Realtime reconstruction
Machine learning in reconstruction at LHC

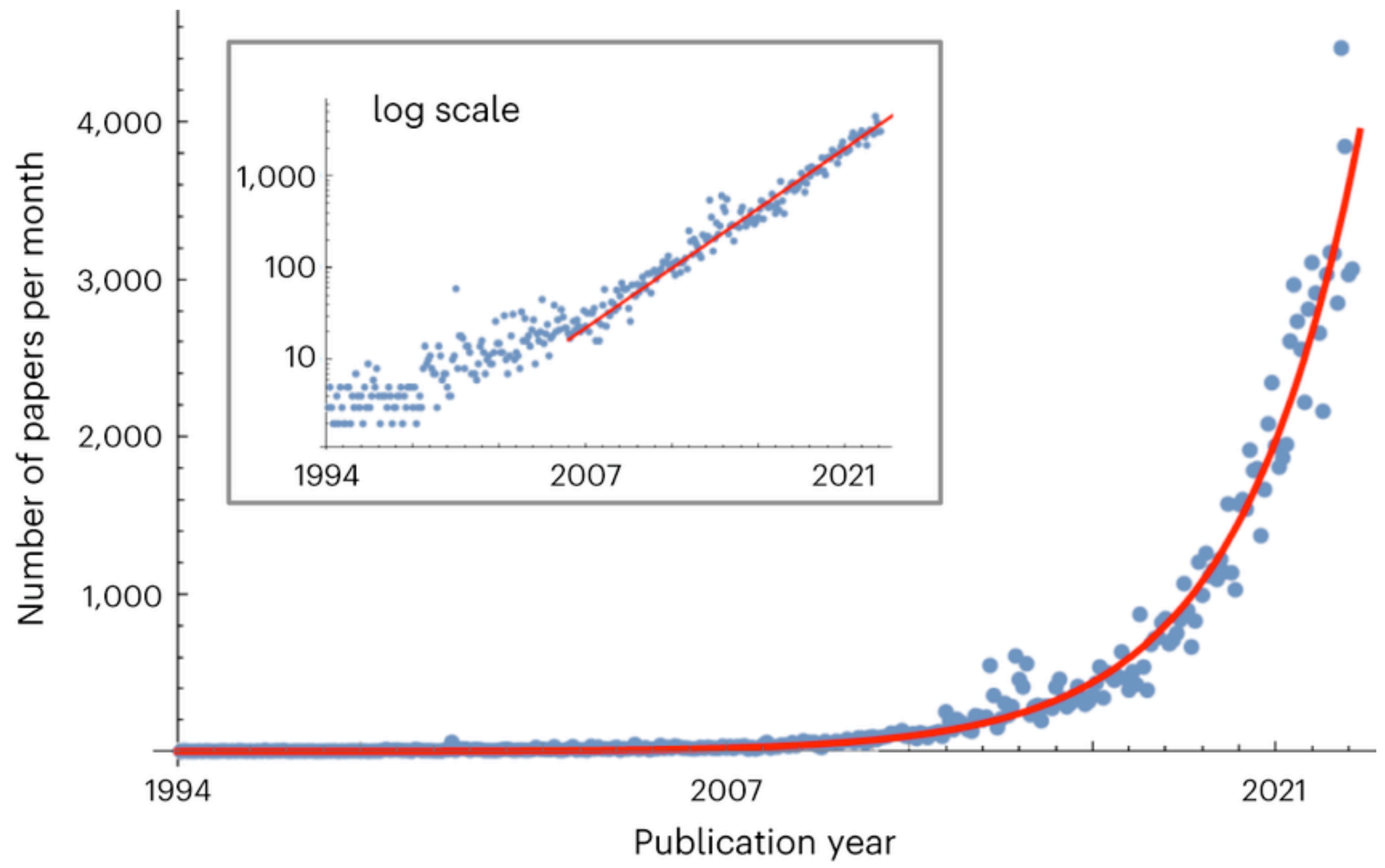
ML4Jets workshop 2024

Simon Akar
Laboratoire de Physique de Clermont Auvergne

ML in HEP

- **Usage of ML algorithm in HEP is not new…**
  - BDTs as an Alternative to ANN for Particle Identification (MiniBooNE) (2004) [*arXiv:physics/0408124*]
  - Tagging heavy flavours with BDTs (2007) [*arXiv:physics/0702041*]

- **…but, like everywhere else, since several years, ML is getting constantly increasing attention, diversity of applications and refined models**



Nature Machine Intelligence | Volume 5 | November 2023 | 1326–1335

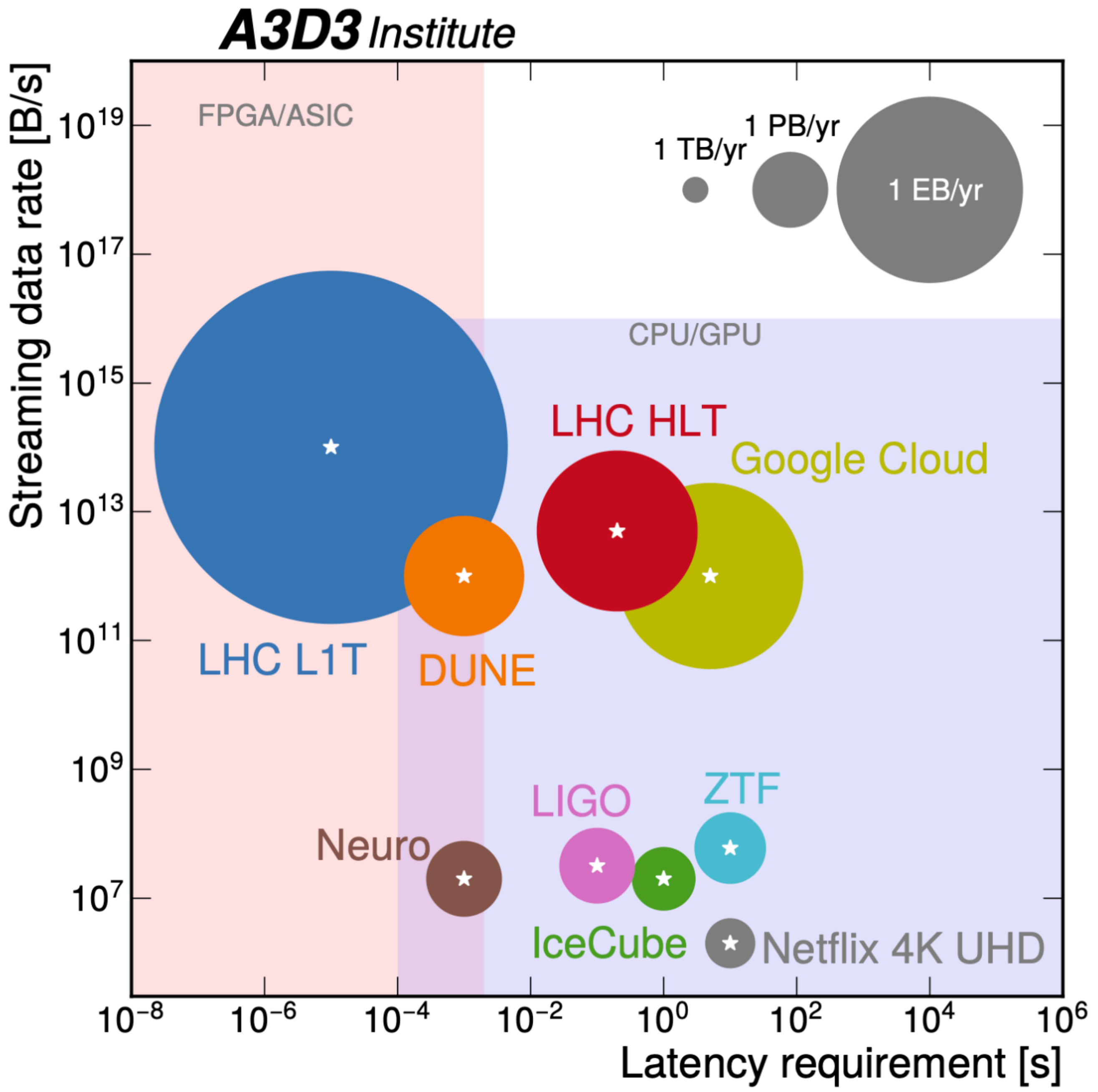
Doubling rate of arXiv papers in categories of AI and ML per month is roughly 23 months

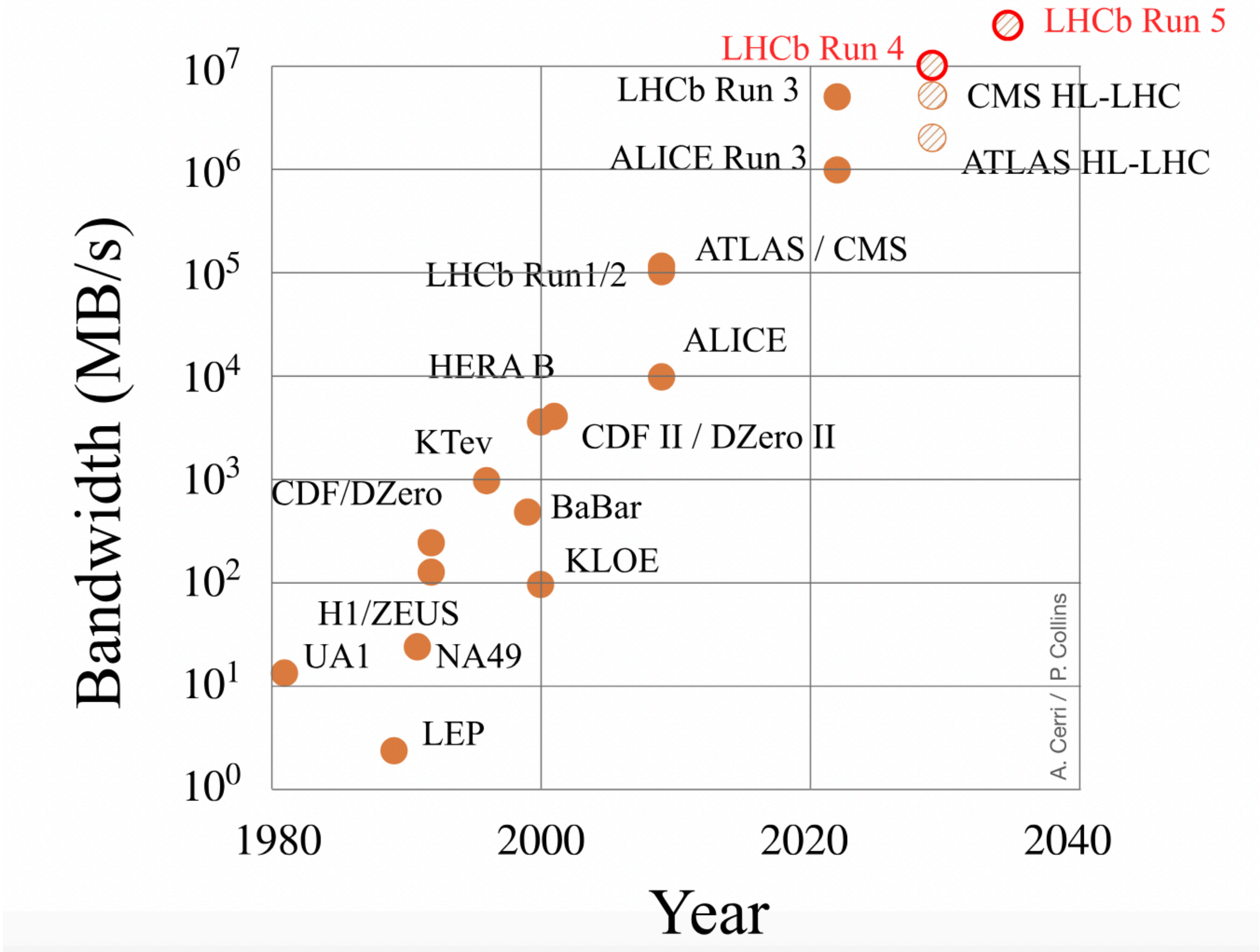
ML in realtime reconstruction

- **What is "realtime"? → Online as opposed to Offline**

1. **Online**: Algorithms, or sequences of algorithms, executed on events read out from the detector in near-real-time as part of the software trigger, typically on a computing facility located close to the detector itself.

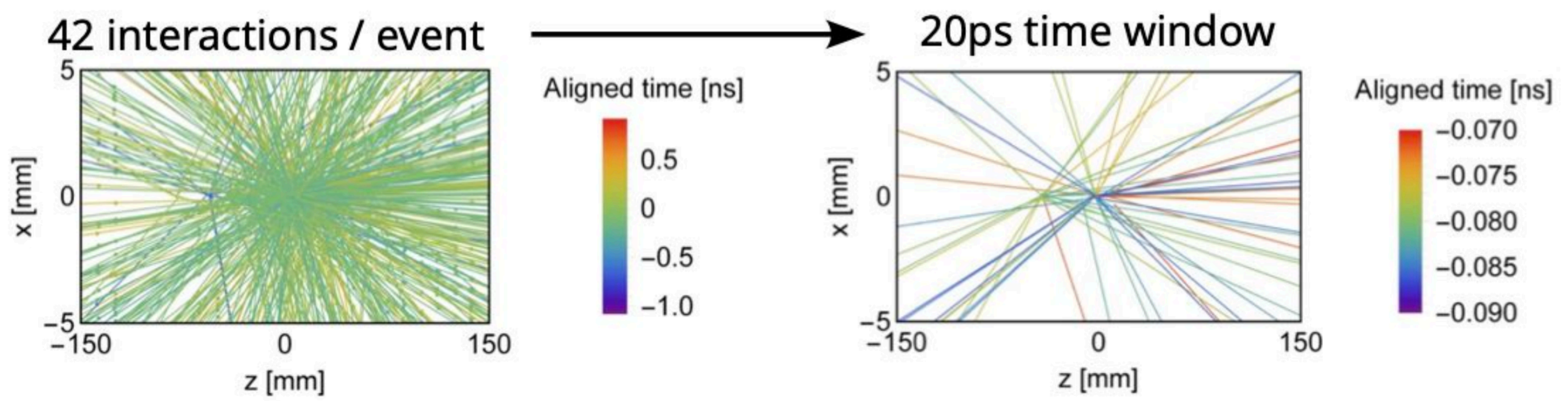
- Depending on the experimental environment "realtime" will have different meaning, hence different constraints:

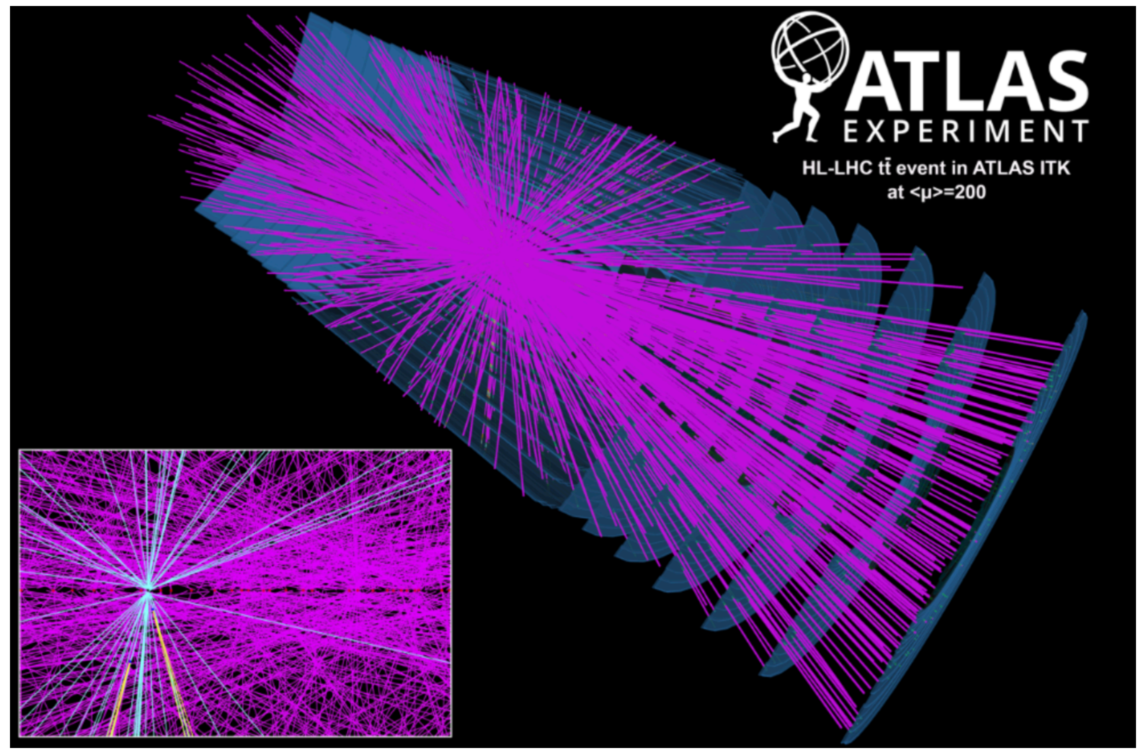




Why using ML in realtime reconstruction

- **Aiming to push our capabilities to search for BSM to the limit**
  - **Continuously increasing number of interactions per event → more objects to reconstruct**
    ↳ Timing will help mitigating this effect
    ↳ ML to further improve performances
  - **Continuously increasing detector granularity → more data to handle**
    e.g. CMS High Granularity Calorimeter
    with ~6.5 M readout channels
  - **Continuously improving ML model architectures**
    Most of the ML developments done outside of HEP
    ↳ **need for a strong community of experts including engineers!**

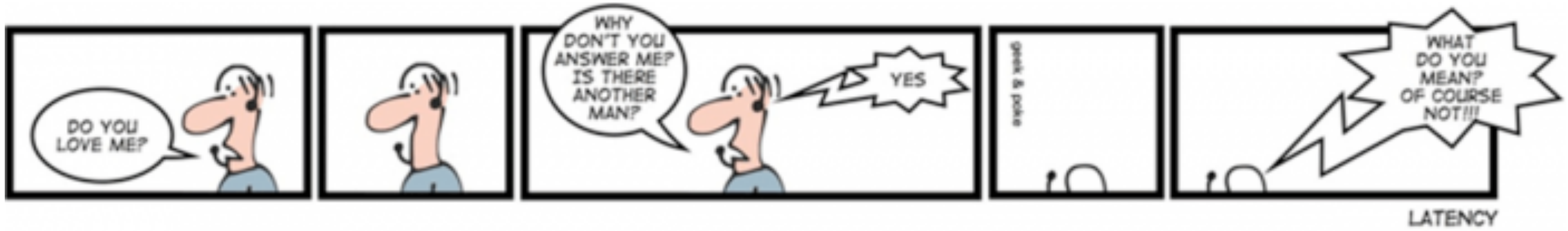




Increasing interest in using ML @ reconstruction and / or trigger levels

Stringent requirements

- Realtime reconstruction is probably the area for ML application in HEP facing the most important challenges



No coming back from discarded data in triggers

- **(Very)-Low latency**
  Capable to cope with the high rates
  ATLAS/CMS: 40 MHz @ L1 (FPGA, ASIC) | ~100 kHz @ HLT (GPU/CPU)
  LHCb: 30 MHz @ HLT (GPU/CPU)

- **Reliability/Flexibility**
  Be able to quickly adapt to detector performance evolution (radiation damage, dead zones, …)

- **Maintainability**
  Ensure the underlying librairie / software can be maintained for relatively long period *O*(10 years)

Common tools

- **From development to running ML algorithms, many common tools currently used in HEP applications:**
  - **ML model librairies** (current "standards")

    PyTorch TensorFlow Keras & more...
  - **Inference engines (GPU/CPU)**

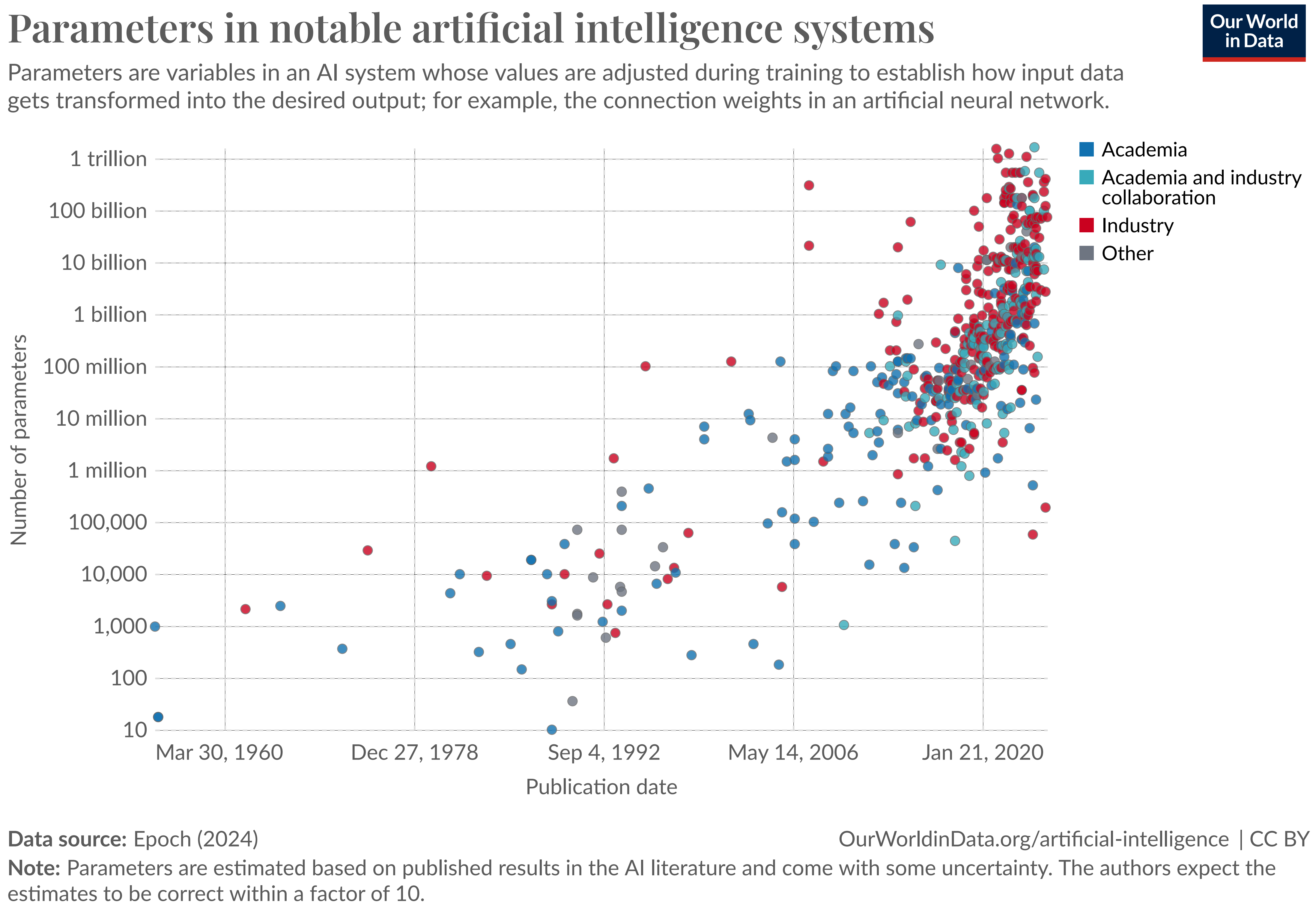
    ONNX RUNTIME TensorRT TMVA/SOFIE & more...
  - **Running model in FPGA/ASIC**

    hls4ml FINN Conifer & more...

    arxiv:1804.06913 *arxiv:1612.07119* *S. Summer @ FDF24*

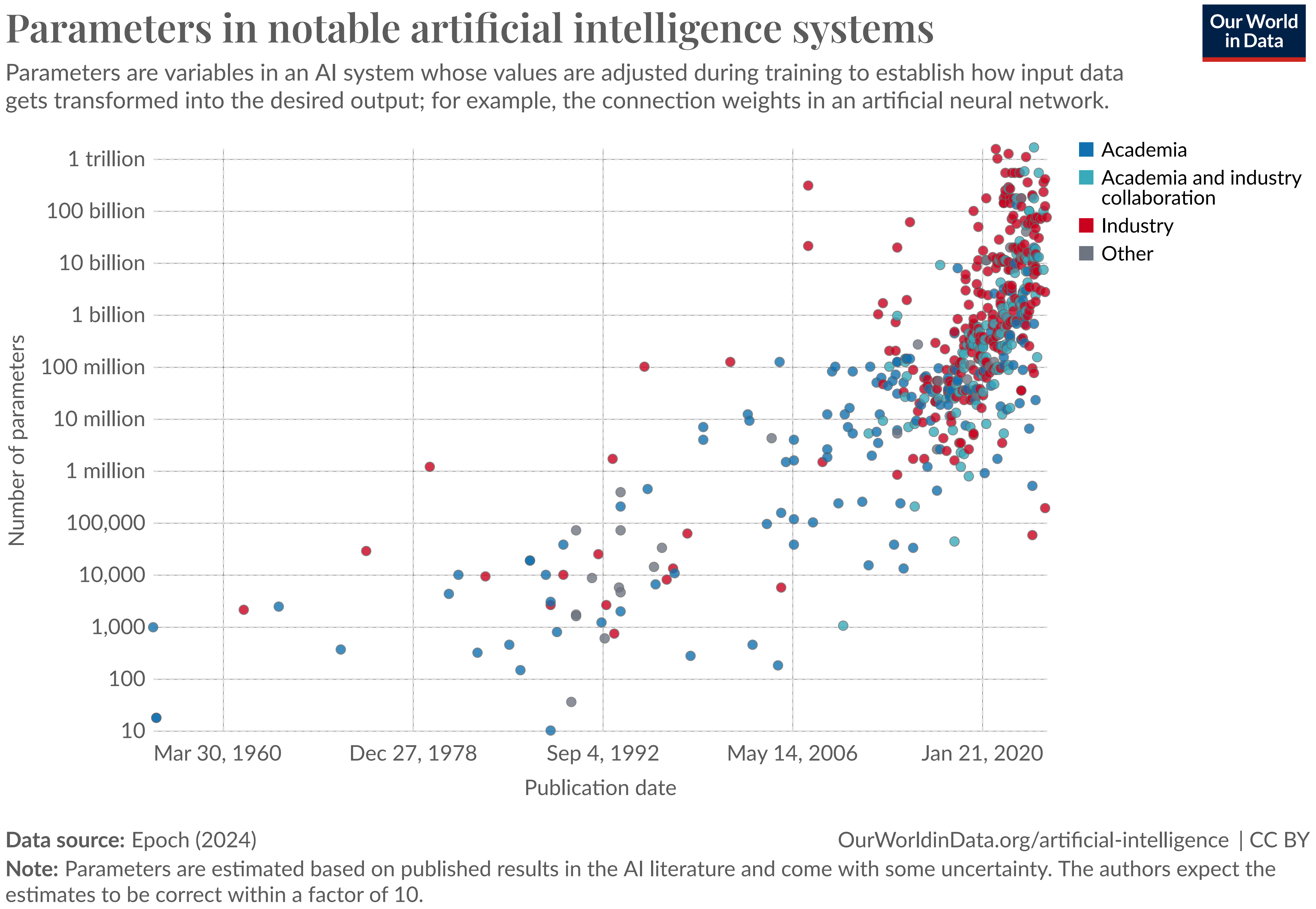
Common challenges

- **Current paradigm in industry, with the advent of LLMs, is toward continuously larger models (~10^{12} parameters in GPT4)**



Common challenges

- Current paradigm in industry, with the advent of LLMs, is toward continuously larger models (~10^{12} parameters in GPT4)



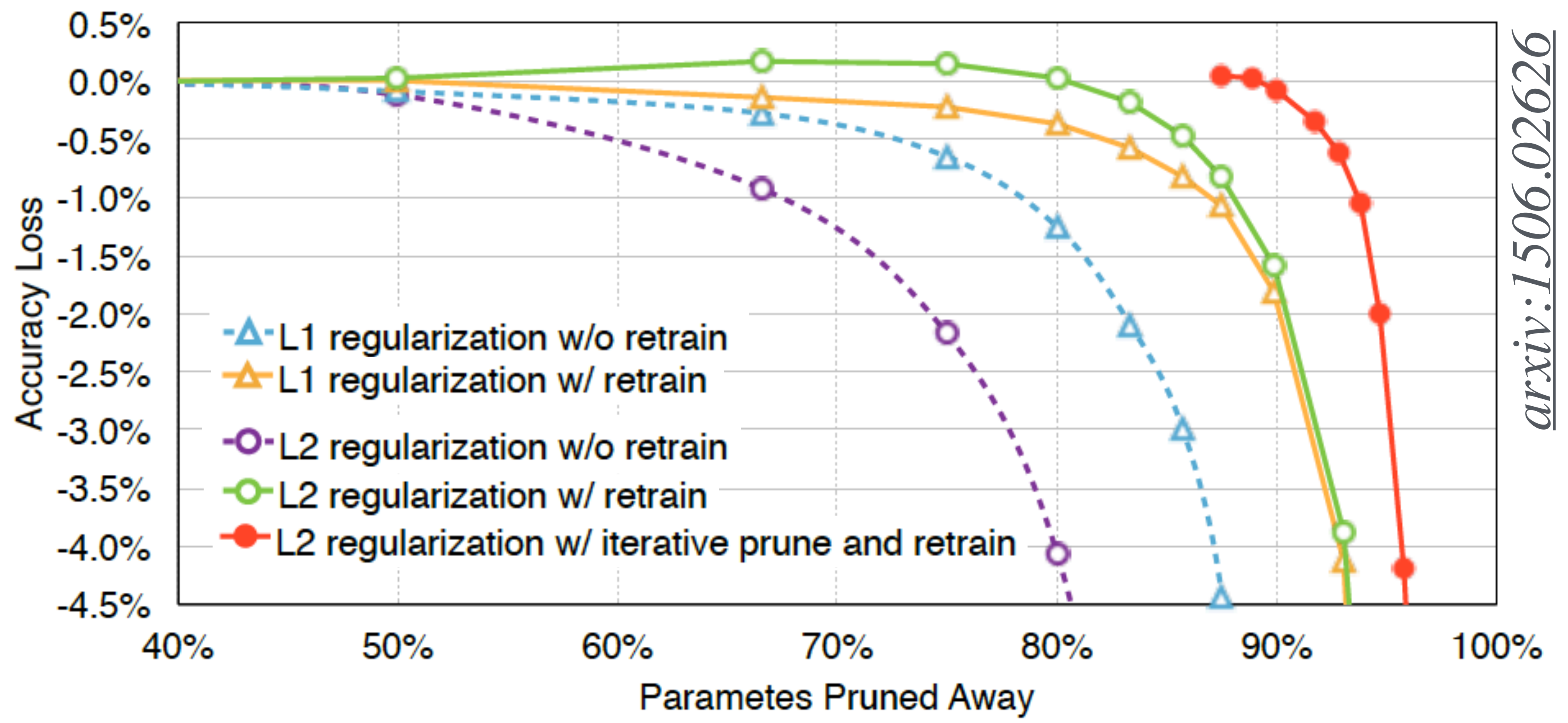
but...

Realtime requires to be more subtle to cope with limited device sizes

Common challenges

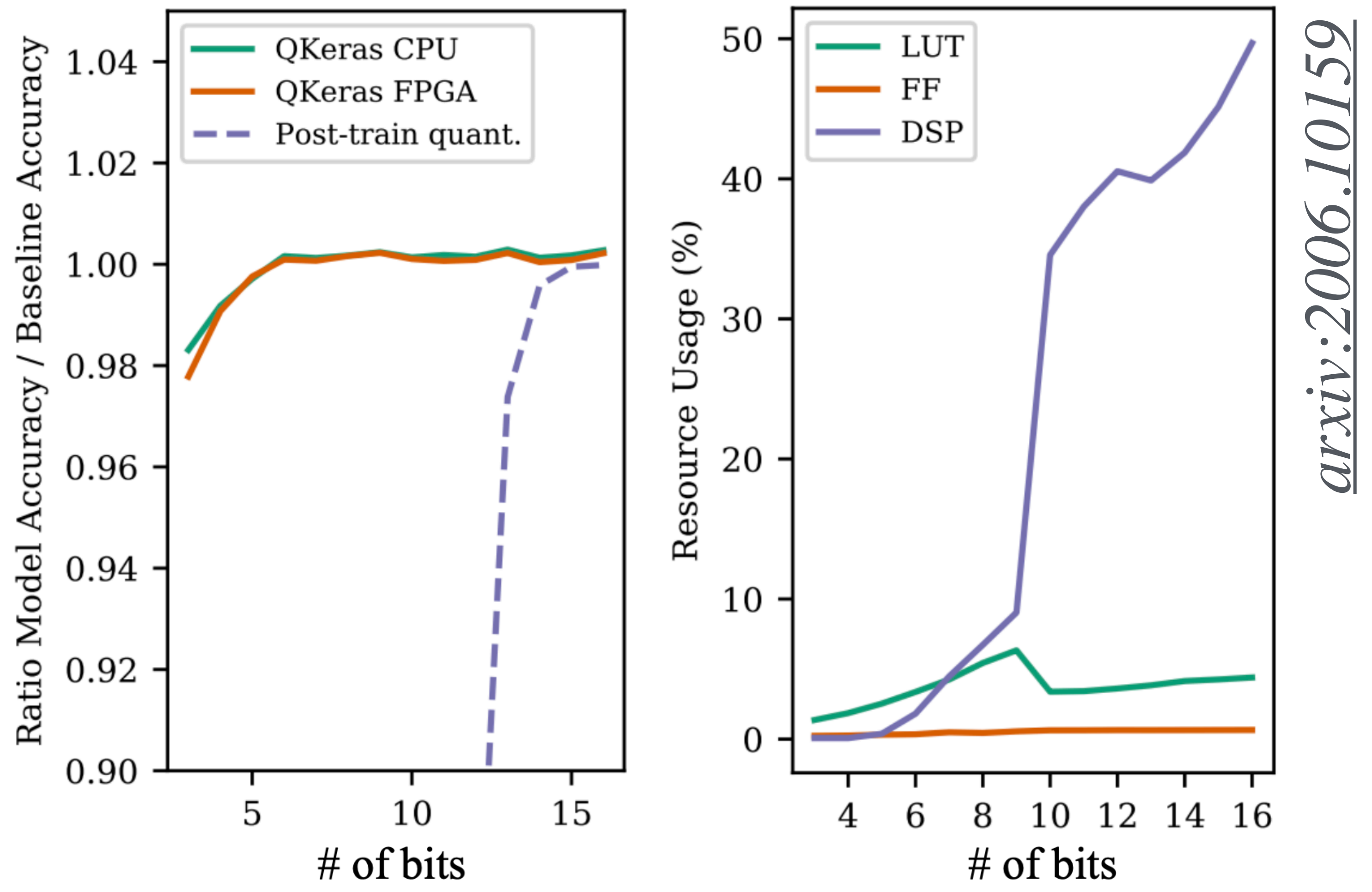
- **Two main approaches to cope with online devices requirements while maintaining a satisfactory level of performances:**

Pruning



- Reduce number of "nodes" and/or "links", typically by setting small weights to zero
- Multiplications by 0 can be completely removed from FPGA design

Quantization

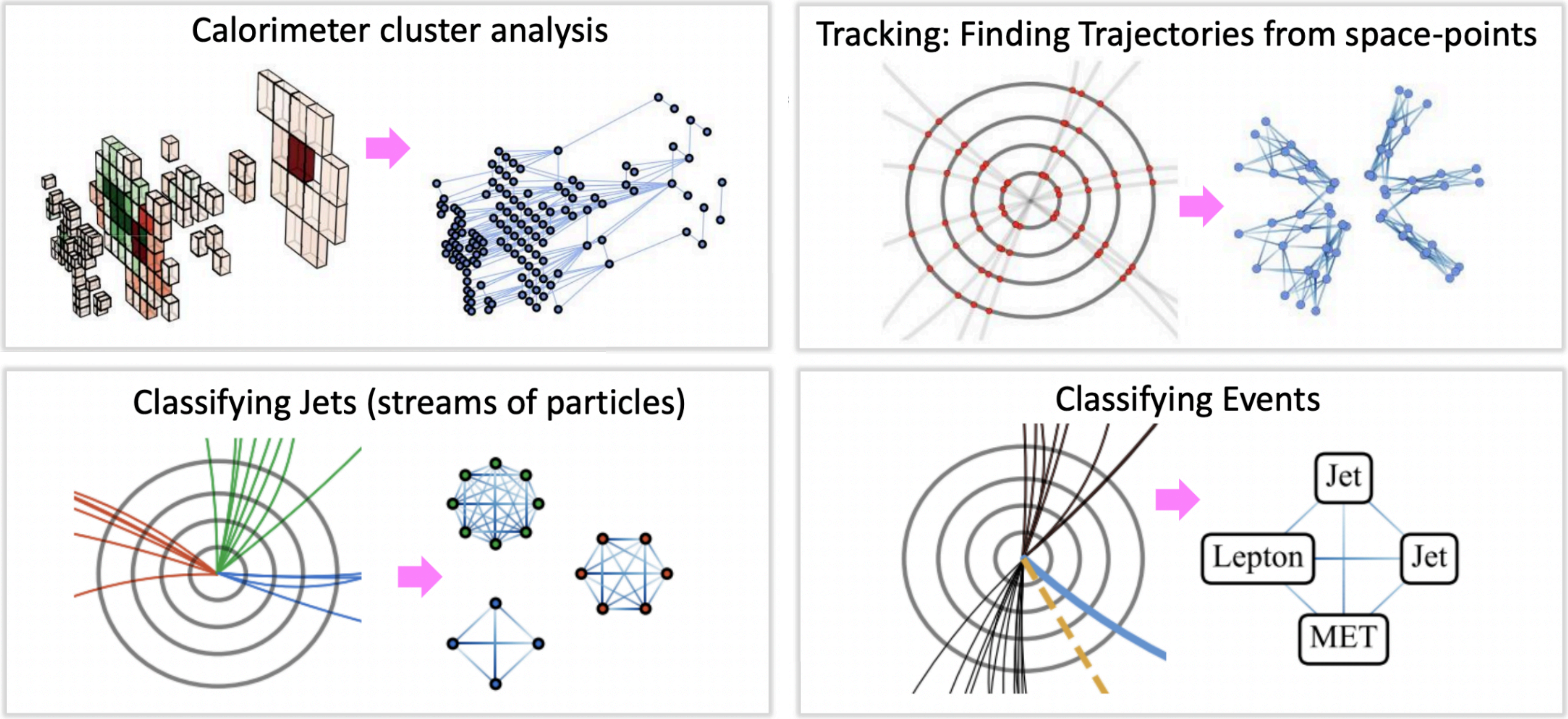


- Generally, performing quantisation-aware training achieves better performance
- Particularly well suited for FPGA (large gain in the multiplier units)

Online ML @ LHC

A selection of ML applications, in operation or in development, for online reconstruction

(very much non exhaustive!)

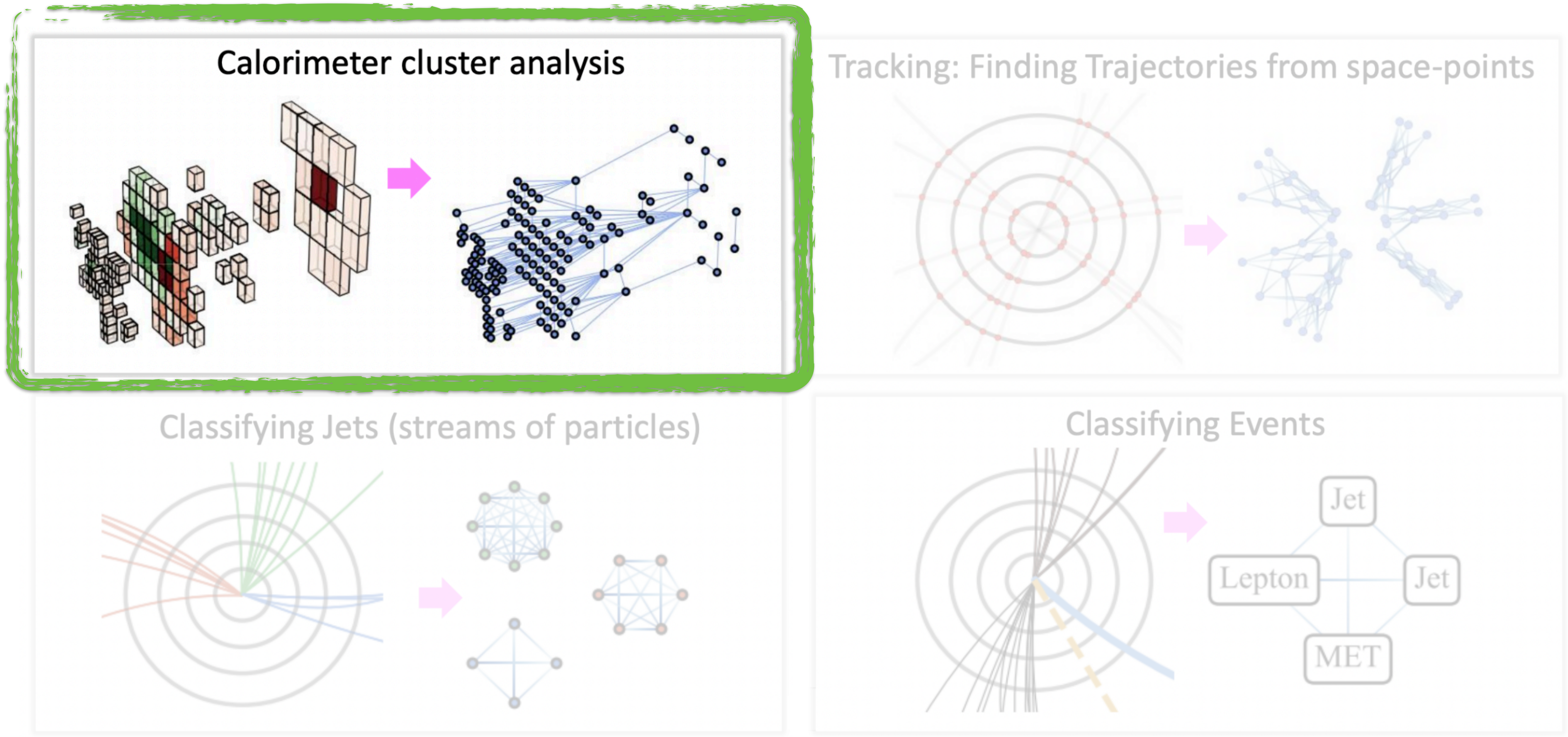


From D. Rankin (FastML for Science Conference 2024)

Online ML @ LHC

A selection of ML applications, in operation or in development, for online reconstruction

(very much non exhaustive!)

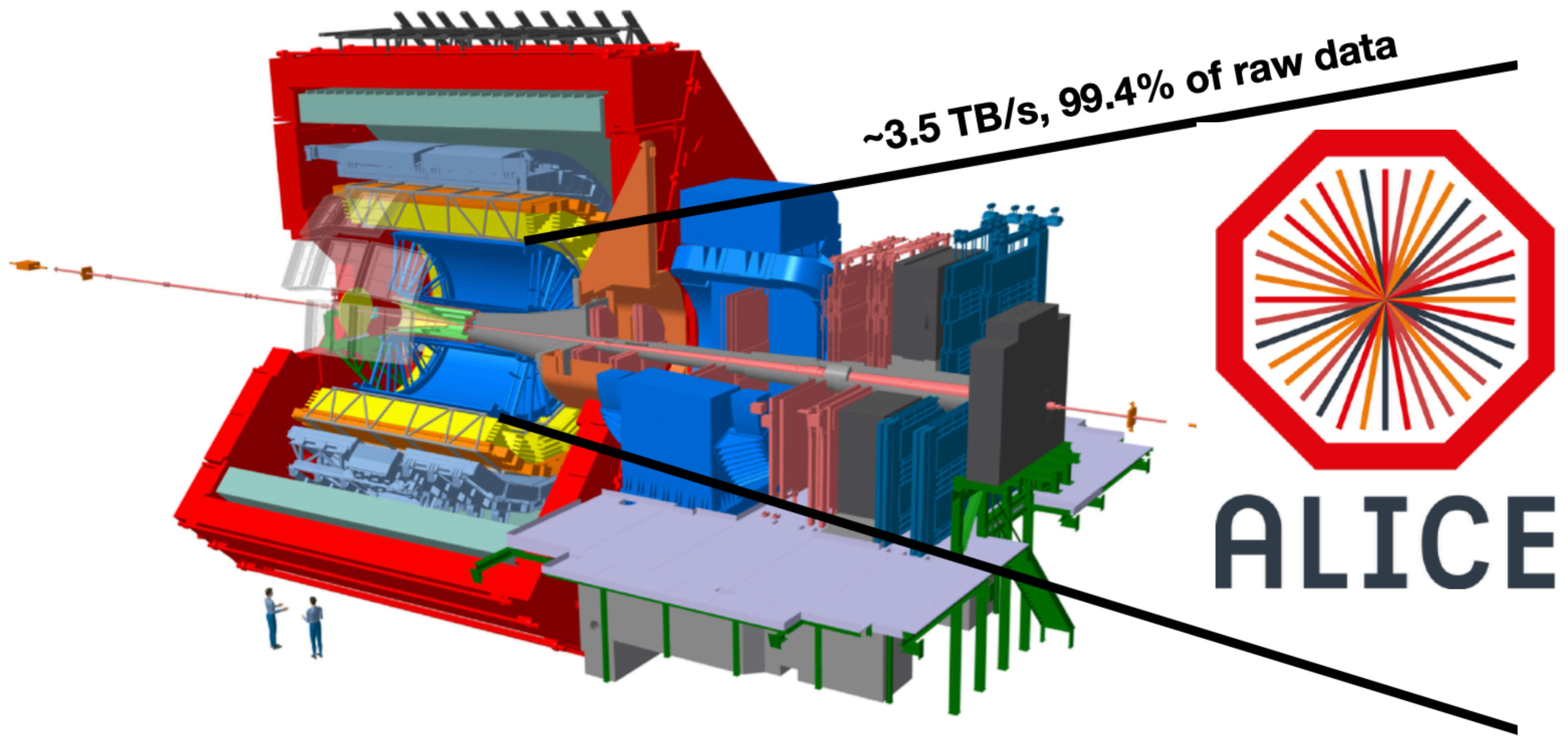


From D. Rankin (FastML for Science Conference 2024)

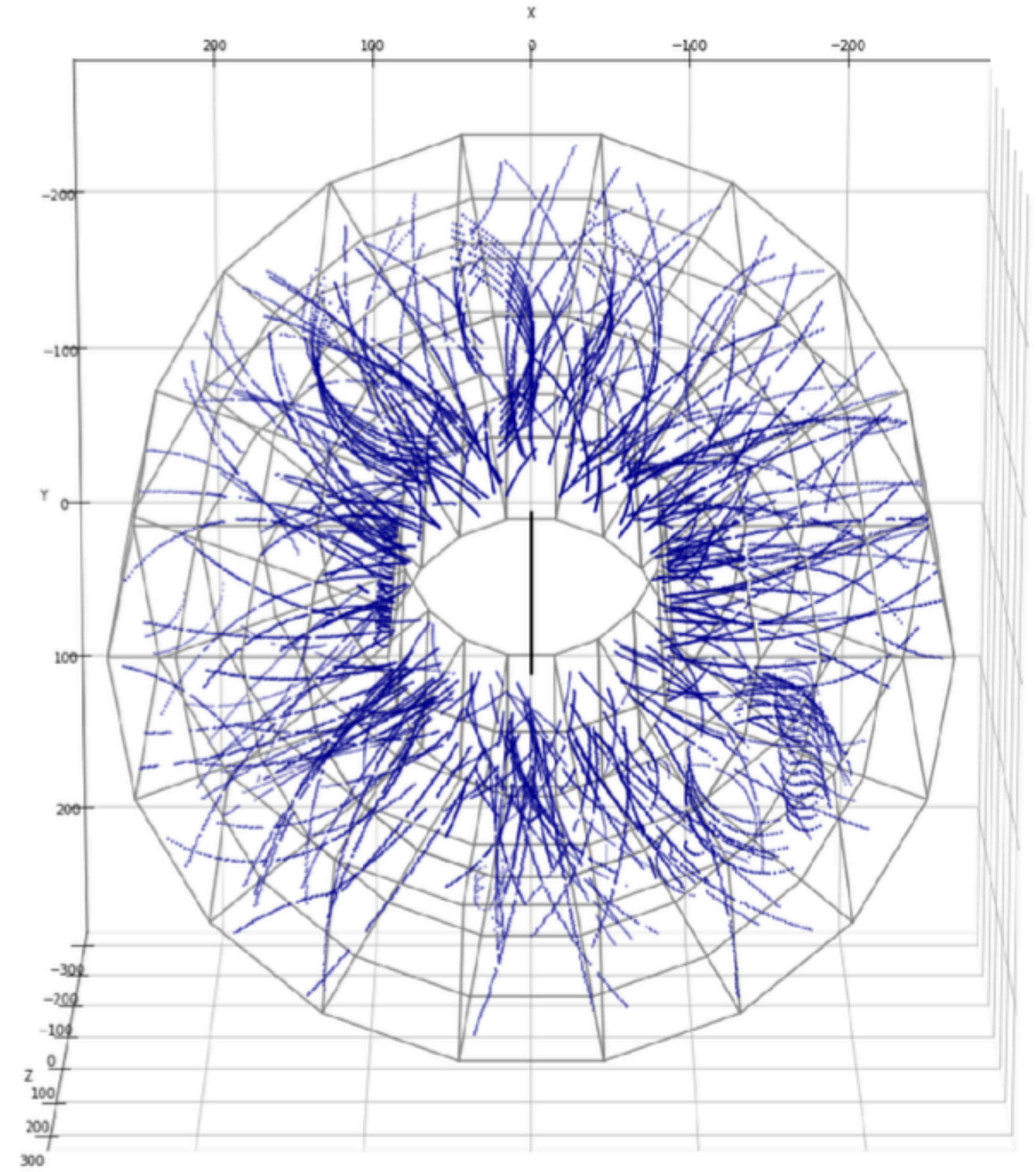
DNN for online clusterisation

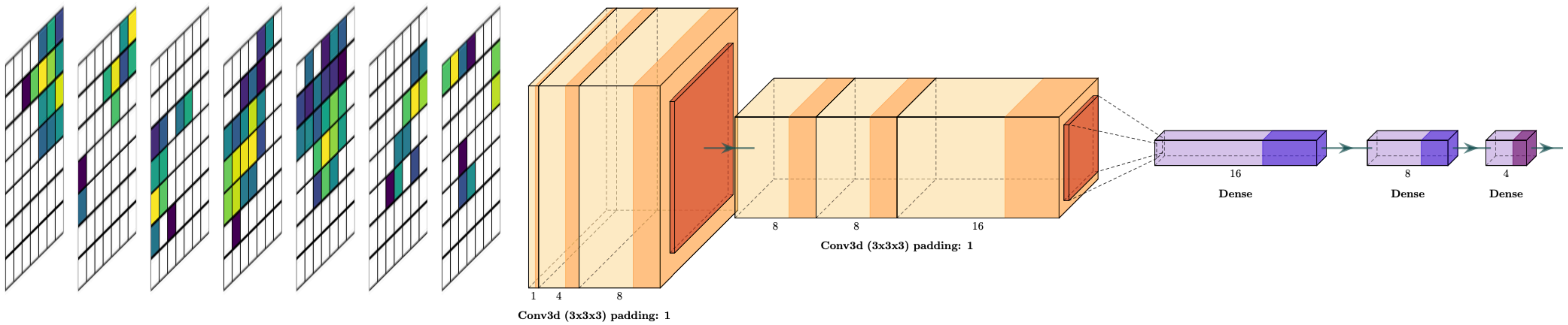
Large amount of data to store

- **Major upgrade during LS2**
  ➞ In Run3 reading full detector @ 50kHz in Pb-Pb collisions with ~3.5 TB/s mostly from TPC (~99% of raw data)
- **Tracking from clusters (number of clusters ≈ stored data size)**
  ➞ reducing the amount of saved clusters while maintaining good tracking performances?
- **Promising studies for cluster classification using DNN online (GPU farm)**
  ➞ different architectures studied (fully connected, 2D or 3D CNN)



The ALICE TPC

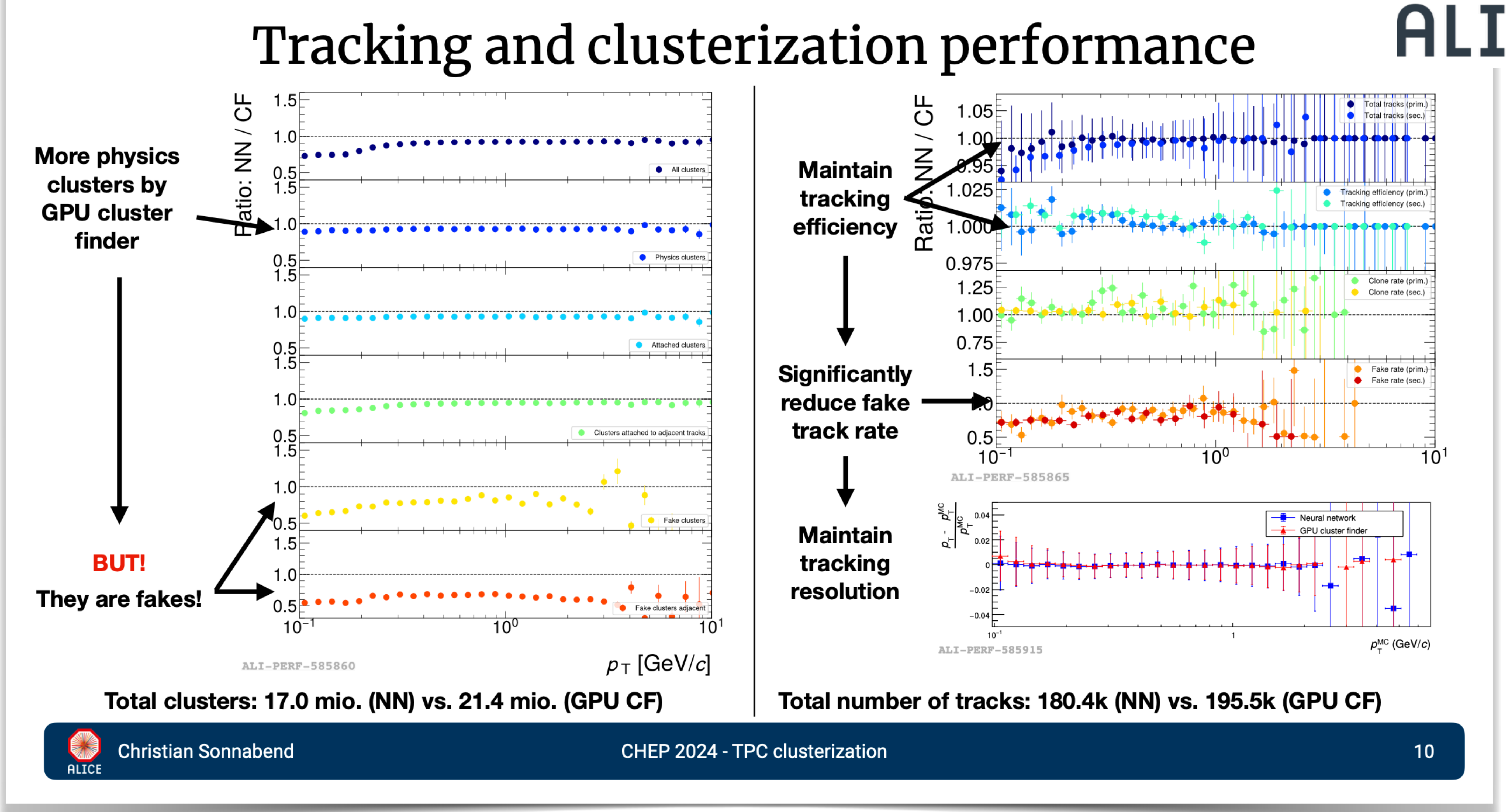




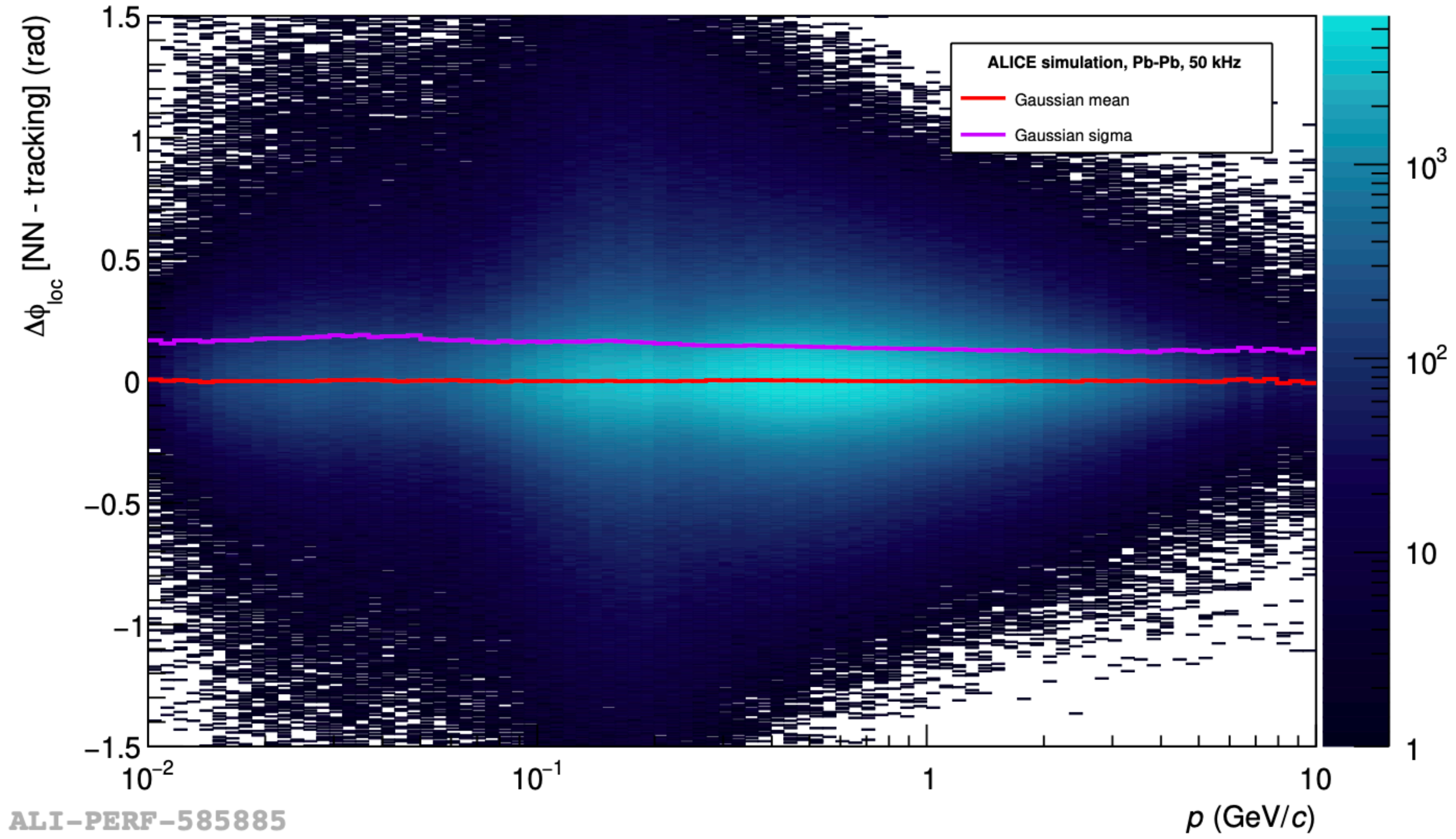
DNN for online clusterisation

Large amount of data to store

C. Sonnabend's @ CHEP24



⊕



NN provides good estimate of local momentum vector, useful for track seeding

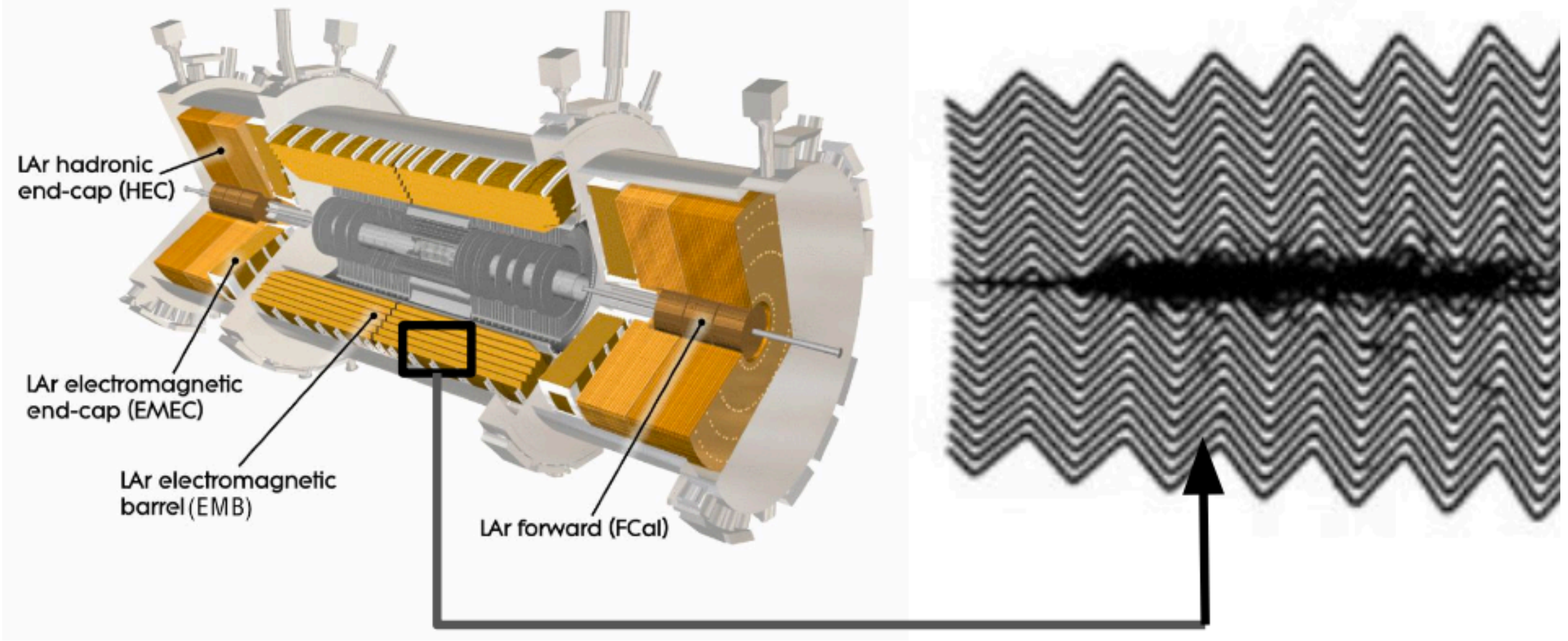
Successfully rejects clusters that are not used in tracking

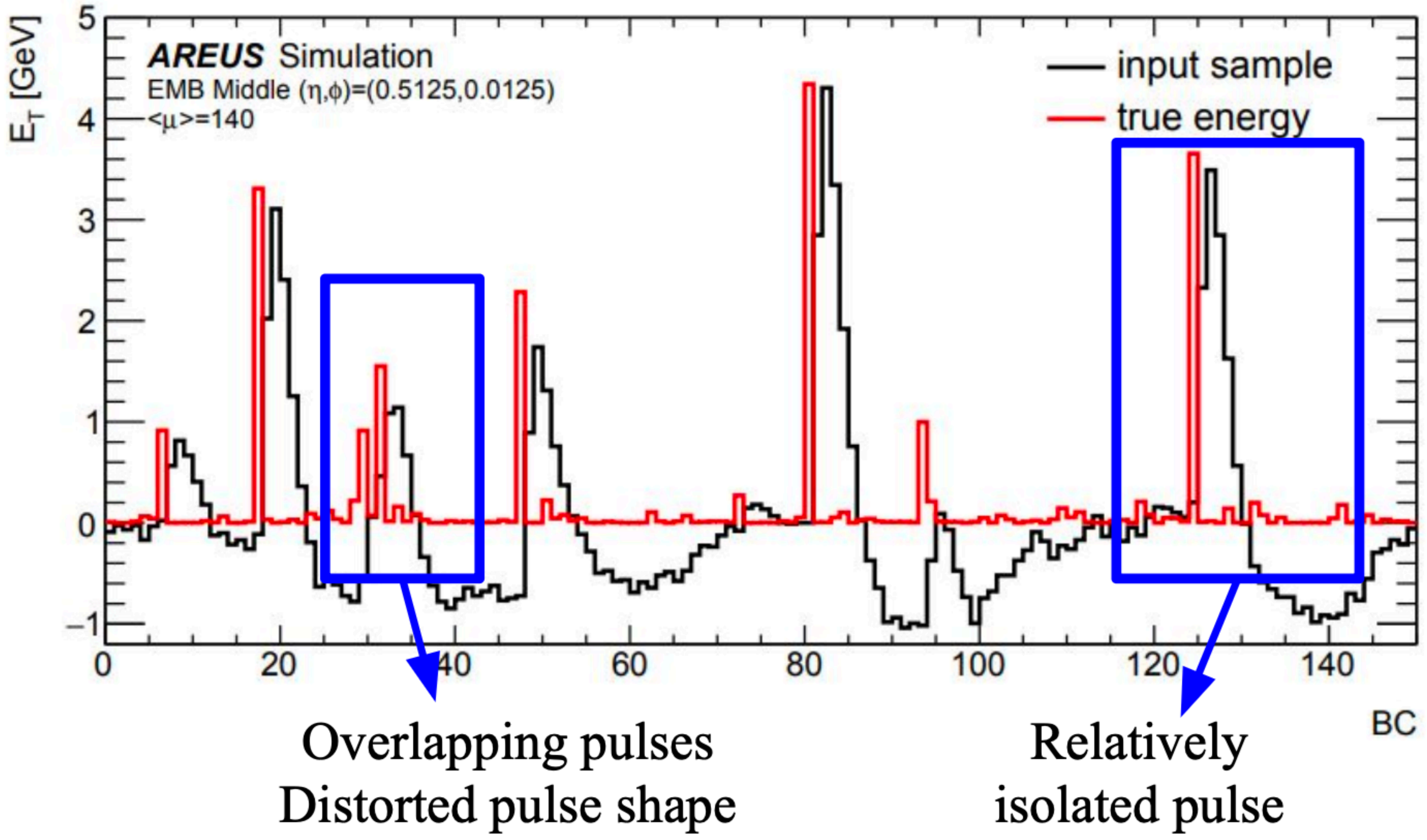
→ potential reduction of effective data-size by ~20% while maintaining / or even improving tracking performance!

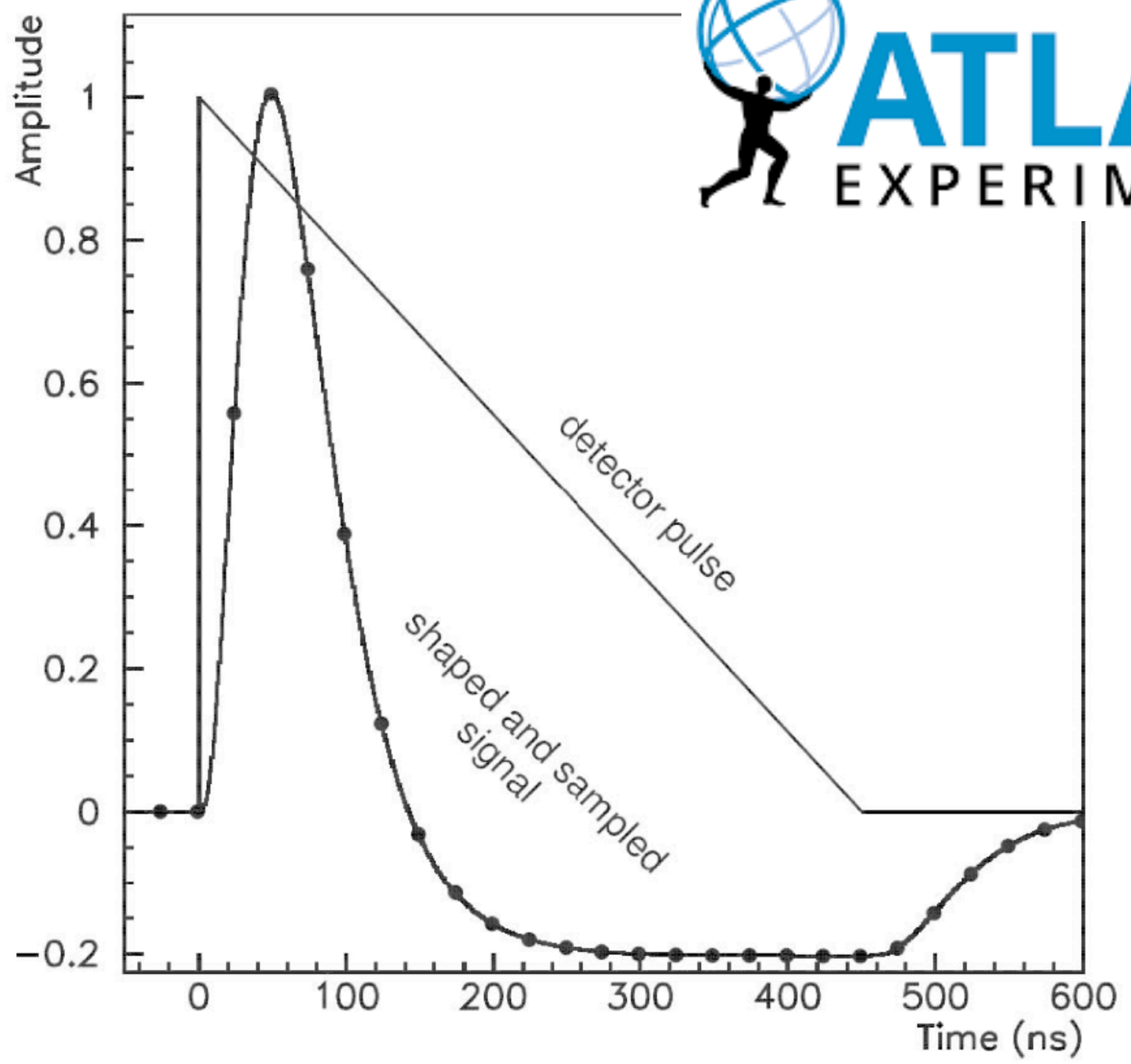
DNN in FPGA

For ATLAS calorimeter in HL-LHC

- **Liquid argon (LAr) calorimeter readout electronics replaced** → FPGAs to compute the energy deposited in the calo
- **Overlapping pulses difficult for heuristic algorithm due to distorted pulse shapes** → Can a NN running in the FPGA improve performances ?





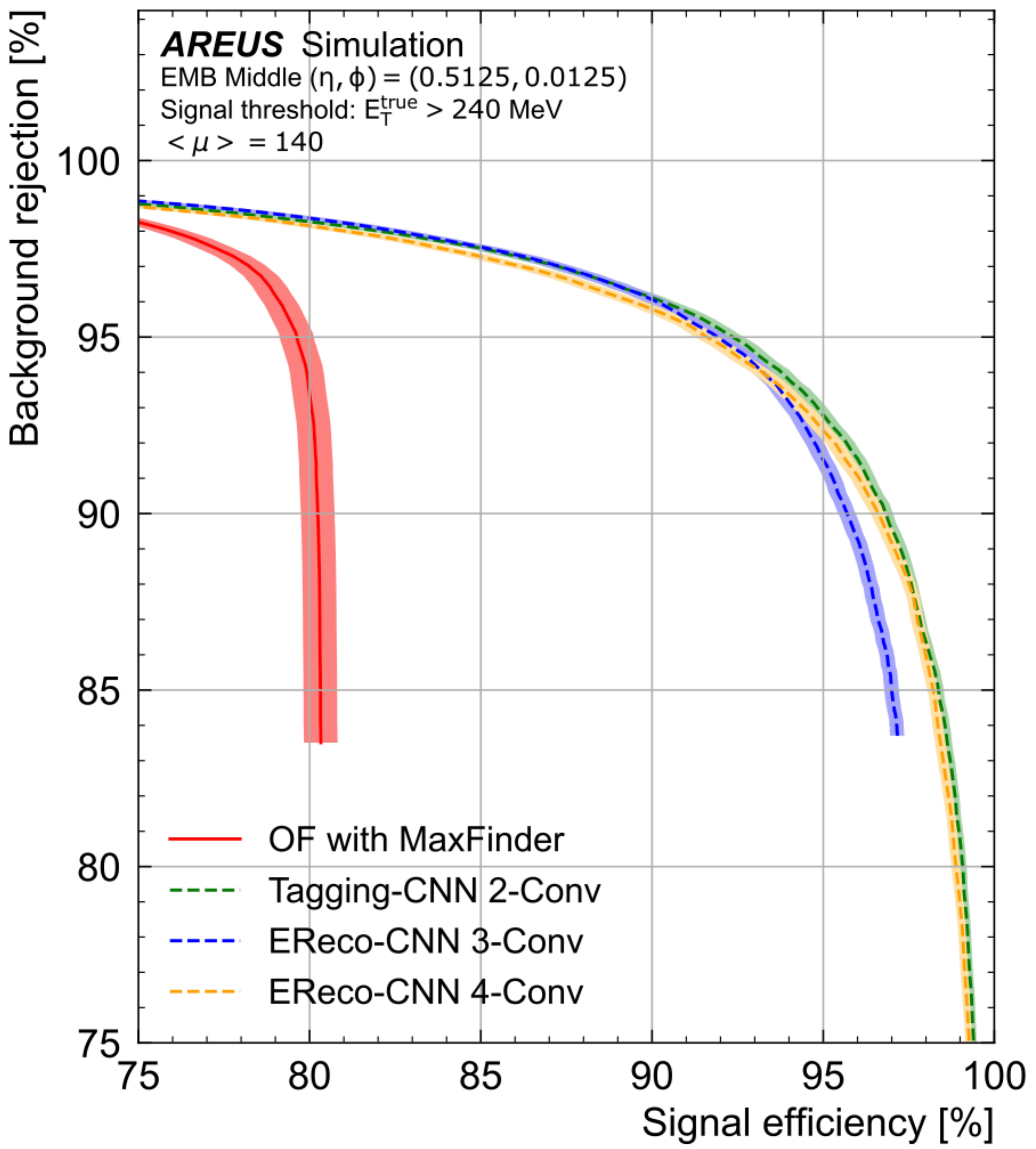


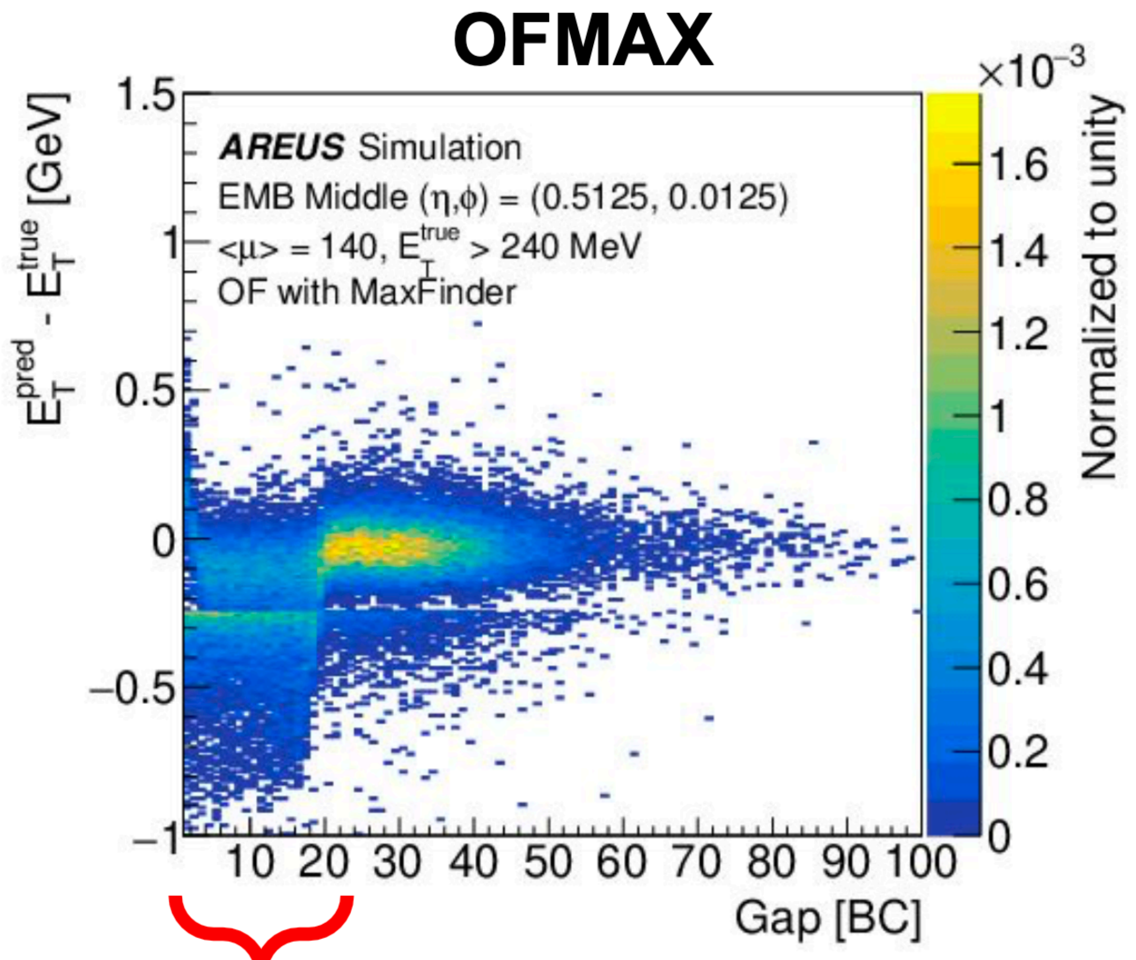
Comput.Softw.Big Sci. 5 (2021) 1, 19

DNN in FPGA

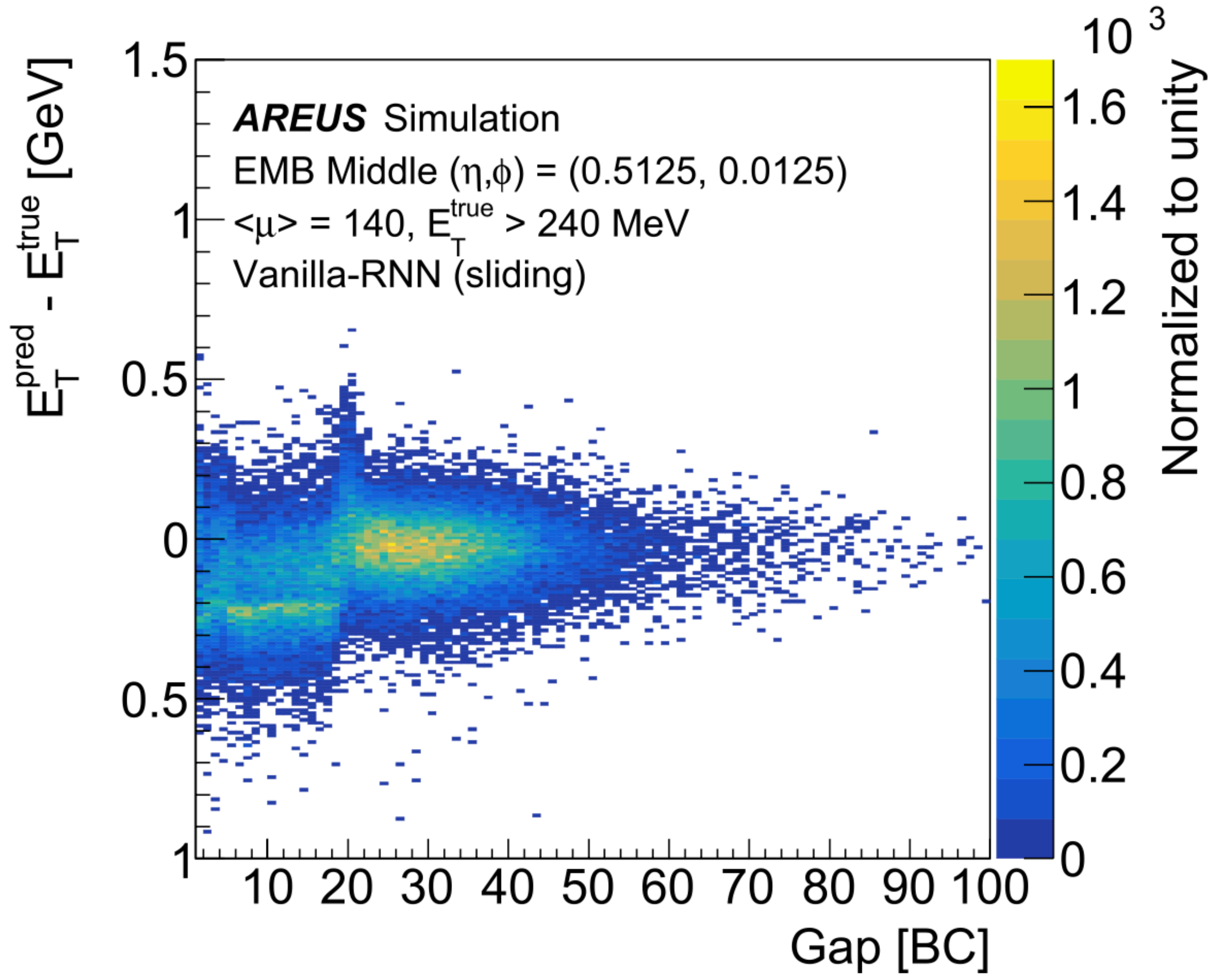
For ATLAS calorimeter in HL-LHC

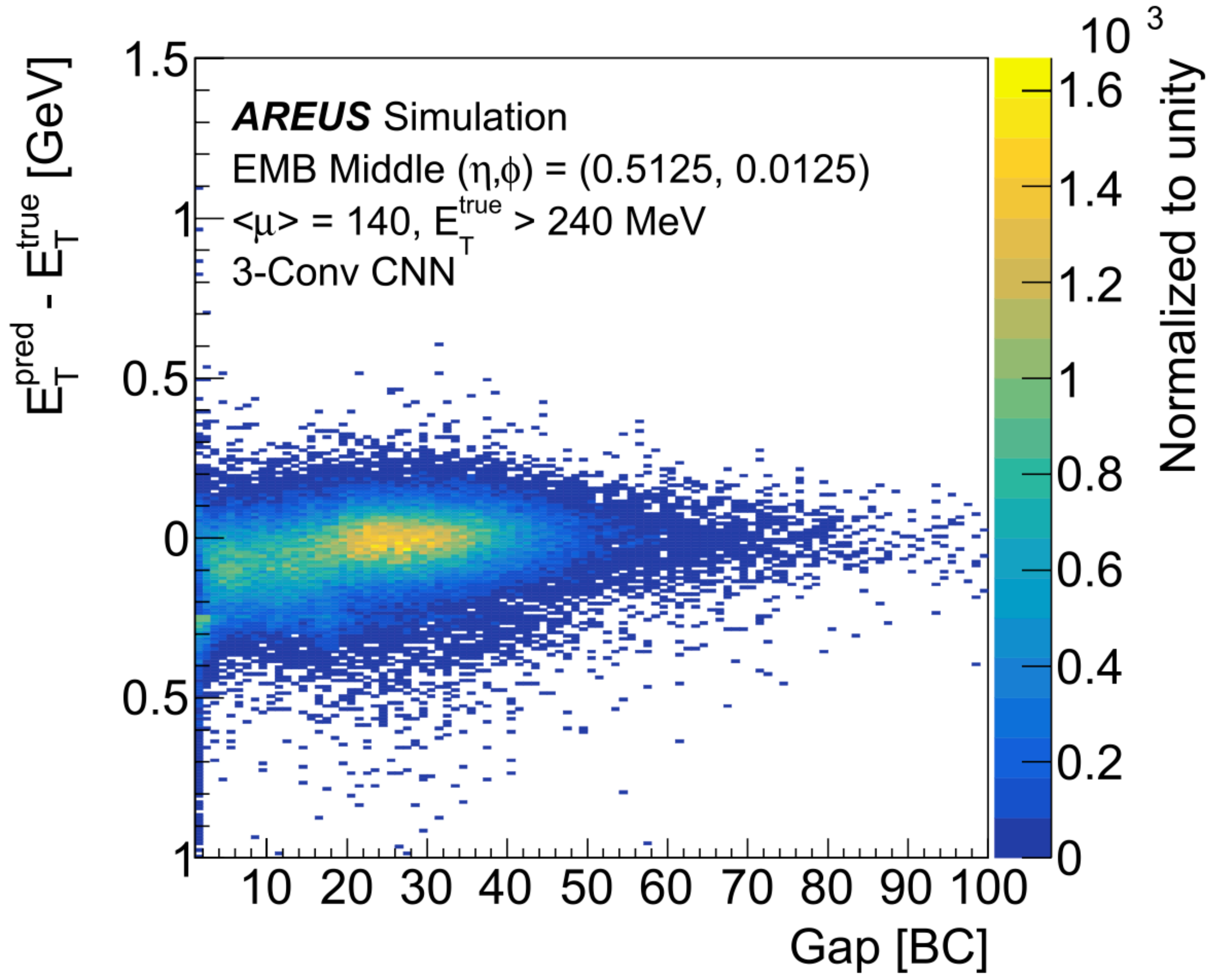
- **Signal efficiency & resolution improved when using DNN** (further improvements needed to match FPGA requirements)





Overlapping signals region





Comput.Softw.Big Sci. 5 (2021) 1, 19

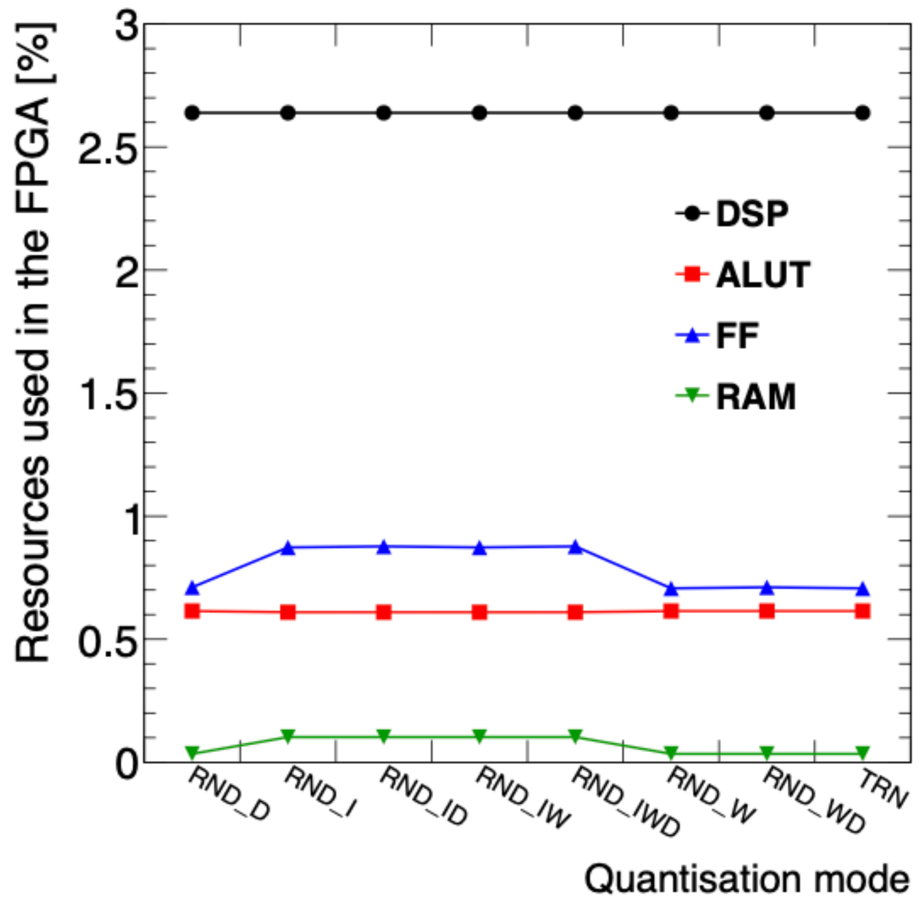
DNN in FPGA

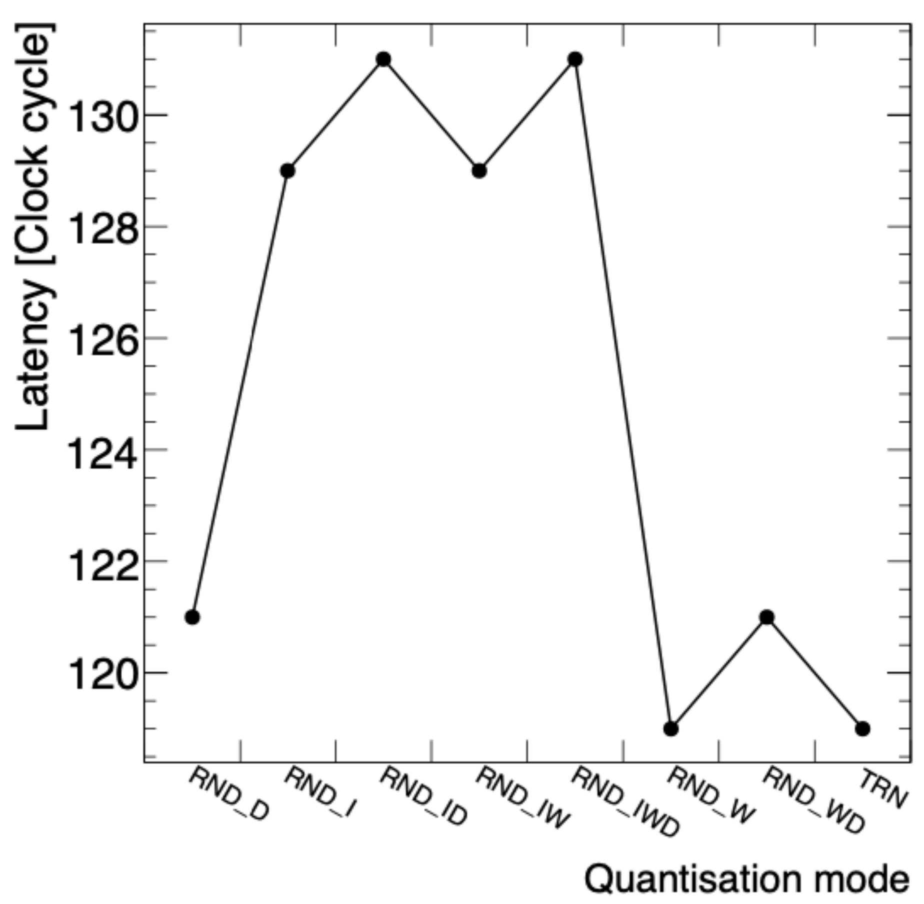
For ATLAS calorimeter in HL-LHC

Recent work on RNN optimisation for inference on Stratix10 from Intel

- **Detailed quantisation studies**
  truncation(TRN) vs rounding (RND) at different steps
- **Using multiplexing**
  384 channels/FPGA → 28 instances of the RNN @ 560 MHz with a multiplexing of 14 and a latency of 65 clock cycles (116ns)
- **Common High level synthesis (HLS) language not sufficient to meet FPGA requirements**
  fine optimisation possible with Very High-Speed Integrated Circuit Hardware Description Language (VHDL)

G. Aad et al 2023 JINST 18 P05017

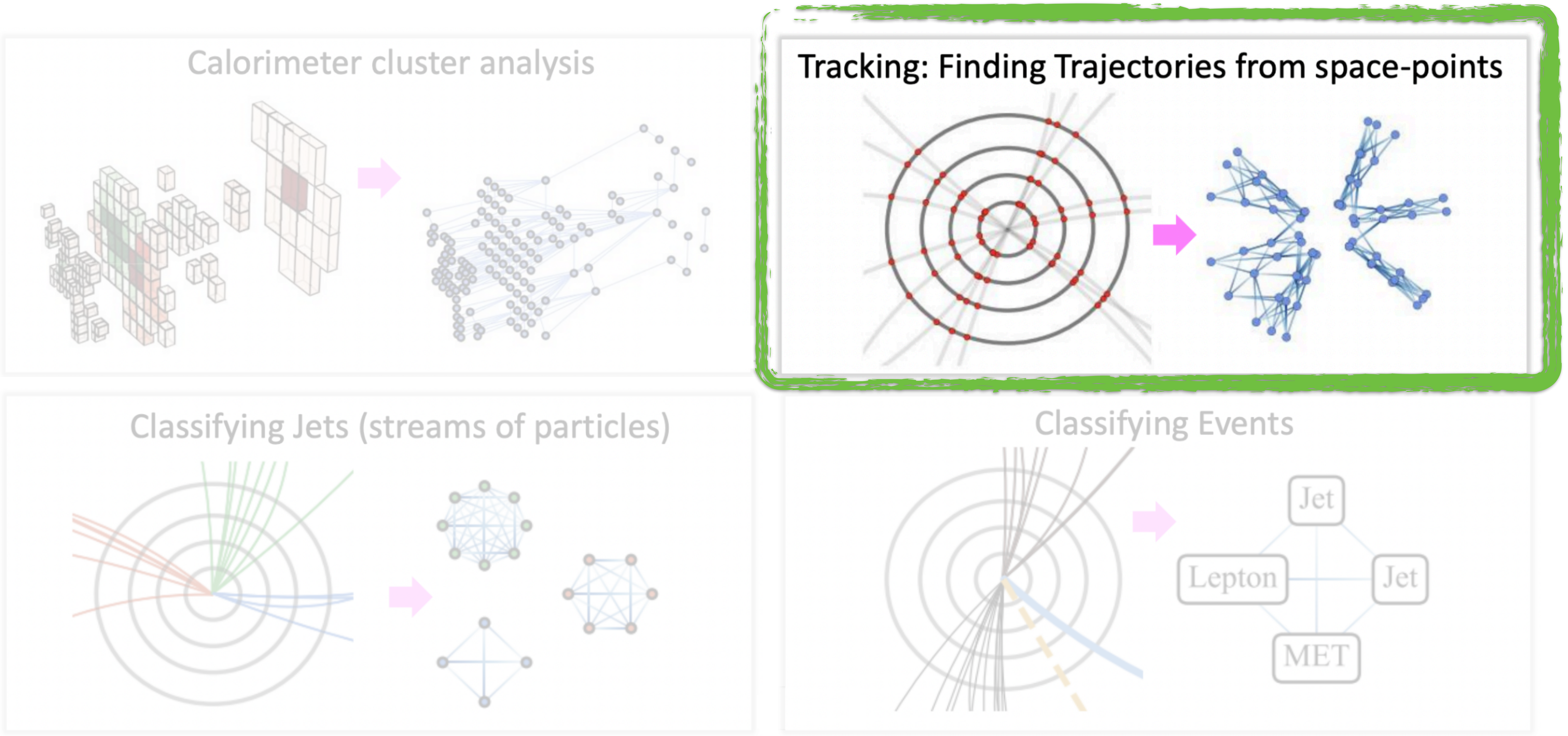




	N networks x multiplexing	ALM	DSP	FMax	latency
Target	**384 channels**	**30%***	**70%***	**Multiplexing x 40 MHz**	**125 ns**
"Naive" HLS	384x1	226%	529%	-	322 ns
HLS optimized	37x10	90%	100%	393 MHz	277 ns
VHDL optimized	28x14	18%	66%	561 MHz	116 ns

A selection of ML applications, in operation or in development, for online reconstruction

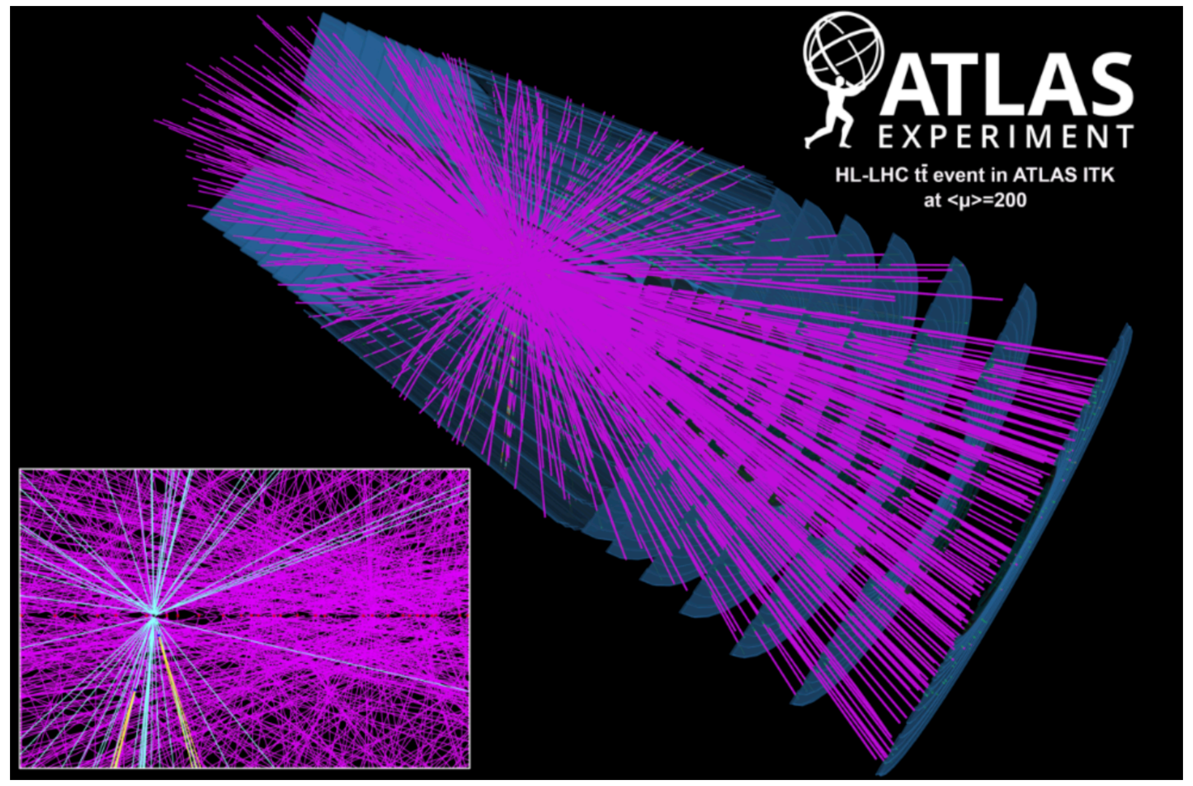
(very much non exhaustive!)



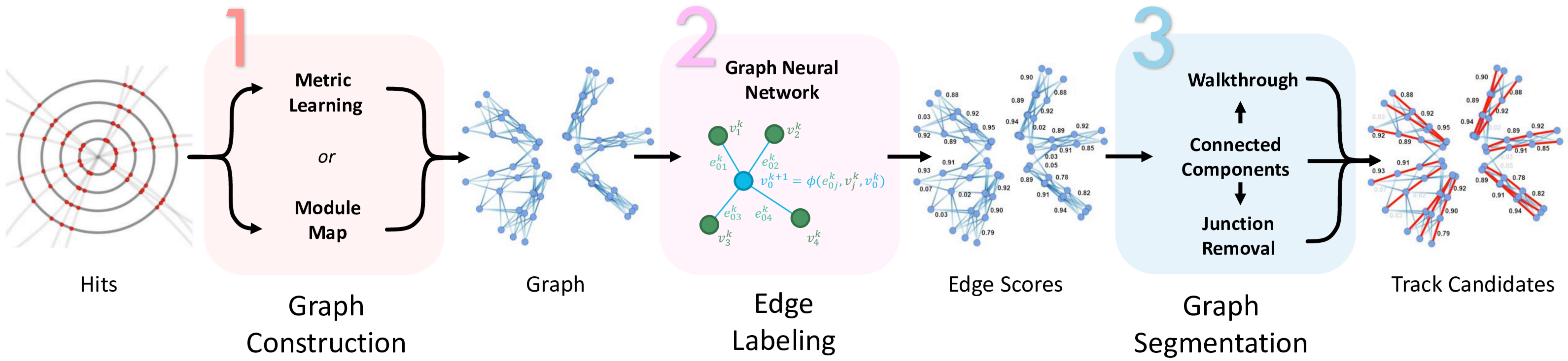
From D. Rankin (FastML for Science Conference 2024)

From hits to tracks

- **For ATLAS ITk in HL-LHC:**
  <μ> ~ 200 ➞ ~ 300k hits/evt



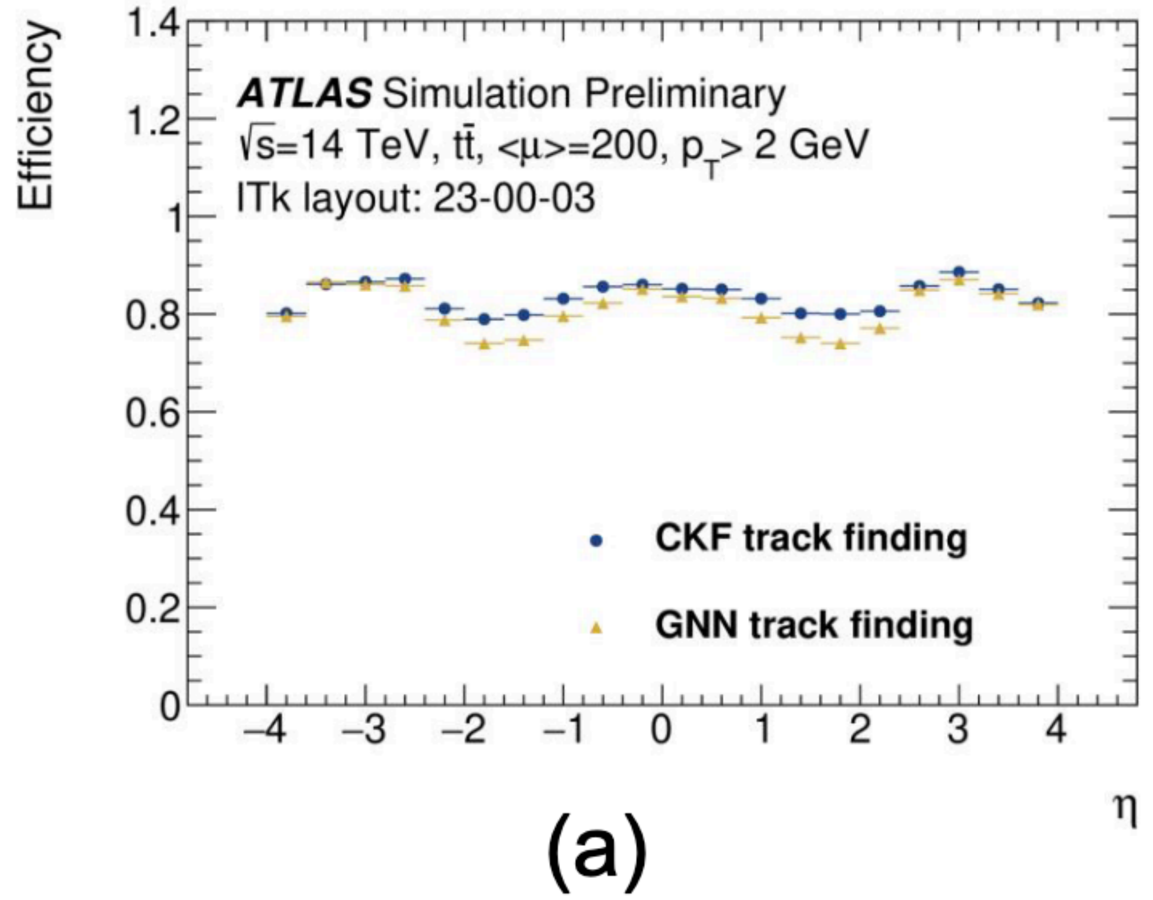
  - **Perform tracking using GNN model**

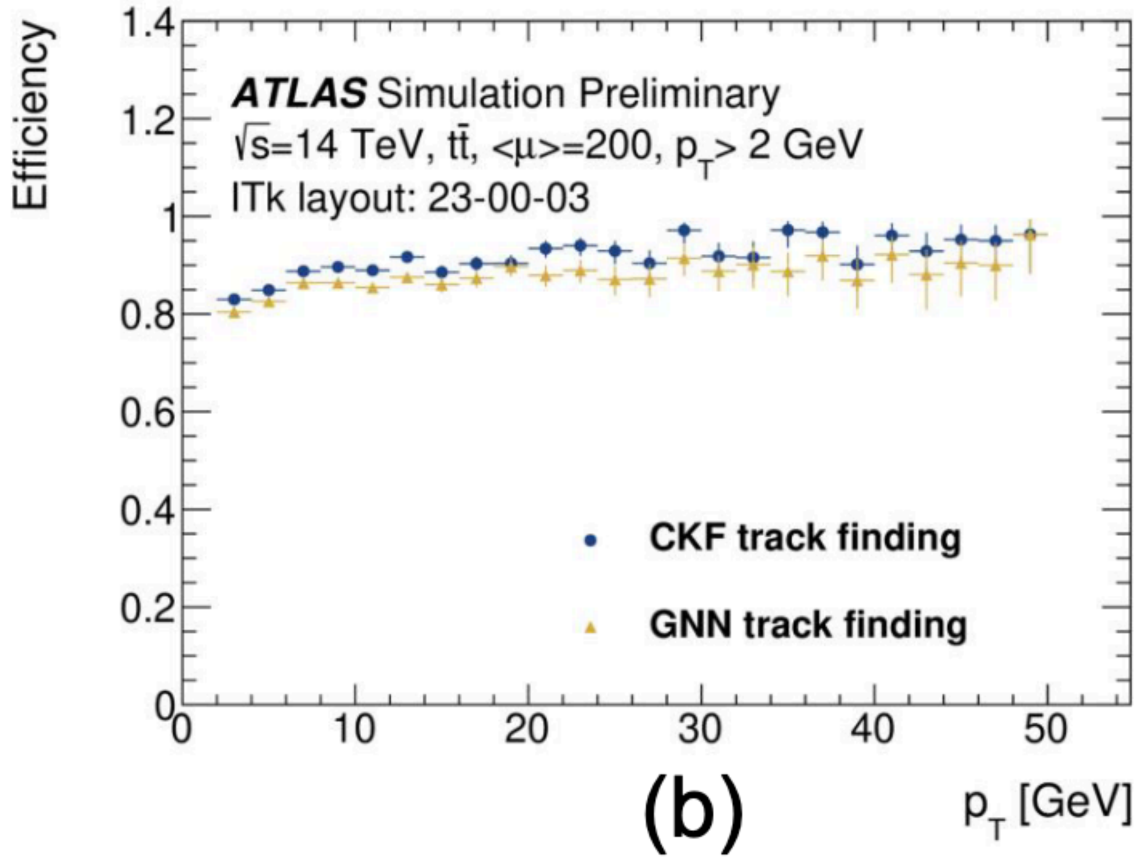


  - **Perform tracking using GNN model** was demonstrated to achieve similar performances than heuristic algorithm

  ATL-SOFT-PROC-2023-047

  J. Stark @ EuCAIFCon24





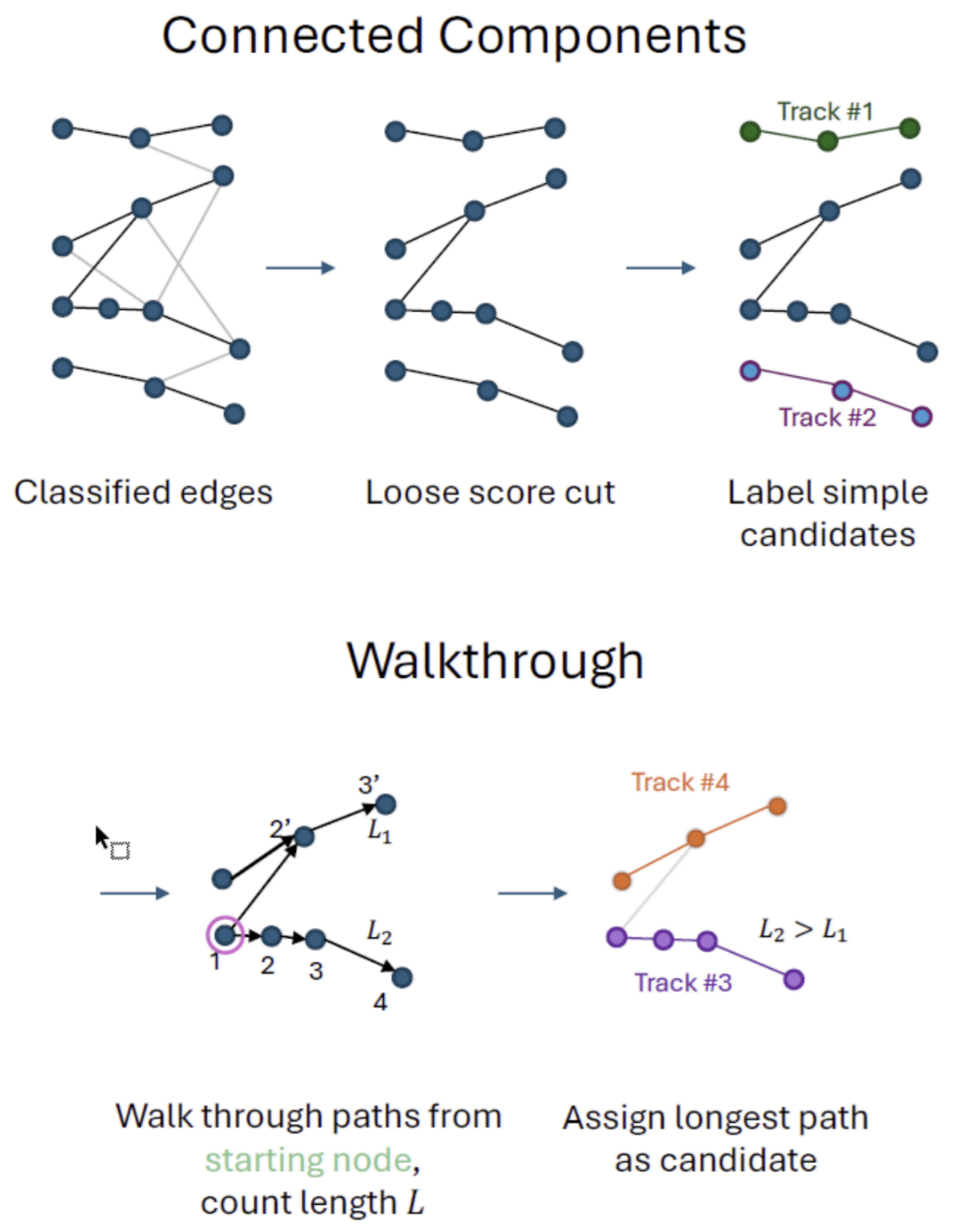
From hits to tracks

- **For ATLAS ITk in HL-LHC: GNN inference in GPU/CPU**
  - **<μ> ~ 200 → ~ 300k hits/evt → fully connected graph ~ O(10^{11}) edges**
  - **Recent optimisations** *A. Lazar @ CHEP24* *ATL-PHYS-PUB-2024-018*

Stage	Pipeline	
	Metric Learning (ms)	Module Map (ms)
1. Graph Construction	505	69
2. Edge Classification	108	323
3. Graph Segmentation	118	118
Sum	**731**	**510**

i.e. Track building

Stage	Efficiency (Relative Difference, %)	Running Time (ms)
CTD23 Walkthrough	—	42,000
FastWalkthrough	+0.53	120
CC	−1.33	6.0
CC+JR	+0.93	40

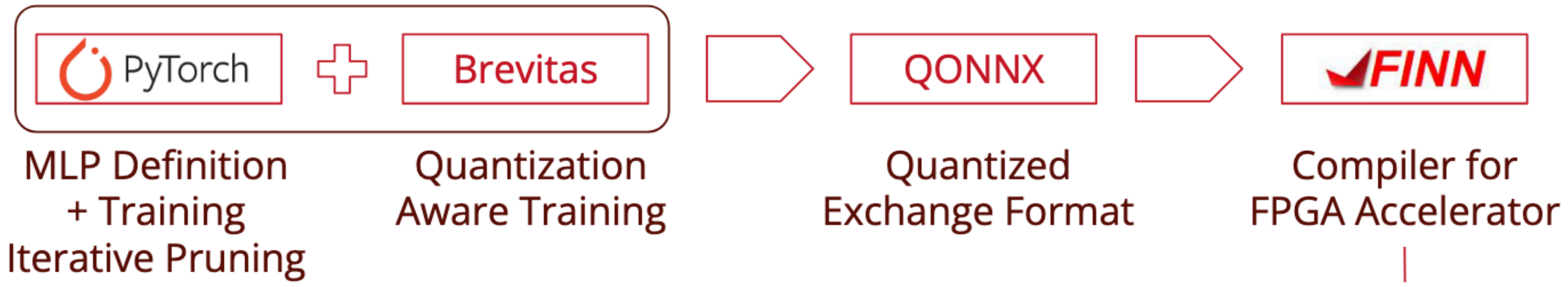


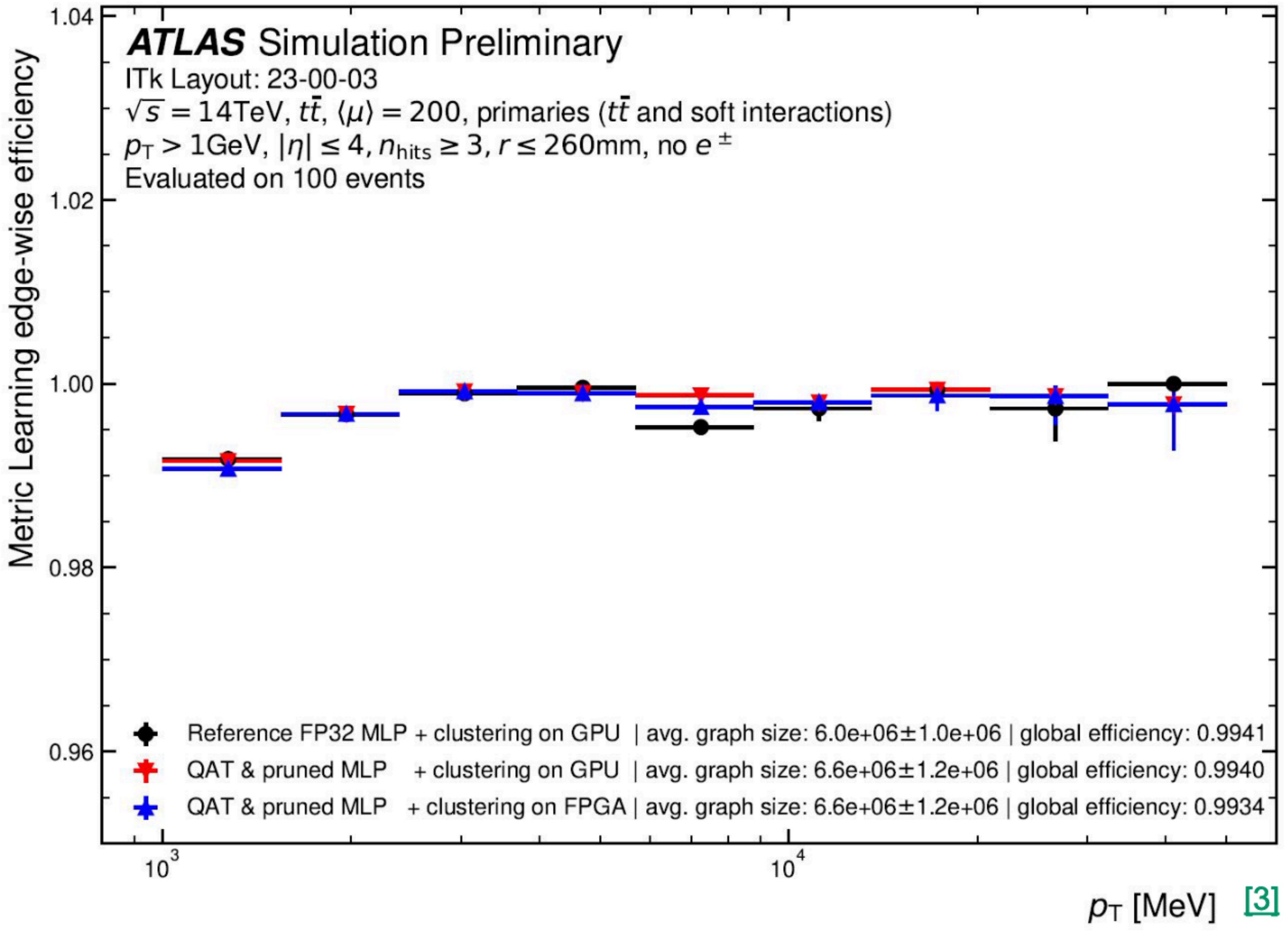
From hits to tracks

For ATLAS ITk in HL-LHC: GNN inference in FPGA

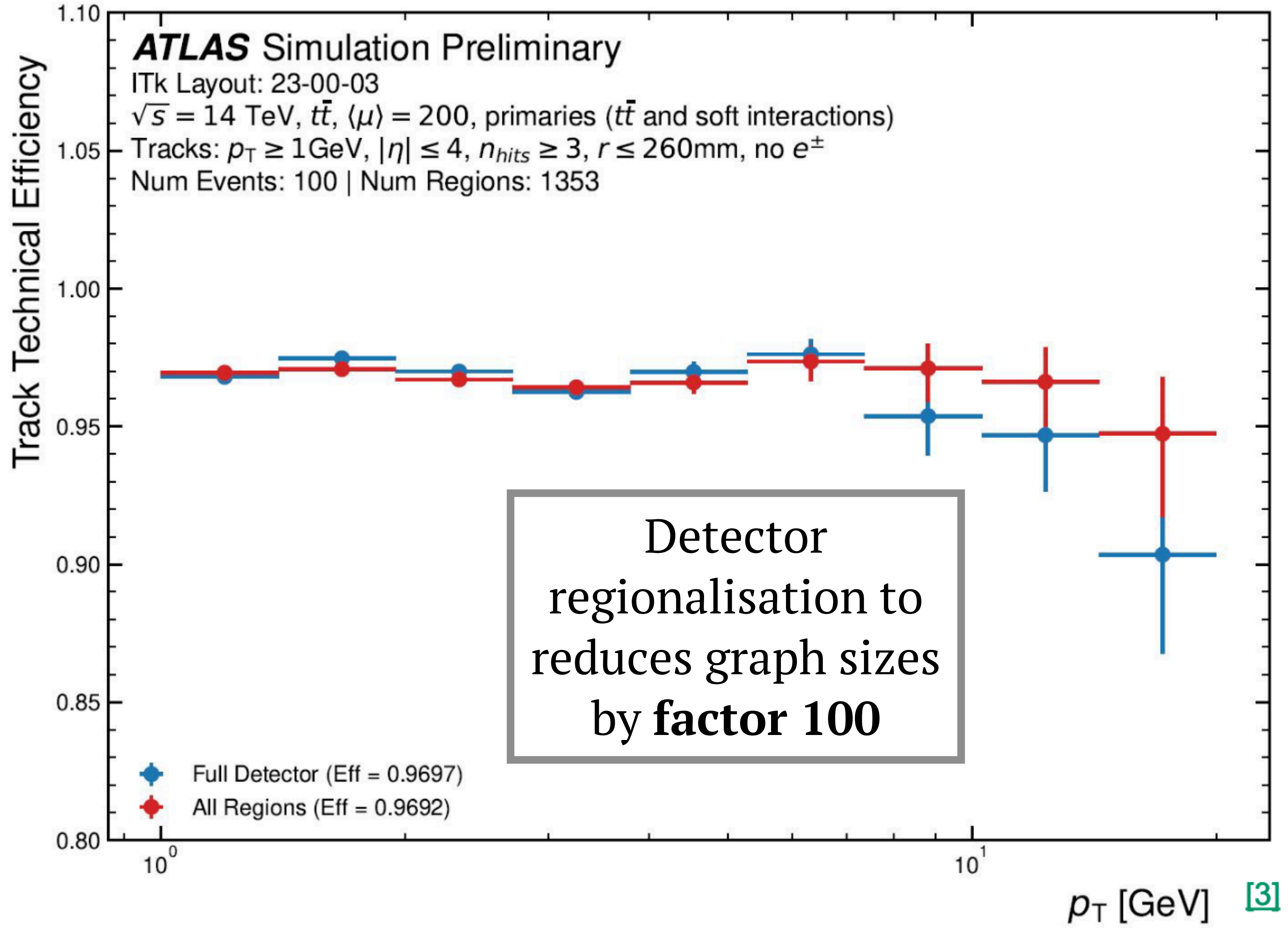
S. Dittmeier @ CHEP24

- Benefit from potential heterogeneous online computing farm for the ATLAS Event Filter at HL-LHC
- Preliminary standalone implementations of FPGA algorithms for graph construction and segmentation





[3]



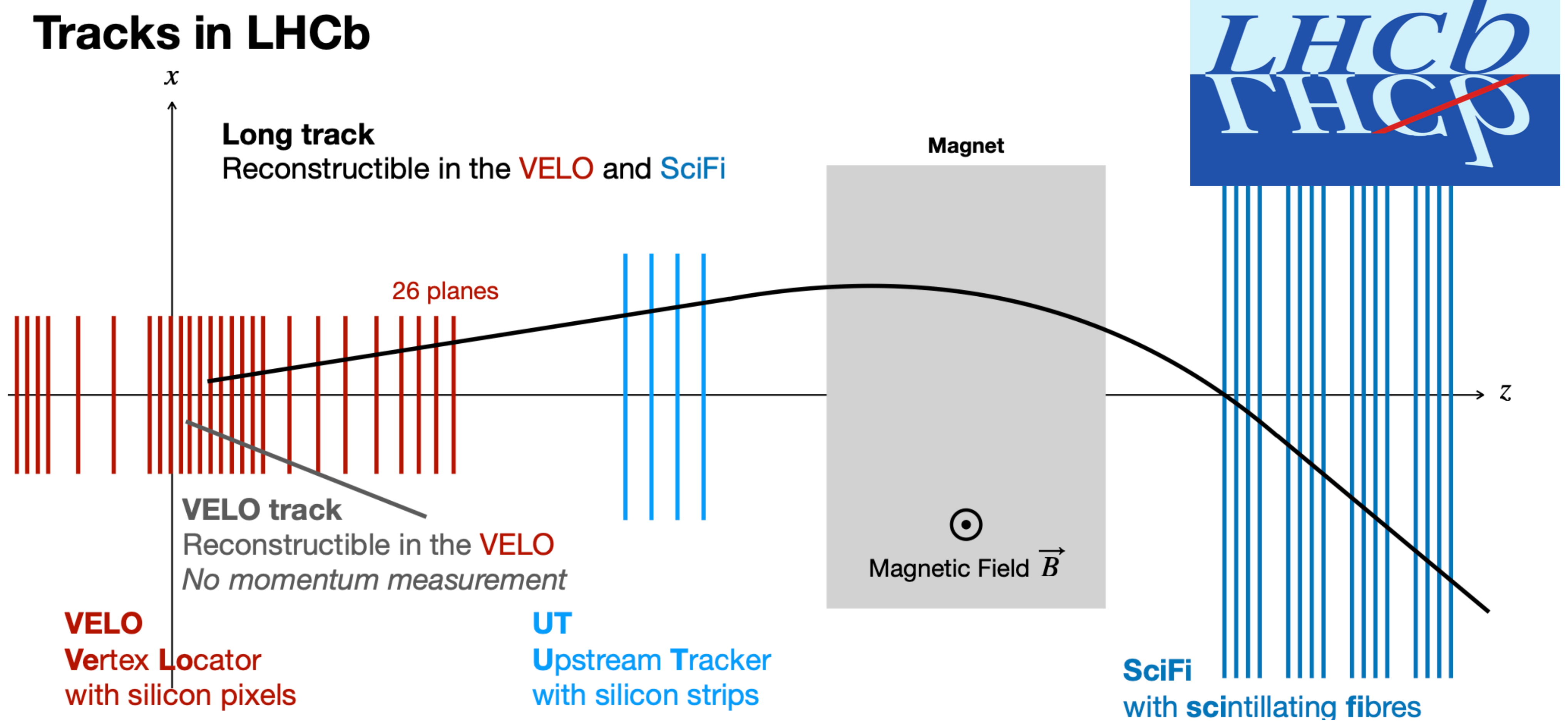
[3]

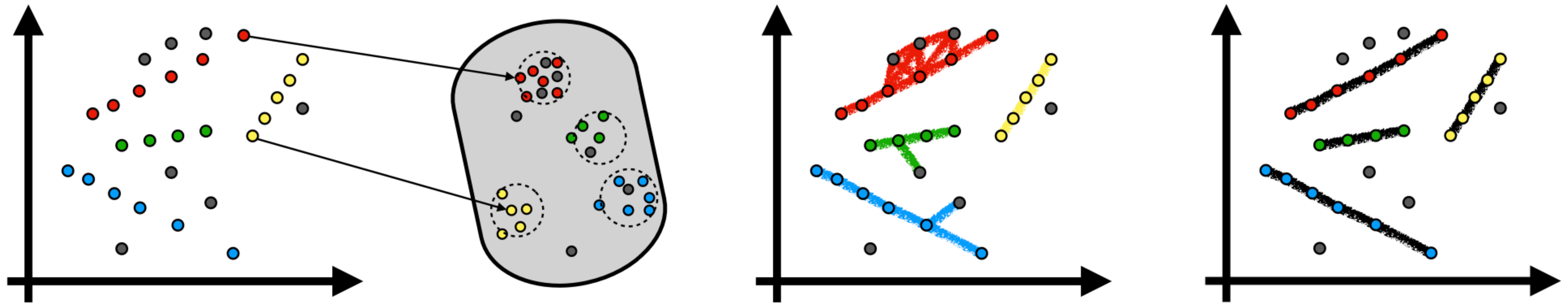
Detector regionalisation to reduces graph sizes by **factor 100**

From hits to tracks

- **Similar approach studied in LHCb:**
  - **Since Run3, LHCb benefits from full software trigger** performing partial event reconstruction & coarse selection and running on a farm of 500 GPUs NVIDIA RTX A5000
  - **GNN-based tracking** (**ETX4VELO**) has been demonstrated to **outperform heuristic algorithm** physics performance in a low-p_T environment with special care for electrons (challenging due to material interaction)

A. Correia @ CTD23 *F. Giasemis @ ICHEP24*



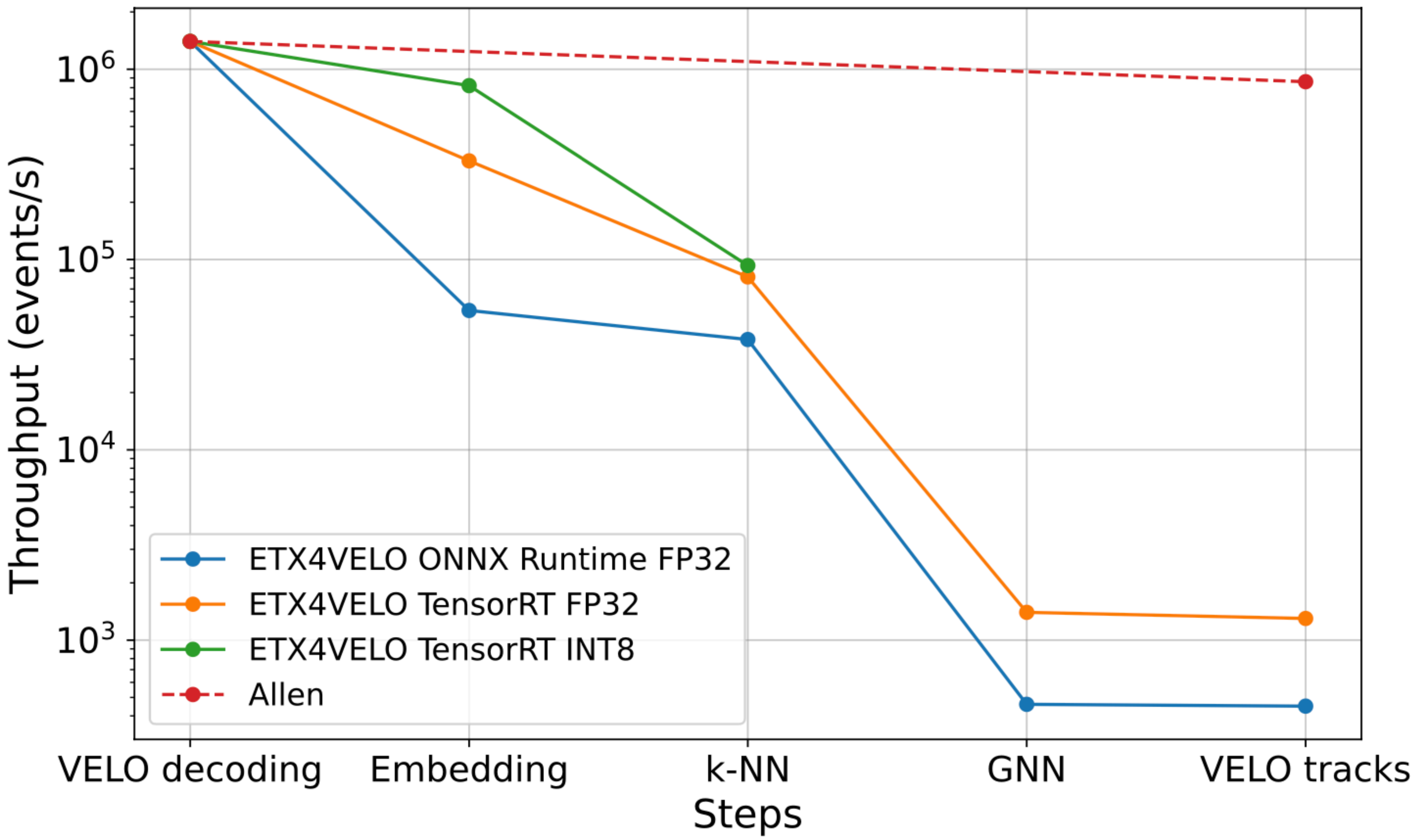


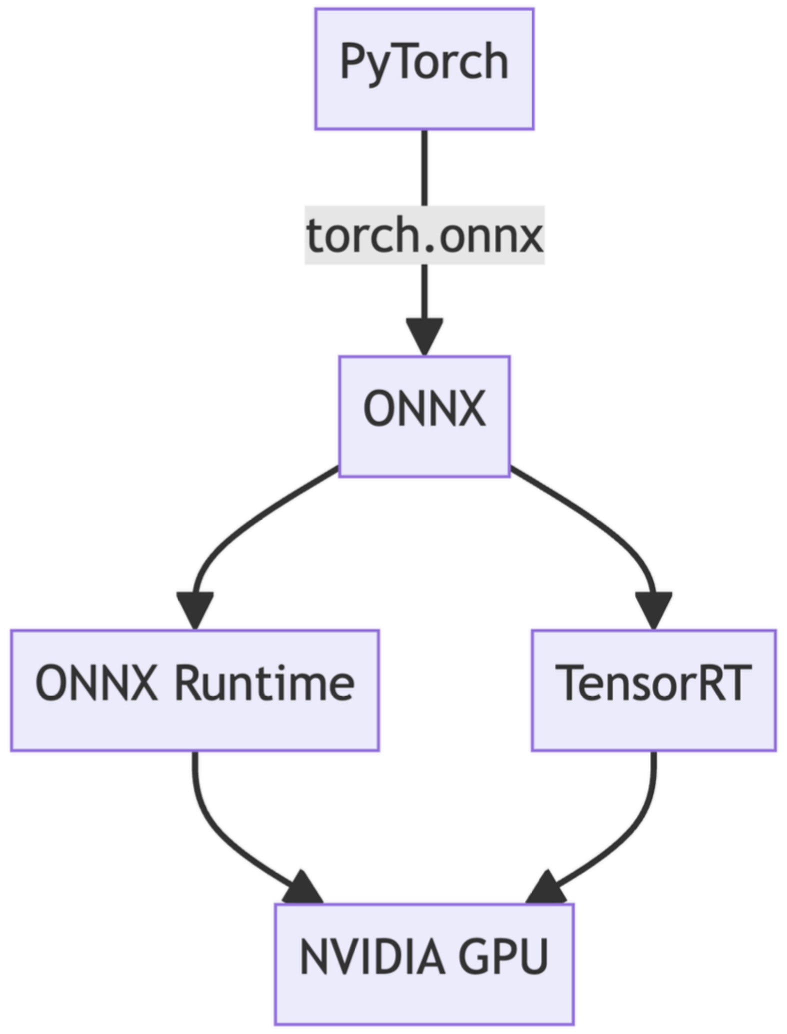
Proportion of	ALLEN	ETX4VELO
Reconstructed particles	99.06 %	99.23 %
Duplicate tracks	2.63 %	1.37 %
Fake tracks	2.17 %	1.04 %

From hits to tracks

- **GNN-based tracking in LHCb:**
  - **Recent studies towards a realistic algorithm → high throughput required**
  - **Tested two inference engines**
  - **Dedicated pipeline steps directly implemented in CUDA**
    ↳ kNN
    ↳ Connected components
  - **Ongoing pipeline optimisation as well as quantisation**

arXiv:2407.12119

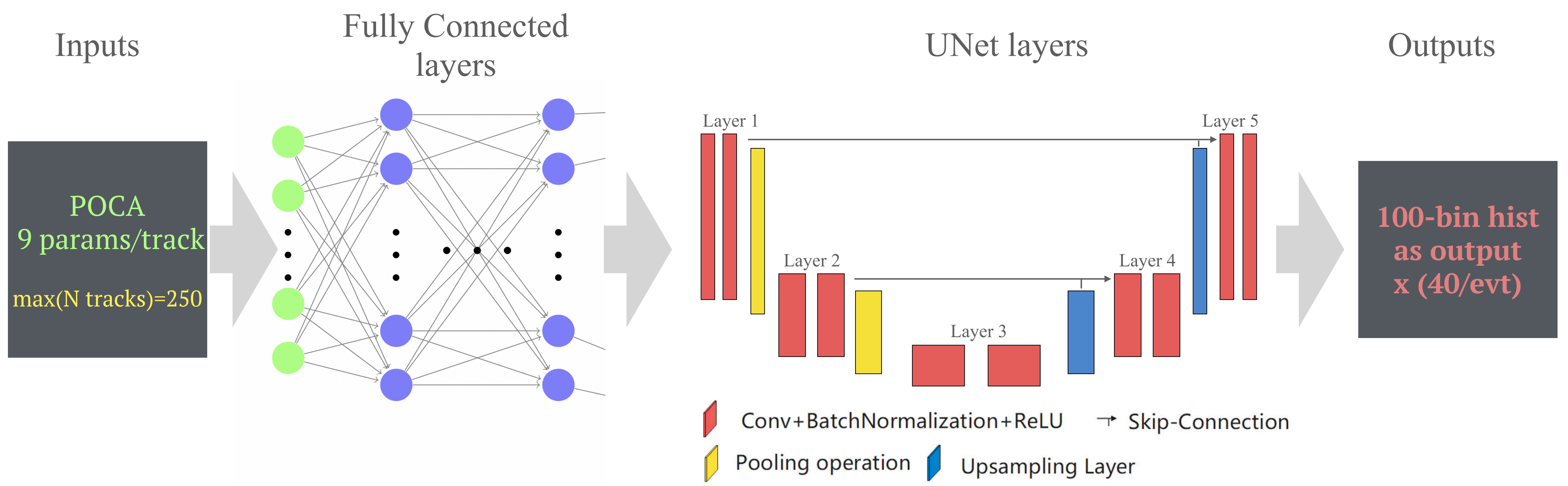


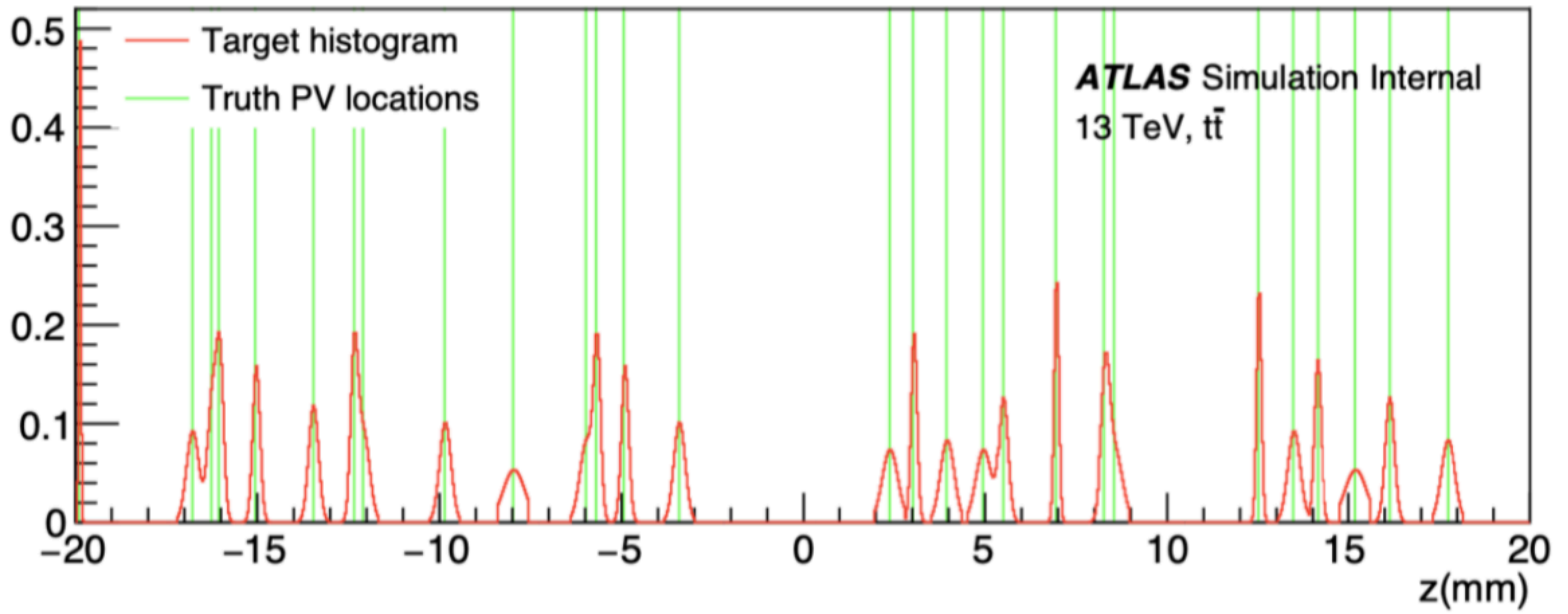


From tracks to Primary Vertex (PV)

PV finding with a hybrid model:

- **Originally developed in LHCb, extended toward ATLAS**
- **Hybrid model: Fully Connected + UNet**
- **Inputs :** Tracks parameters
  Target : Gaussians with heights and widths reflect the expected PV resolutions

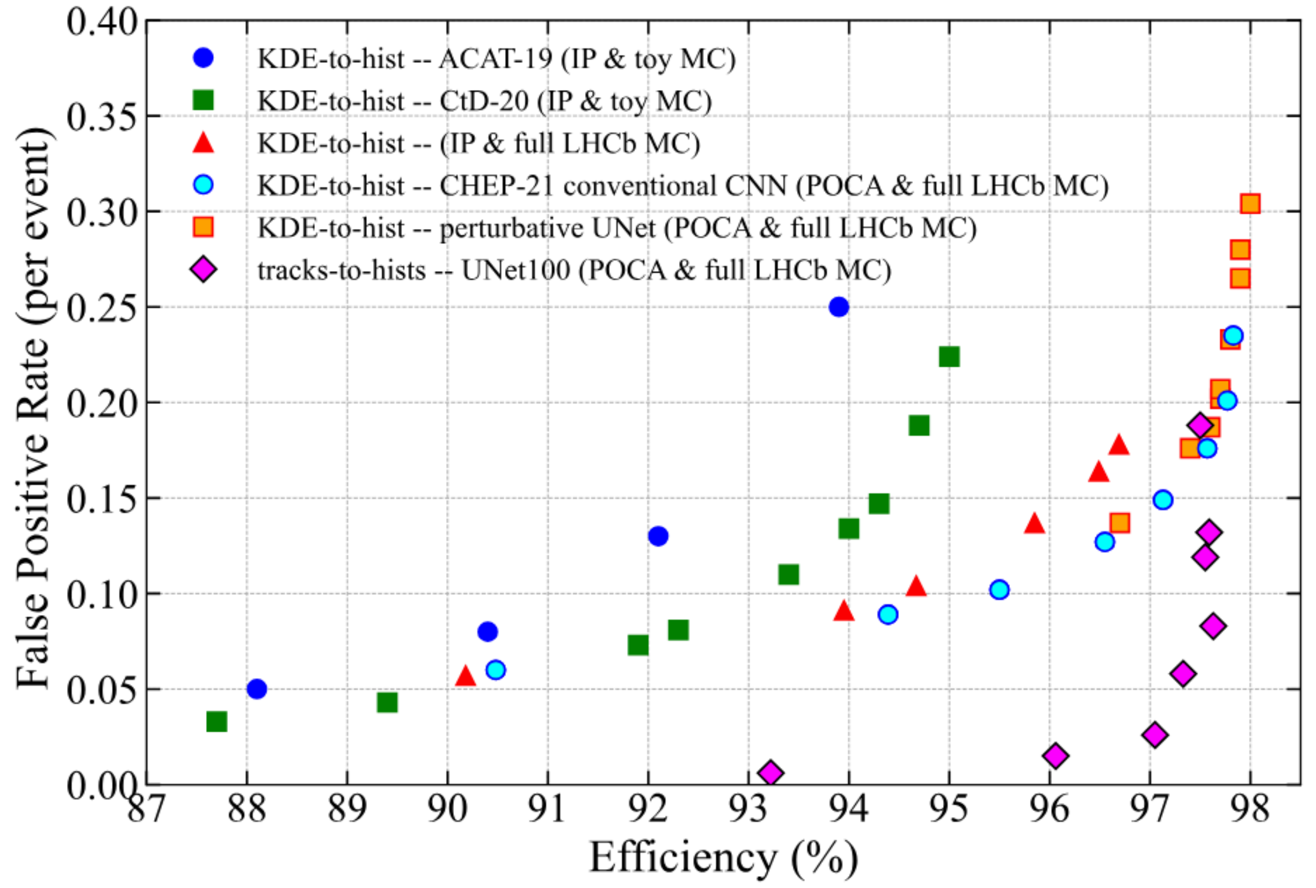


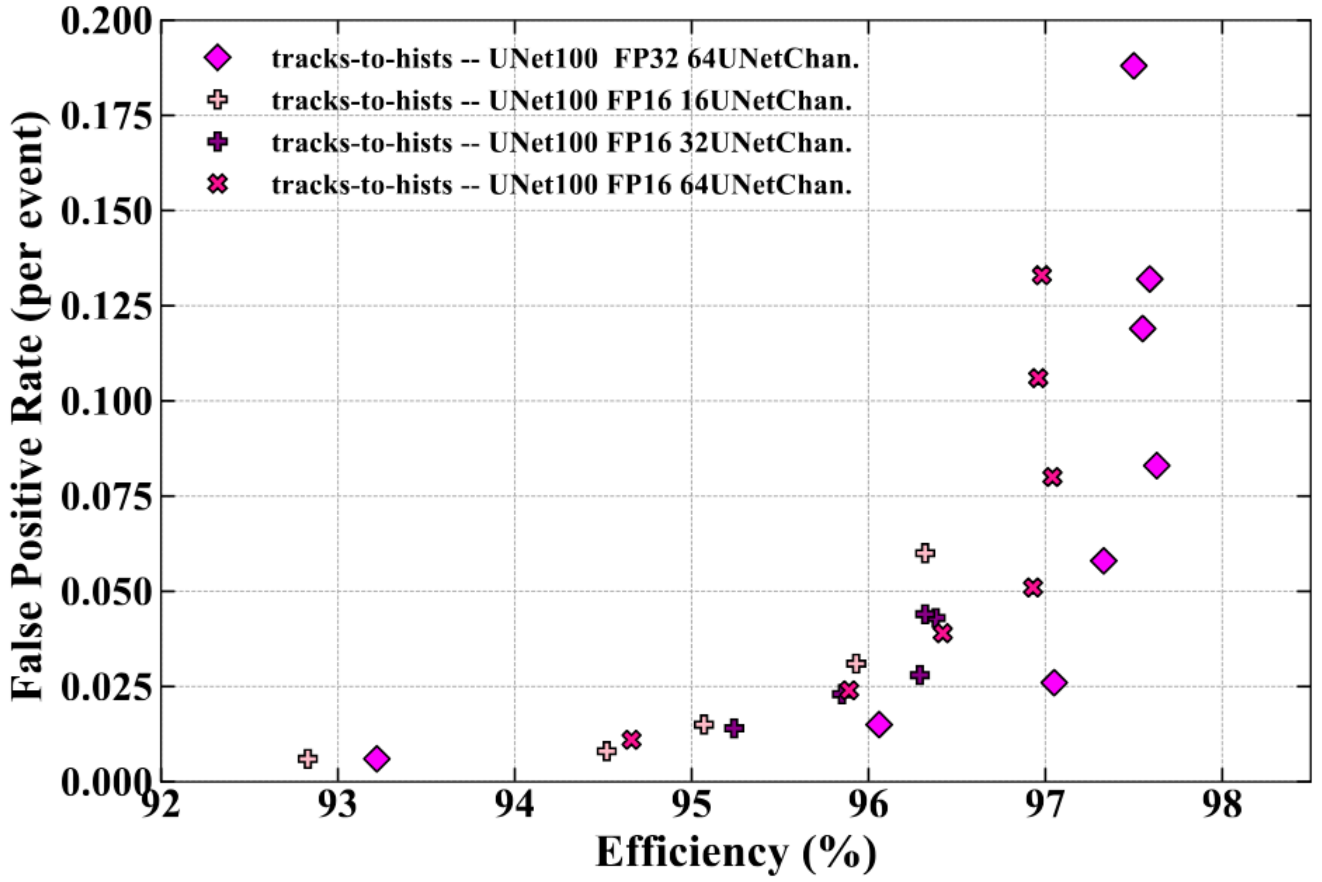


From tracks to Primary Vertex (PV)

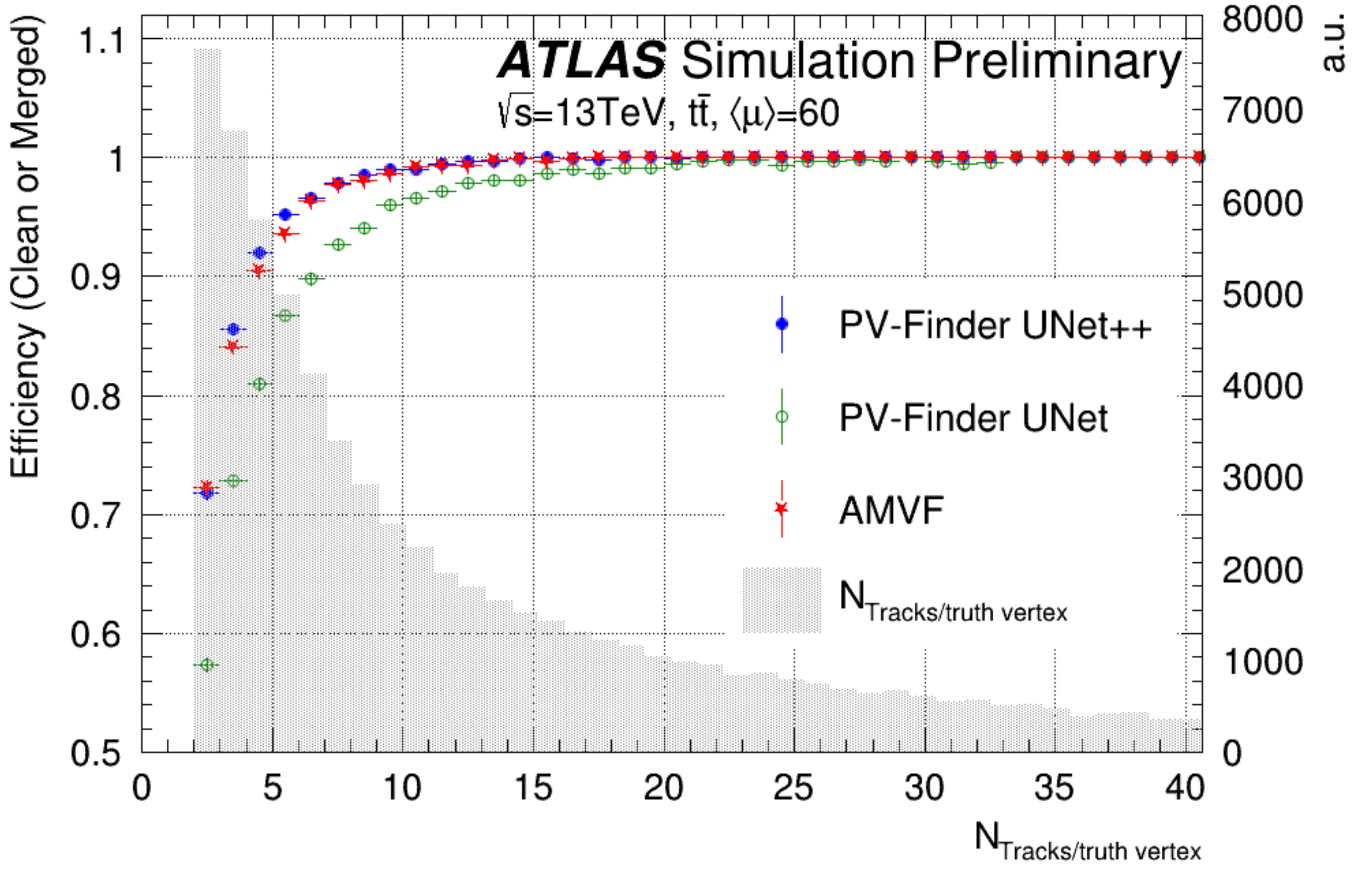
PV finding with a hybrid model:

arXiv:2407.12119





ATL-PHYS-PUB-2023-011



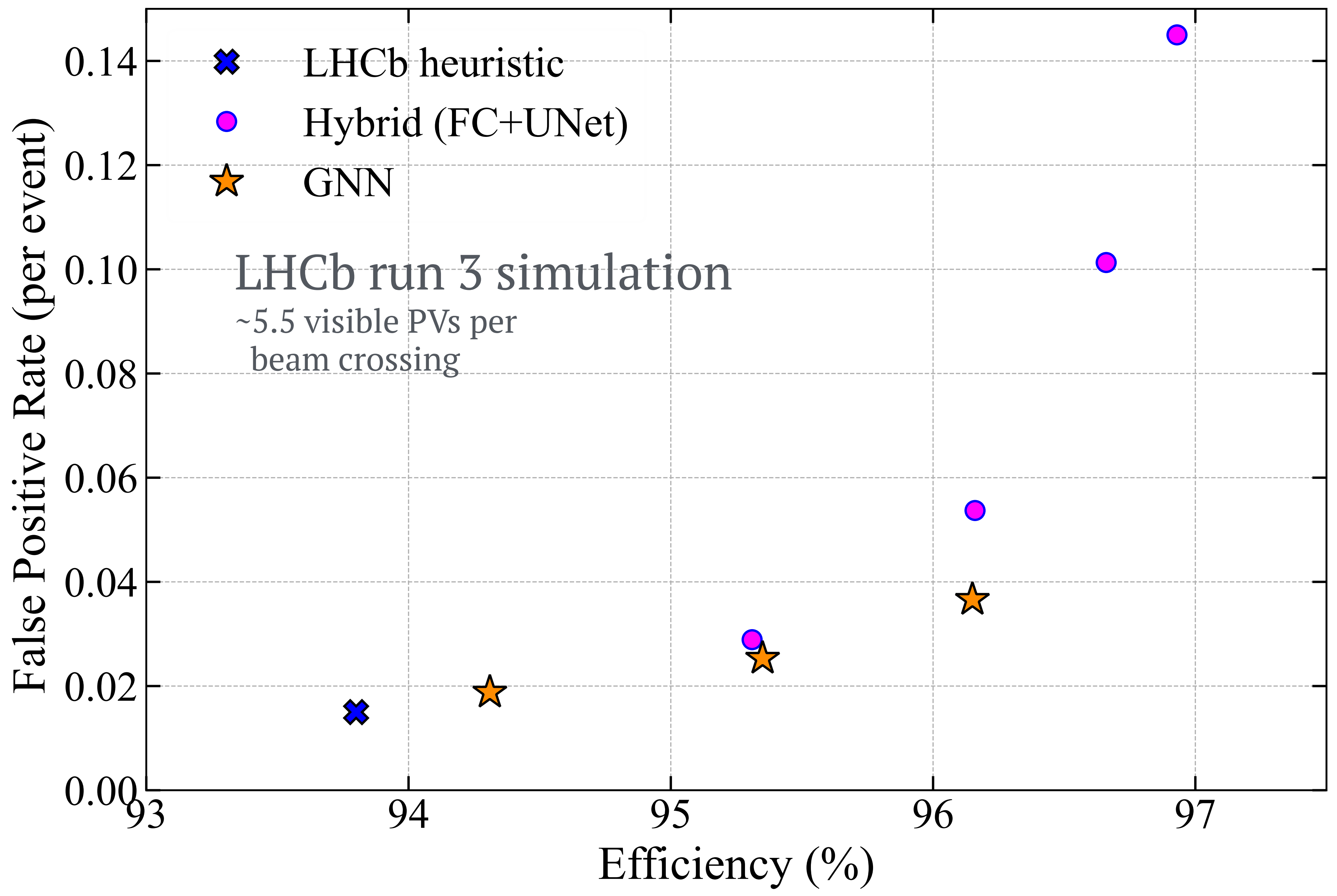
- **Iterative design improvement** with increased performance
- **Pruning & reduced precision studies** towards speeding up the inference, for application in HLT1 (next step)

- **Proof-of-concept** without hyper parameter optimisation
  - ↳ **2x better vertex resolution**
  - ↳ **Similar efficiency and false positive rates**

From tracks to Primary Vertex (PV)

PV finding with a hybrid model:

- **Recent alternative approach using GNN model** *S.A. @ EuCAIFCon24* (based on ETX4VELO)
  - ↳ track ↔ PV association by construction
  - ↳ improved physics performance

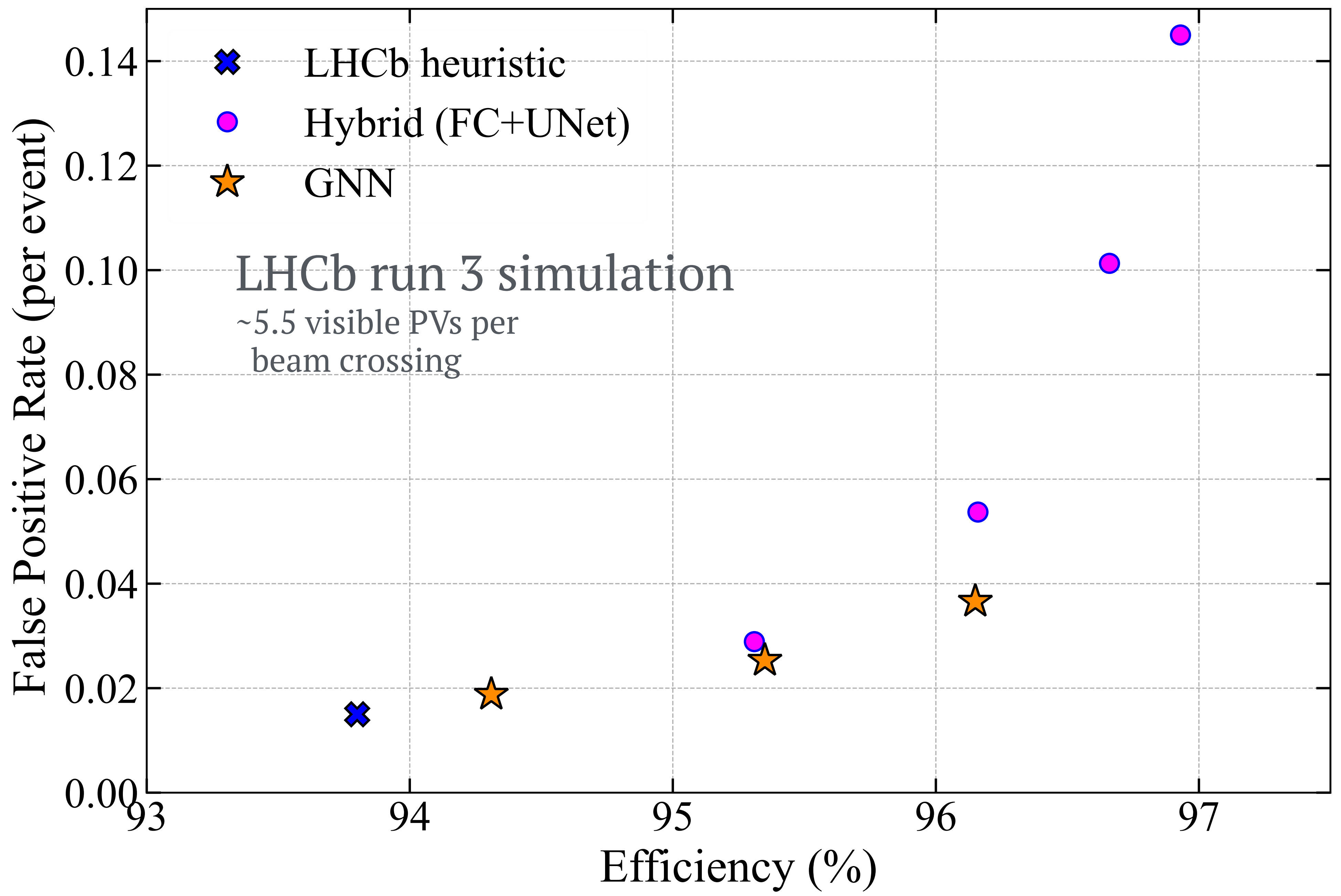


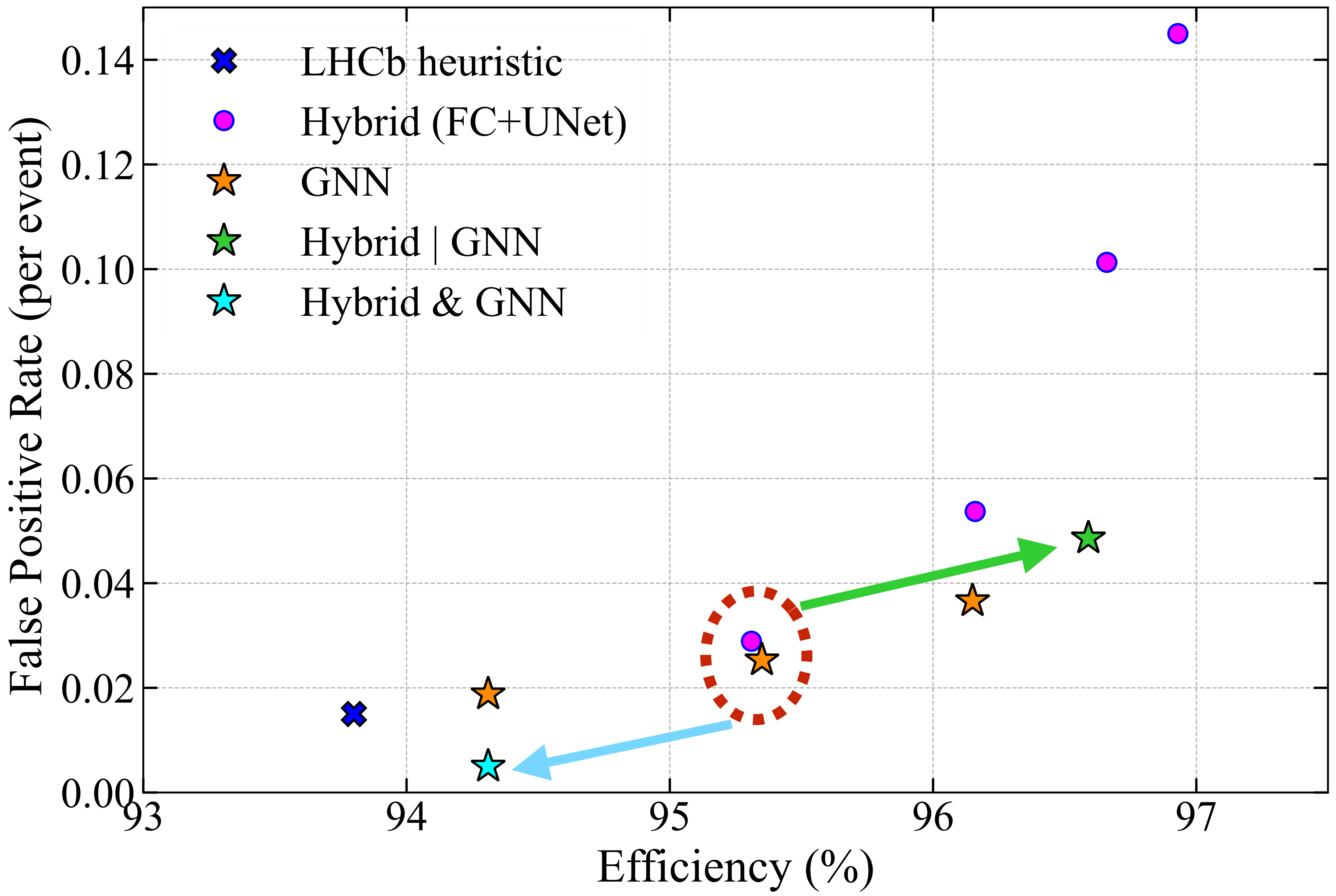
From tracks to Primary Vertex (PV)

PV finding with a hybrid model:

- **Recent alternative approach using GNN model** *S.A. @ EuCAIFCon24* (based on ETX4VELO)
  - ↳ track ↔ PV association by construction
  - ↳ improved physics performance

↳ GNN and hybrid model (trained on same input data and features) learned different representations

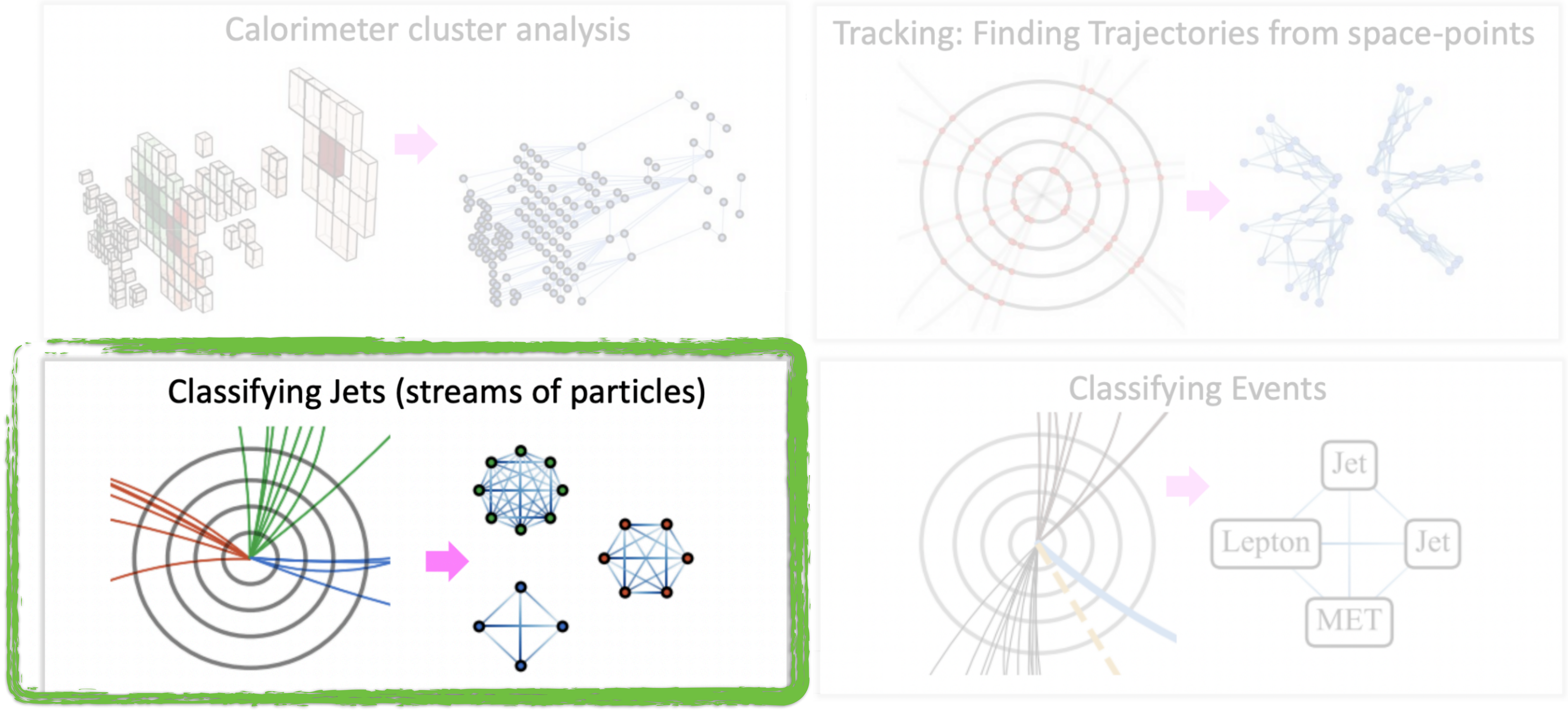




Online ML @ LHC

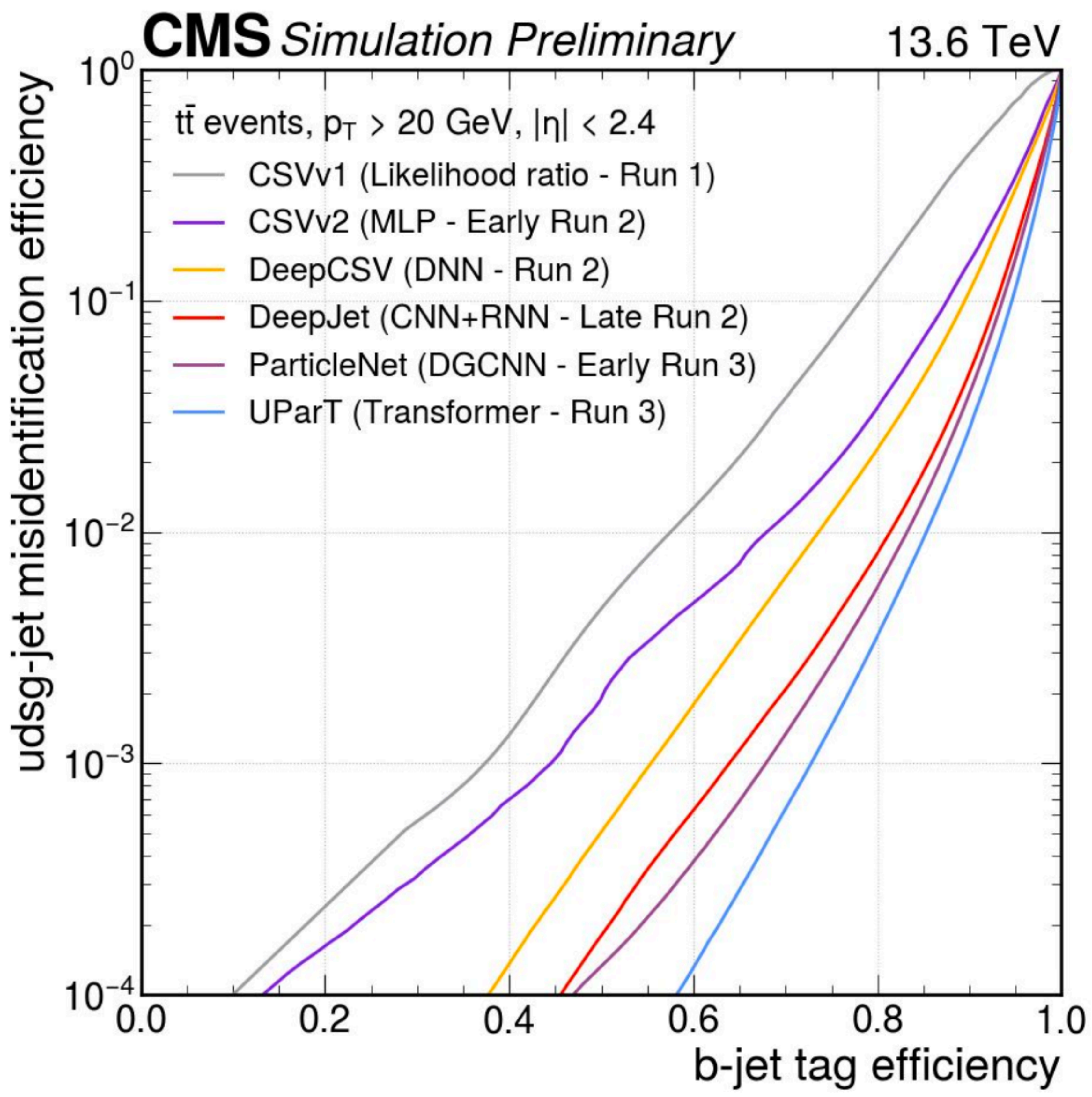
A selection of ML applications, in operation or in development, for online reconstruction

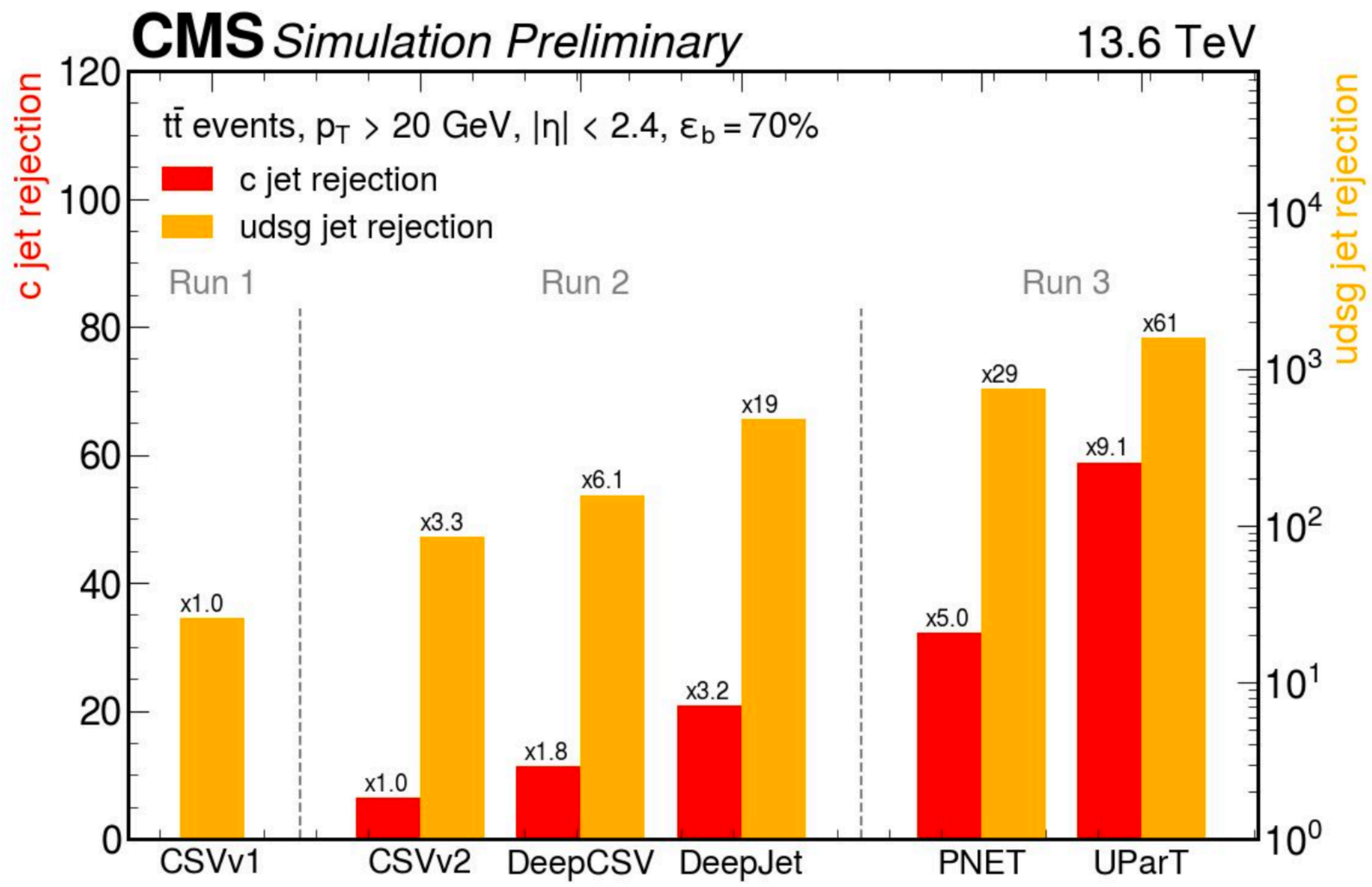
(very much non exhaustive!)



From D. Rankin (FastML for Science Conference 2024)

- **Flavour jet tagging:**
  - **Steady progress over the years for heavy-flavour tagging** *CMS-DP-2024-066*



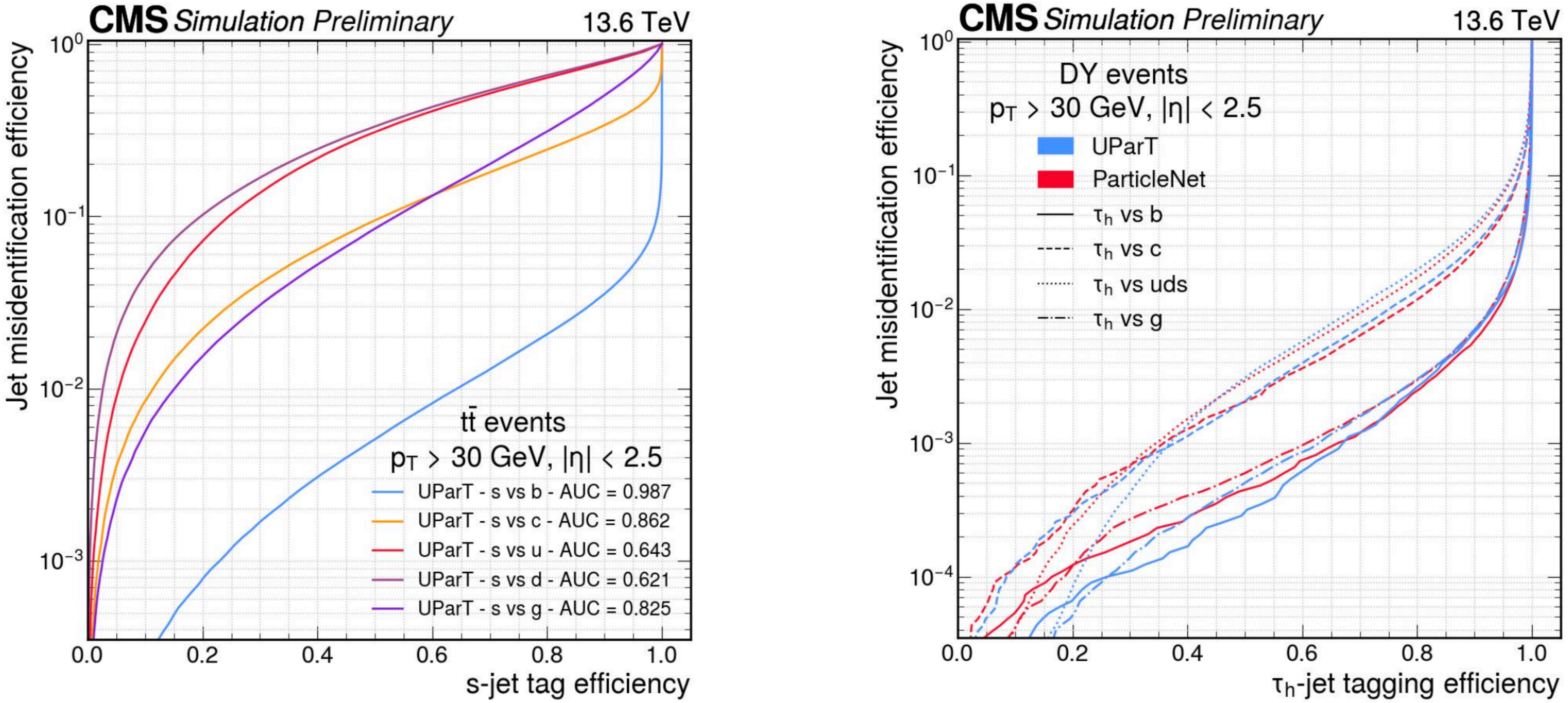


Enhanced sensitivity to HH → 4b
(as well as many other modes)
CMS-DP-2023-050

Classifying jets

Flavour jet tagging:

- Recent models (GNN or Transformer) enable s-jet (pioneer) as well as τ-tagging (improved) *CMS-DP-2024-066*



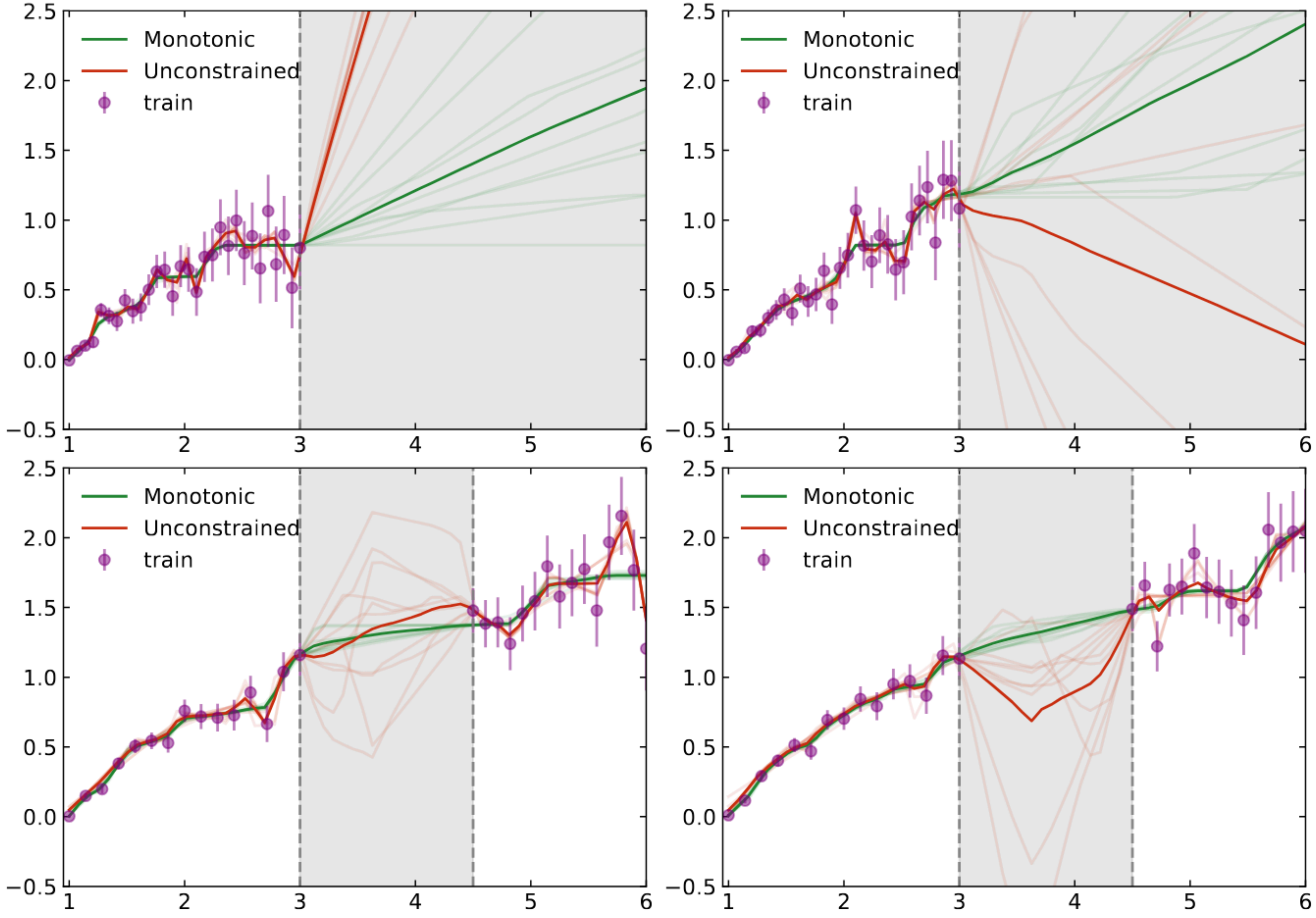
- Transformer model appears to be much more computationally efficient: ~7 improvement in inference speed from (larger) ParticleNet to (smaller) UParT

Lipschitz neural networks

- **Trigger algorithms require:**
  - **Robustness**
    against detector instabilities and simulation inaccuracies
    ↳ weight-normalisation scheme during training
  - **Monotonicity**
    in certain features for out-of-distribution
    ↳ addition of a residual connection to the network

Monotonic Lipschitz neural networks impose desired constraints in the behaviour of the network by construction

arXiv:2112.00038

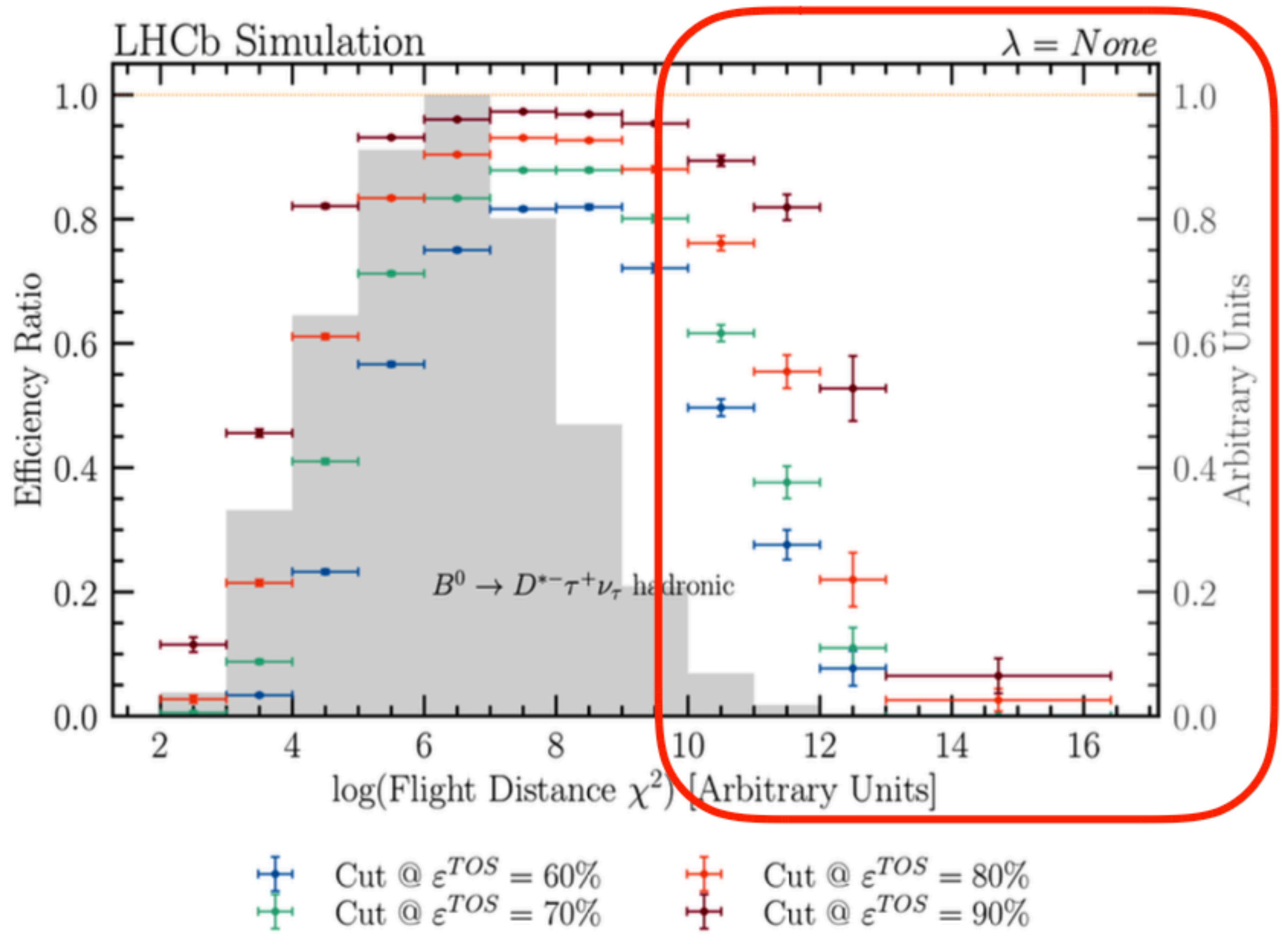


Lipschitz neural networks

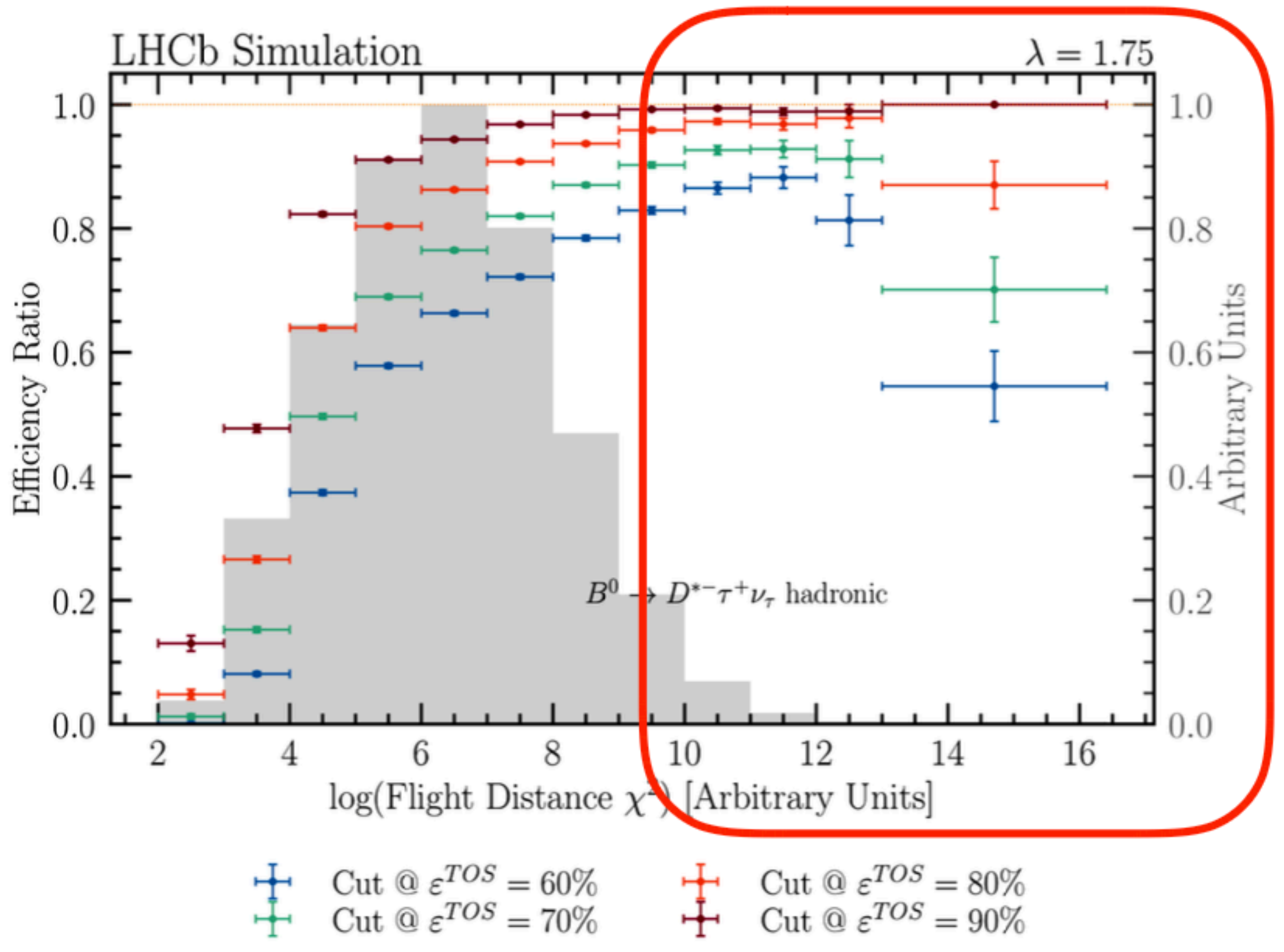
- **Currently in LHCb's trigger:**
  - **Topological triggers in HLT2** *arXiv:2312.14265*

    ↳ Monotonicity imposed in the IP-χ^2 and the p_T

    ↳ Enhanced sensitivity to long-lived candidates



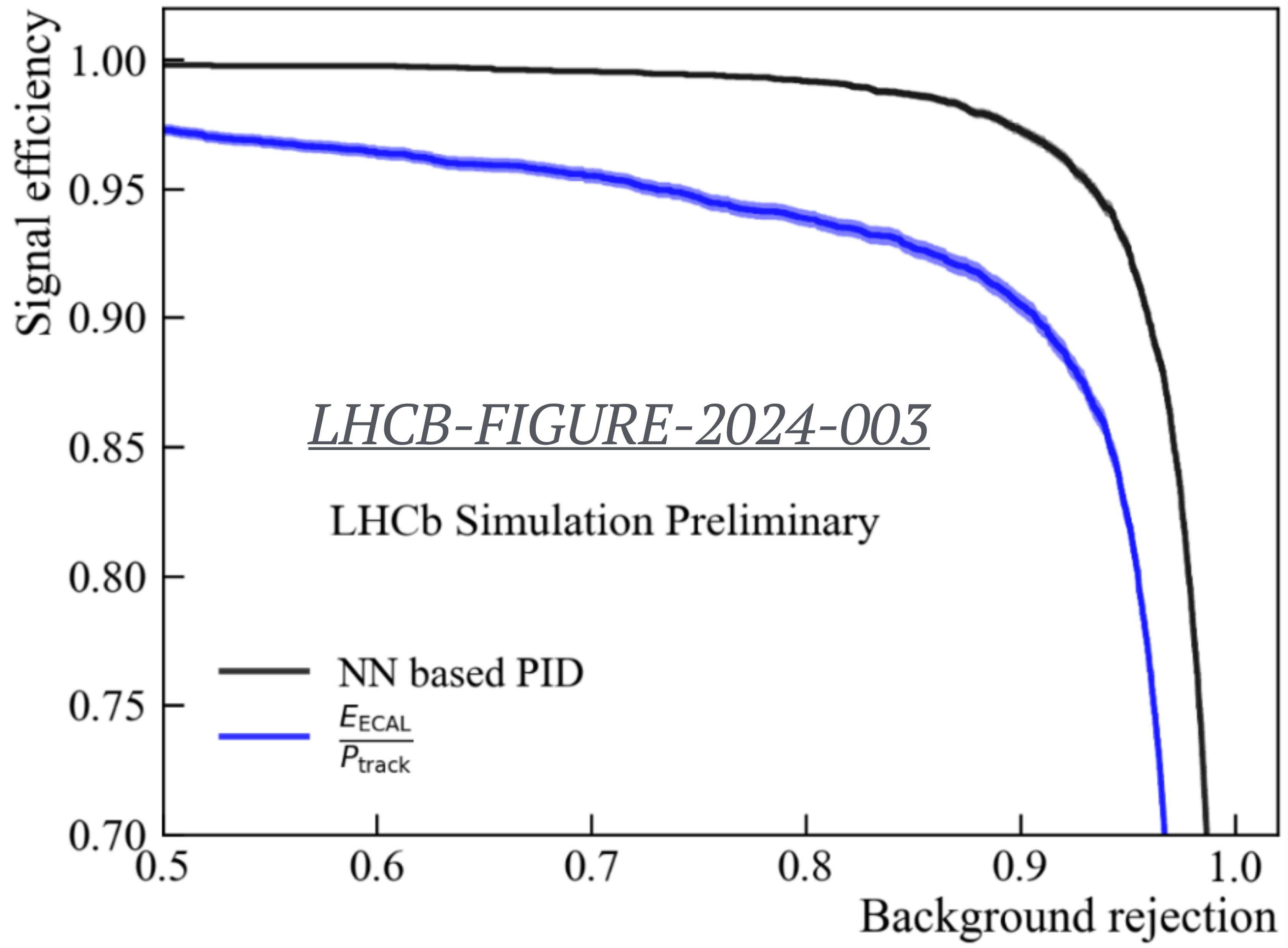
Unconstrained NN



Lipschitz monotonic NN

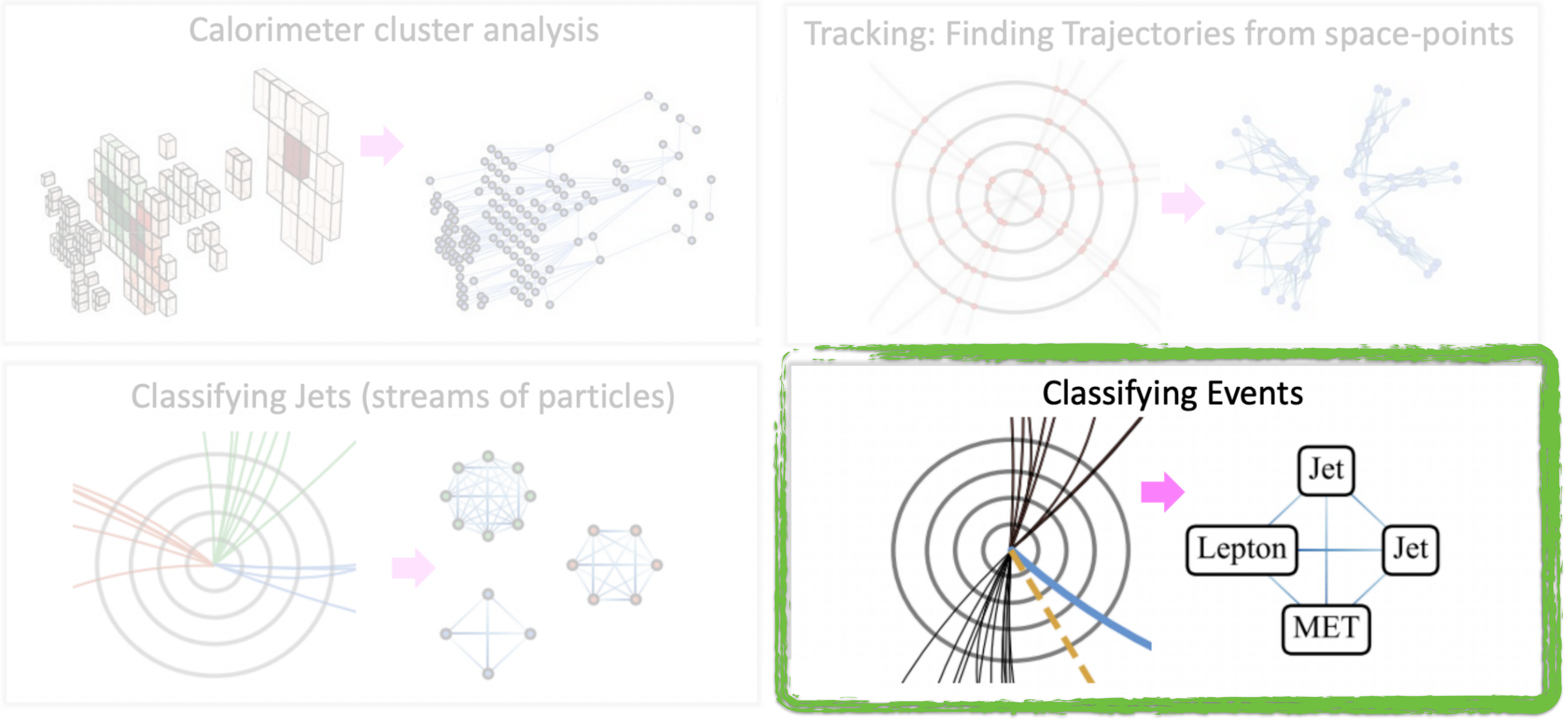
  - **Electron ID at the HLT1 (GPU)**

    ↳ Large improvement with respect to the conventional (not ML based) algorithm



A selection of ML applications, in operation or in development, for online reconstruction

(very much non exhaustive!)



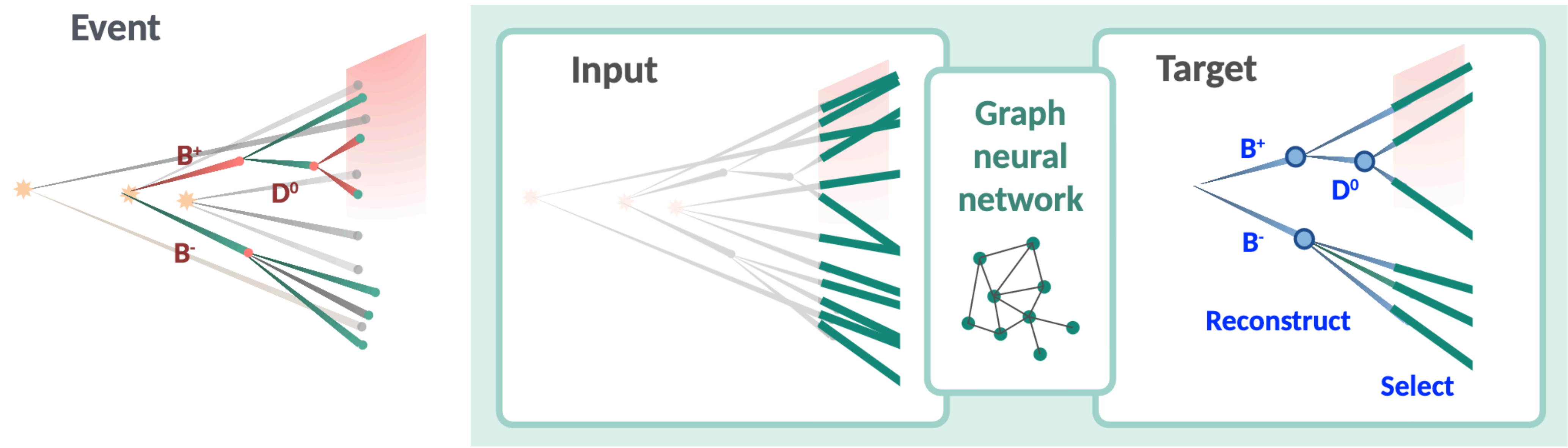
From D. Rankin (FastML for Science Conference 2024)

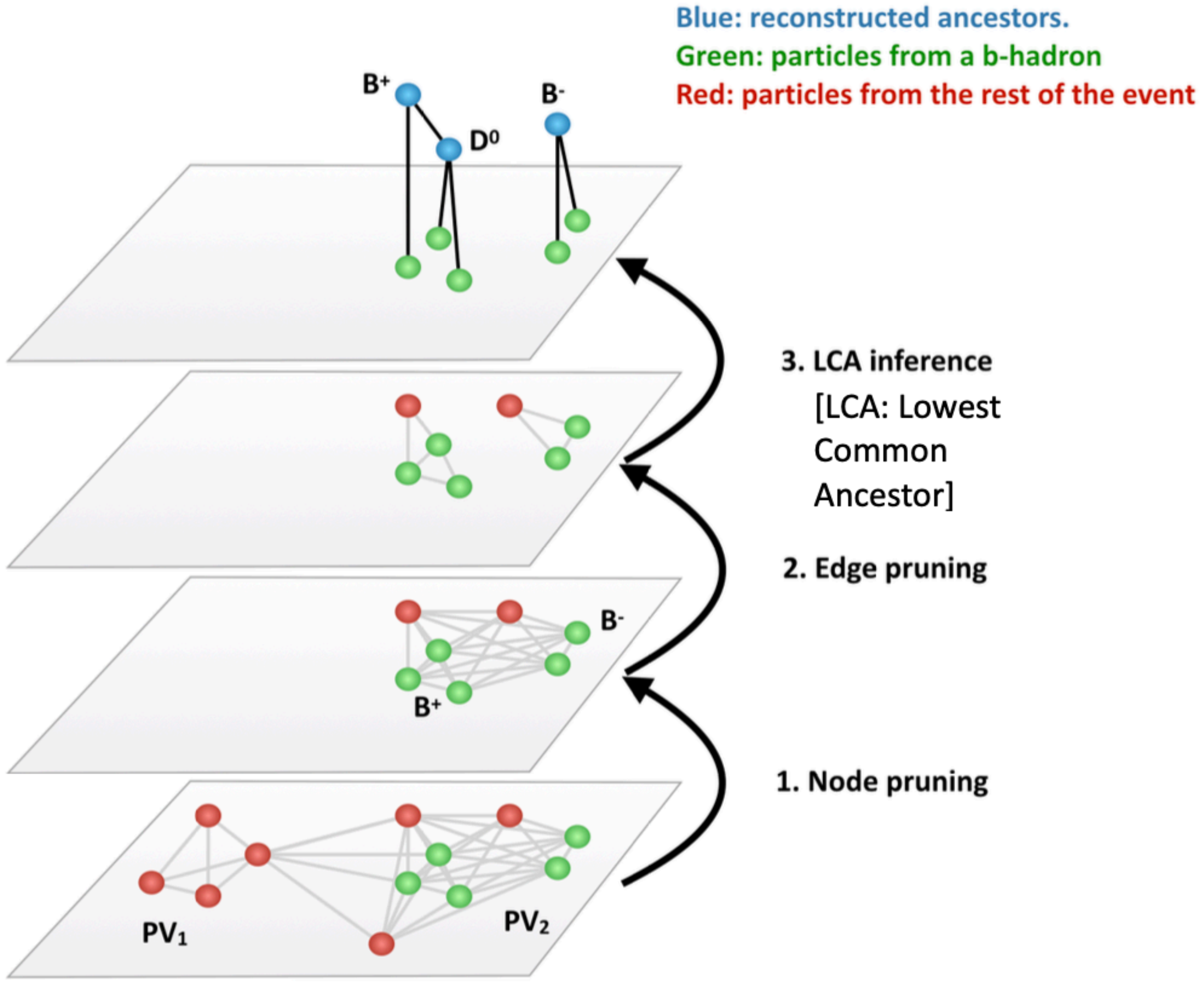
Deep-learning based Full Event Interpretation

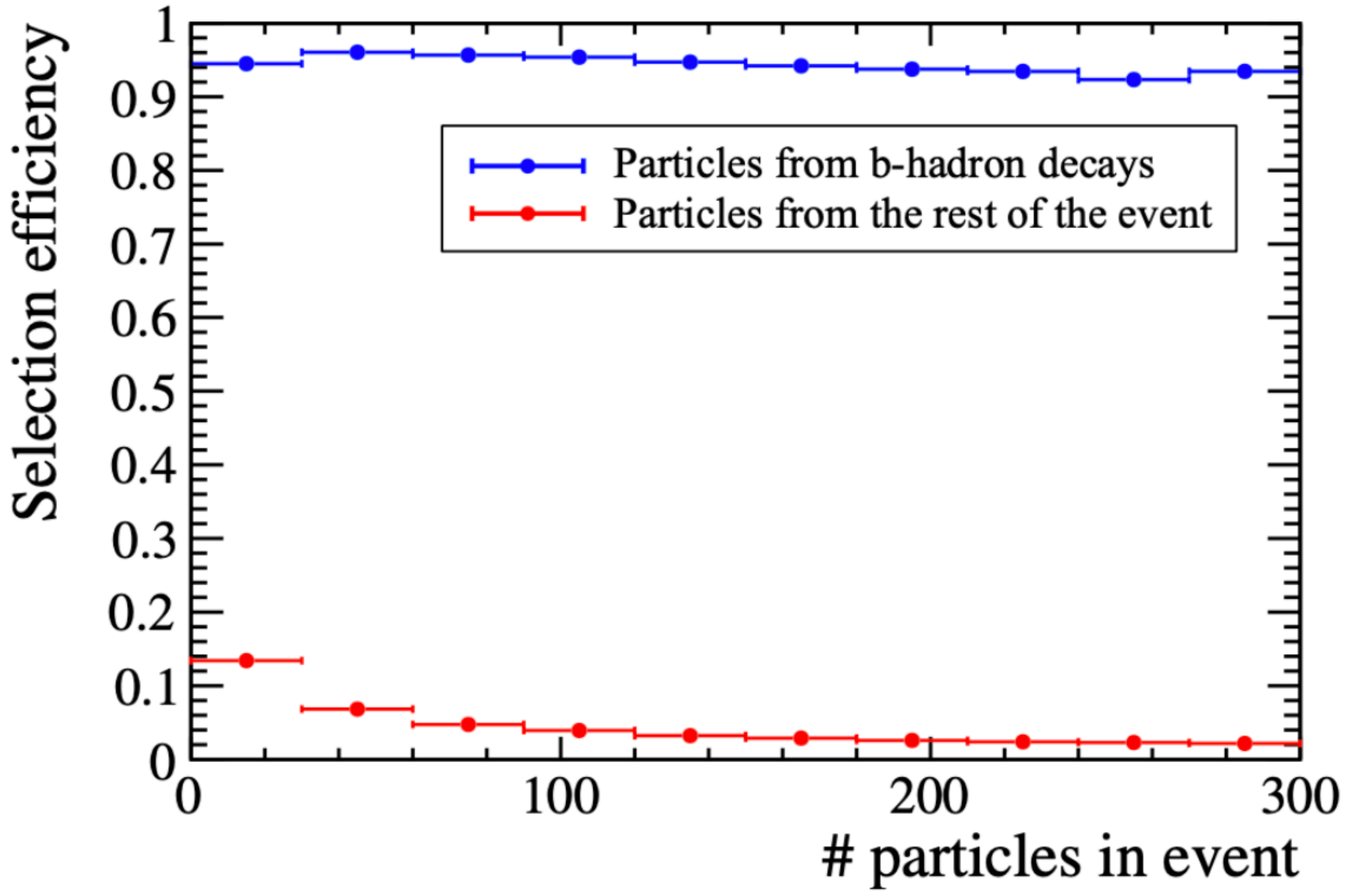
One-go inclusive multi-signal reconstruction + pileup suppression, for optimal event filtering

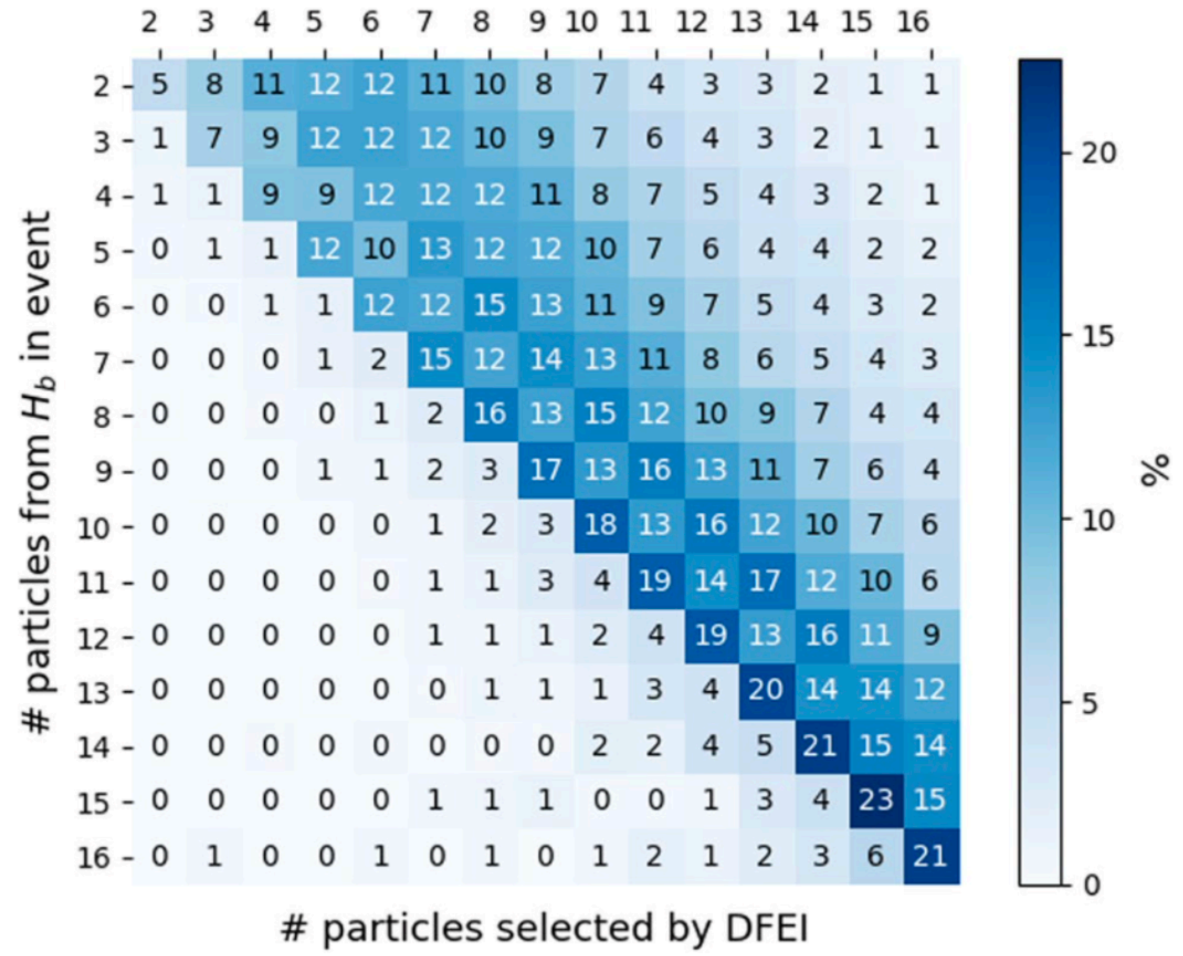
- Based on three sequential GNN modules

Comput Softw Big Sci 7, 12 (2023)









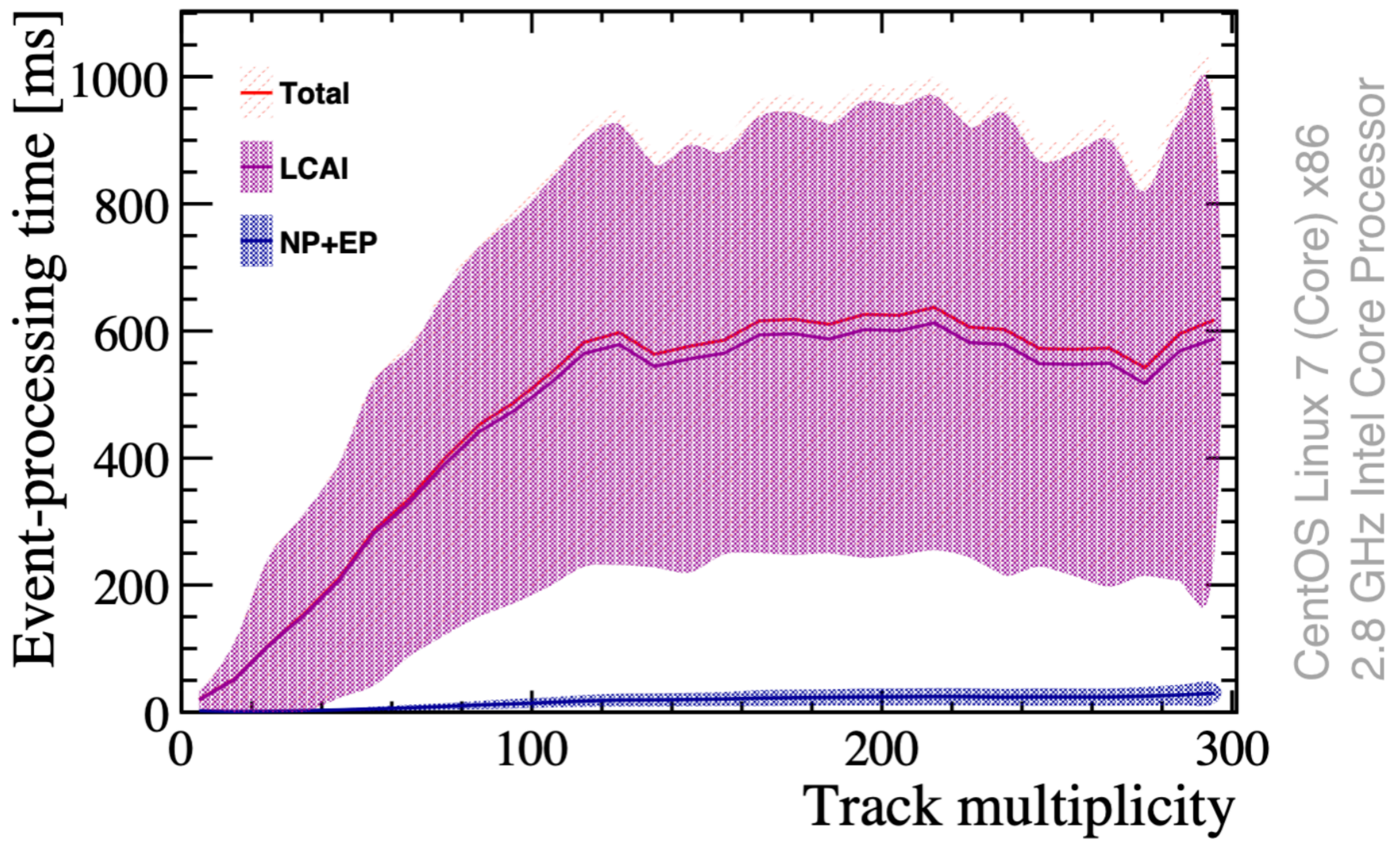
Powerful event-filtering irrespectively of the particle multiplicity,
as found in inclusive b-hadron simulation.

Deep-learning based Full Event Interpretation

- **One-go inclusive multi-signal reconstruction + pileup suppression, for optimal event filtering**

  Comput Softw Big Sci 7, 12 (2023)

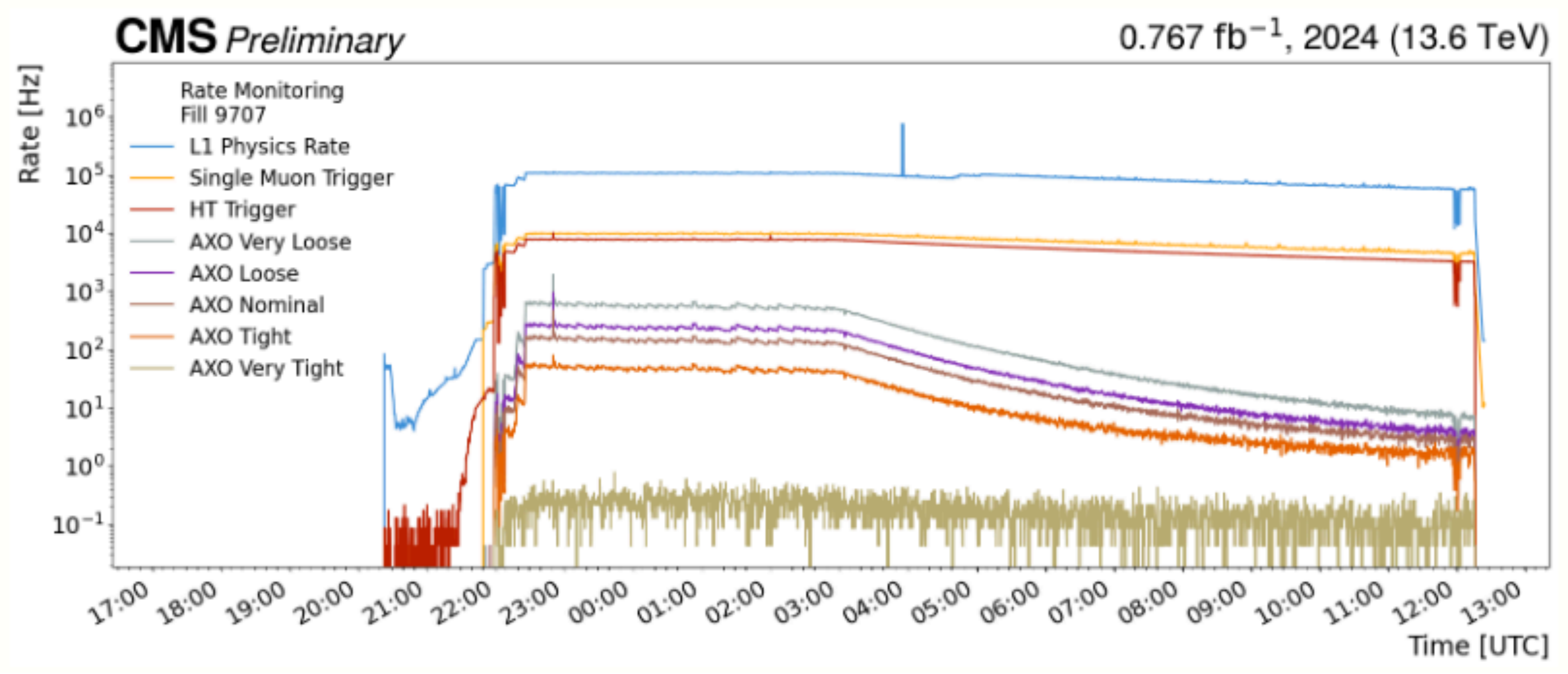
  - **Recent improvements to model inference** *F.L. Souza De Almeida @ ACAT24*
    seconds/evt on CPU with first prototype

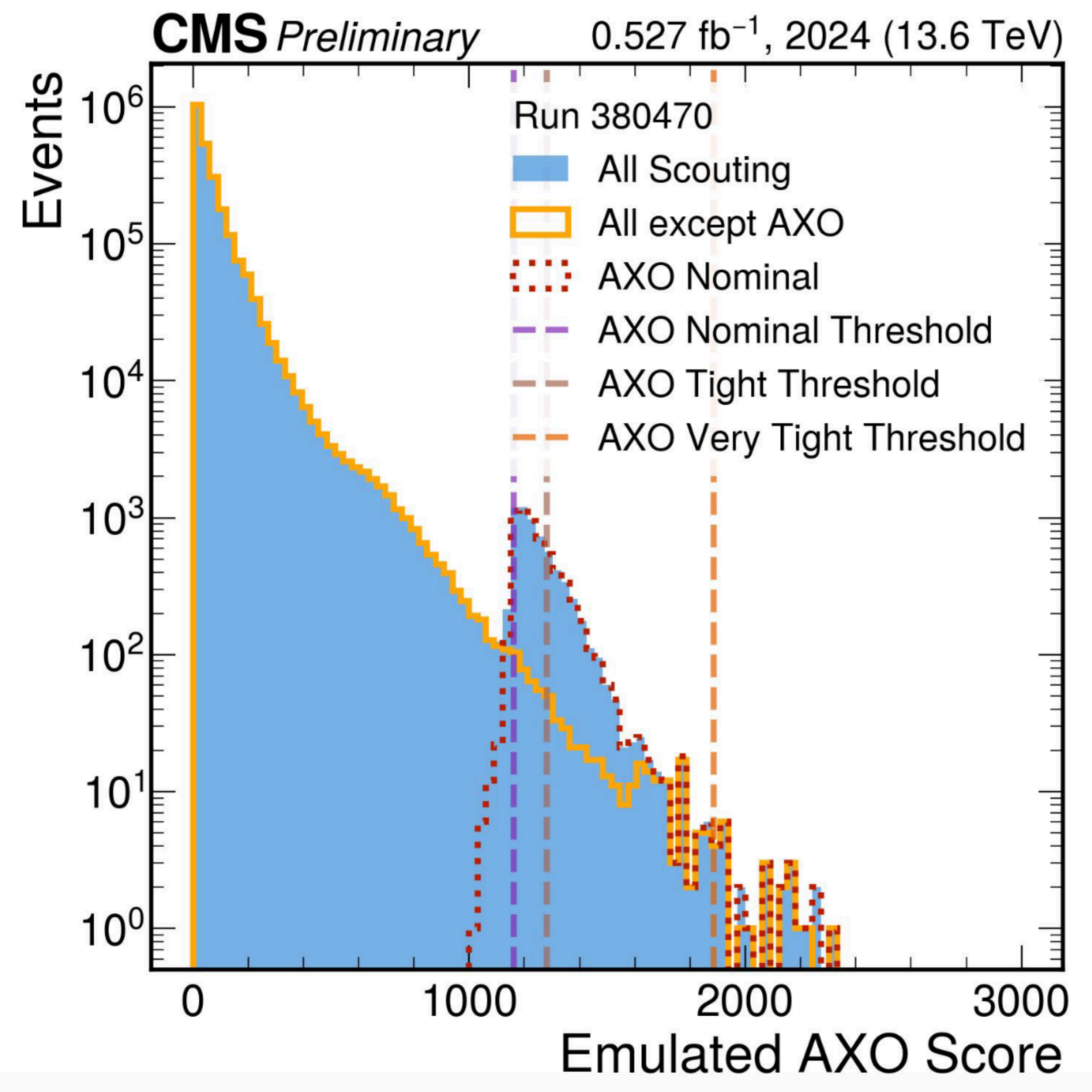


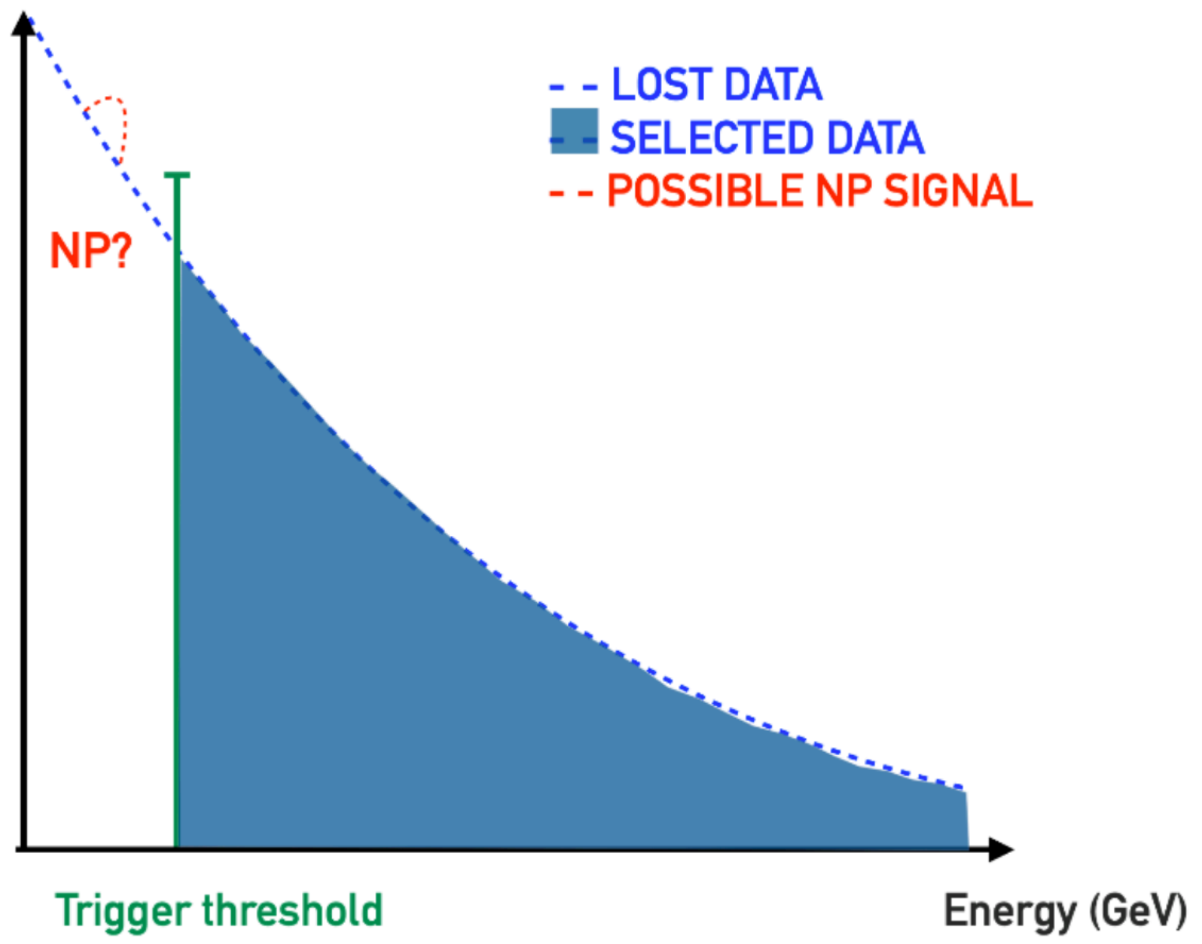
- **Full inference pipeline in C++**
  LCAI GNN converted using TMVA::SOFIE
- **Replacement of Node and Edge Pruning steps (NP & EP) with BDTs**
- **Overall timing now dominated by LCAI**
  (ongoing optimisation of this step)

AXOL1TL: *CMS-DP-2023-079* *CMS-DP-2024-059* *N. Zipper @ FalstML24*

- Variational Auto-encoder (VAE) based algorithm to select anomalous (NP?) events in real-time in L1 physics trigger (40 MHz)
- FPGA integration through hls4ml+vivado toolchain
- Running in safe mode and deployed in the Global Trigger Test Crate in 2023
- Integrated into L1 in 2024

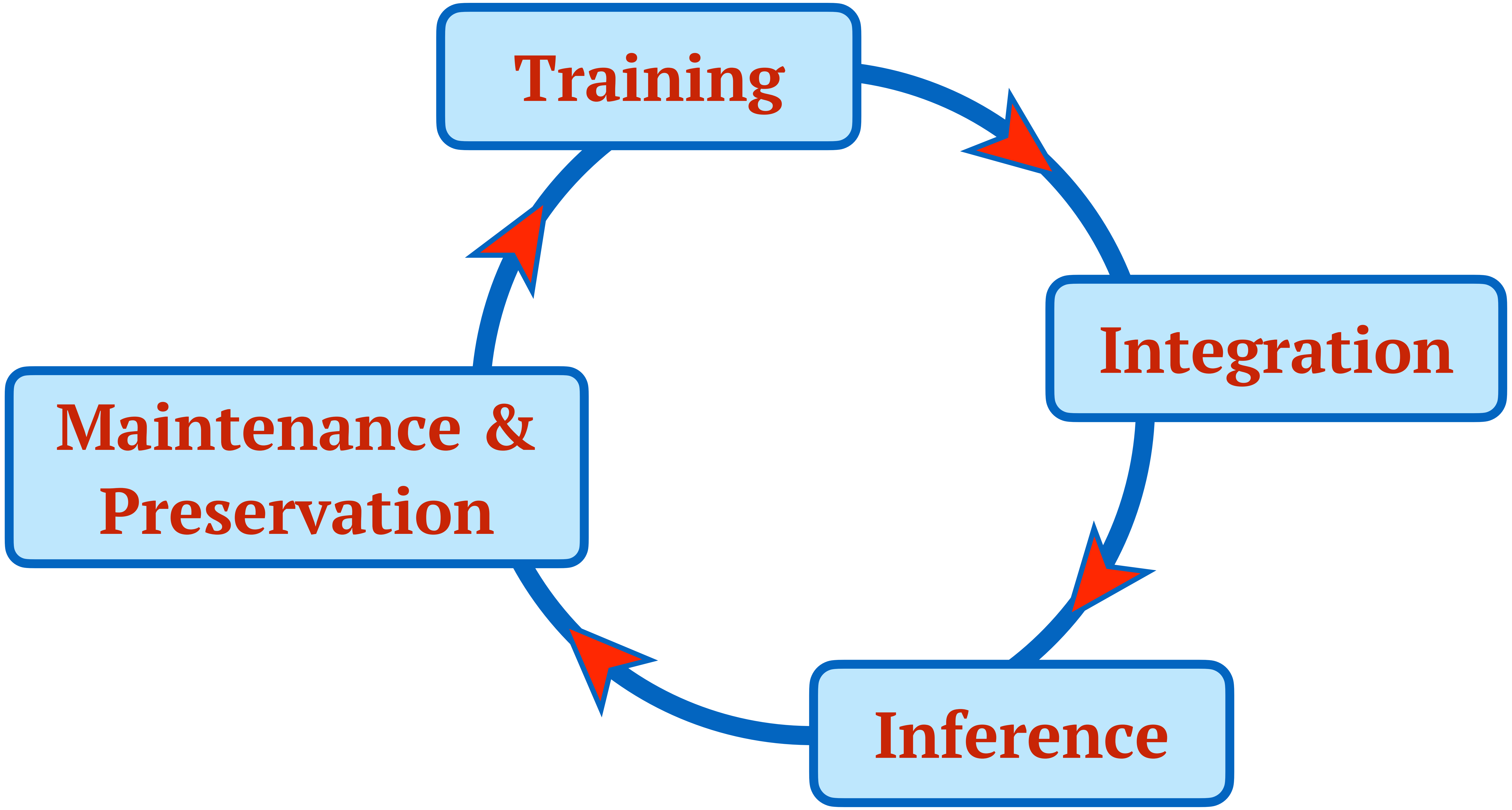






Network preferentially identifies large multiplicity events, potentially large gains in new physics acceptance

Common challenges



Efficient and sustainable exploitation of ML presents challenges at various steps

Common solutions among CERN collaborations is paramount!

Common challenges

- **ML is developing at an incredible pace**

arXiv:2407.12119

Machine Learning in High Energy Physics Community White Paper

May 17, 2019

2.2 Brief Overview of Machine Learning Algorithms in HEP

There are different types of DNN used in HEP: fully-connected (FCN), convolutional (CNN) and recurrent (RNN). Additionally, neural networks are used in the context of Generative Models, where a Neural Network is trained to reproduce the multidimensional distribution of the training instances set. Variational AutoEncoders (VAE) and more recent Generative Adversarial Networks (GAN) are two examples of such generative models used in HEP.

State-of-the-art new model architectures (GNN, transformers) already in use in 2024

Efficient ML integration into reconstruction requires very specific domain knowledge

Need for permanent engineers positions

Final remarks

- **Increasing state-of-the-art ML algorithms in reconstruction**
  - **Offline ML techniques are shifting toward online applications** to increase physics reach, but ultimately detector capabilities will drive are ability to perform good physics
  - Increasing focus on **long-term maintainability** of ML solutions and development of **common pipelines**

- **Infrastructure & technical expertise will be key**
  - **Centralised training infrastructure**
  - **Support for heterogeneous architectures**
  - **On-chip inference optimisation**
  - **Maintain and develop collaboration with industry**