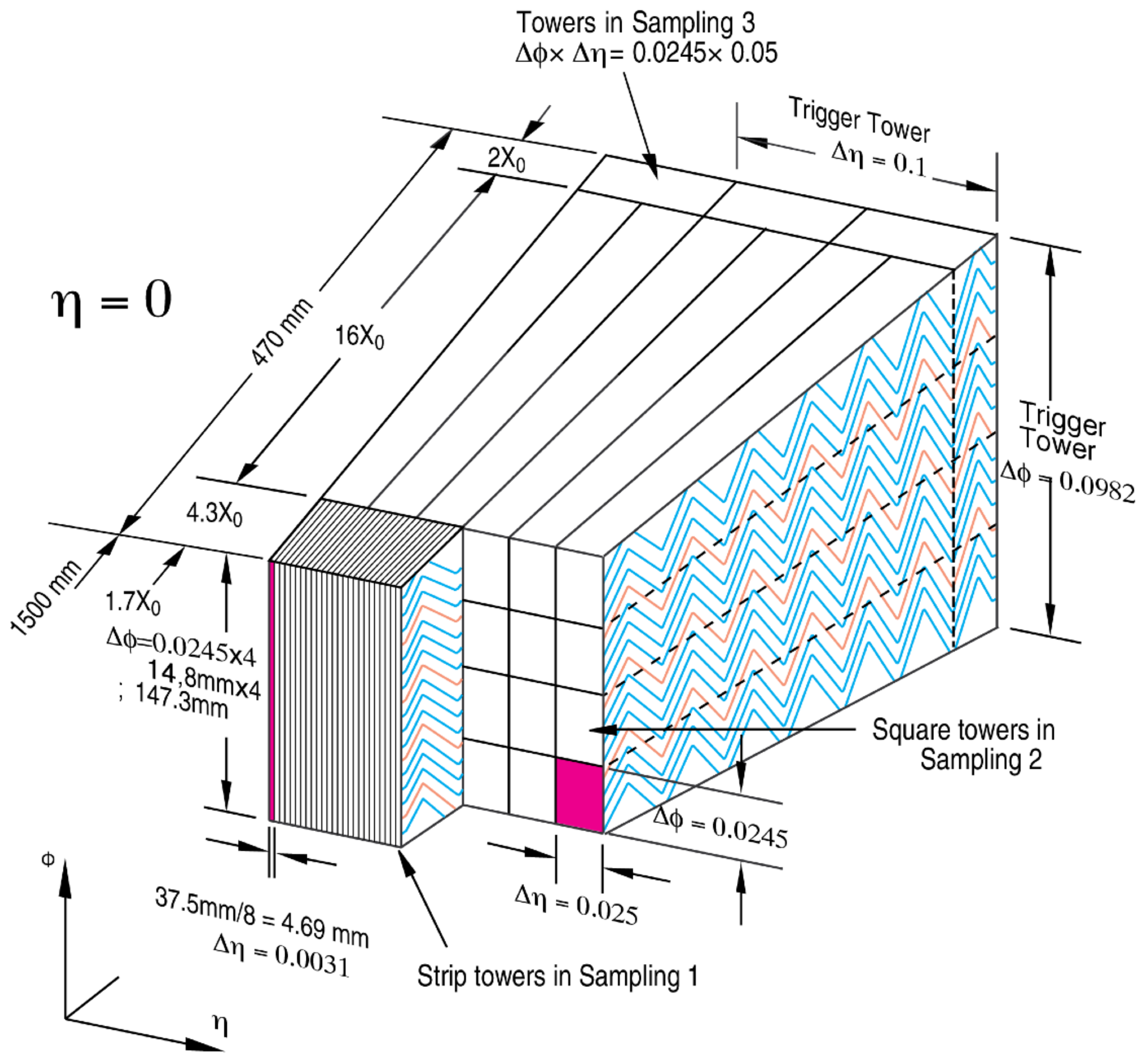
Large scale tests confirm most of the expecations and prove validity of the approach in production

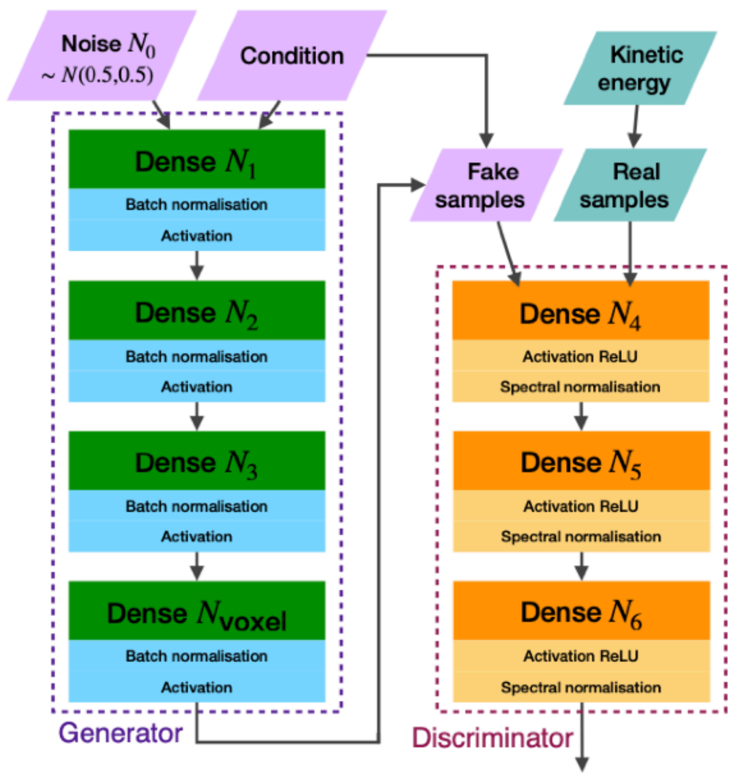
ATLAS - AtlFast3 simulation suite

Fast simulation

GEANT is a common simulation tool but very compute intensive -> Calorimeter GAN simulation

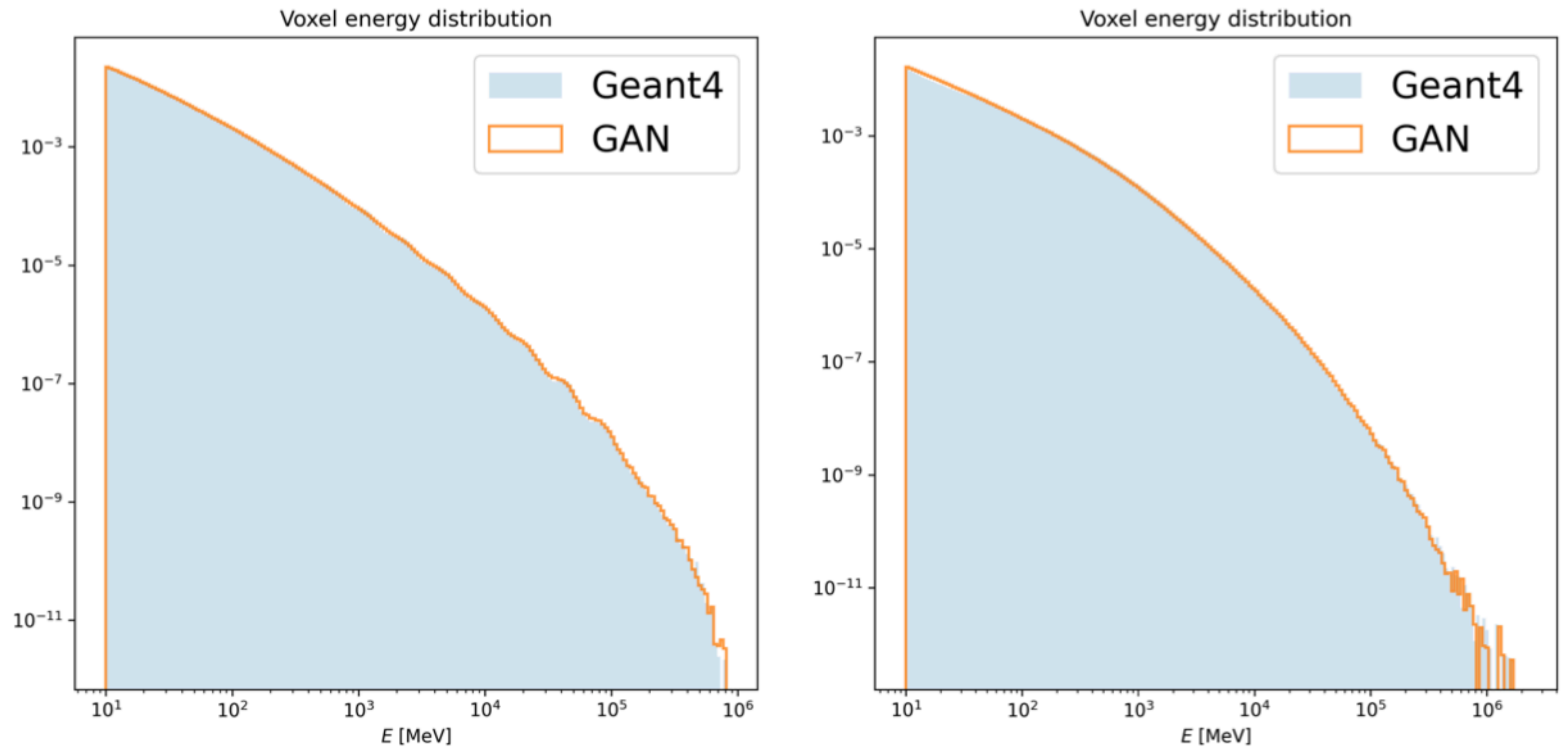
AtlFast3: FastCaloSimV2 (parameterized model) + FastCaloGAN (ML)

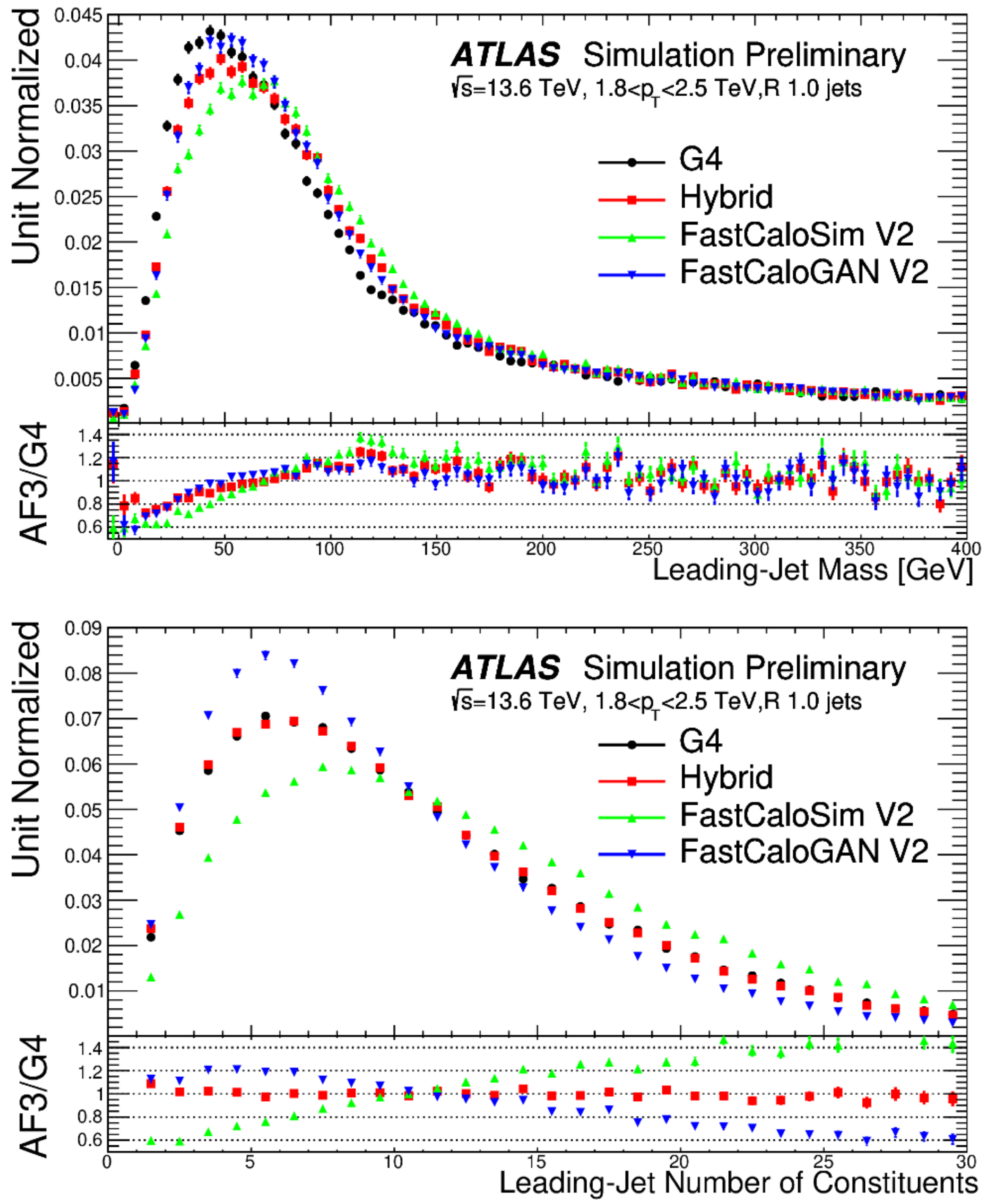




Batch size	Time per batch [ms]
1	6.3(3)
10	6.8(15)
100	8.2(15)
1000	14.4(21)
10000	70.8(48)

Per-voxel energy deposit!

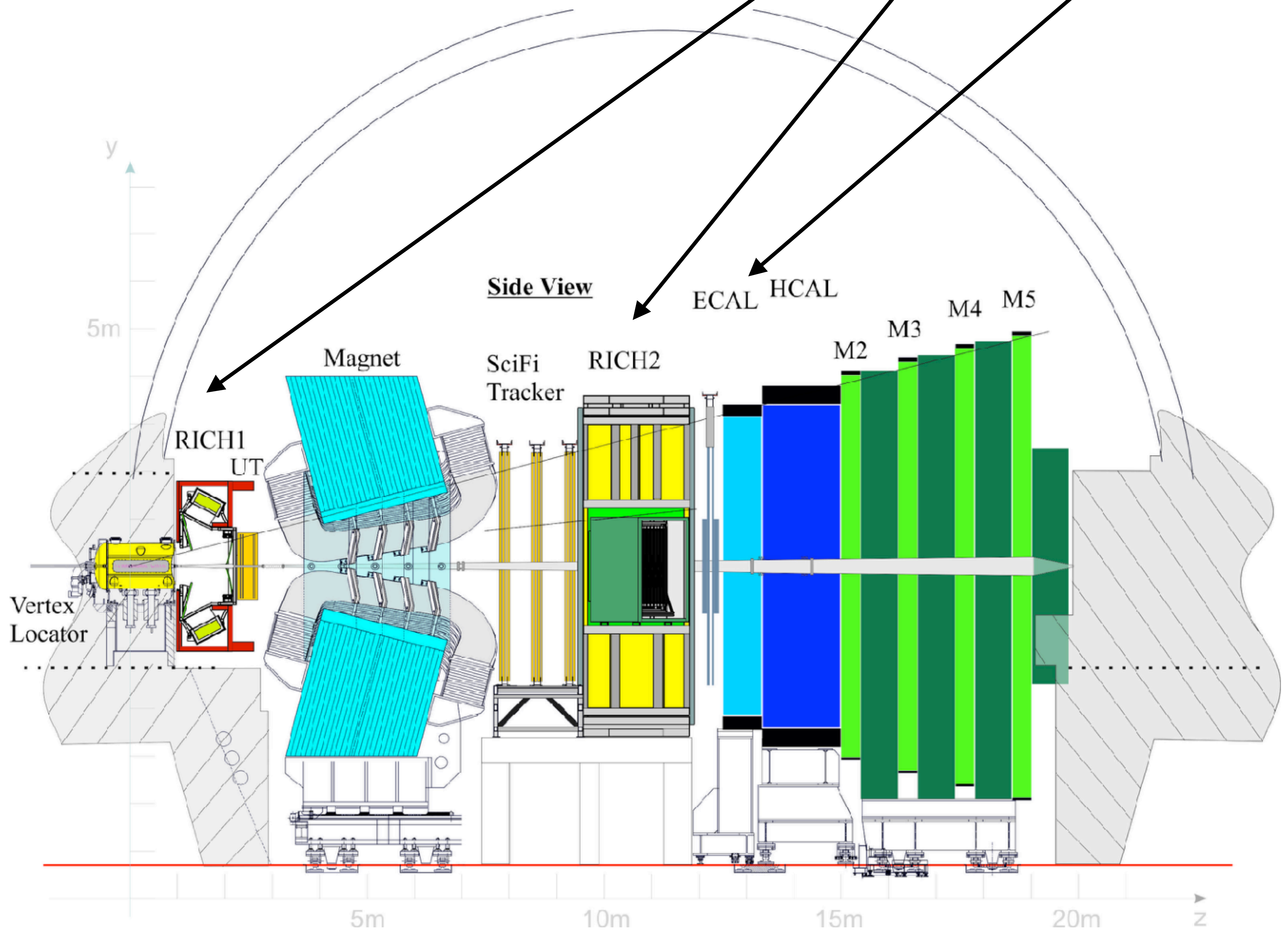




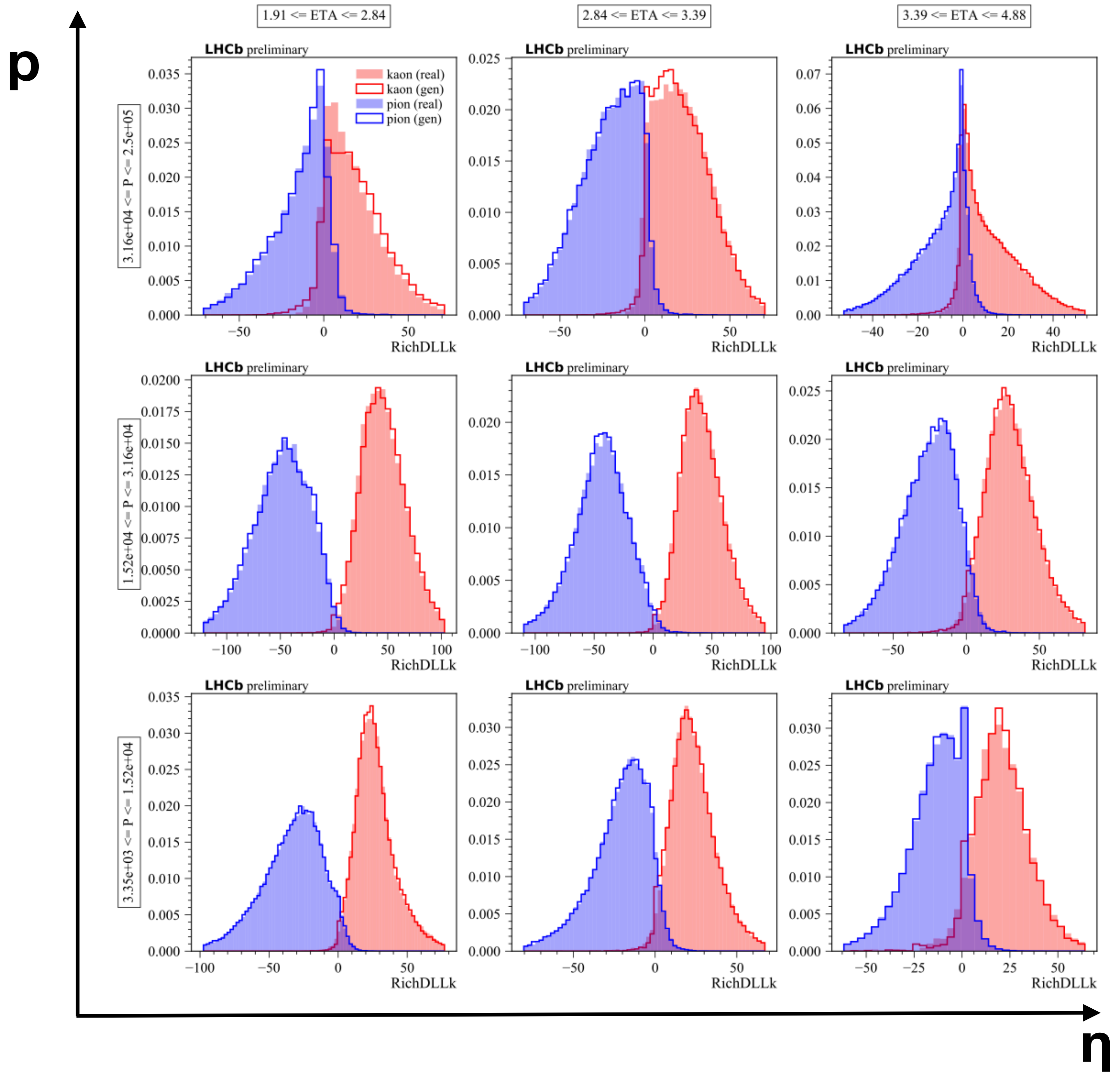
LHCb – RICH and ECAL simulation

Fast simulation

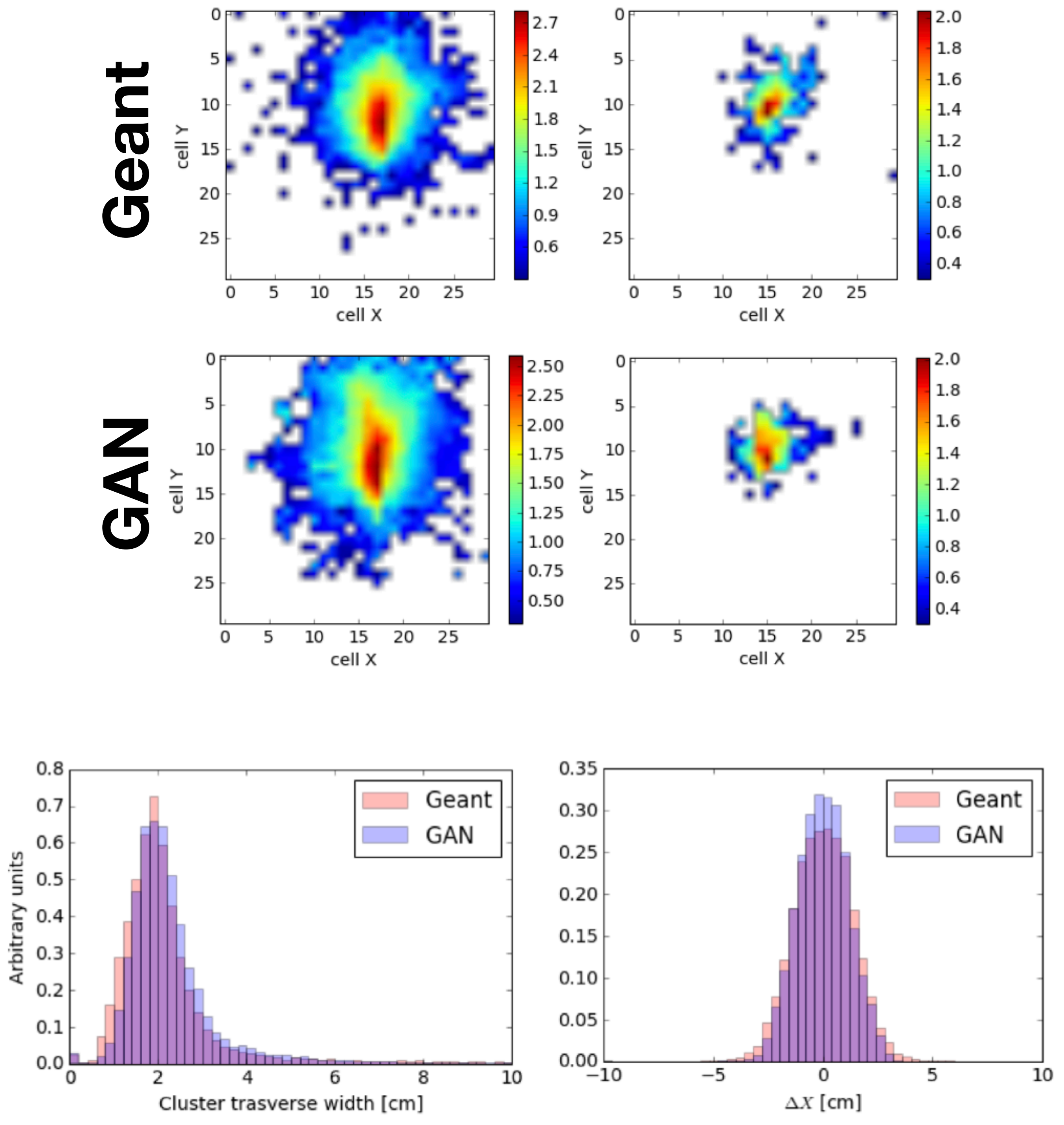
Demanding simulations for LHCb: RICH & ECAL



RICH PID simulation



ECAL simulation



ALICE – ZDC fast simulation

ZDC is located over 100 meters away from IP2 -> GEANT simulation is slow

- Simulation via DC-GAN
- Auxiliary regressor to fine-tune loss on maximum photon deposit
- Post-processing via scaling on Wasserstein distance