Realtime Machine Learning in the CMS Level-1 Trigger

Artur Lobanov
Universität Hamburg
Institut für Experimentalphysik

Workshop on Realtime Machine Learning | Gießen | 10.4.20224

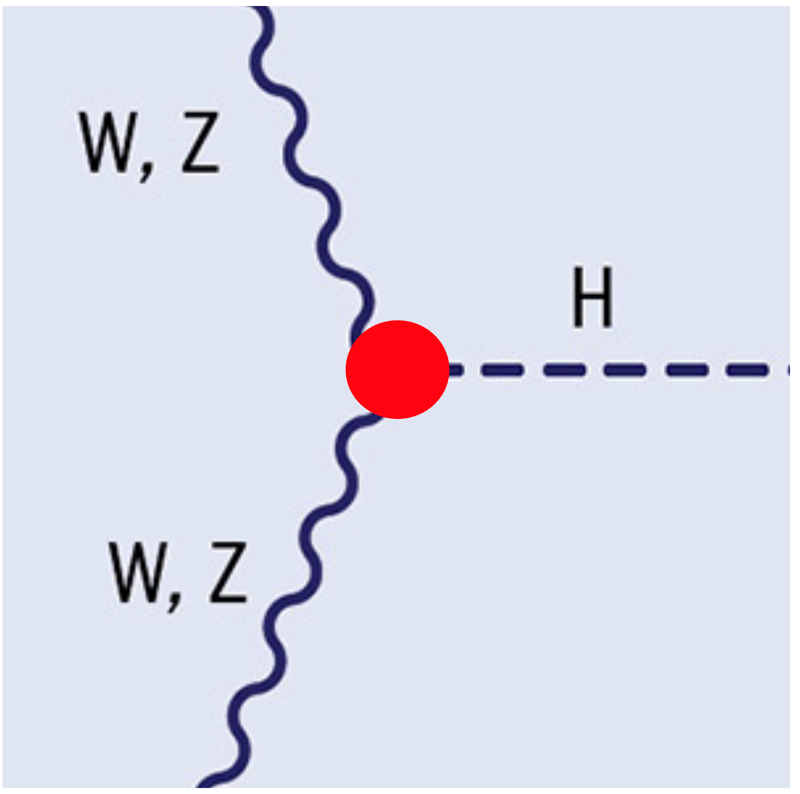
CLUSTER OF EXCELLENCE
QUANTUM UNIVERSE

CMS Germany BMBF FSP-CMS

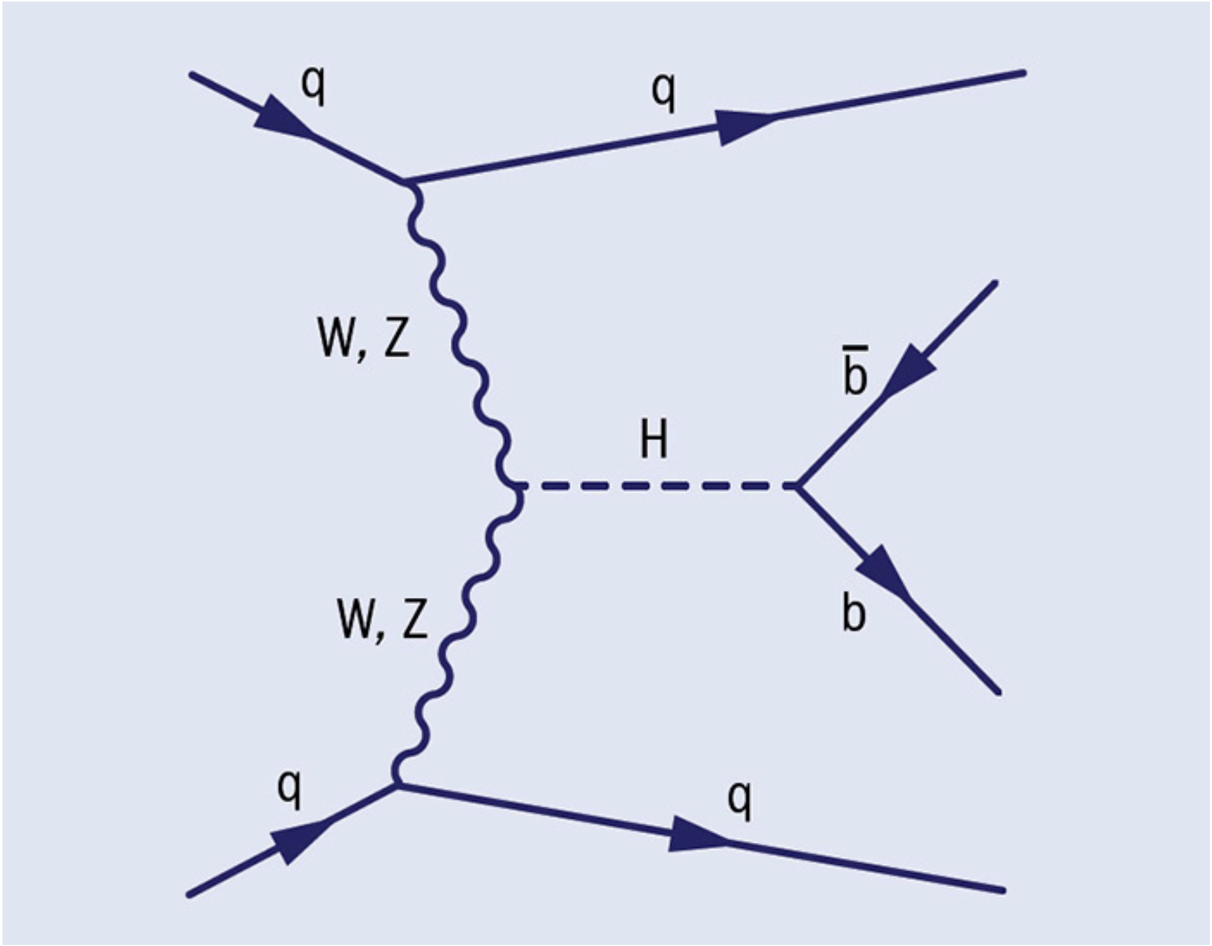
Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

What we do today @ the Large Hadron Collider (LHC)

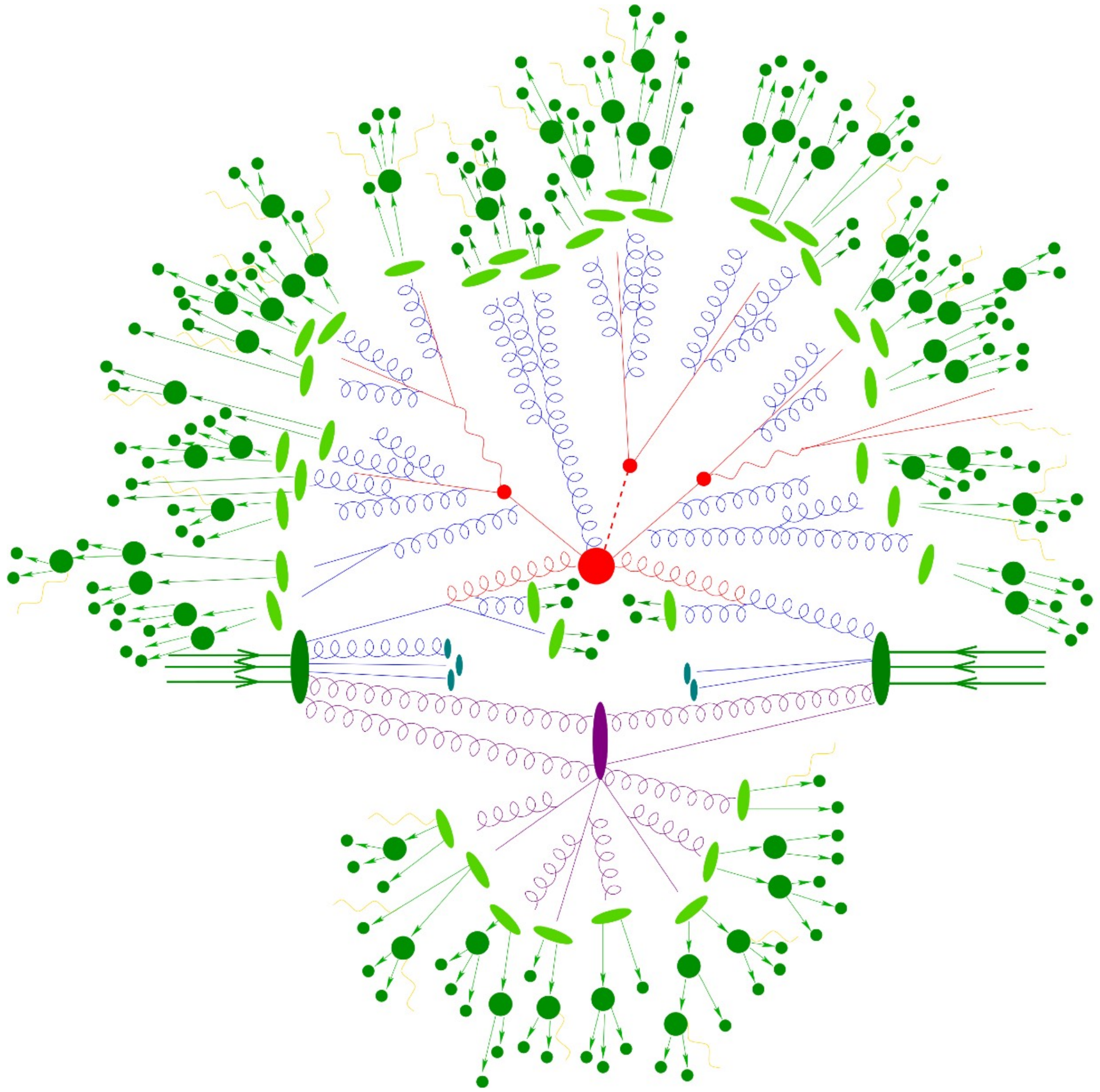
What we want to study



How collisions help us



What actually happens



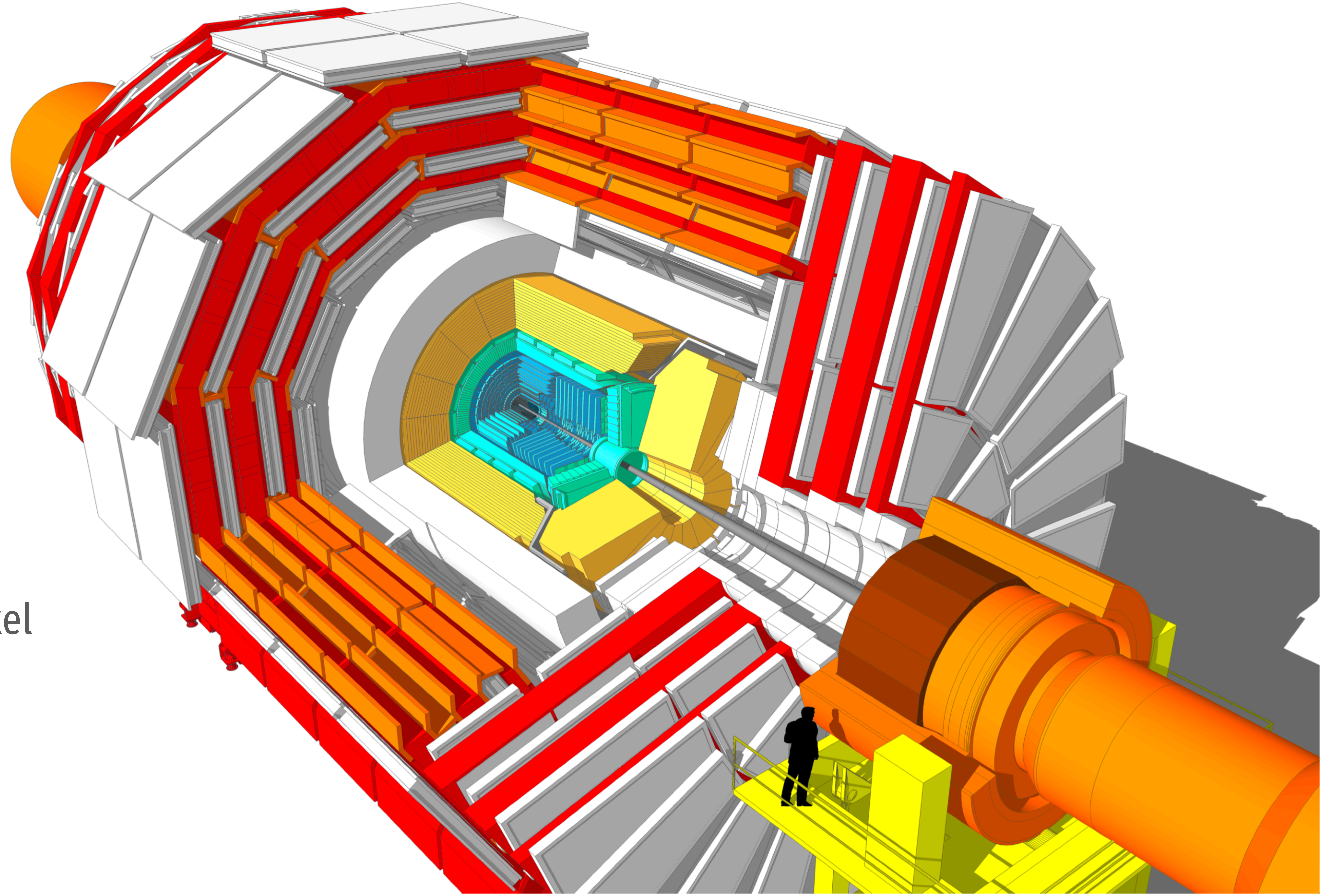
Production of a Higgs boson (H) through Vector Boson Fusion (W/Z)

Partons and hadronization

The CMS experiment at the LHC

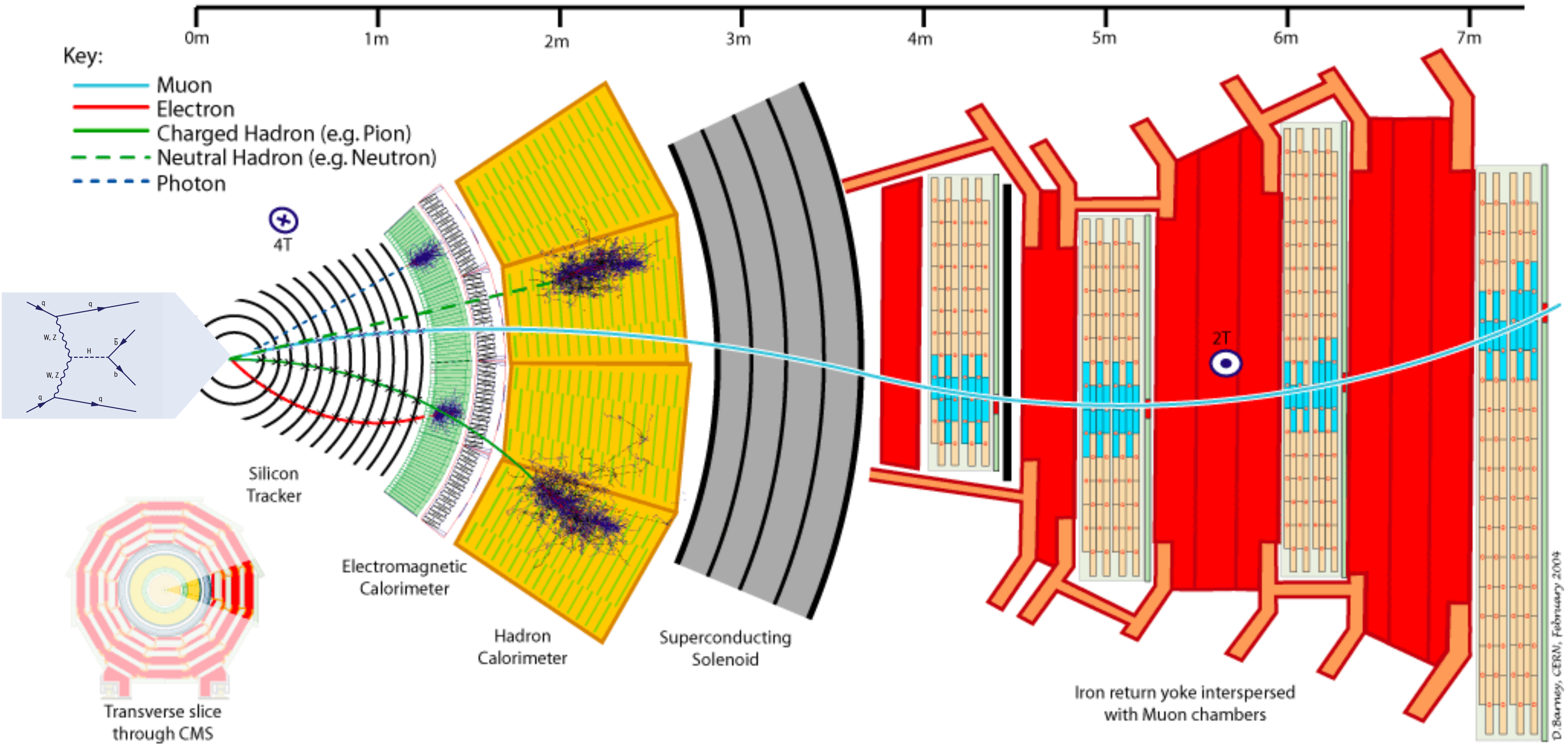
The CMS experiment:

LHC camera with 100 Mpixel



How CMS sees particles

Different particle types can be measured with different detectors



MACHINE LEARNING IN CMS / HEP EXPERIMENTS

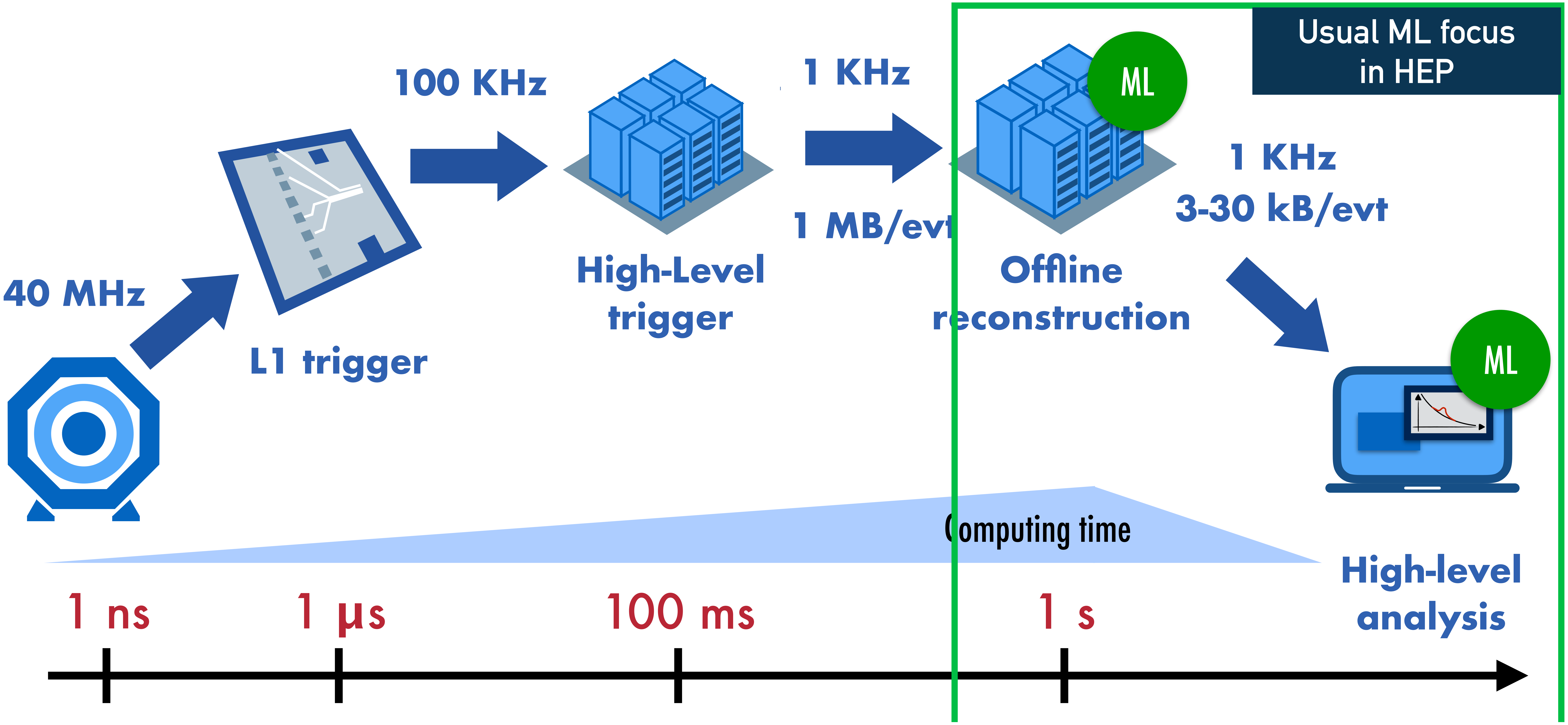
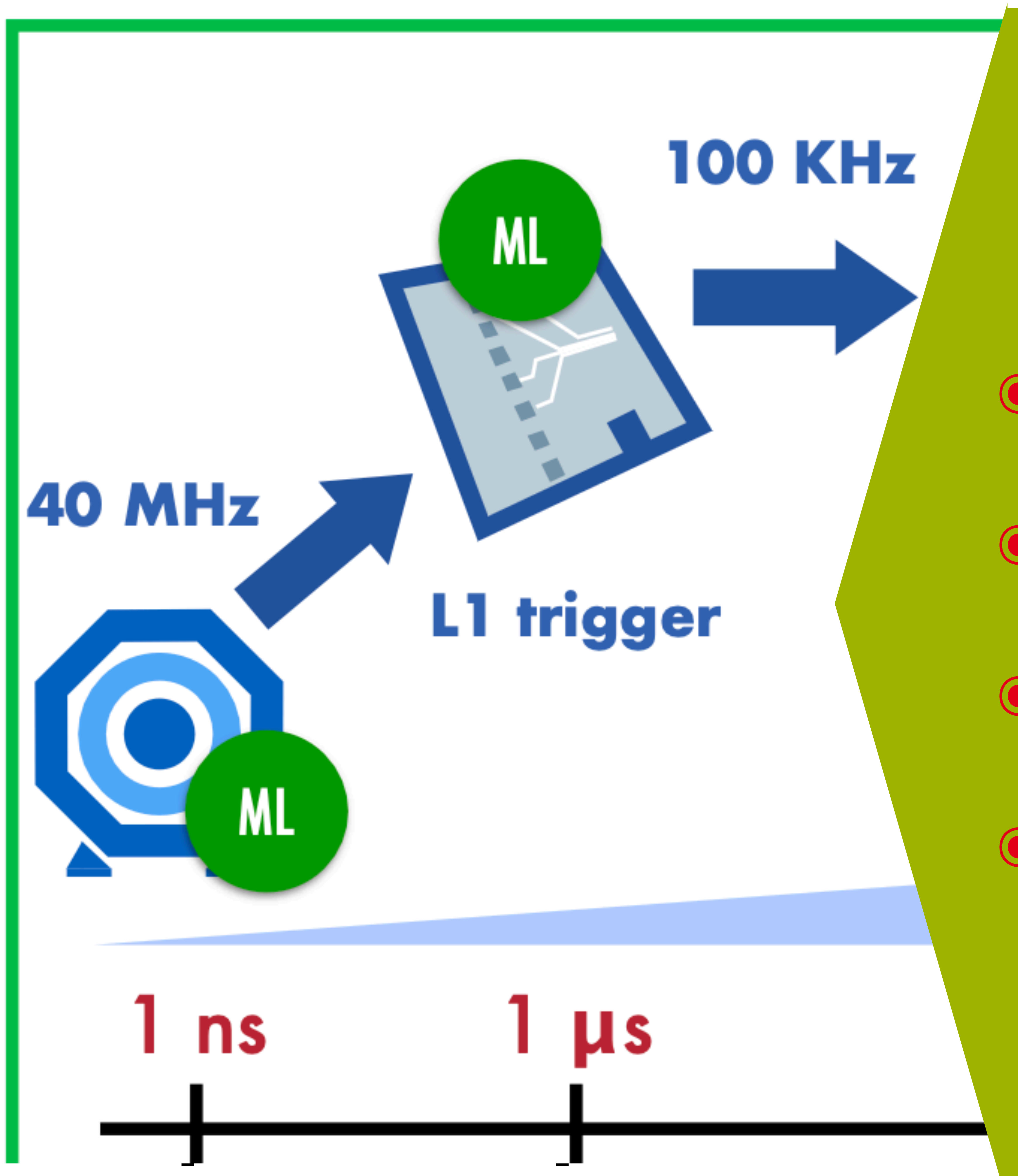


Image from Sioni Summers, Jennifer Ngadiuba et al.

ML in HEP Experiments



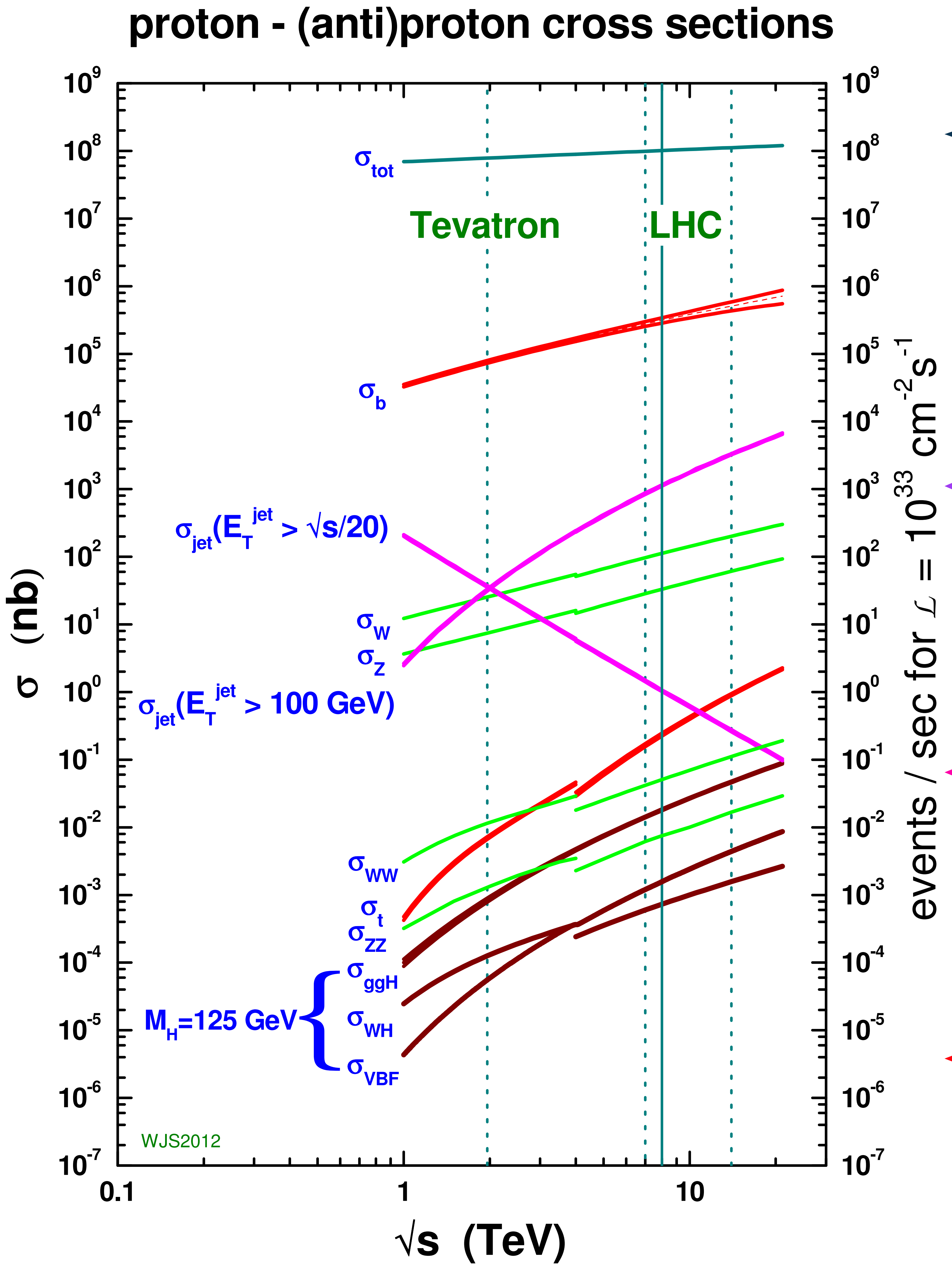
This talk!

- CMS Level-1 Trigger
- ML for Reconstruction
- ML for Event selection
- Other Fast ML: ASIC, Scouting

Event selection: Trigger

Searching for the needle in the LHC Haystack

Probability decreasing --->



LHC: 40 Million proton collisions per second

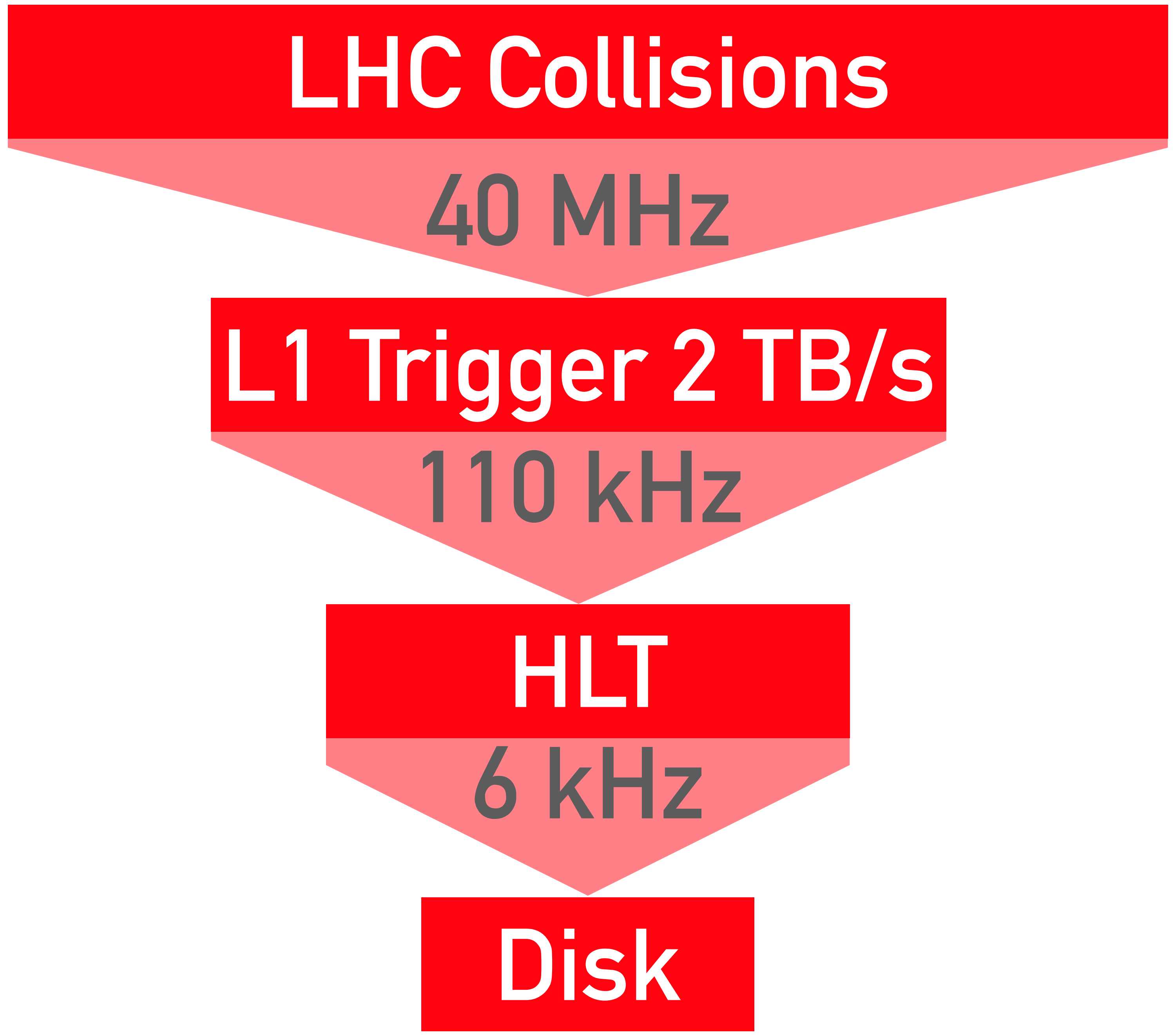
1000 W/Z bosons produced / second

1 Higgs boson is produced / second

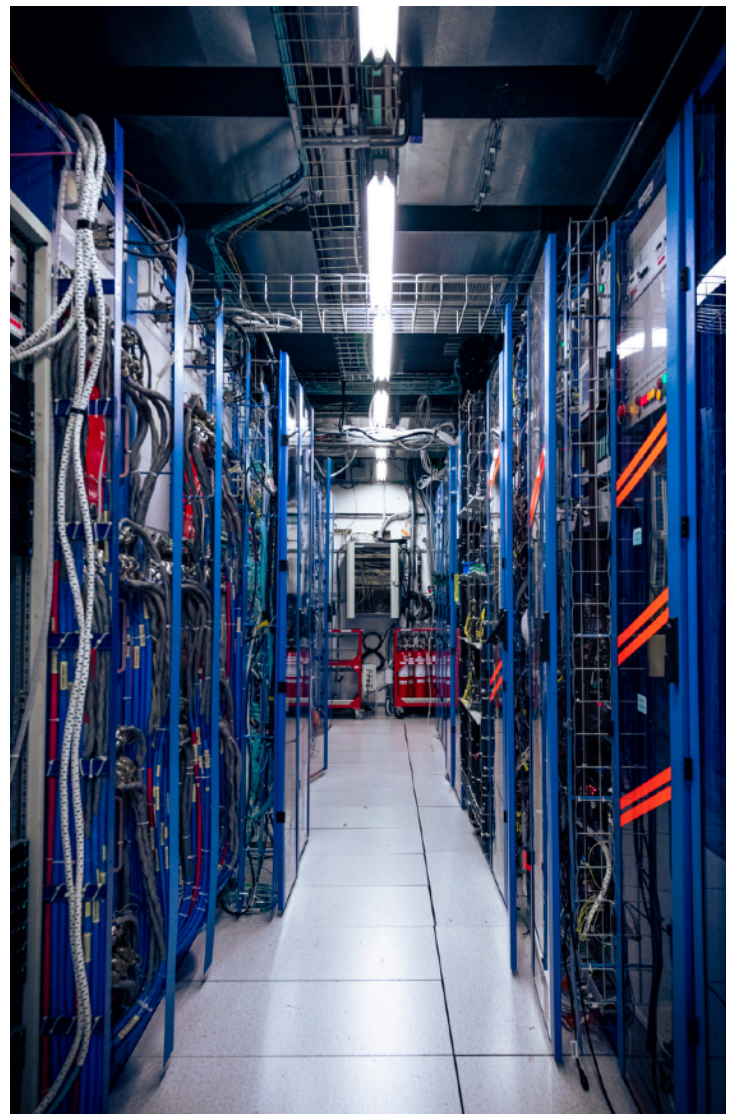
New physics (= Anomalies) hiding here?

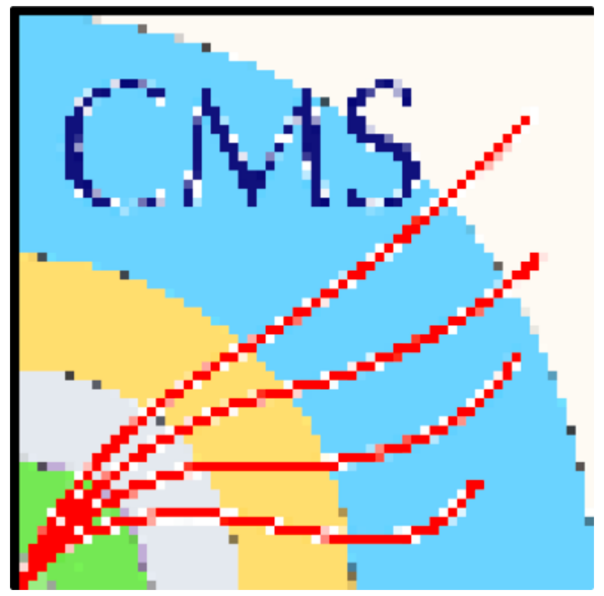
* LHC values from 2010 -> now higher luminosity

The CMS Trigger System



- Cannot record 40 MHz of collision data!
- **CMS** exploits **a two-level trigger (filter)**:
  1. **Level-1 Trigger** (L1T)
     - Implemented in **hardware** on **FPGAs**
     - Receives **coarse detector data**
     - **Decision within *O*(μs)**
  2. **High-Level Trigger** (HLT)
     - Uses **CPU/GPUs in a computing farm**
     - Full resolution of detector data
     - **Decision within < 1 second**





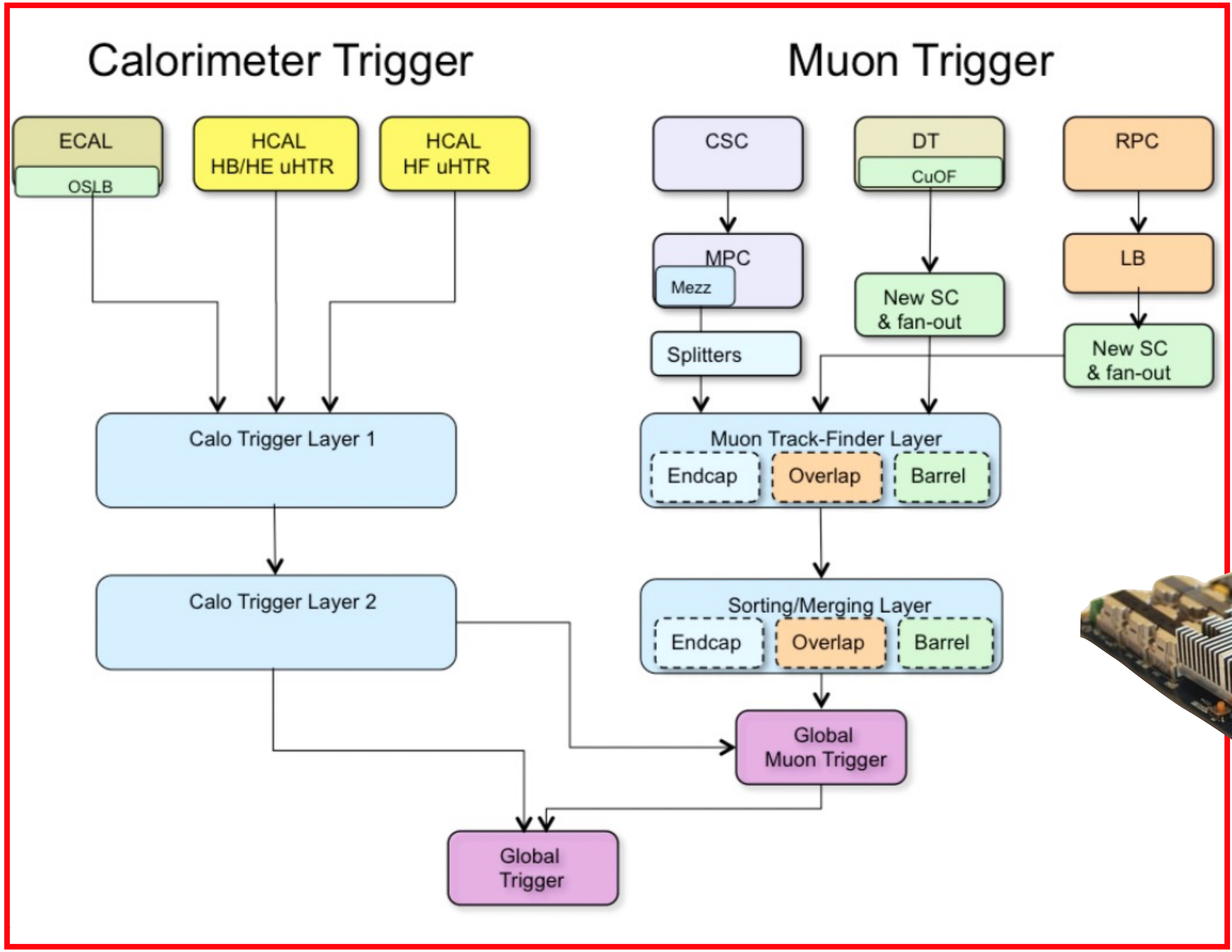
L1 vs HLT resolution

THE CURRENT CMS LEVEL-1 TRIGGER

Raw detector data "in"

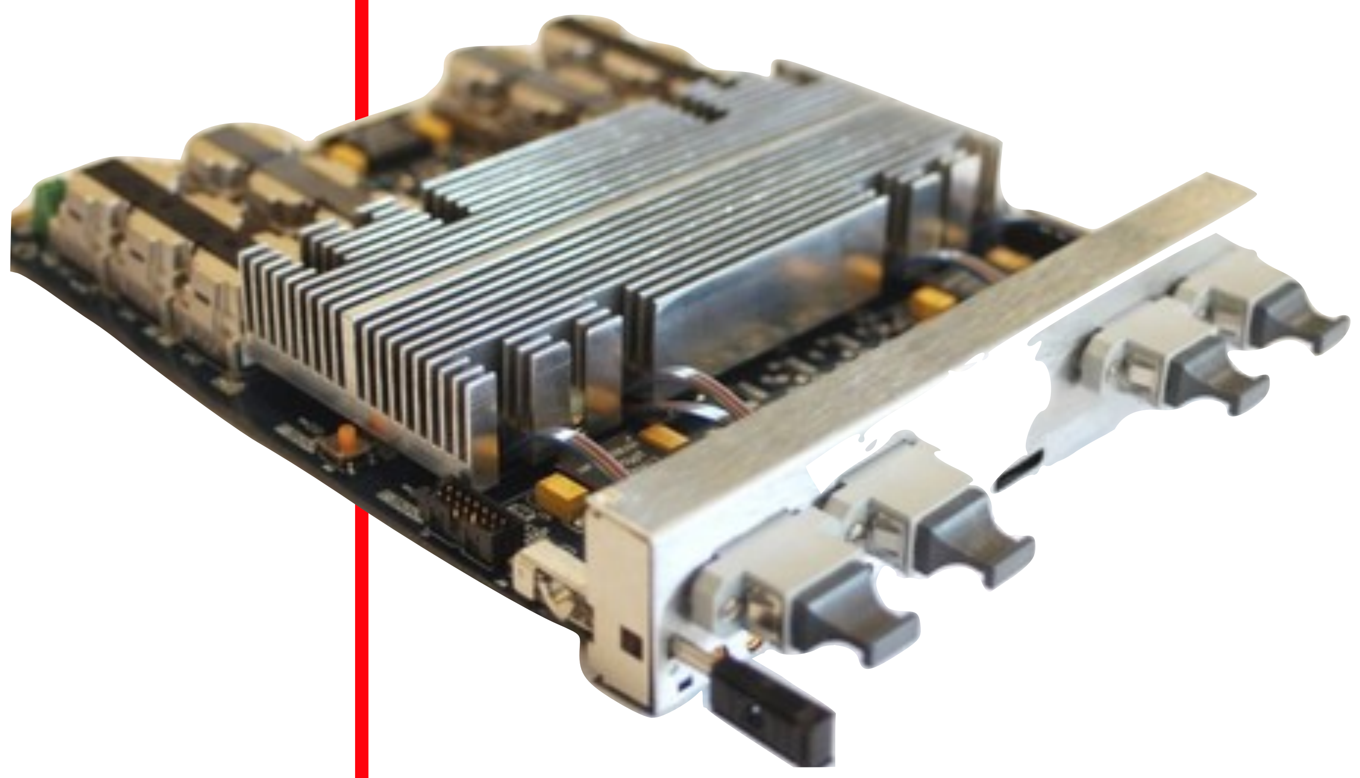
Processing data and reconstructing physics objects

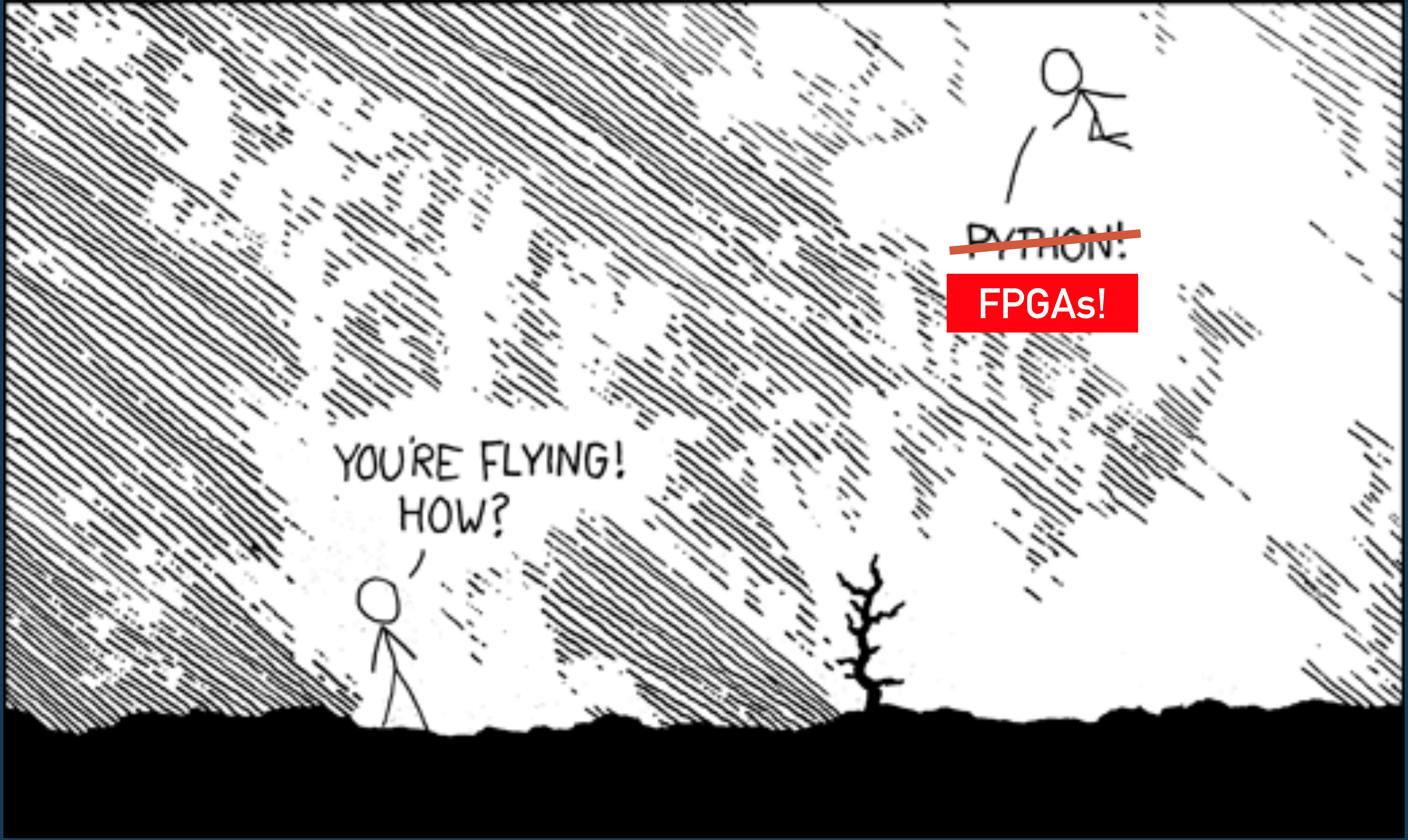
Final decision



Current FPGA Workhorse: Xilinx Virtex7

on custom host boards (MP7/FC7) in uTCA crates





xkcd “Python”

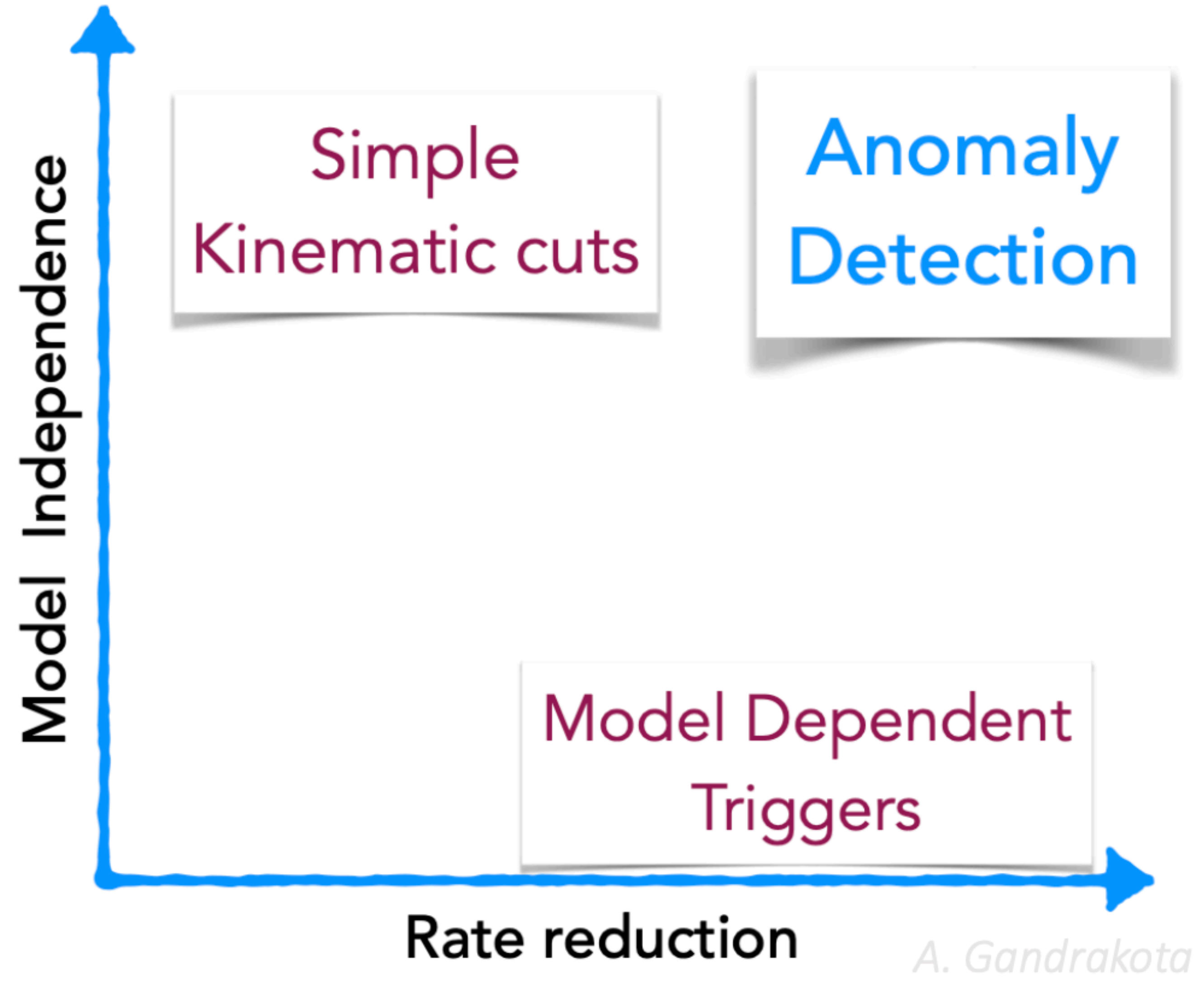
Anomaly detection @ CMS L1 Trigger

Anomaly detection in CMS

- Searching for new physics at the LHC – multiple fronts:
  - **Direct**: e.g. looking for exotic particles (peak or excess searches)
  - **Indirect**: precision measurements of particle parameters (e.g. H couplings)
  - **Anomaly detection** *using recorded data* (*examples at this conference*)

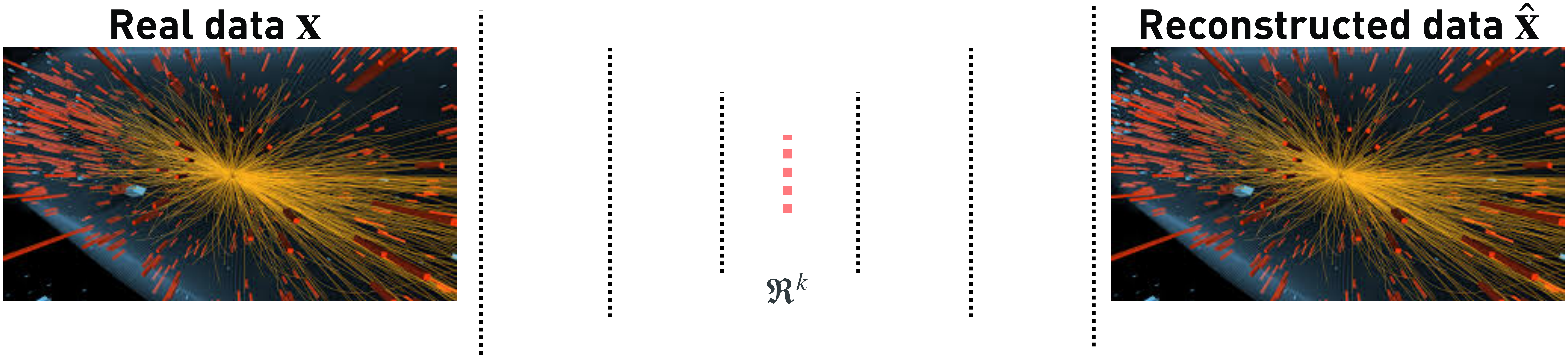
- All rely on existing selection (trigger) algorithms
  –> Model dependent or high energy thresholds

- **What if anomalous collisions are NOT RECORDED?**
  –> Anomaly detection at trigger level!



ANOMALY DETECTION WITH AUTO-ENCODERS

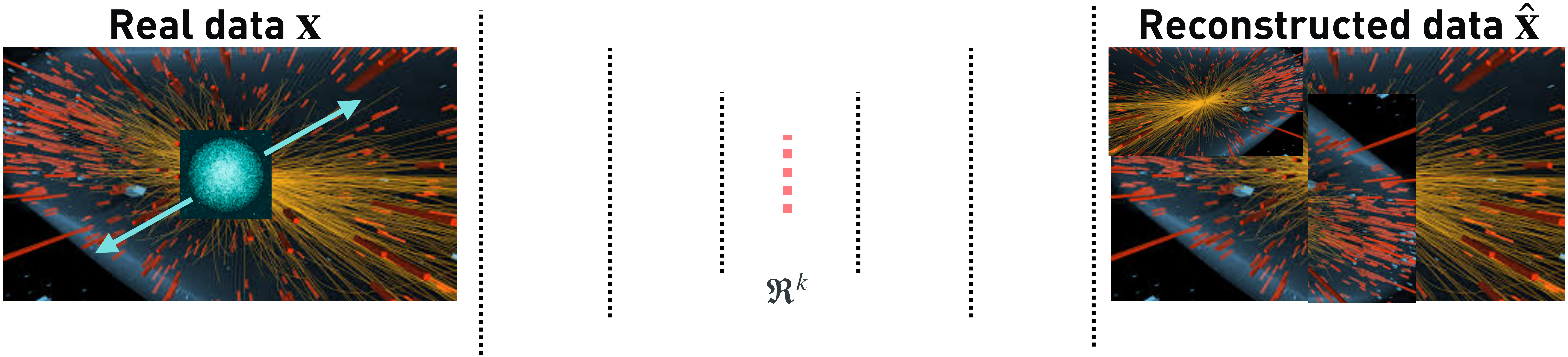
- **Autoencoders train unsupervised on data**
  - Learn to compress and to reconstruct the data
  - Difference $\hat{x} - x$ = "degree of abnormality"



Anomaly detection with auto-encoders

- **Autoencoders train unsupervised on data**
  - Learn to compress and to reconstruct the data
  - Difference $\hat{x} - x$ = "degree of abnormality"

➤ **If trained on "background" –> "signal" is anomalous!**

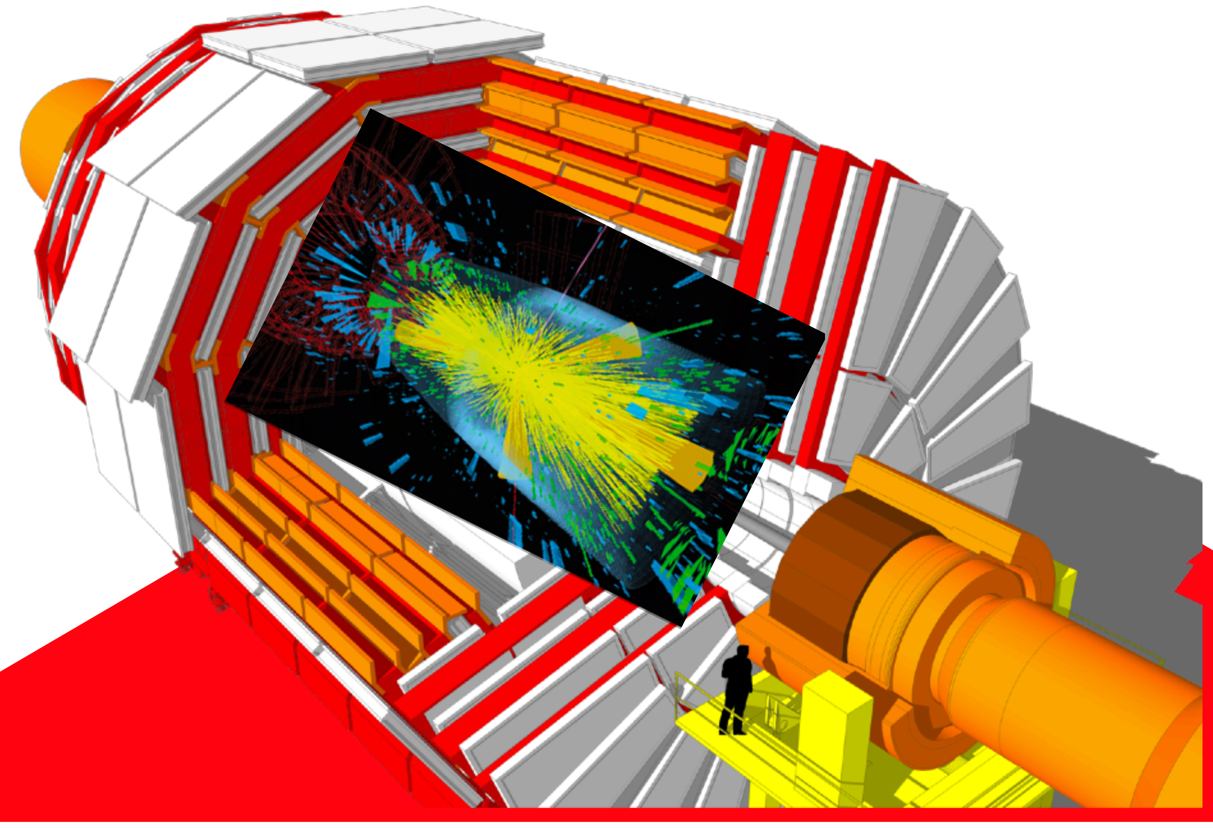


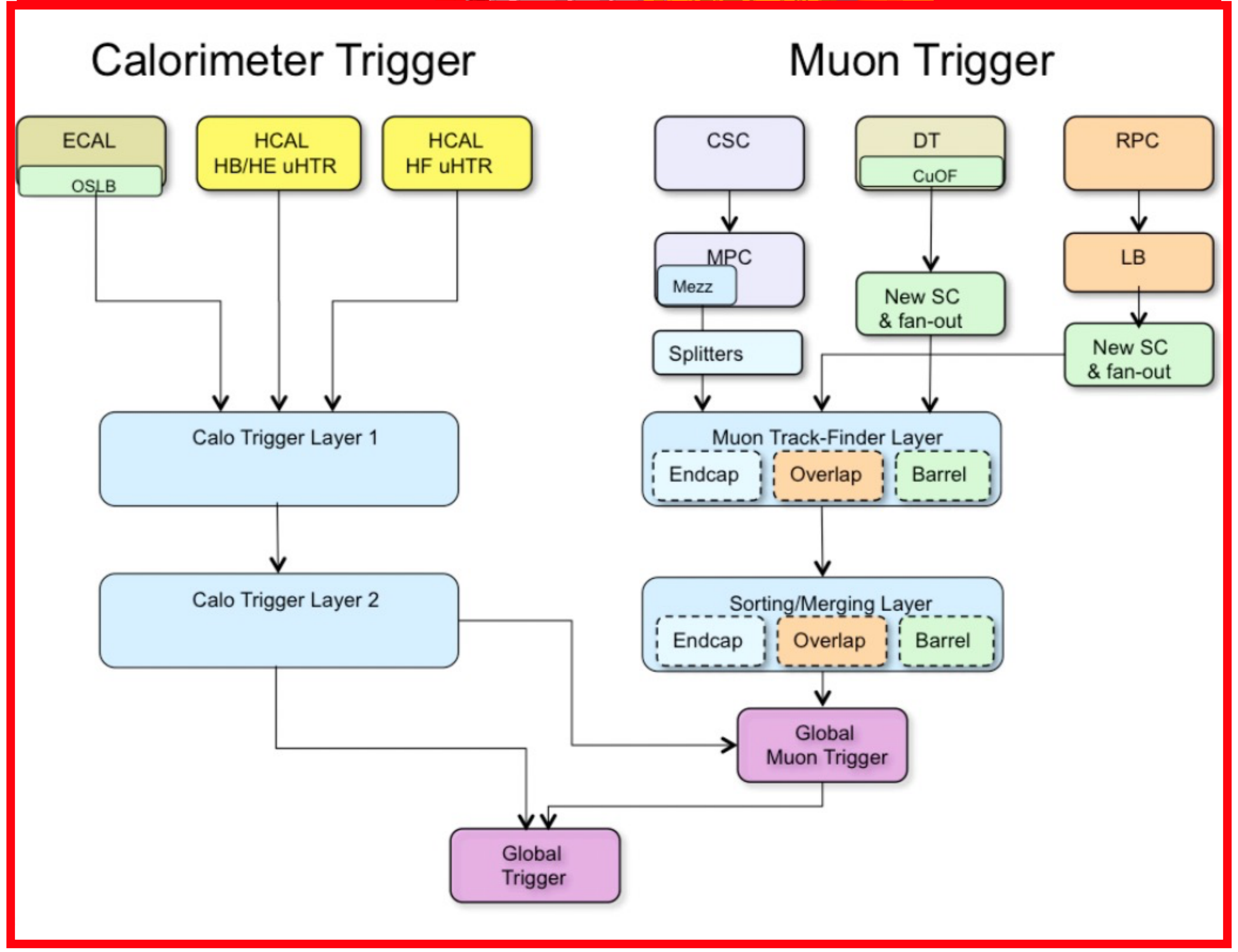
ANOMALY DETECTION @ CMS LEVEL-1 TRIGGER

Raw detector data "in"

Raw detector images: CICADA

Reconstructed objects: AXOL1TL

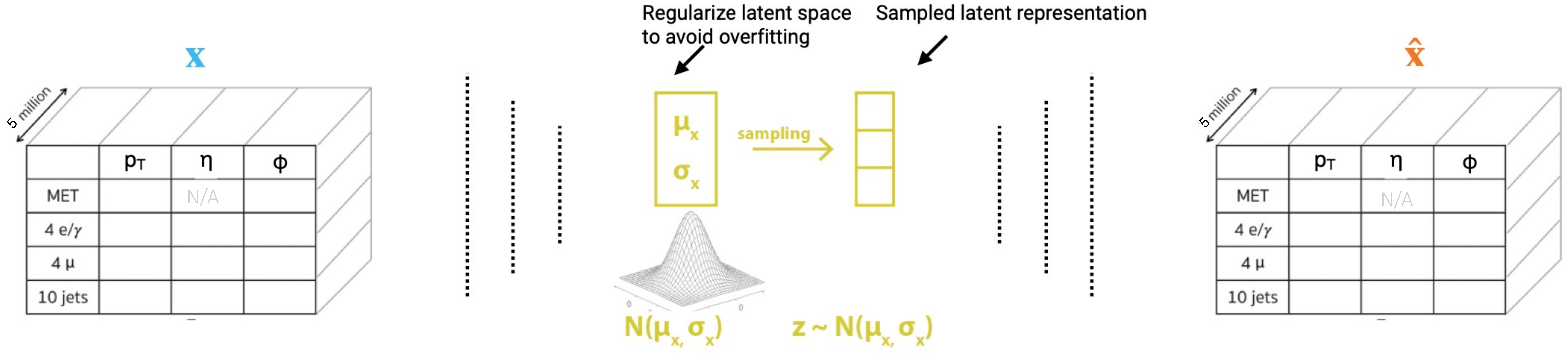




High-level inputs: AXOL1TL

AXOL1TL: ANOMALY DETECTION WITH OBJECT TOPOLOGY

- AXOL1TL (**A**nomaly e**X**traction **O**nline **L**evel-**1** **T**rigger a**L**gorithm) **is a variational auto-encoder:**
  - Encodes input as a distribution over the latent space
  - Add regularisation term in loss: KL divergence, how different is distribution from Gaussian
- **Inputs: L1 trigger objects 4-vectors** (pT, η, ϕ)
  - Most energetic 4 electron/photons, 4 muons, 10 jets and missing transverse energy (MET)



$$\text{loss} = ||x - \hat{x}||^2 + \text{KL}[\,N(\mu_x,\sigma_x), N(0, I)\,]$$

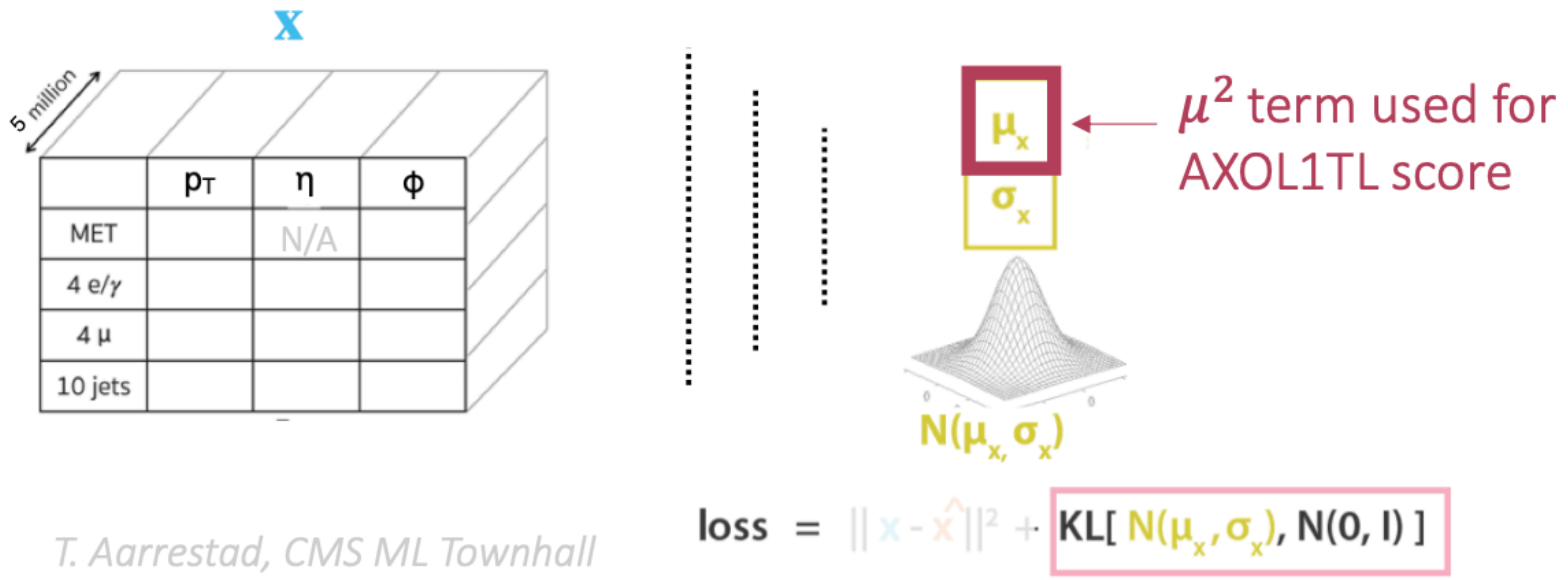
CMS-DP-2023-079

T. Aarrestad

AXOL1TL: ARCHITECTURE OPTIMISATION

CMS-DP-2023-079

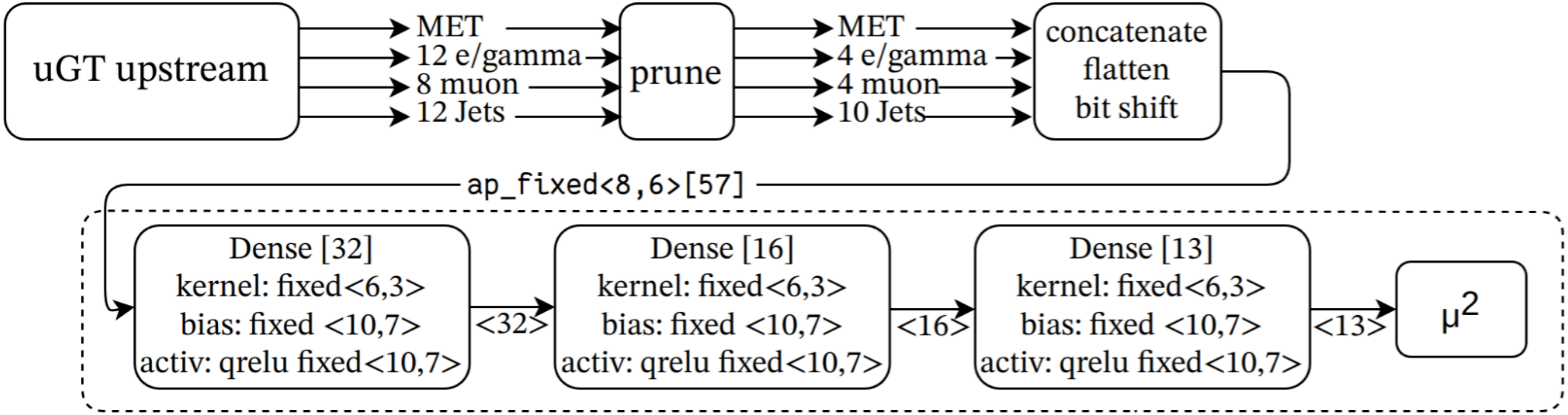
- Full NN architecture does not fit the L1/FPGA constraints
  - **–> only use encoder half of the network**
    - Compute degree of abnormality from latent space directly
    - No need to use inputs for anomaly score computation
    - **Half network size and latency!**



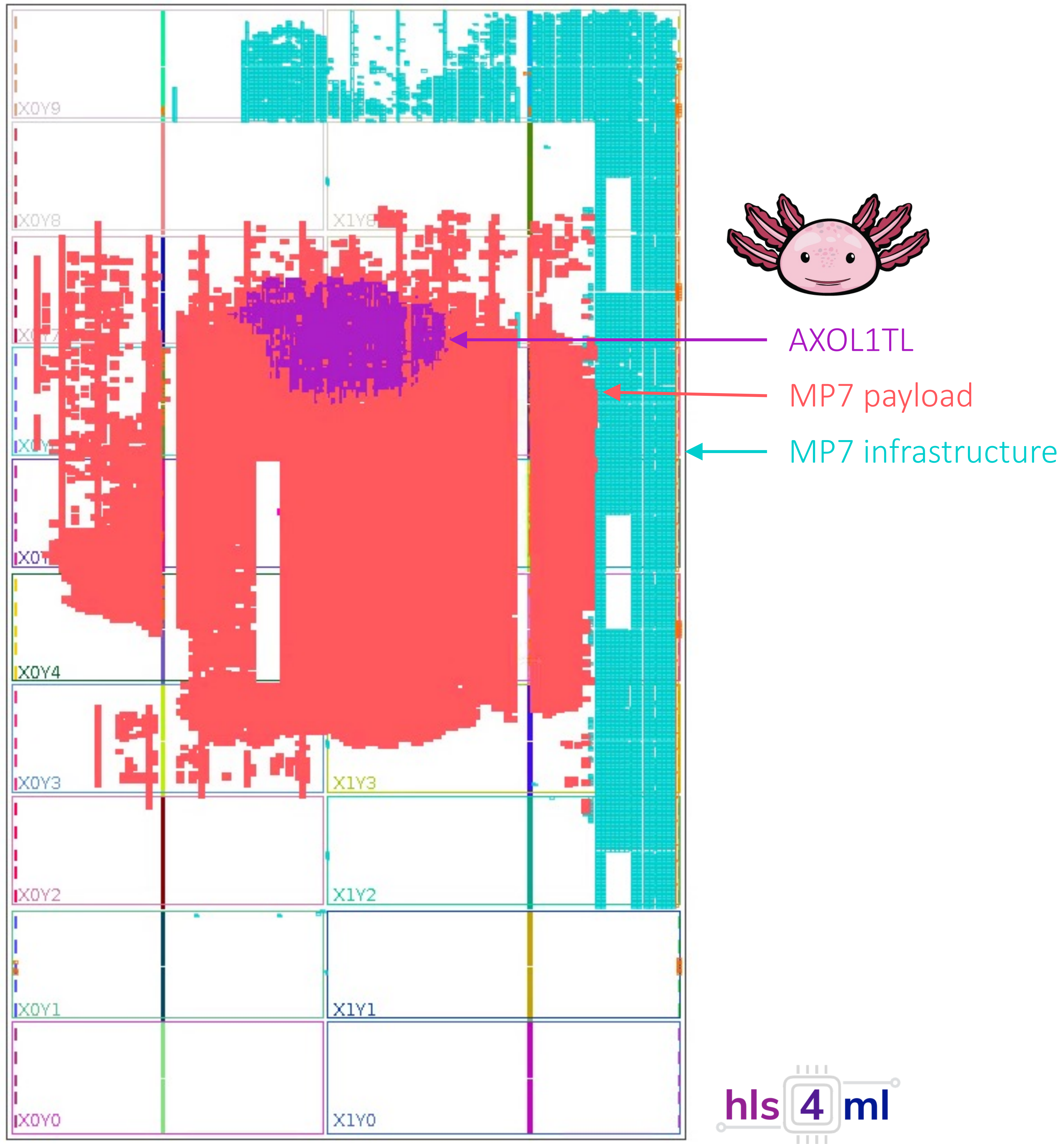
T. Aarrestad, CMS ML Townhall

AXOL1TL: COMPRESSION

- **Quantization-aware training with QKeras and FPGA adaptation with HLS4ML**
  - Narrow, shallow model, aggressively quantised
- Output is one vector [13,1], corresponding to μ part of [μ,σ] KL loss (dropping σ as it is small -> reduces processing time)
- **Anomaly score: sum squared of the μ vector**



AXOL1TL: FPGA Implementation



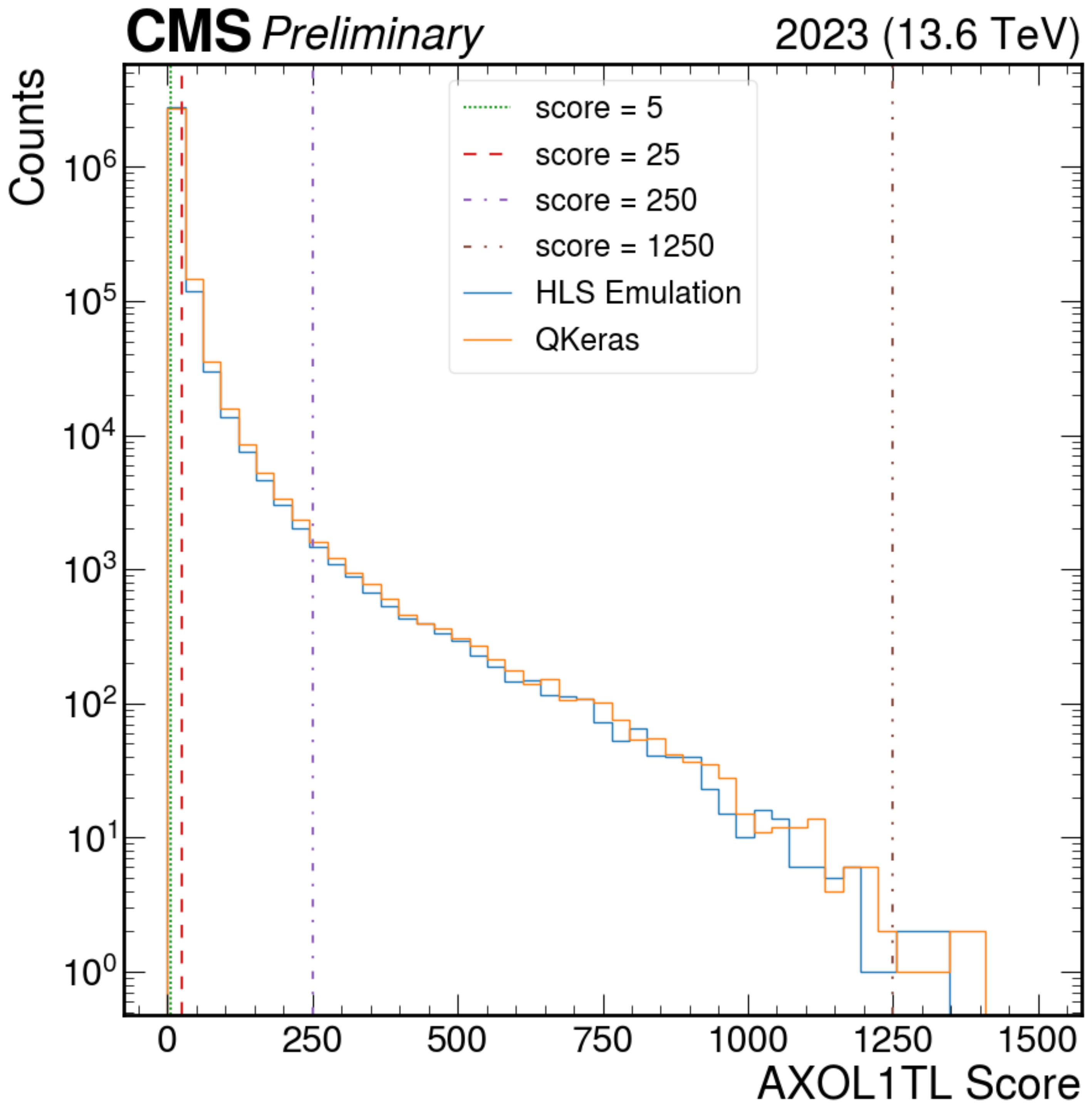
- Implemented on Xilinx Virtex-7 XCVU9P FPGA
- **Met requirements on latency and resources**
  - **50 ns latency & ~1% resources**

Resource utilization of Virtex-7 FPGA chip on Imperial College MP7 µGT board

	Latency	LUTs	FFs	DSPs	BRAMs
AXOL1TL	2 ticks 50 ns	2.1%	~0	0	0

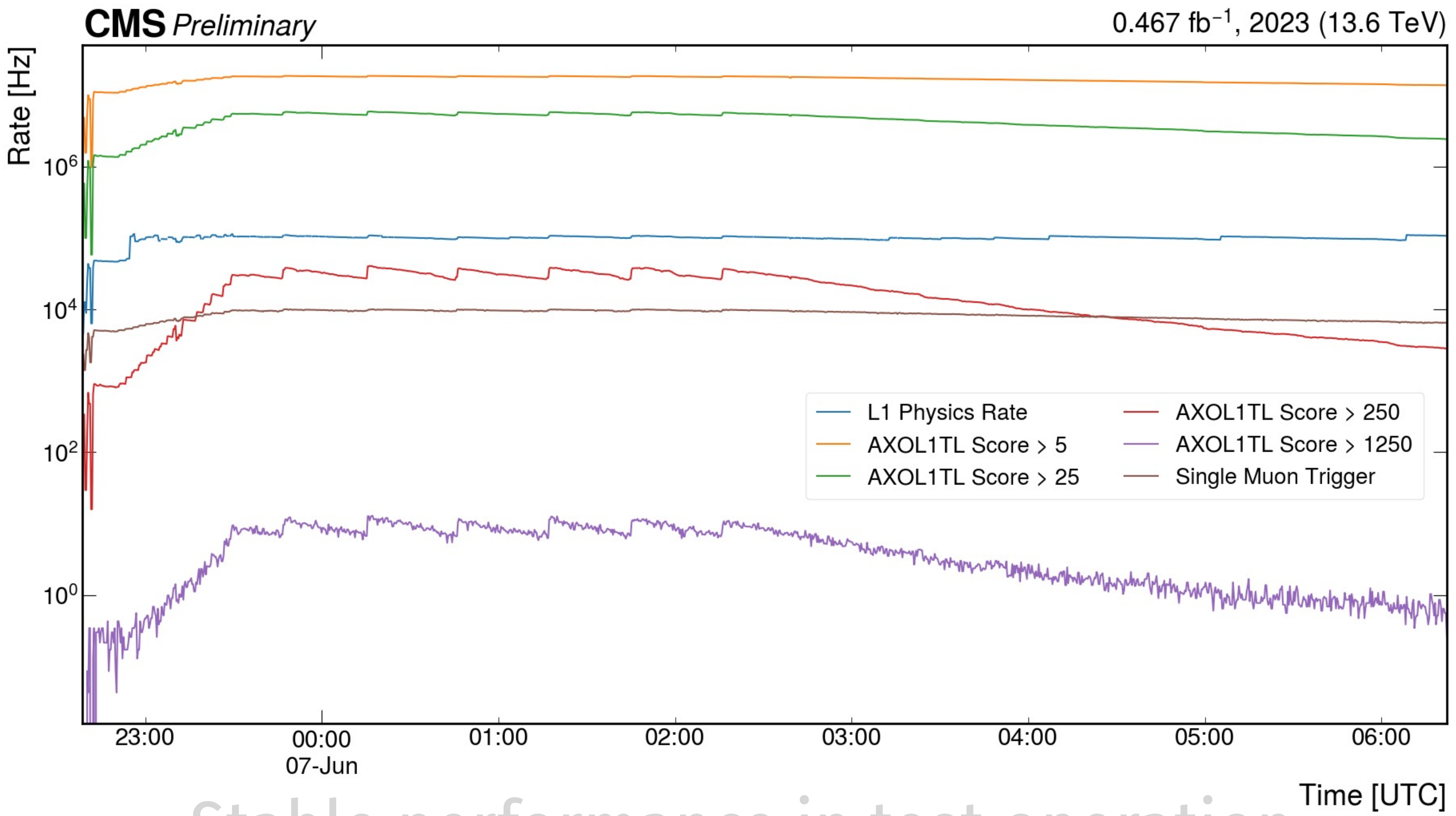
CMS-DP-2023-079

AXOL1TL: Commissioning



Anomaly score distribution for unbiased (random) LHC collision data

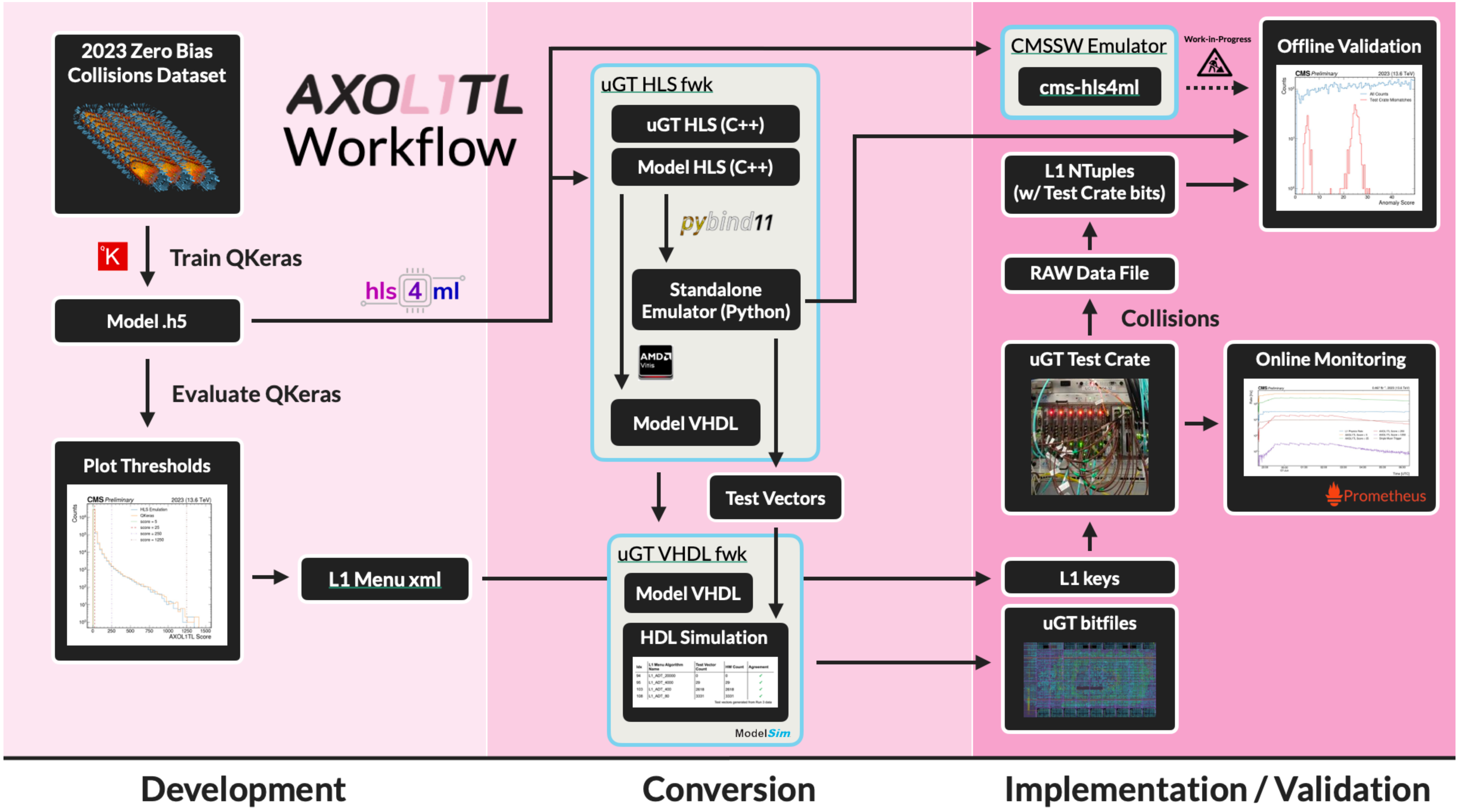
- AXOL1TL is trained with unbiased data collected by CMS during 2023 with √s=13.6 TeV
  - 10.5 million events (50/50% for training/testing)
  - Selected 5 test scores in firmware
- **Commissioned in Global Trigger Test Crate during proton collisions** in 2023 –> stable as standard triggers



Stable performance in test operation

CMS-DP-2023-079

AXOL1TL: Pipeline



AXOL1TL: Firmware Validation with Test bench

- Trigger bits for the L1 menu including 4 anomaly detection thresholds: scores >1250, >250, >25, and >5 from top to bottom
- Test vector column: generated from inference results of a standalone C++ emulator
- HW count: comes from standard global trigger firmware simulation workflow using ModelSim. Perfect bit agreement is observed

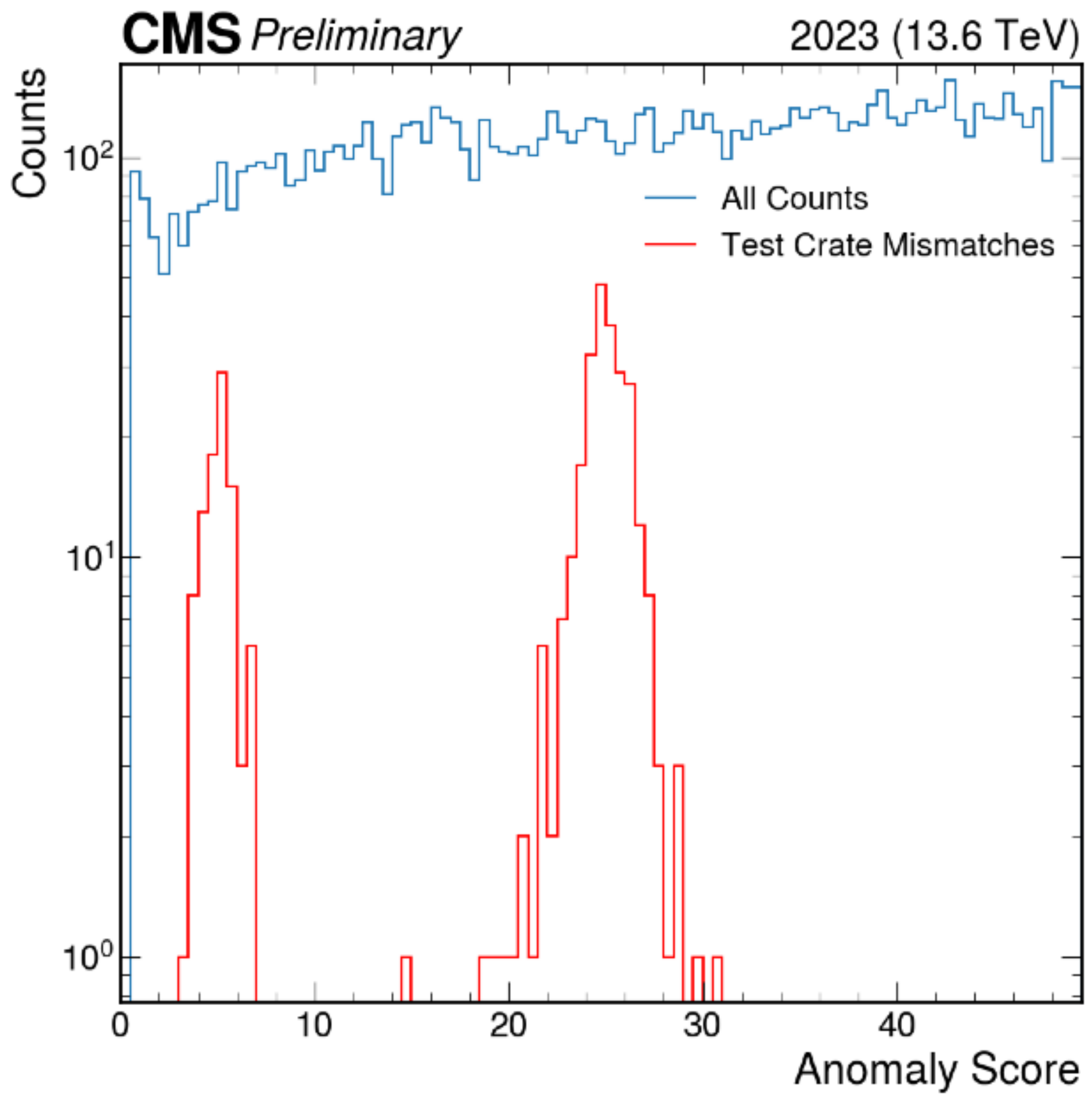
Idx	L1 Menu Algorithm Name	Test Vector Count	HW Count	Agreement
94	L1_ADT_20000	0	0	✓
95	L1_ADT_4000	29	29	✓
103	L1_ADT_400	2618	2618	✓
108	L1_ADT_80	3331	3331	✓

Test vectors generated from Run 3 data

AXOL1TL: FW VALIDATION

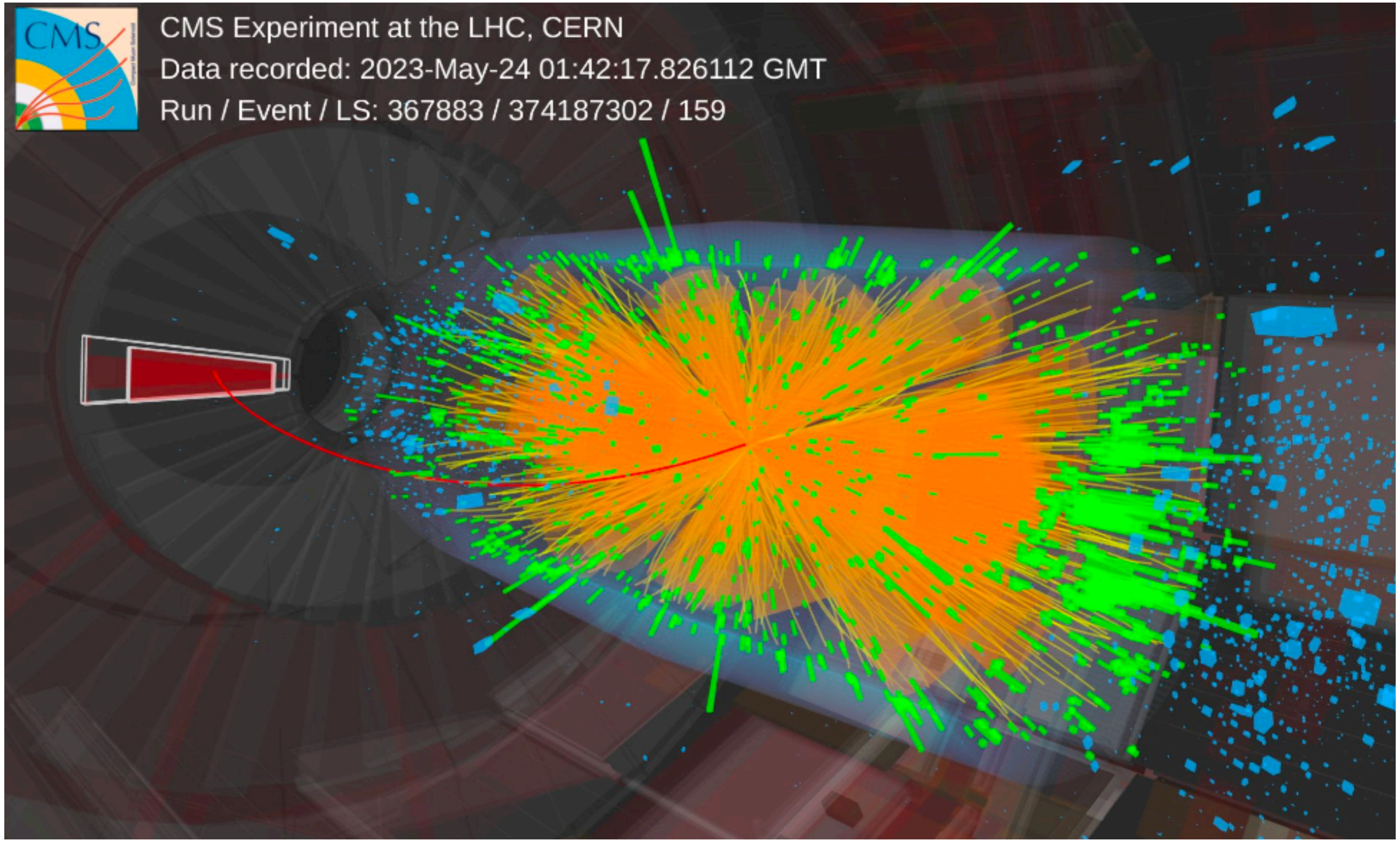
Test Crate Validation

L1 Menu Algorithm Name	Test Crate Count	Standalone Emulator Count	Mismatches
L1_ADT_20000	1	1	0
L1_ADT_4000	742	741	19
L1_ADT_400	21236	21229	253
L1_ADT_80	25468	25481	93



Anomaly Detection hardware vs. emulation trigger mismatches. Events from promptly reconstructed 2023 Ephemeral ZeroBias data where hardware bits are recorded from configured µGT test crate. In table (left), Test Crate Count shows events triggered in hardware and read out into data and Standalone Emulator Count is evaluated via offline inference with L1 objects. Anomaly score distribution of all events (right): red segments represent mismatches between hardware and emulation. Clustering near decision boundaries implies issue is due to precision/rounding problem. Minimal mismatches in hardware vs. emulation (≤ 1%) observed.

AXOL1TL: EVENT DISPLAY



- Example of an anomalous event during 2023 pp collisions (from random trigger dataset)
  - **Highest anomaly score event not triggered by L1**
- L1 objects:
  - 11 jets with pT > 20 Gev
- Offline objects:
  - 7 jets with pT > 15 GeV from the same vertex
  - 75 identified vertices

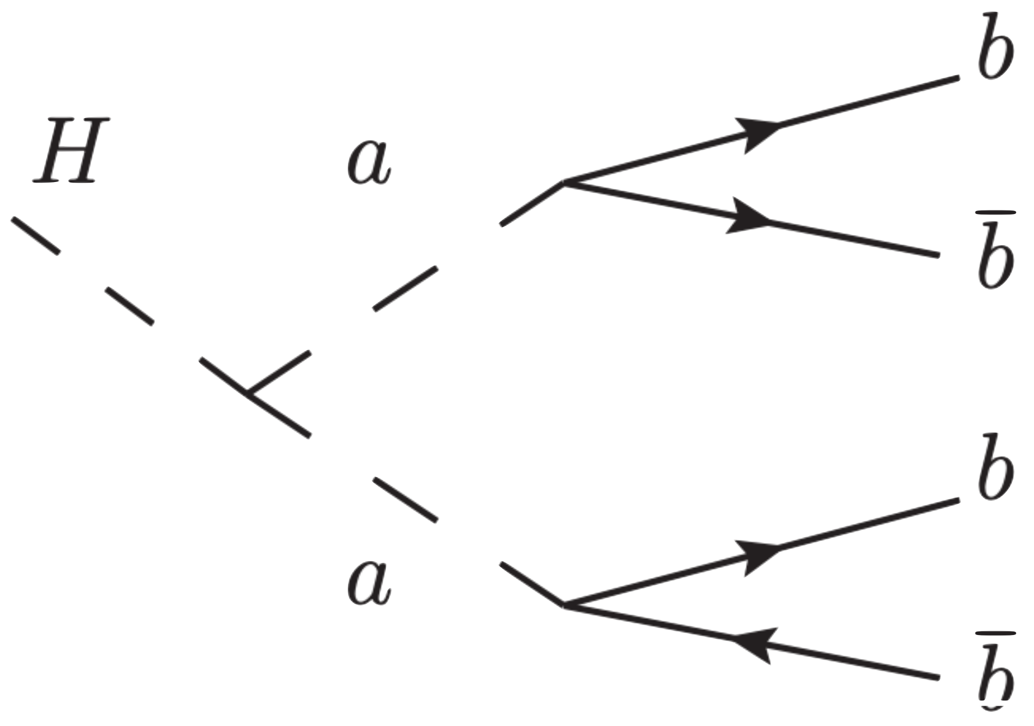
CMS-DP-2023-079

AXOL1TL: Physics performance

- Use **simulated hypothetical exotic signal as a anomaly candidate**
- Significant **performance improvement on various SM and beyond the SM signals** by adding AXOL1TL to the 2023 trigger menu

$$\text{Improvement} = \frac{\text{L1 Efficiency w/ AXOL1TL@freq}}{\text{L1 Efficiency w/o AXOL1TL}} - 1$$

- Example performance improvement for H->aa[15 GeV]->4b signal:



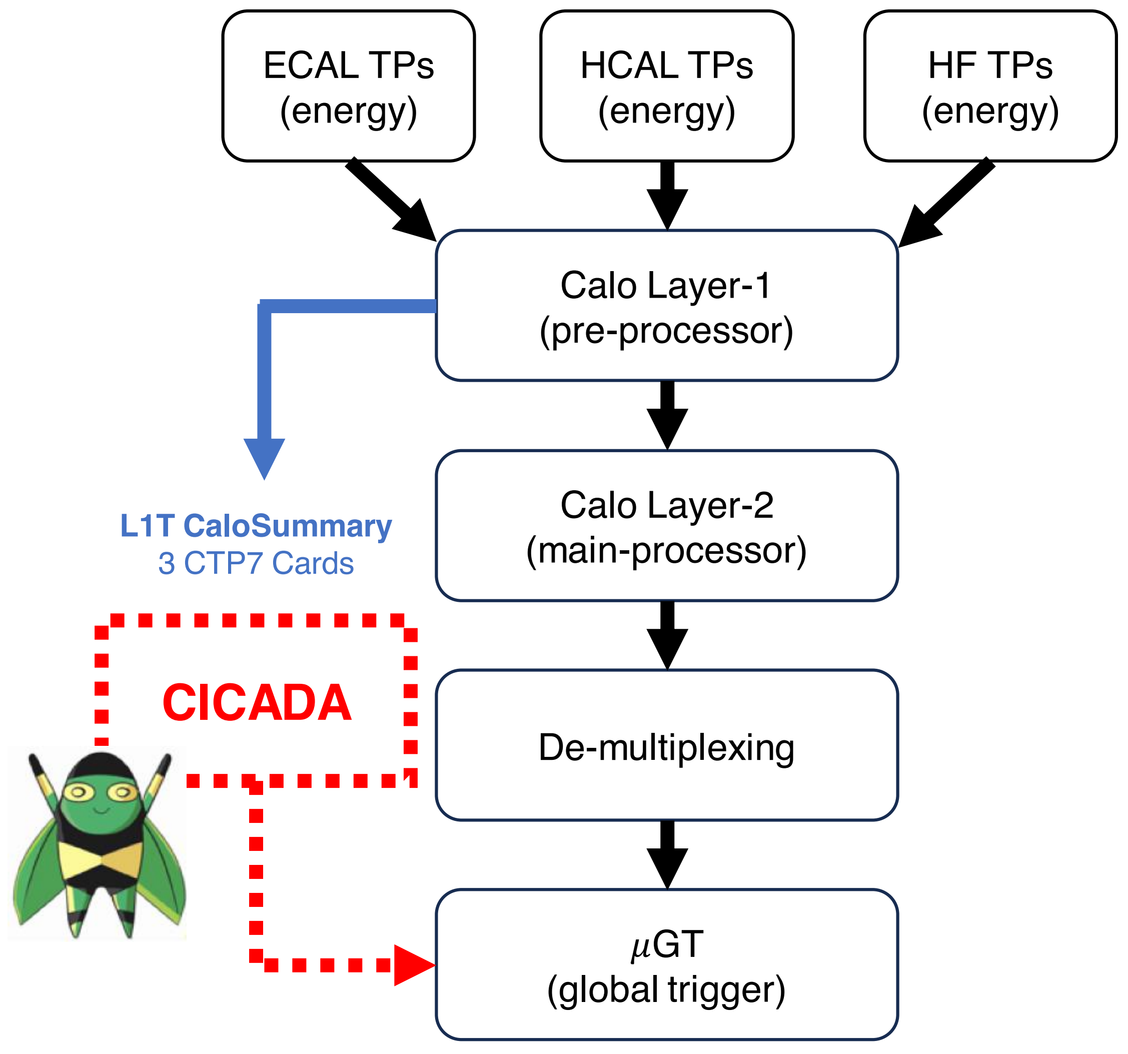
AXOL1TL Rate	1 kHz	5 kHz	10 kHz
Signal Efficiency Gain	46%	100%	133%

- **Starting data-taking with ~*O*(100) Hz L1 rate in 2024 pp collisions soon!**

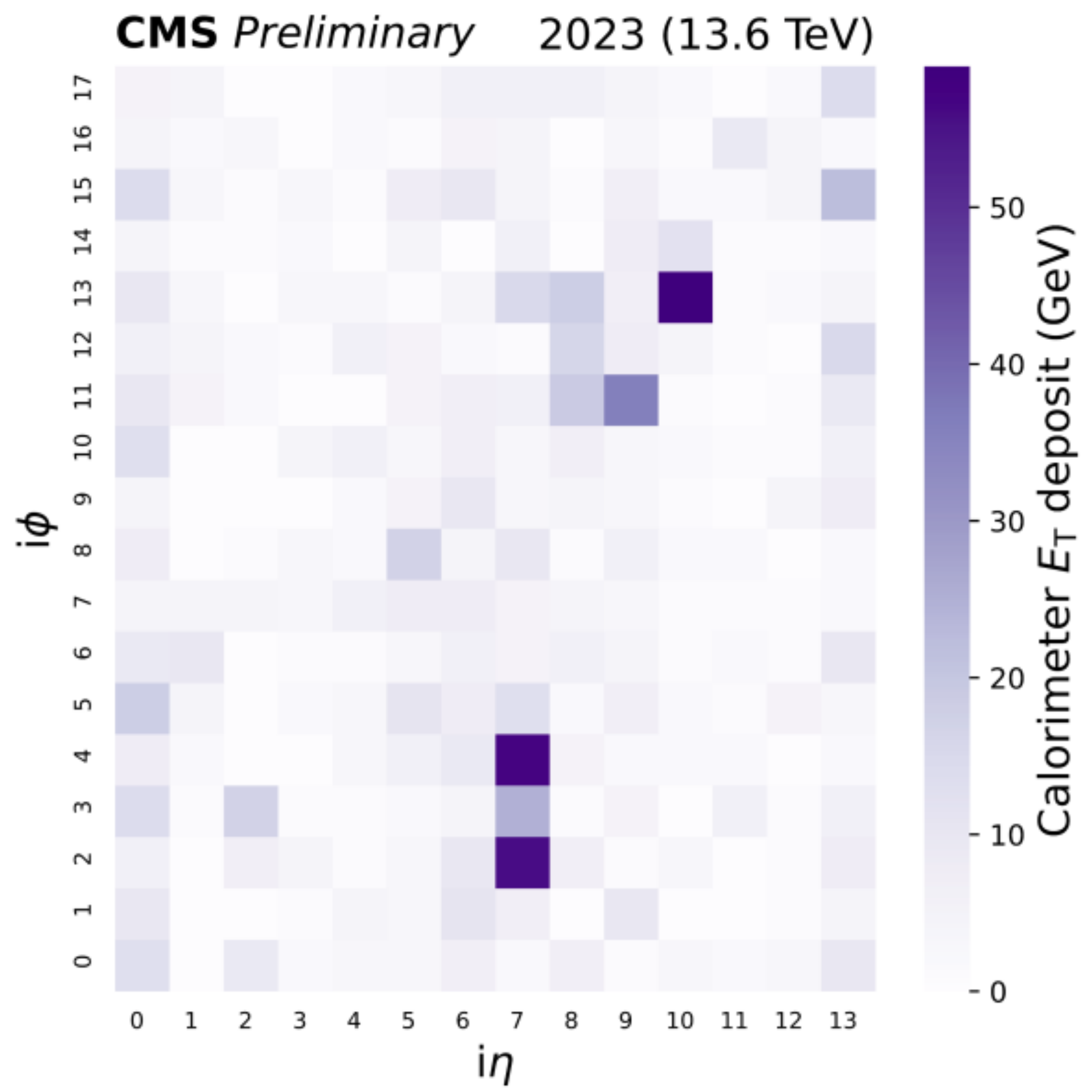
CMS-DP-2023-079

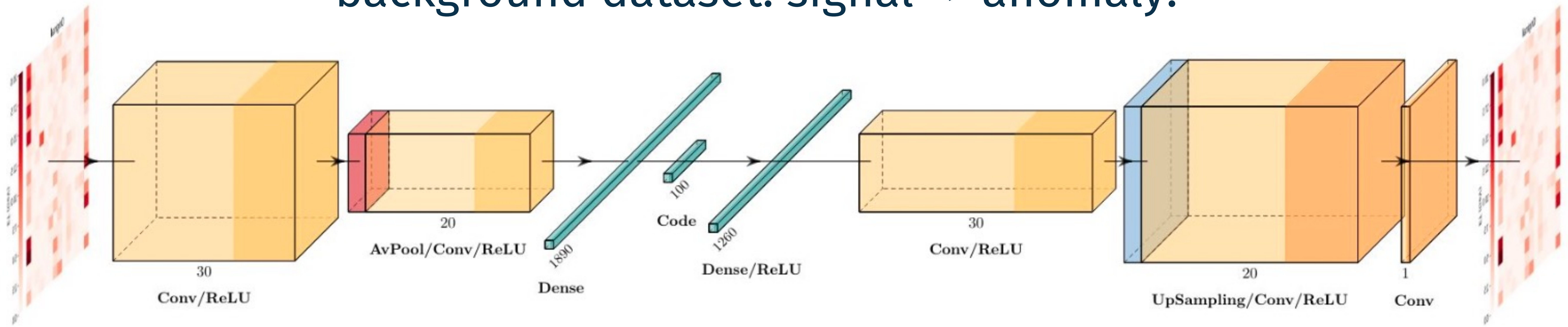
RAW FEATURES: CICADA

CICADA: Anomaly Trigger on Raw inputs



- CICADA (CMS DP-2023/086): **C**alorimeter **I**mage **C**onvolutional **A**nomaly **D**etection **A**lgorithm
- **Using raw inputs of calorimeter:**
  - Image of 18 x 14 energy deposits
  - **Independent of domain knowledge** (standard trigger algorithms)
- Convolutional auto-encoder trained on background dataset: signal -> anomaly!

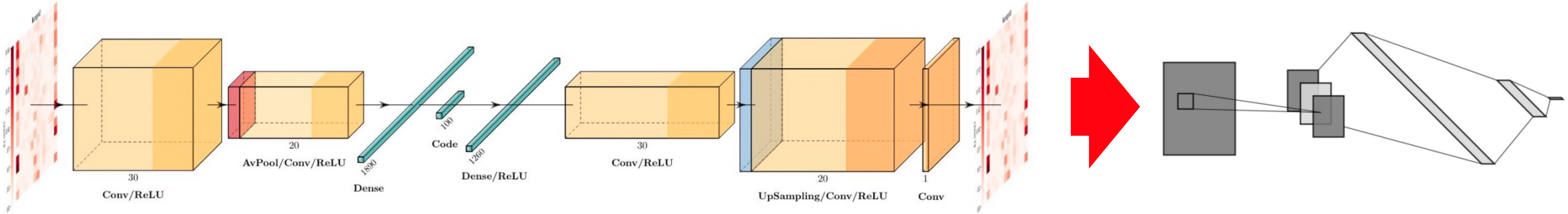




CMS DP-2023/086

CICADA: Knowledge Distillation

- Full CICADA model is too complex for FPGA resources / L1 Trigger requirements
  –> **use Student-Teacher Knowledge Distillation**
  - **Teacher model**: complete encoding and decoding of the original input data
    - **Anomaly score (reconstruction error)**: average of the squared error (predicted – input) in reconstruction for each of the 252 individual energy deposits (Mean Squared Error)
  - **Student model: regresses the anomaly score of the teacher model**
    - Smaller convolutional layer with only 4 filters + few dense layers
      -> 10x faster & less resources -> fits FPGA/L1T requirements

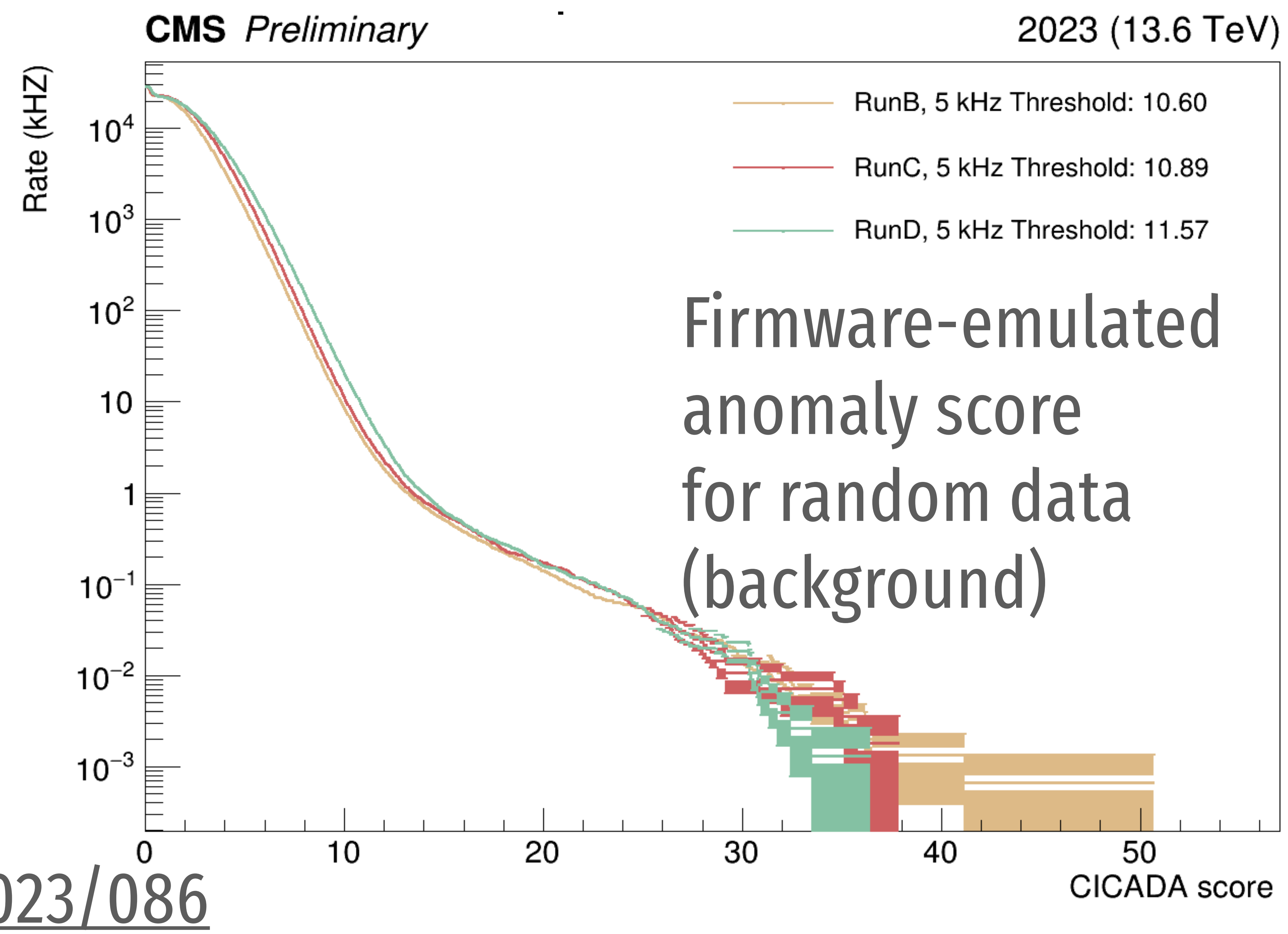


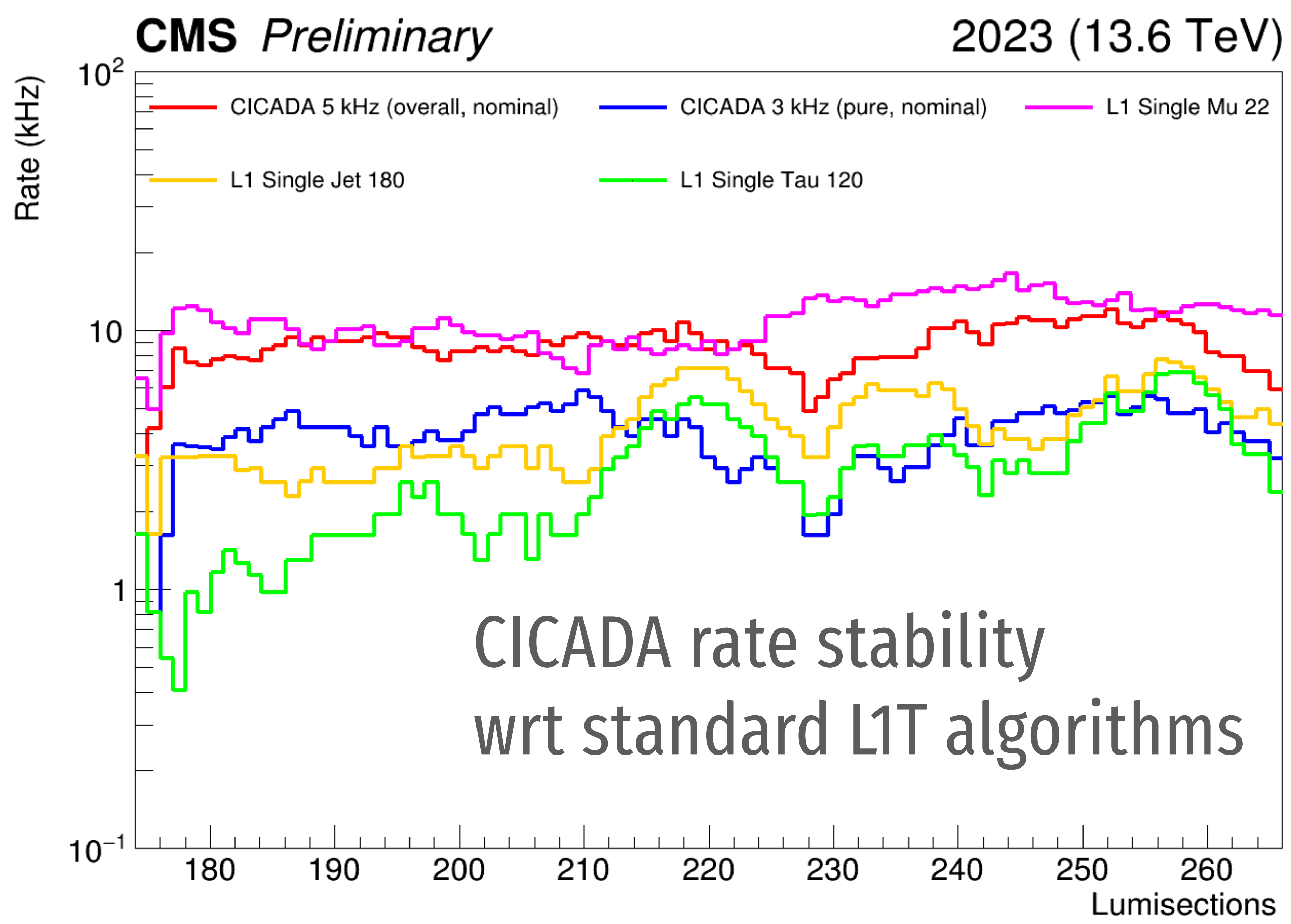
Teacher Model Student Model

CMS DP-2023/086

CICADA: Commissioning

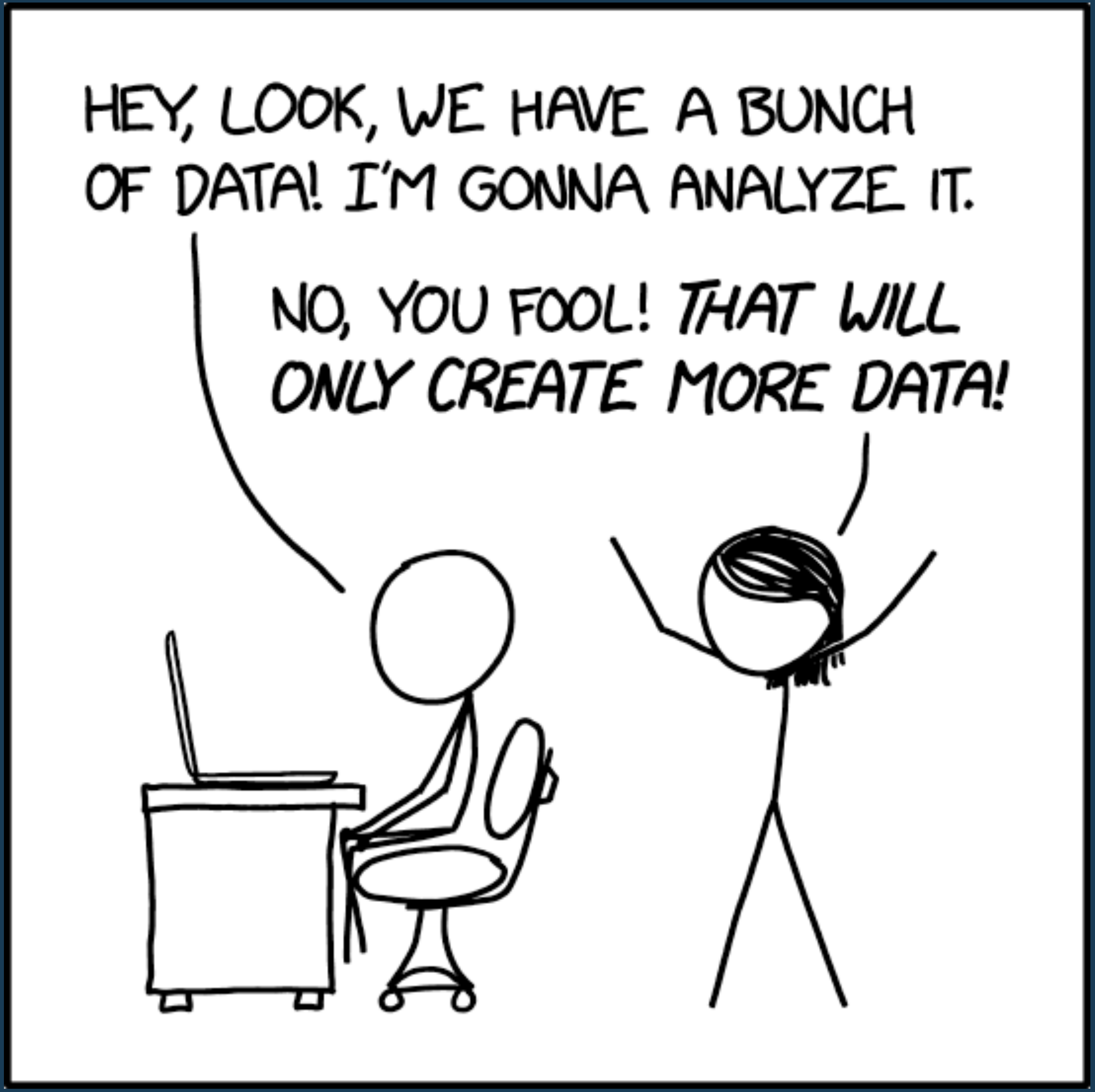
- **CICADA currently being commissioned in the L1 Trigger test system**
  - Software-based emulation based on Firmware (HLS4ML) and validated
  - Preliminary performance estimates promising + operational stability tested
- **This is the first anomaly detection on low-level inputs in a LHC trigger system!**





CMS DP-2023/086

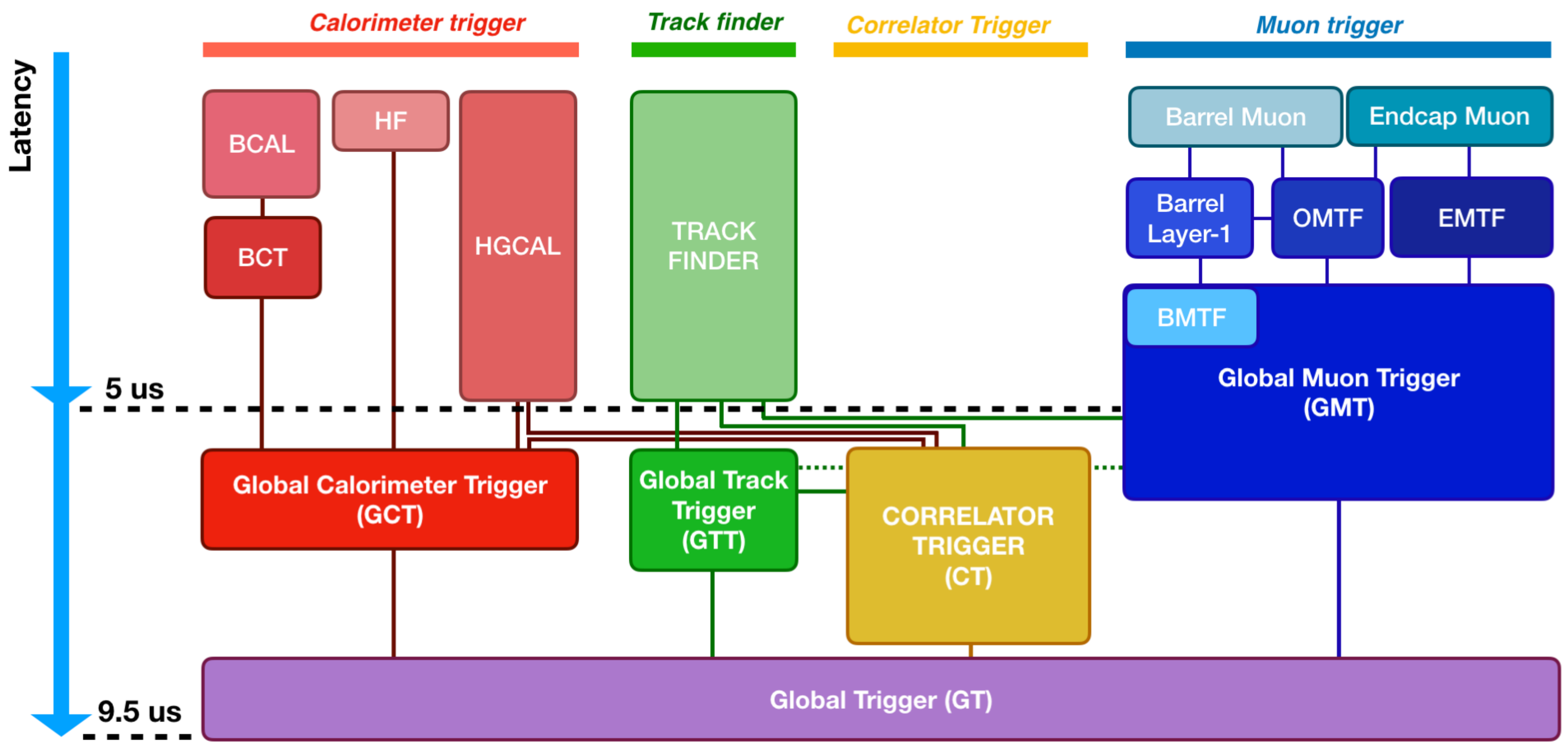
Towards the High-Luminosity LHC



xkcd “Data Trap”

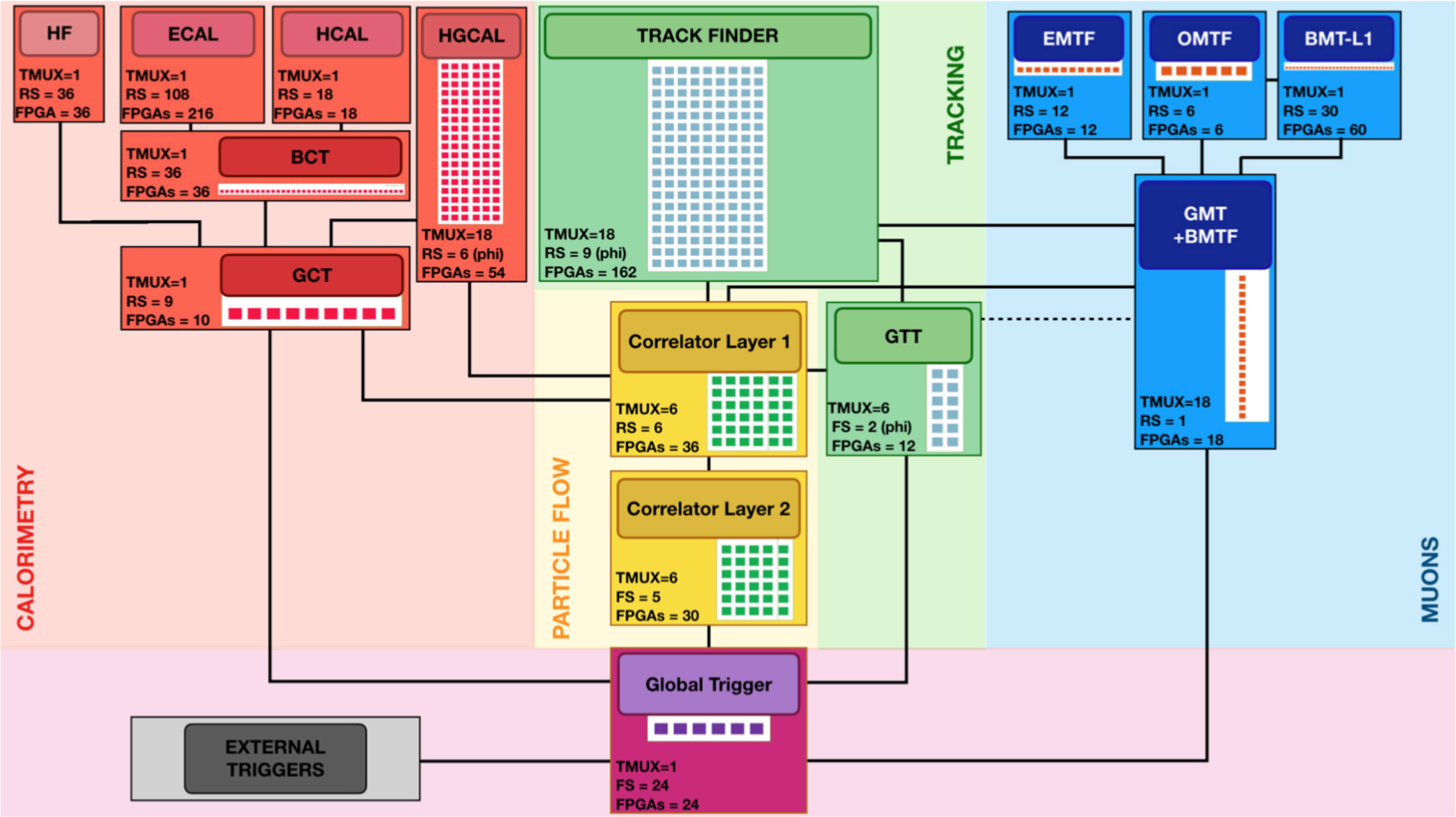
CMS L1 Trigger for the High-Luminosity LHC

- **High-Luminosity phase of the LHC (HL-LHC) will start in 2029:**
  3x higher instantaneous luminosity and pileup wrt current conditions
  - CMS will upgrade most of its detectors, including all (trigger) electronics
- **L1 Trigger for the HL-LHC:**
  - Bandwidth: 2 –> 63 TB/s
  - Output 100 –> 750 kHz
  - Latency: 4 –> 12 us
- **Tracking @ L1T + new processing systems will enable "offline-like" reconstruction**



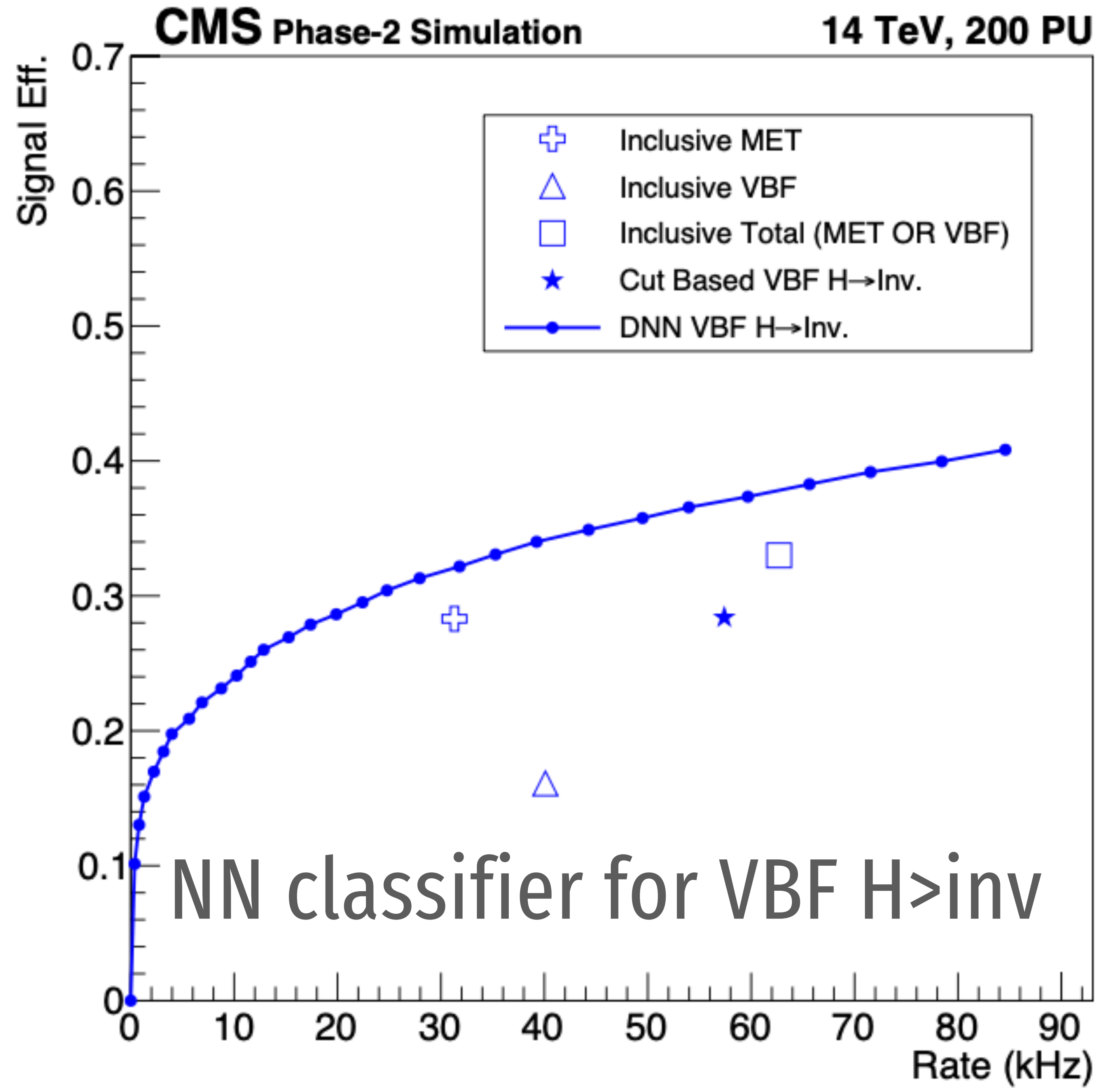
FPGAs: workhorse of the CMS Level-1 Trigger

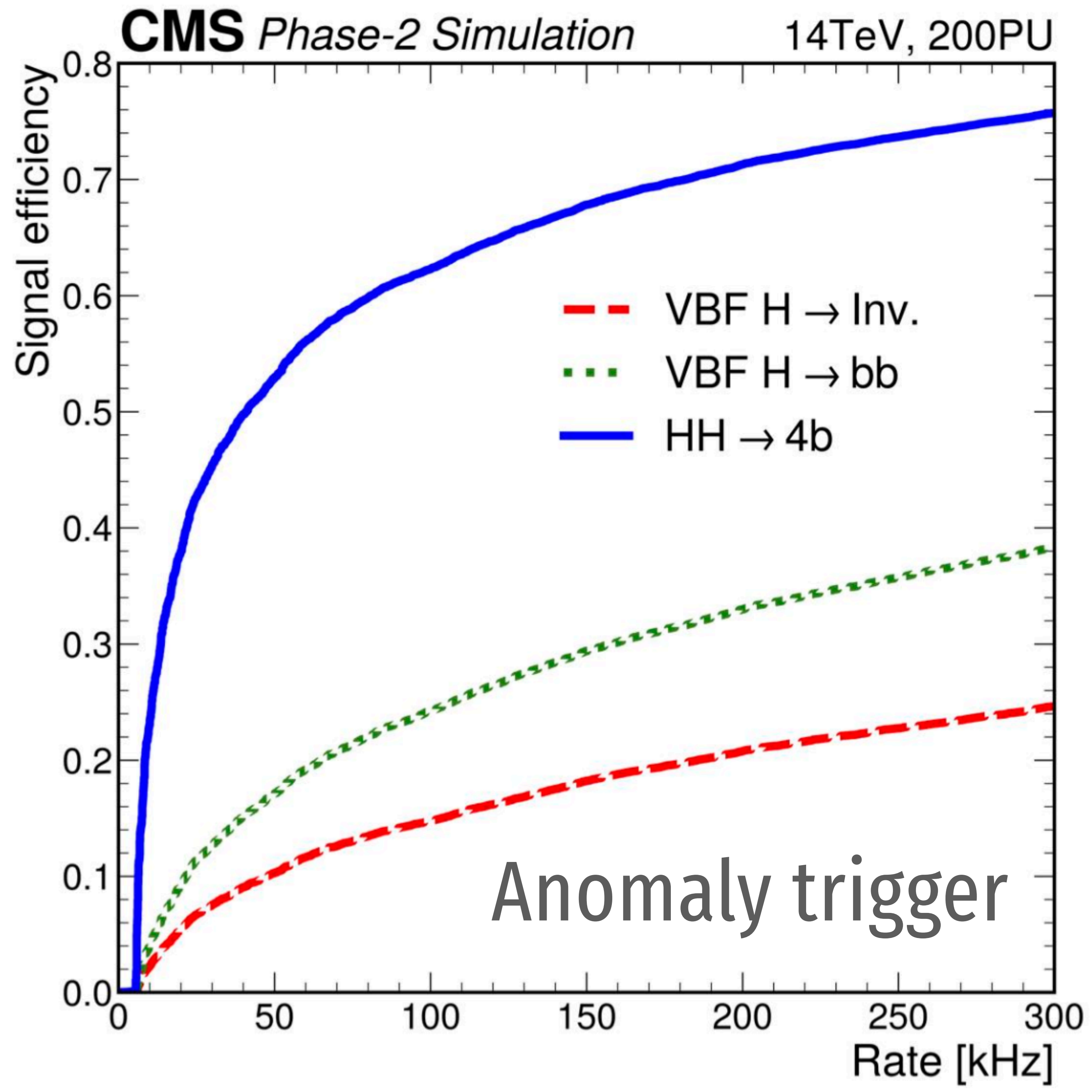
Hundreds of Xilinx VU13P FPGAs



L1 Event classification @ HL-LHC

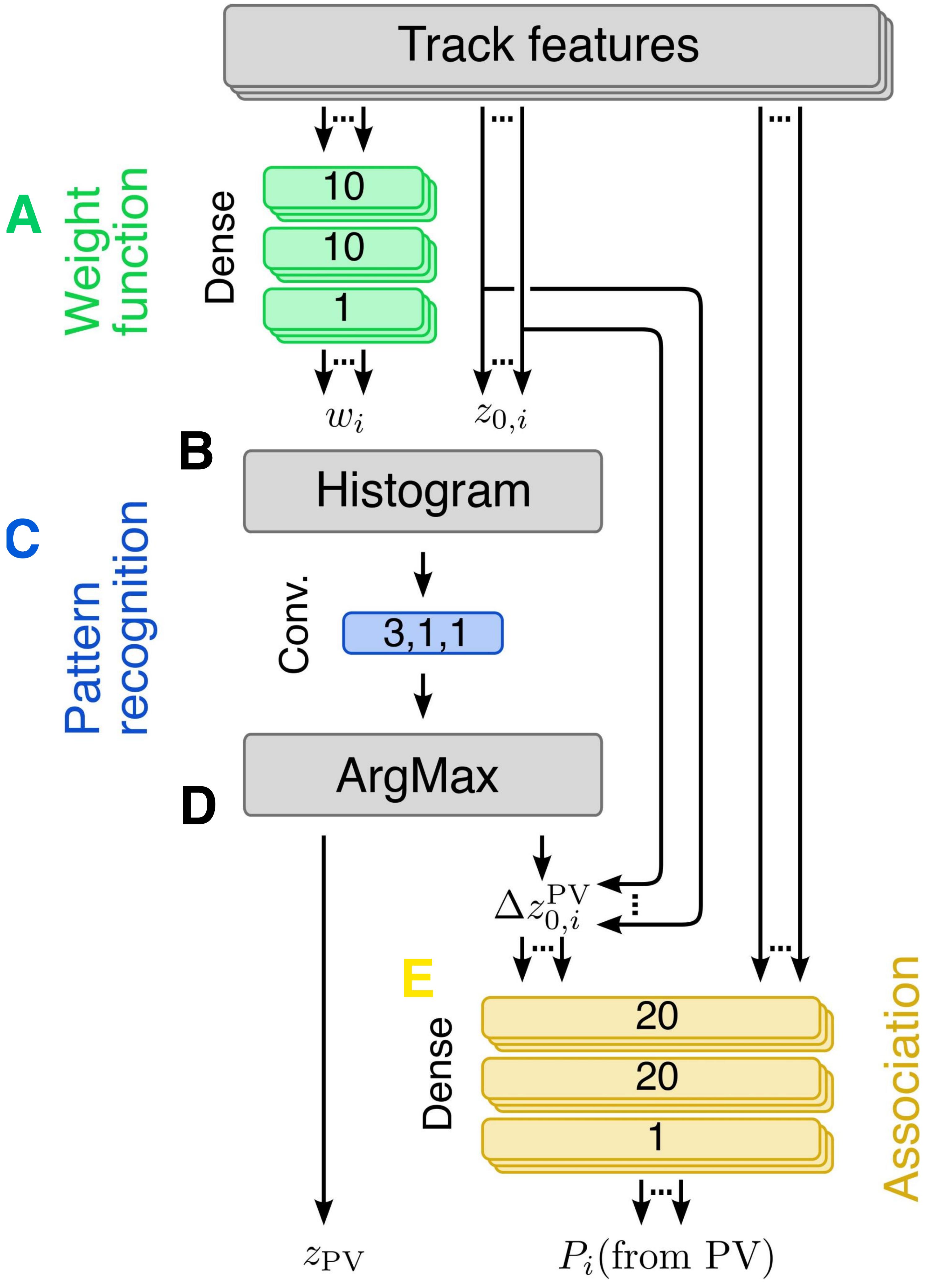
- ML-based triggers proposed in the L1T “TDR” for the High-Luminosity LHC
- **Classifier approach**: binary classifier for known signals trained on simulation (Note)
- **Anomaly detection**: auto-encoder based on L1 trigger objects (as AXOL1TL)
  - Sensitivity at the ~same order as of the classifier approach (e.g. VBF H>inv)





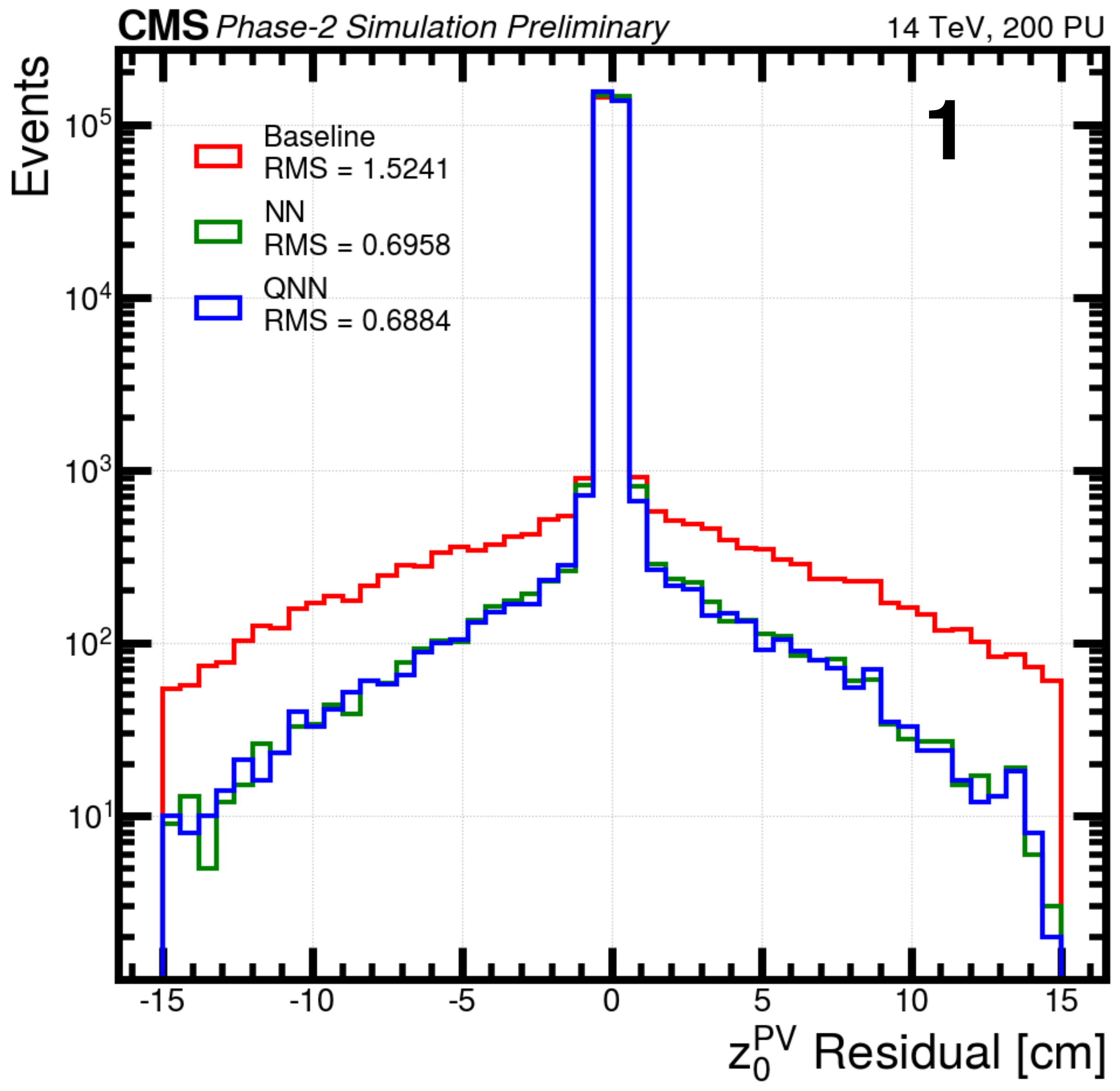
- Tests of AXOL1TL and CICADA pave the way for anomaly triggering at the HL-LHC in CMS!

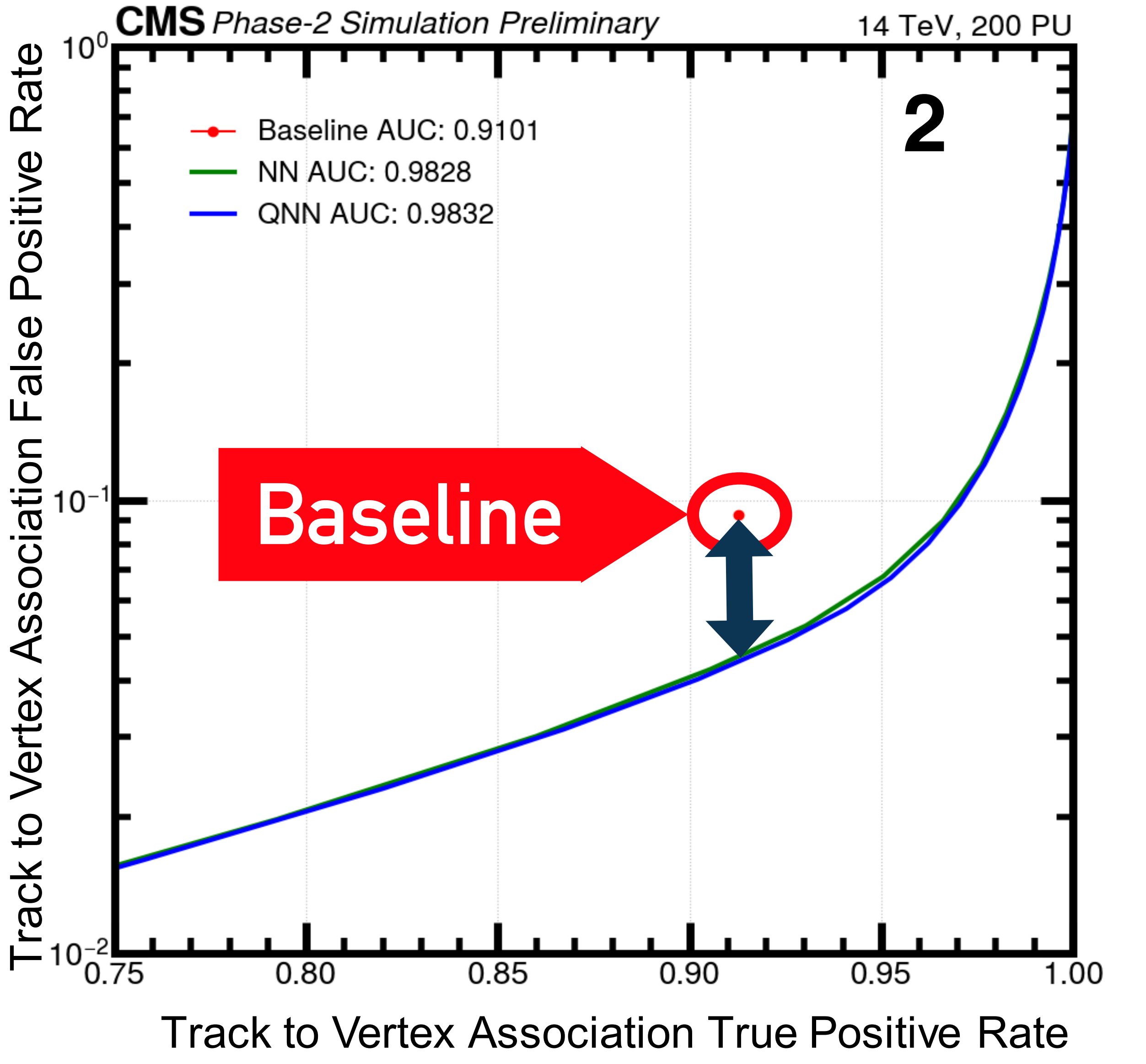
Vertex reconstruction



@ACAT 2021

- Reconstructing the "primary vertex" of the collision
  - The vertex with the highest energy (amongst 100s of other vertices!)
- **Multi-step approach of learning**: feature weighting, pattern recognition, track-to-vertex association
- (Q)NN drastically improves baseline at latency < 500ns!

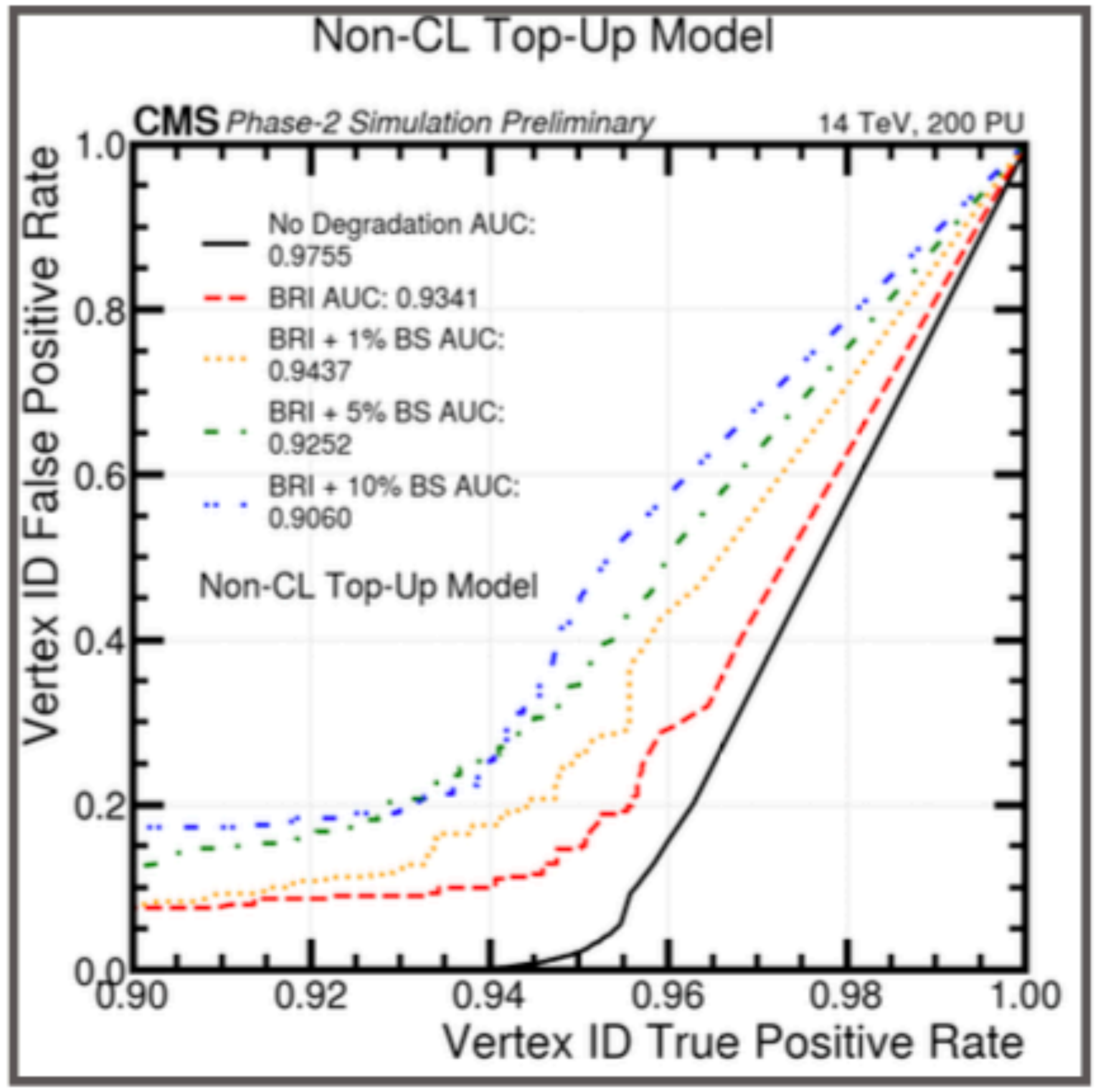


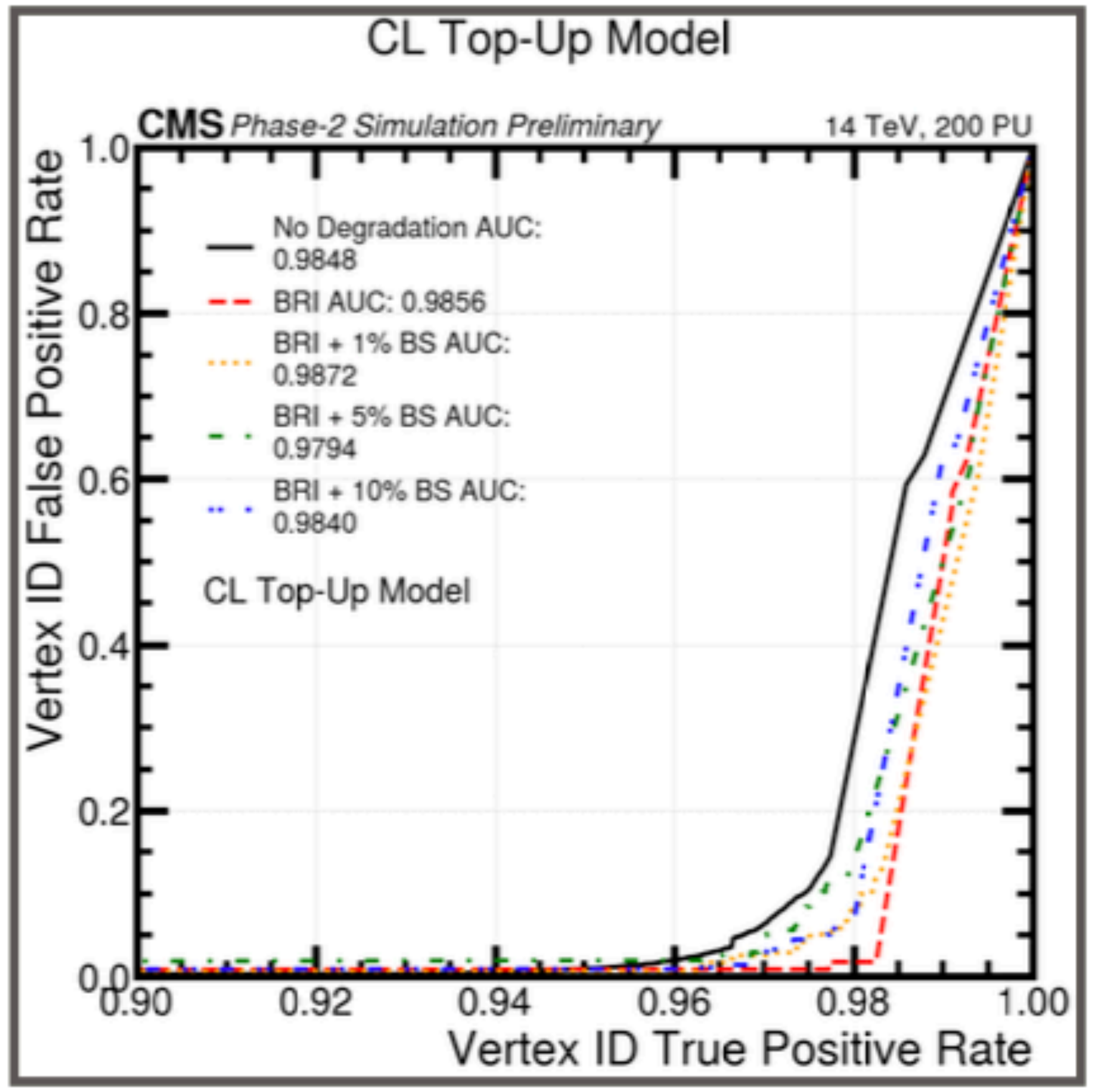


CONTINUAL LEARNING FOR VERTEXING

[Public note](#)
[CHEP23](#)

- Need to deal with changing detector conditions: ageing, noise, LHC conditions, etc
  - Normally do dedicated training/algo optimisation with full dataset
- **Continual Learning: train a model with a continuous stream of data**
  - Learns from a sequence of partial experiences rather than all the data at once
  - Update model to changing conditions without large MC production



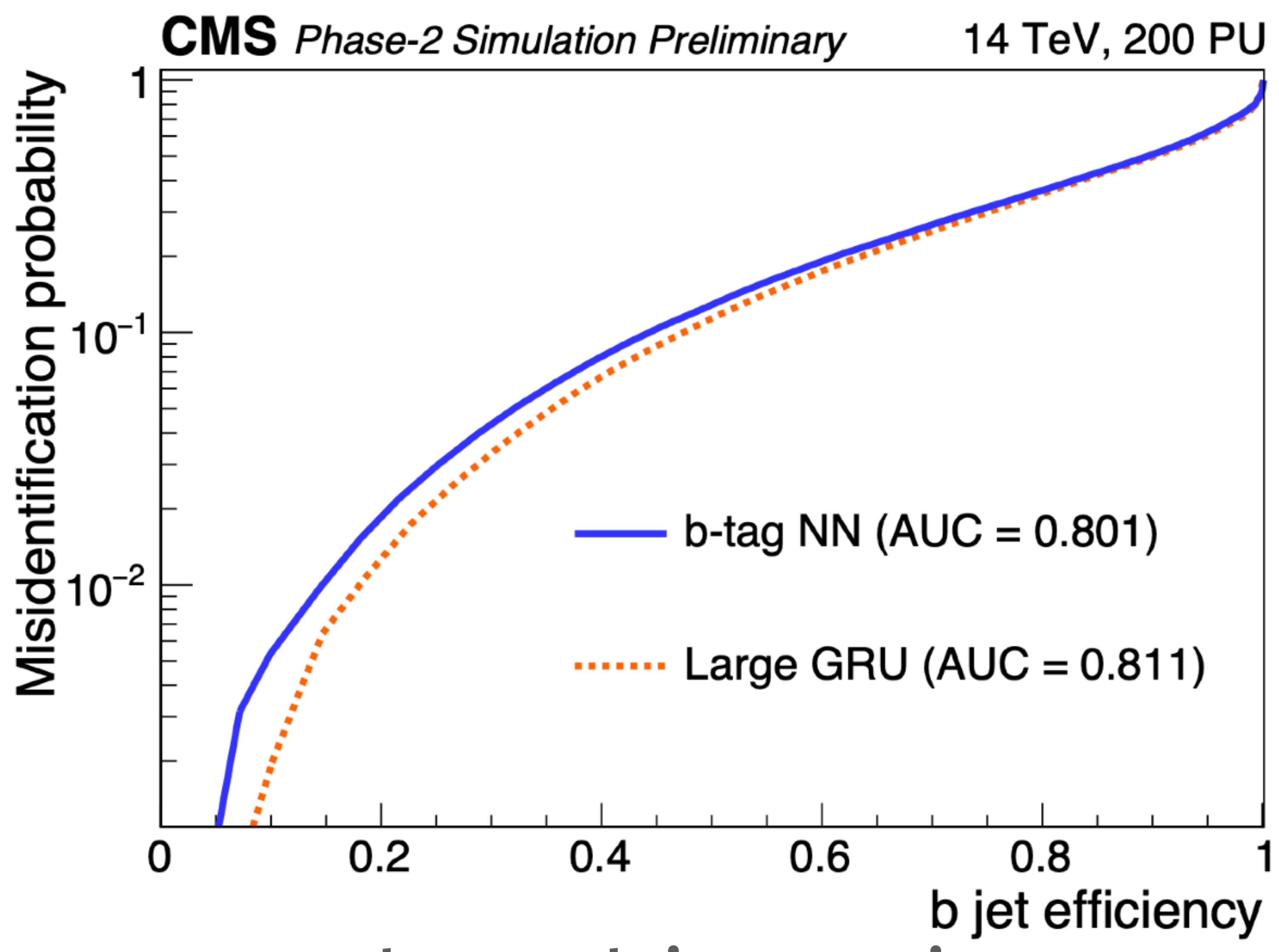


- Method tested on Vertex reconstruction:
  - CL outperforms a simple retrained model when detector defects are applied to the training data

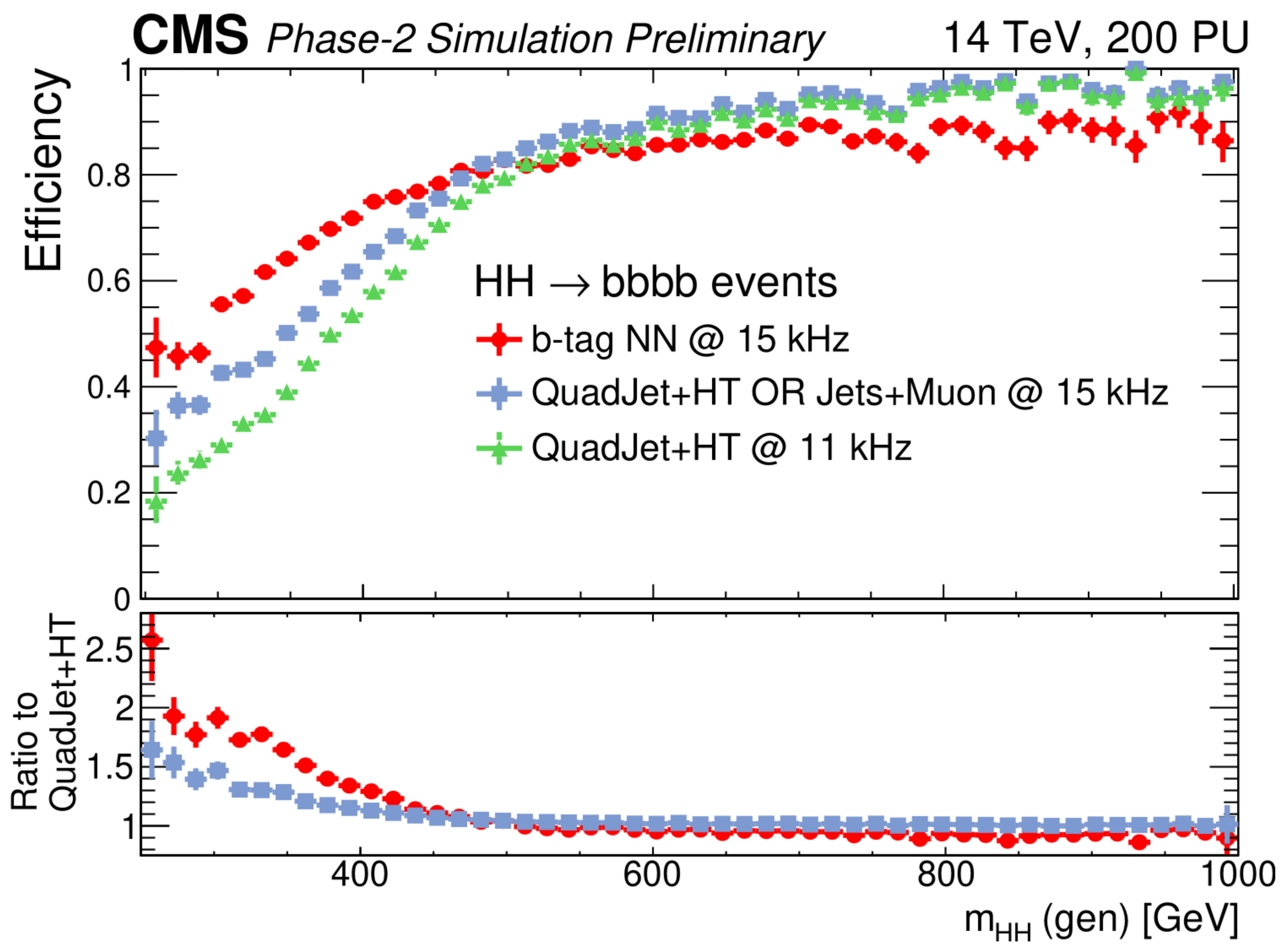
Jet Flavour Classification

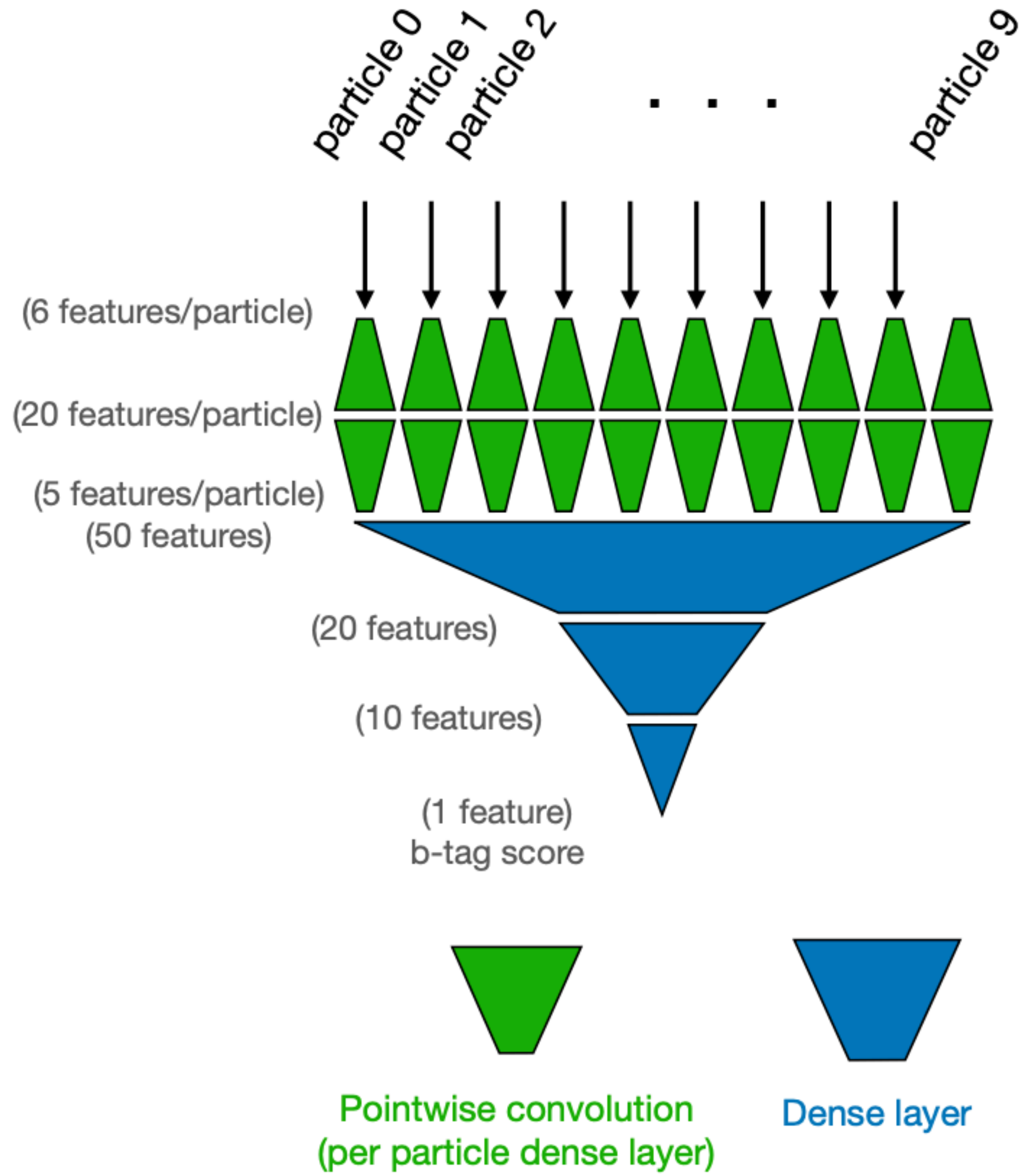
CMS Note DP2022_021

- Jet flavour ML identification successful offline, now online!
- **Different ID problems and architectures being studied for HL-LHC:**
  - Flavour tagging, pileup (background) rejection
  - Investigating Graph/Interaction NNs for this case
- **Proof-of-concept for jet identification shown for L1@HL-LHC!**



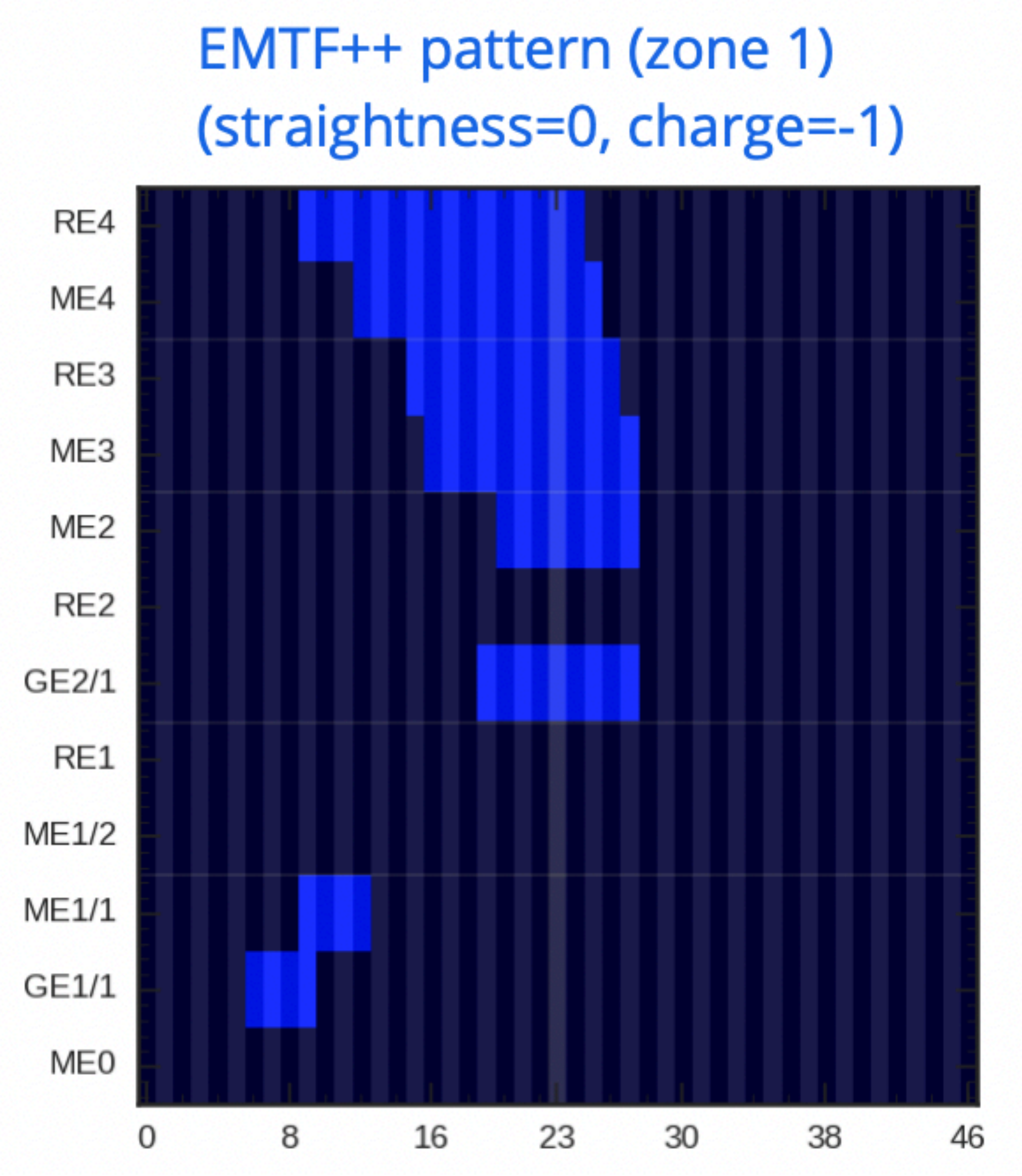
b quark jet tagging



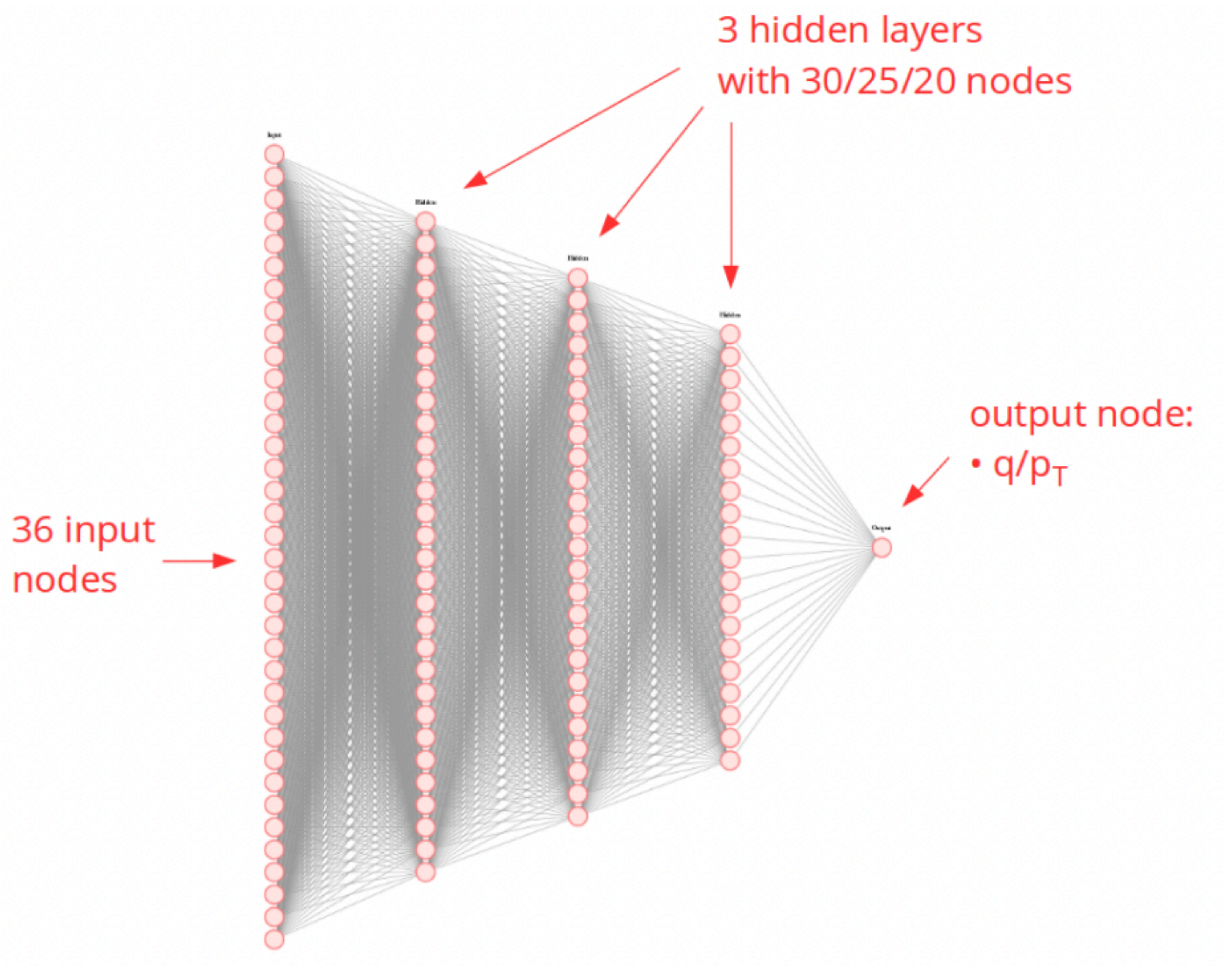


Muon reconstruction

- **Regressing the muon momentum** based on the hits in the muon detectors
- Based on features extracted from previous track finding



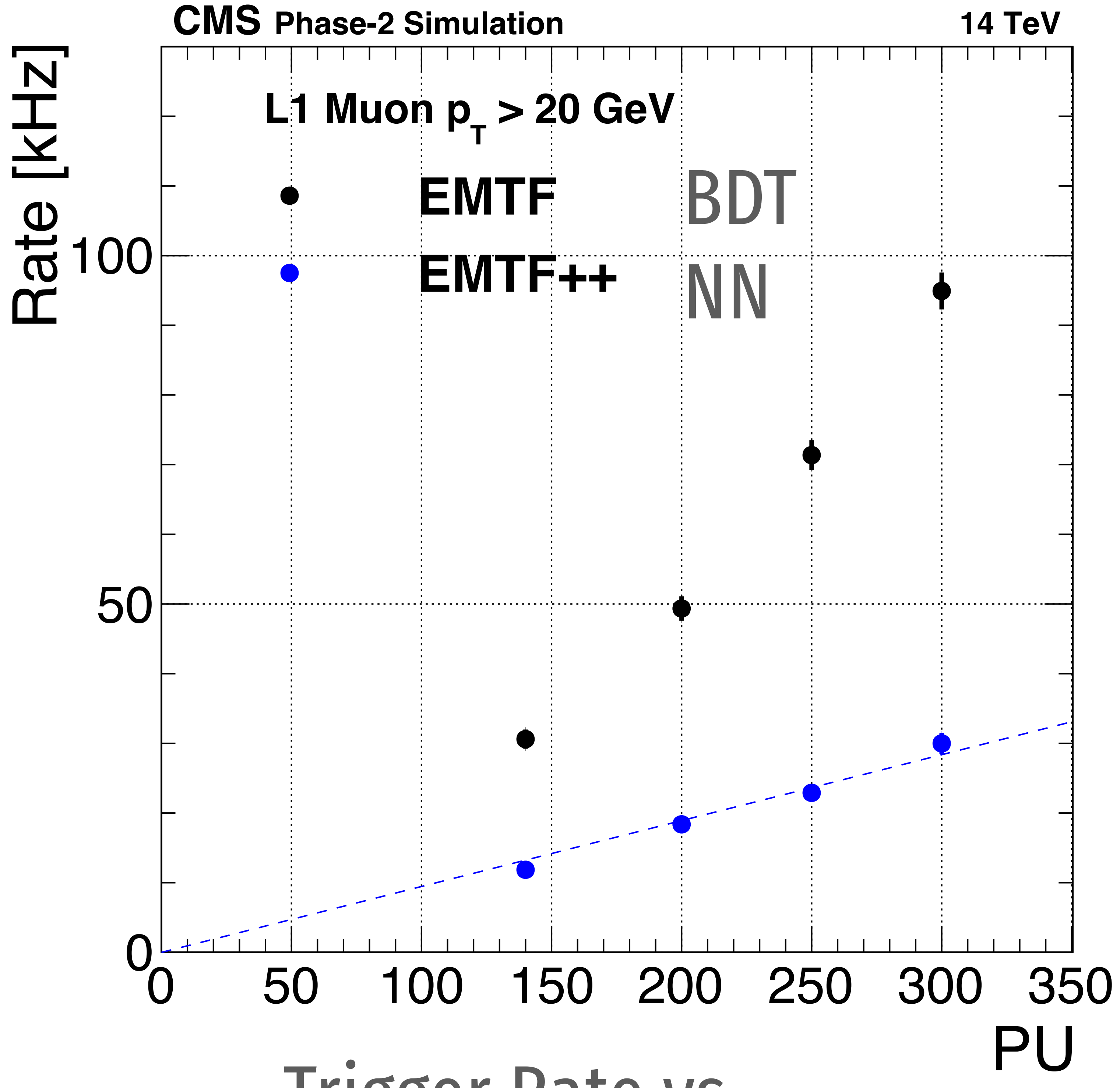
[CPAD 2019]



36 features from several sub tracks

	ME1/1	ME1/2	ME2	ME3	ME4	RE1	RE2	RE3	RE4	GE1/1	GE2/1	ME0
ϕ	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
θ	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
bend	✔	✔	✔	✔	✔							✔
quality	✔	✔	✔	✔	✔							✔
time												

CMS L1T Phase-2 TDR



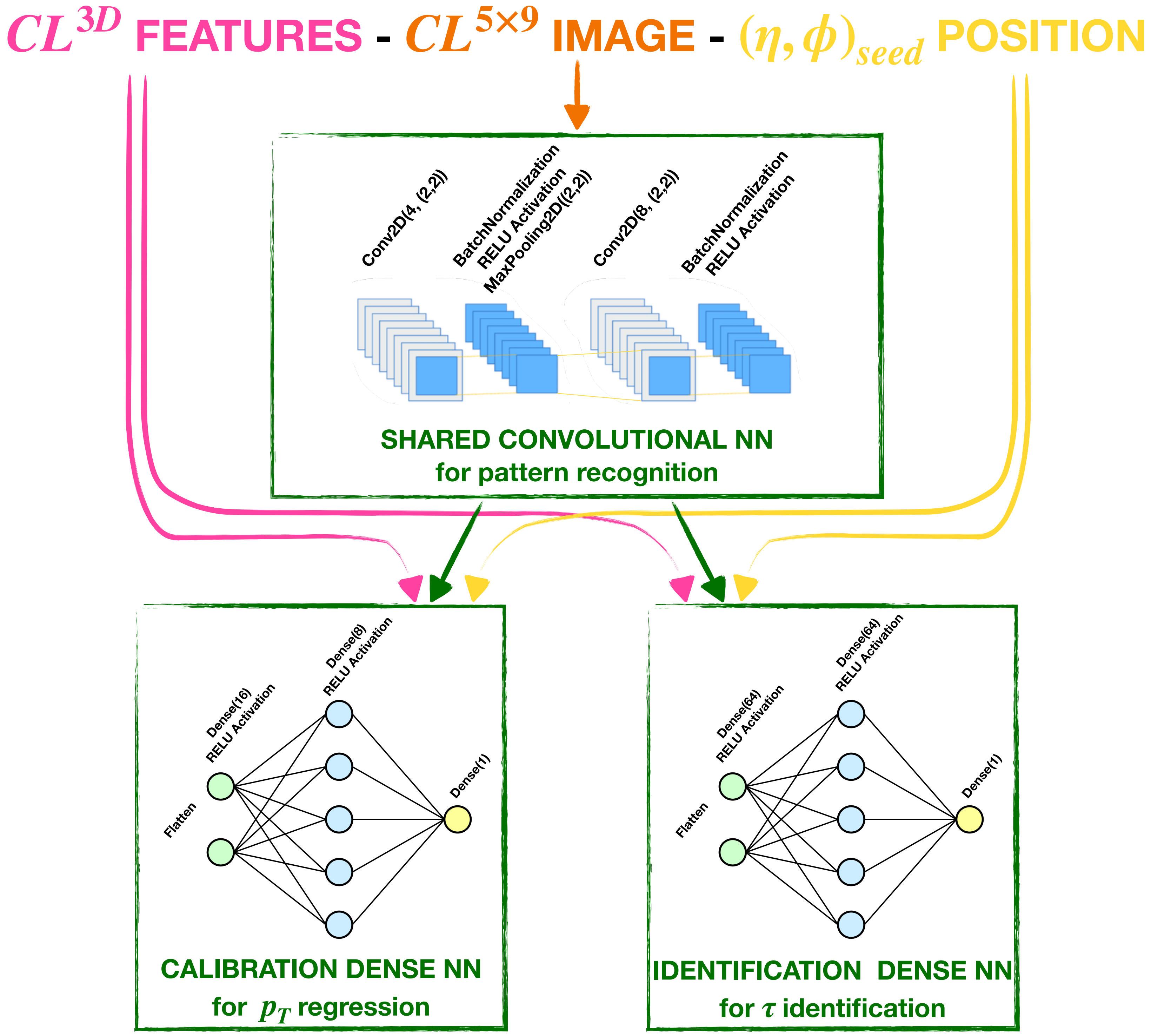
Trigger Rate vs “collision intensity”

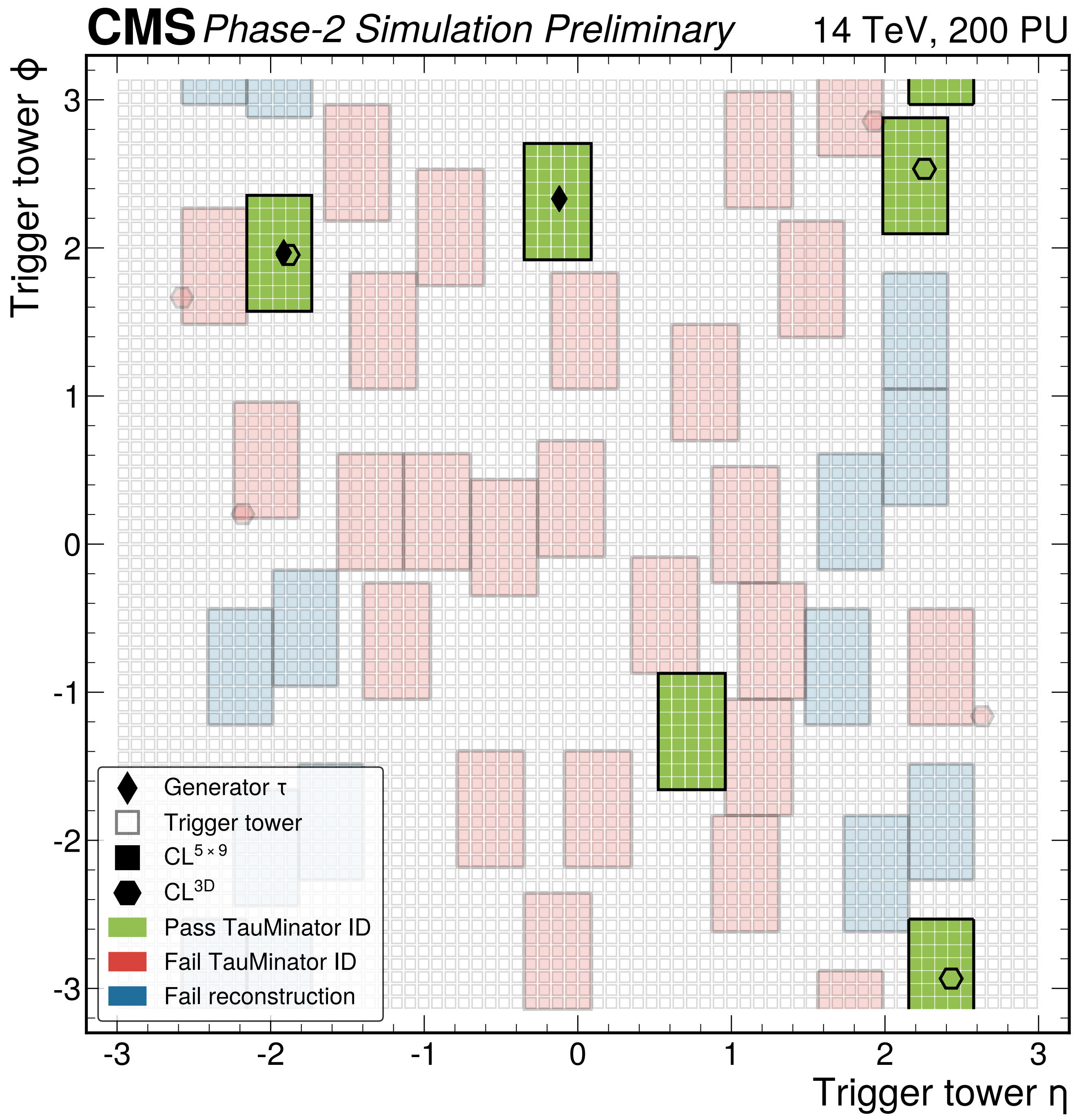
Lower rate with NN!

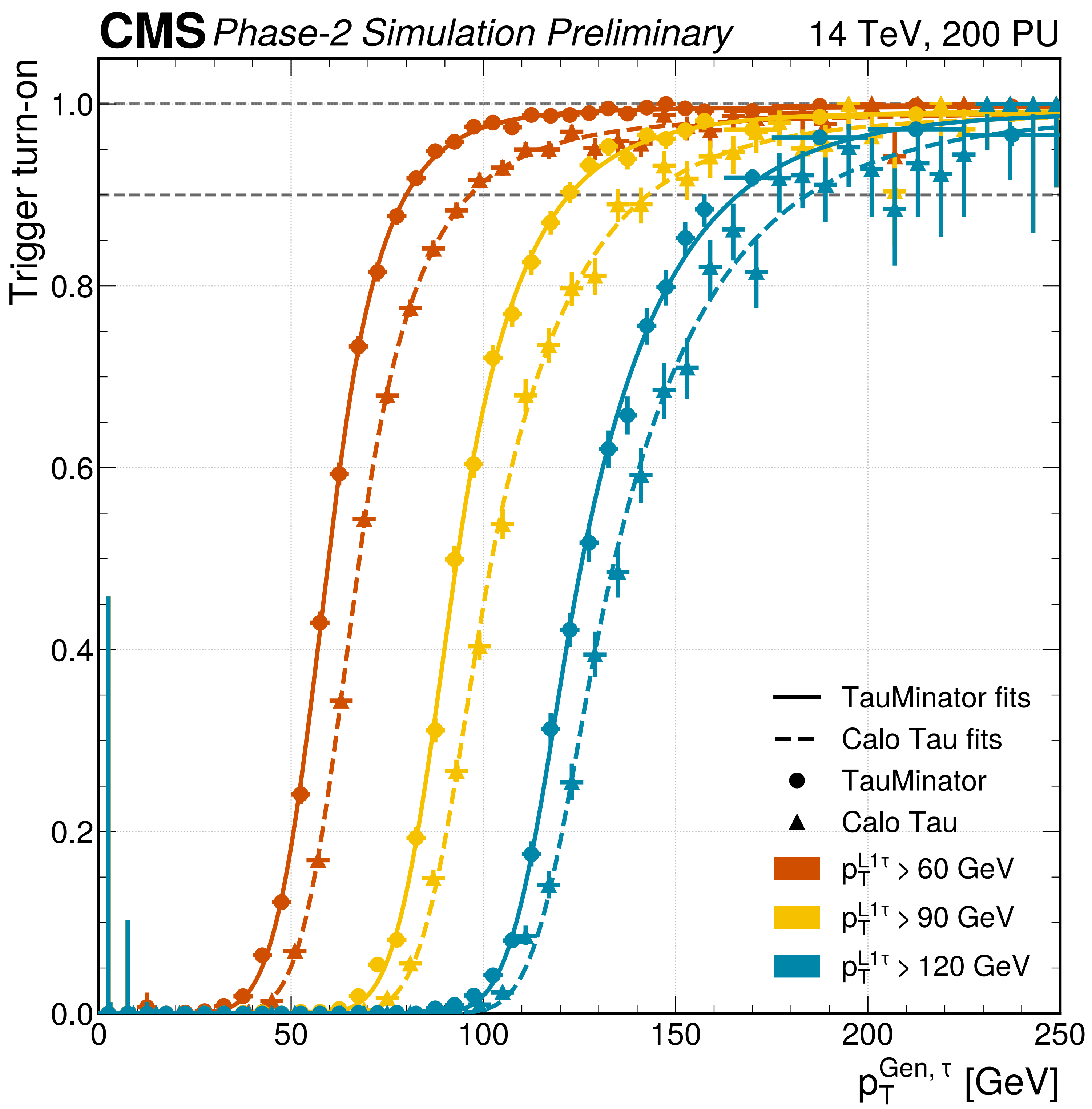
Hadronic Tau reconstruction

CMS DP-2023/062

- Hadronic Tau reconstruction very challenging in the L1 Trigger environment
- New **CNN approach for tau reconstruction with calorimeter-only information** followed by two **NNs for pT regression and ID**
  - Outperforms baseline algorithm: better efficiency and lower fake rate

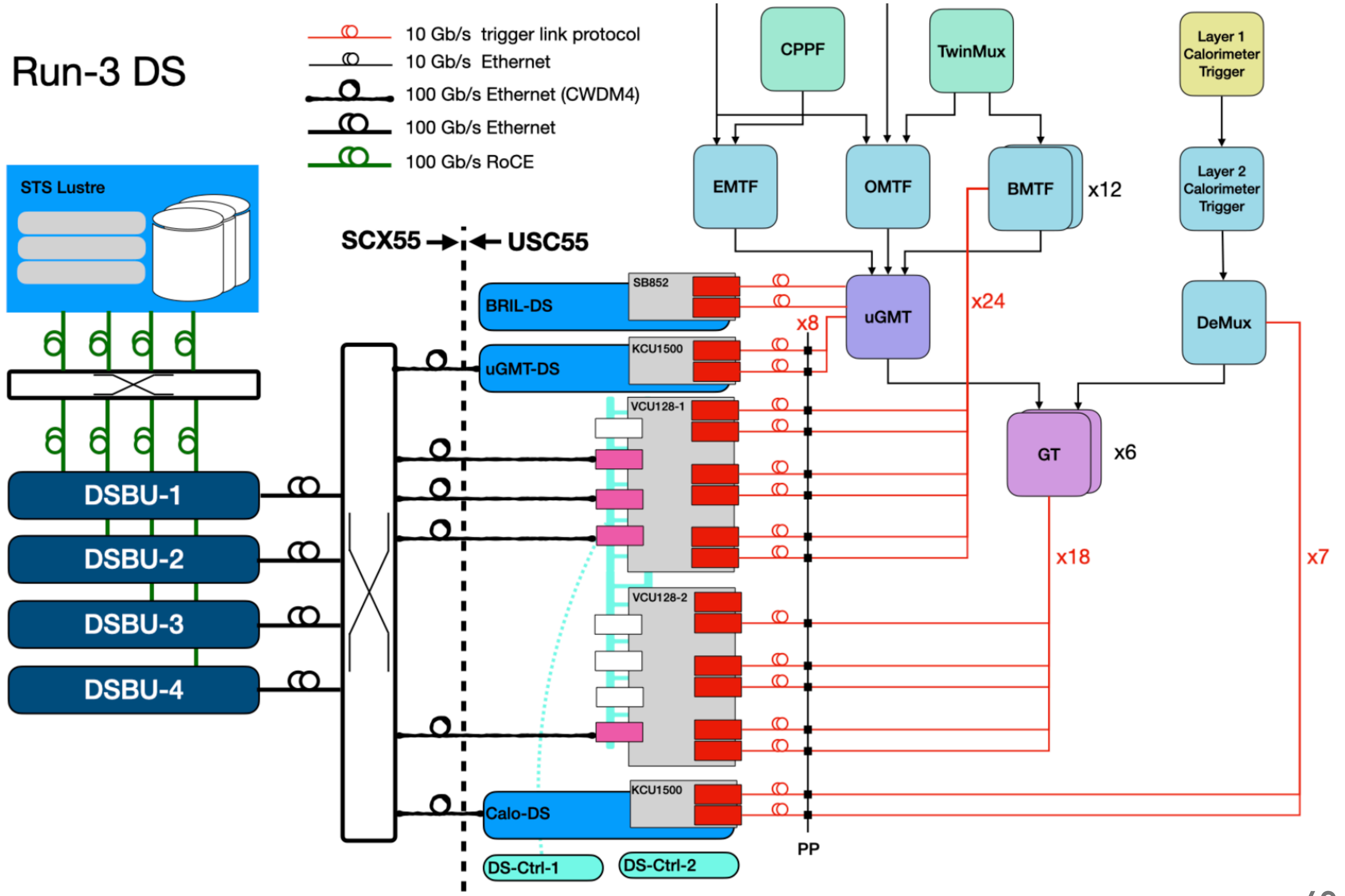






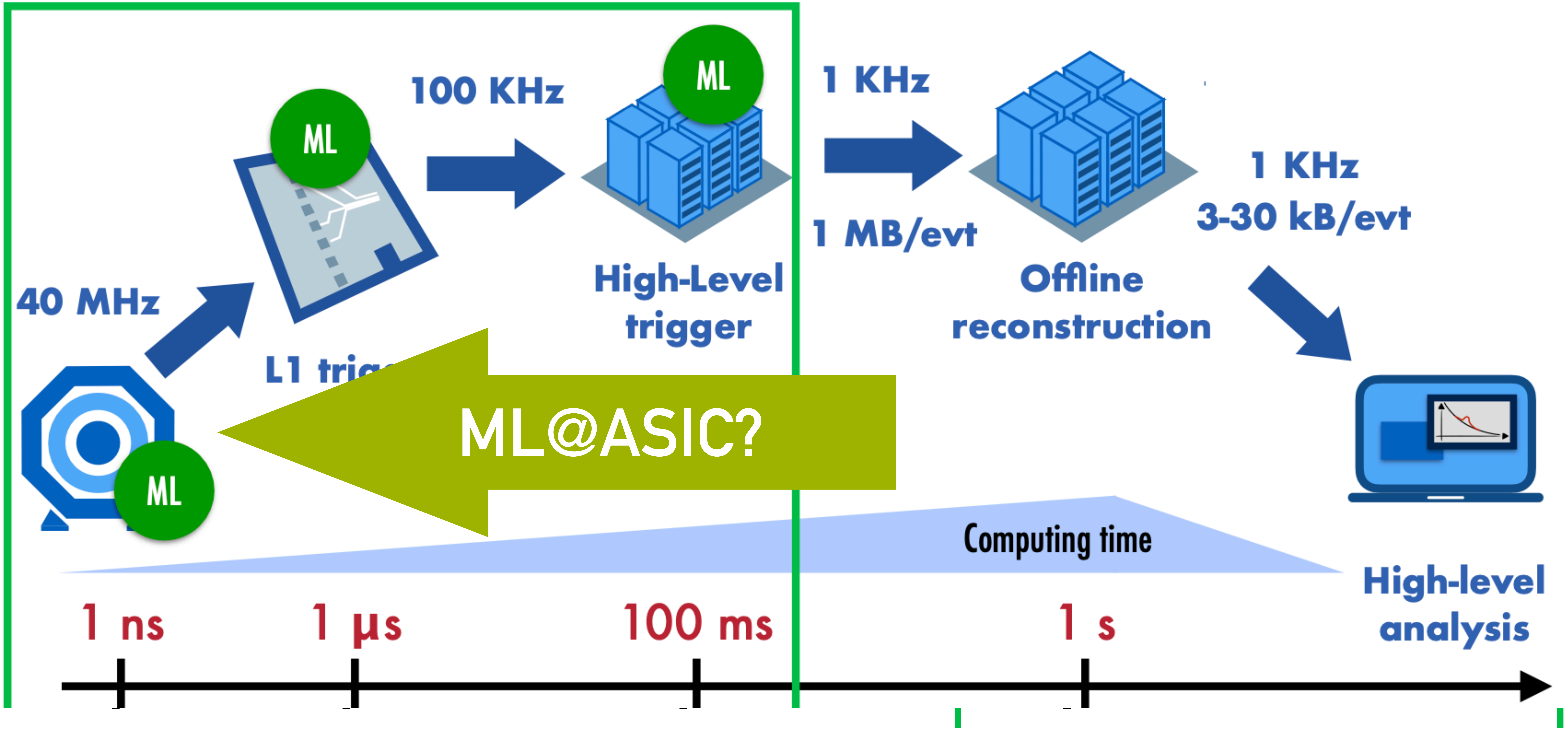
Sneak Peak: ML @ L1 Scouting

- **Scouting: directly store L1 physics objects for downstream analysis @ 40 MHz**
  - Triggerless readout <> lower quality objects
- Run-3 demonstrator system reads L1 objects (muons, calo jets, EG, taus) with very heterogenous system:
  - 3 boards: KCU1500, SB-852, VCU128)
  - Output technologies: DMA, TCP/IP



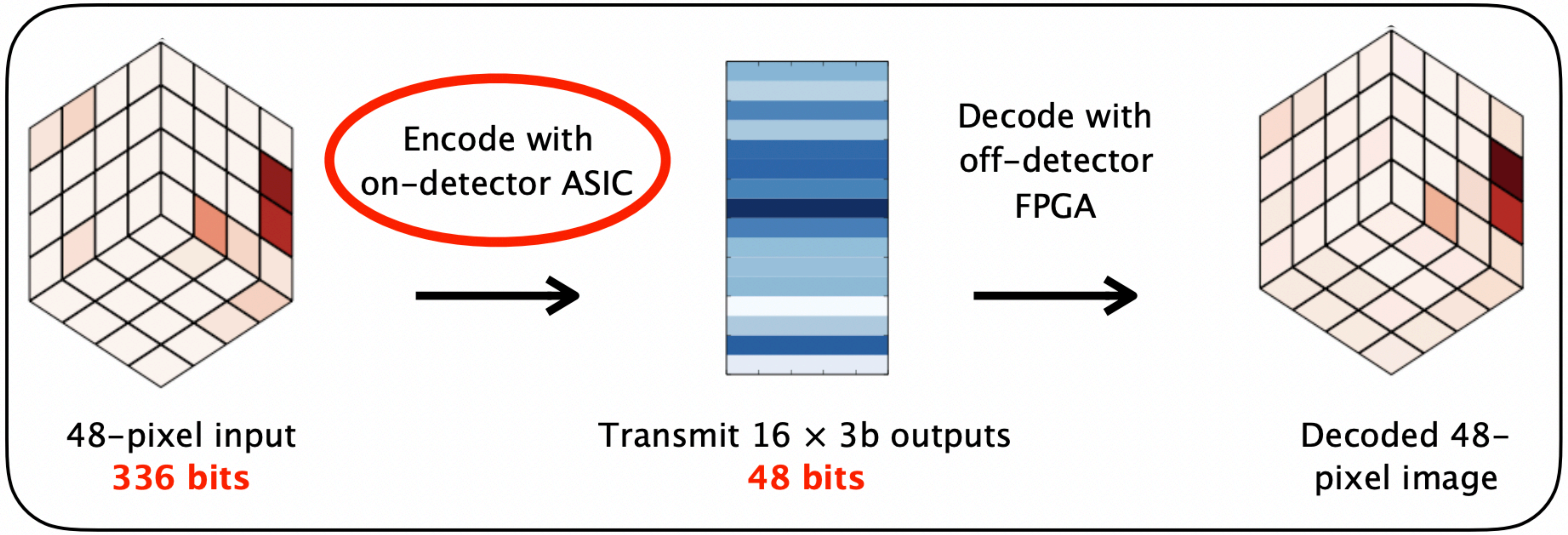
- Studying **ML methods for realtime calibrations with HLS4ML & proprietary SW** (Micron Deep Learning Accelerator)
- **Detailed overview here and here**

ML in Detector Readout ASICs



AUTOENCODER IN CMS CALORIMETER TRIGGER ASIC

- Data volume needs compression at detector readout: done by ASICs (fast/efficient!)
- CMS will be **using an (auto)encoder for compressing the data *intelligently*** in the High-Granularity Calorimeter concentrator ASIC (HL-LHC)
  - **Decoding done on the FPGA** side of the Level-1 Trigger (or direct use in NN@FPGA?)



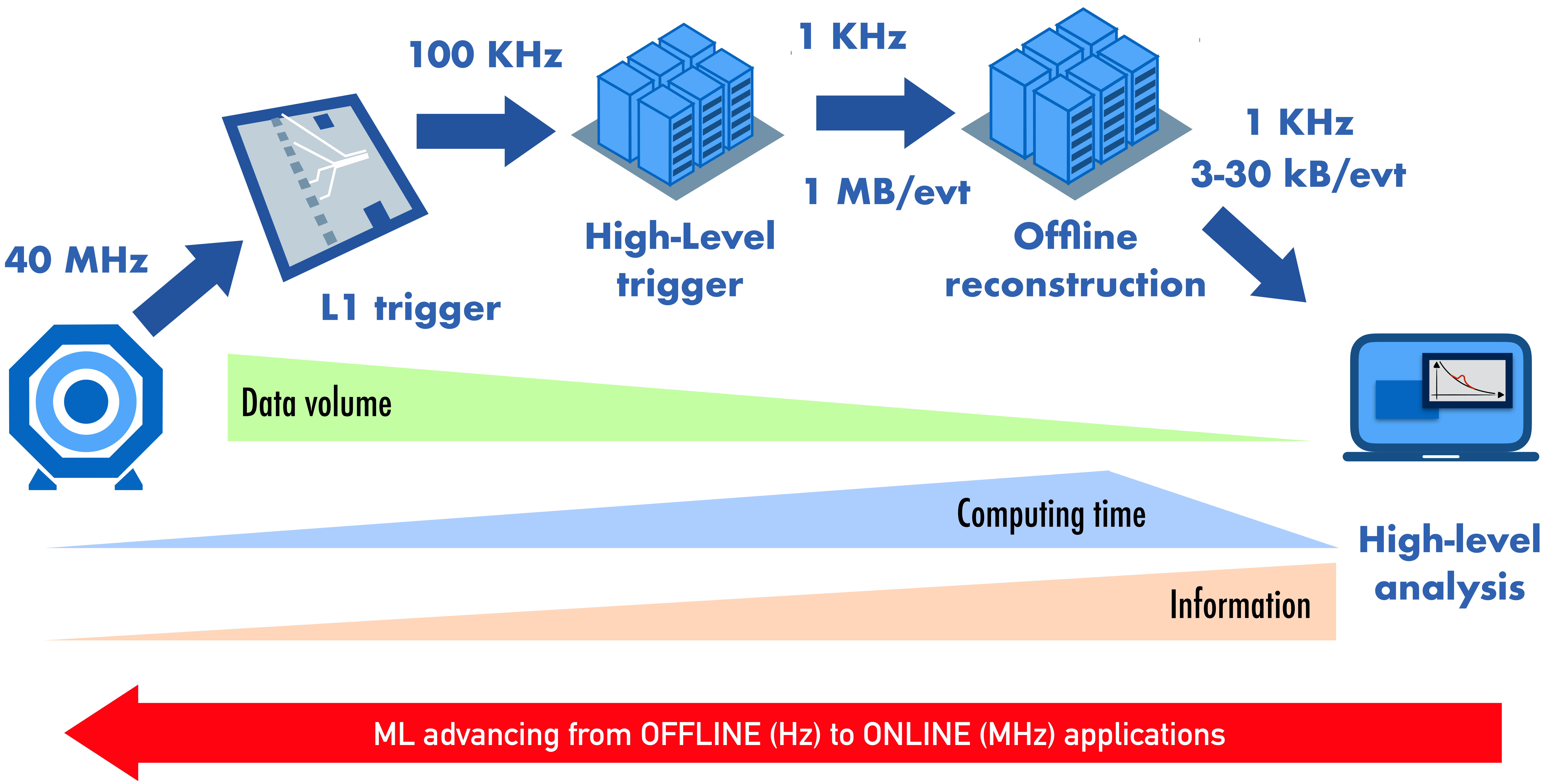
Transmitting more information with limited bandwidth!

CPAD 2021

FPGAs were designed for ASIC prototyping —> used HLS4ML for design!

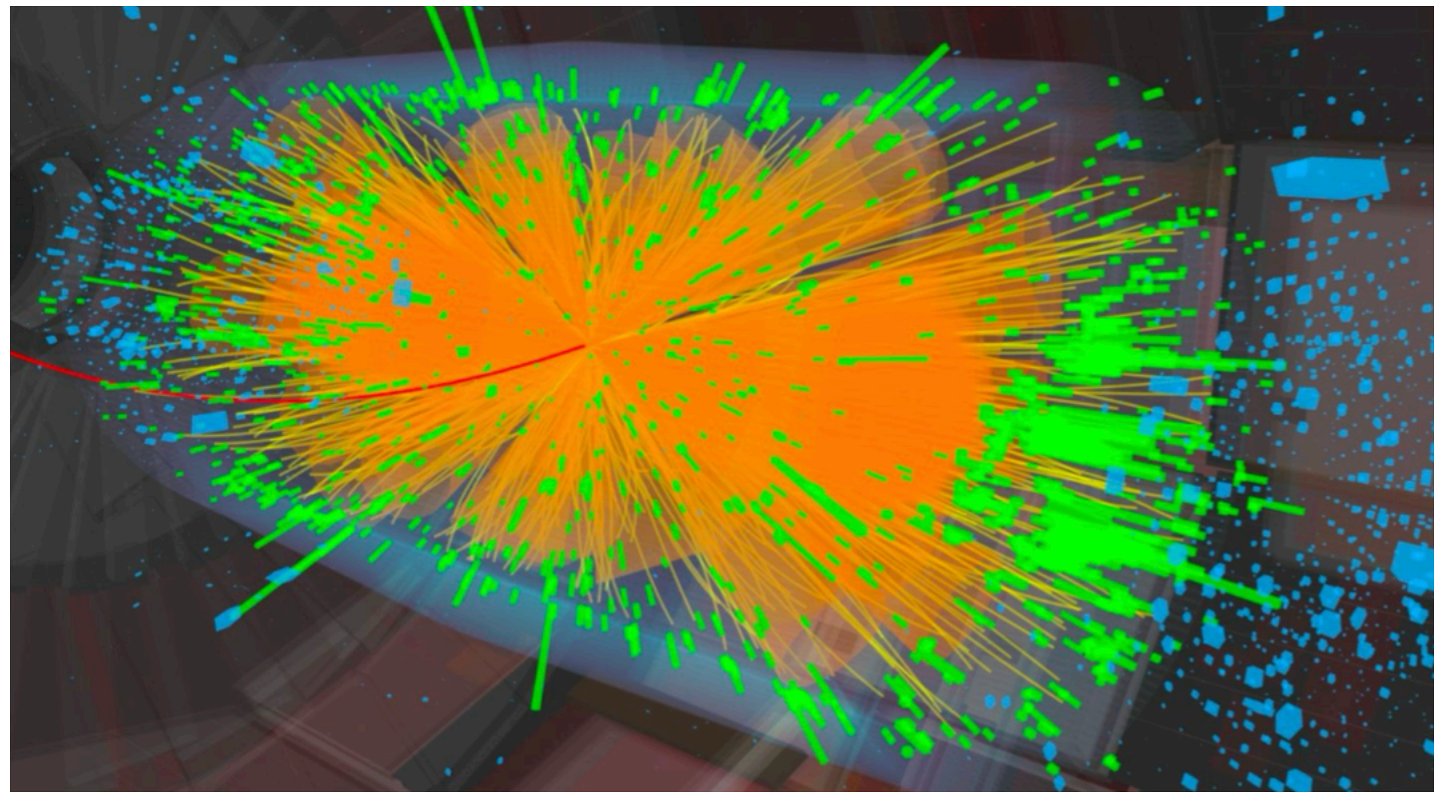
Summary

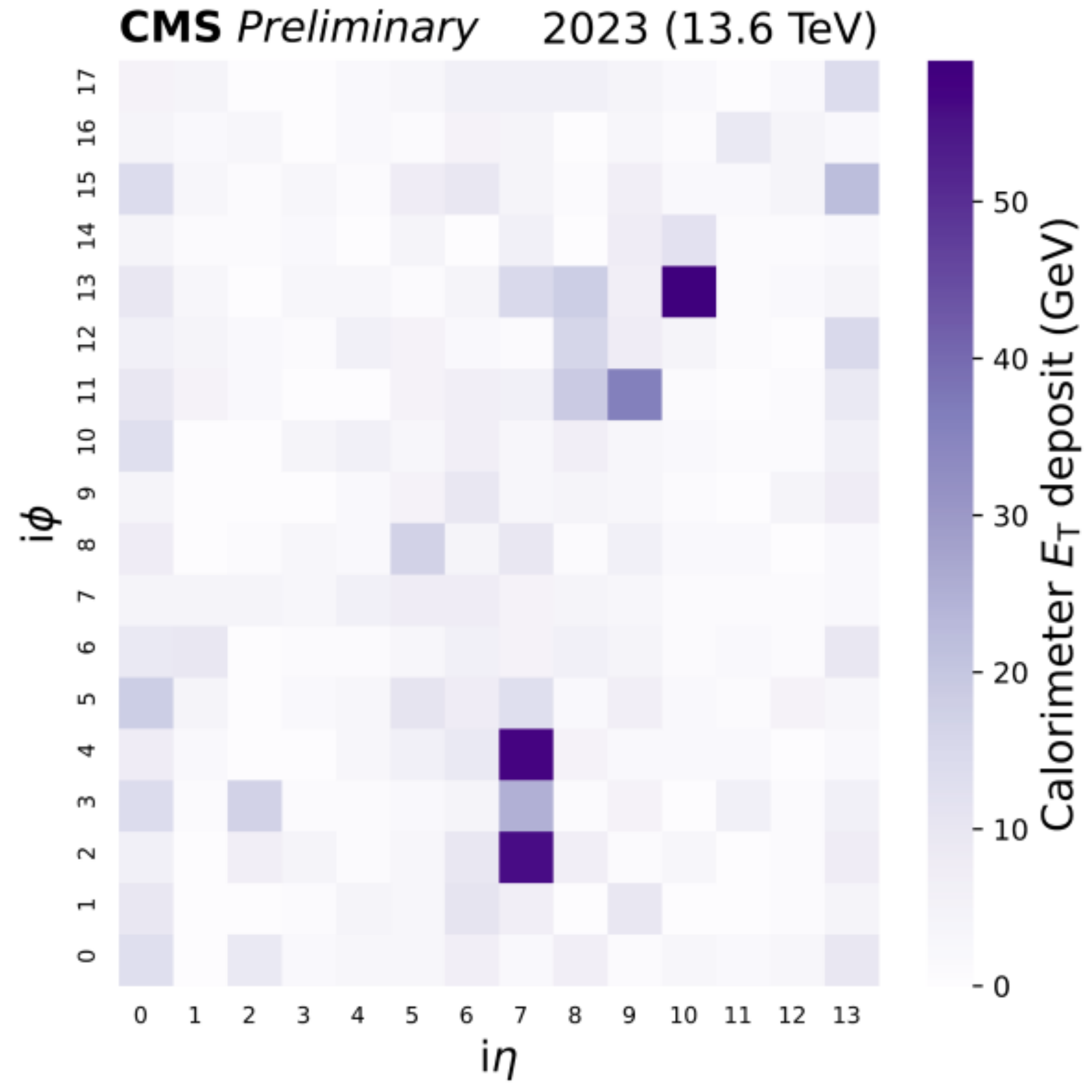
Fast ML in CMS



Anomaly Detection with the CMS Level-1 Trigger

- Various **anomaly searches for new physics performed at the LHC**
- Opening **a new direction:**
  anomaly detection in the CMS Level-1 Trigger
  - **Challenging environment for L1T**:
    - Hardware/FPGAs: restricted resources and latency (ns!)
    - Physics: <60> simultaneous collisions,
      only calorimeter and muon detector data
- **Two auto-encoder approaches being commissioned** in CMS:
  - **AXOL1TL**: using high-level physics objects [CMS-DP-2023-079]
  - **CICADA**: using raw detector data [CMS DP-2023/086]
- **Promising prospects for anomaly triggering in CMS!** [HL-LHC L1T]





ML in CMS L1 Trigger for HL-LHC

- A plethora of L1 components exploring and using ML algorithms for the HL-LHC
  - All based on HLS: HLS4ML (NN) and Conifer (BDT)
- **Current Run-3 algorithms (AXOL1TL/CICADA) are an important test bed for the future**

Each small box is one Xilinx Ultrascale+ FPGA

These have NNs and/or BDTs inside

hls 4 ml   Conifer

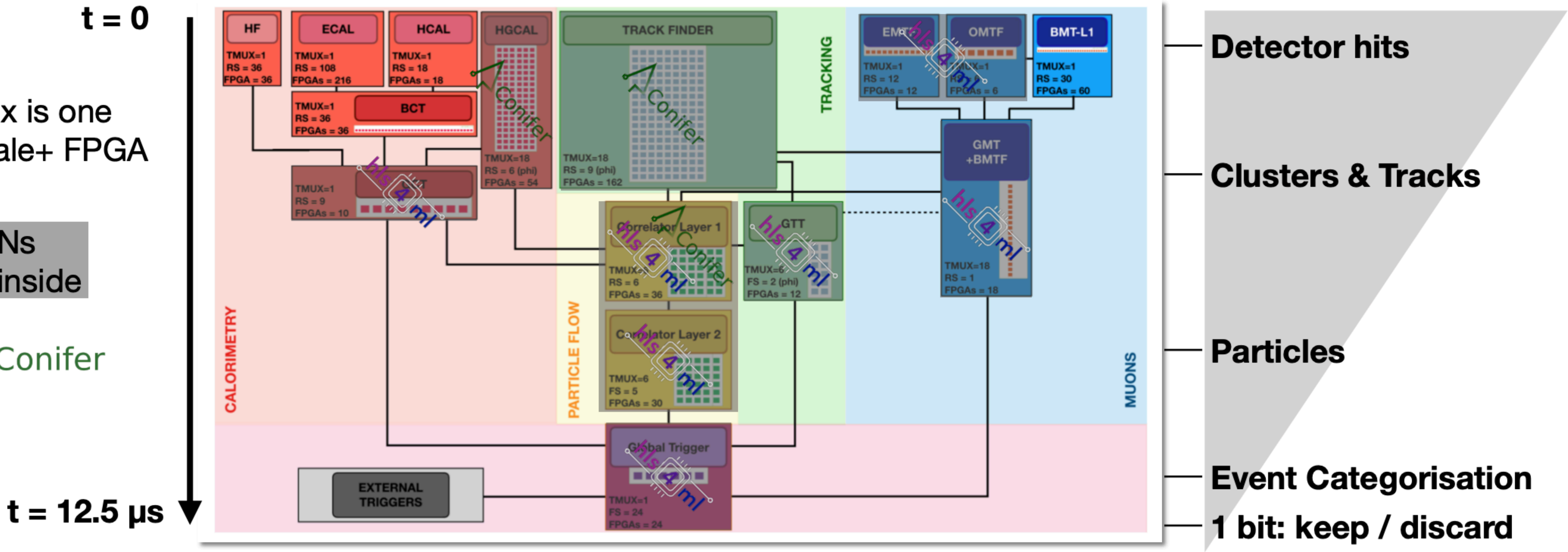
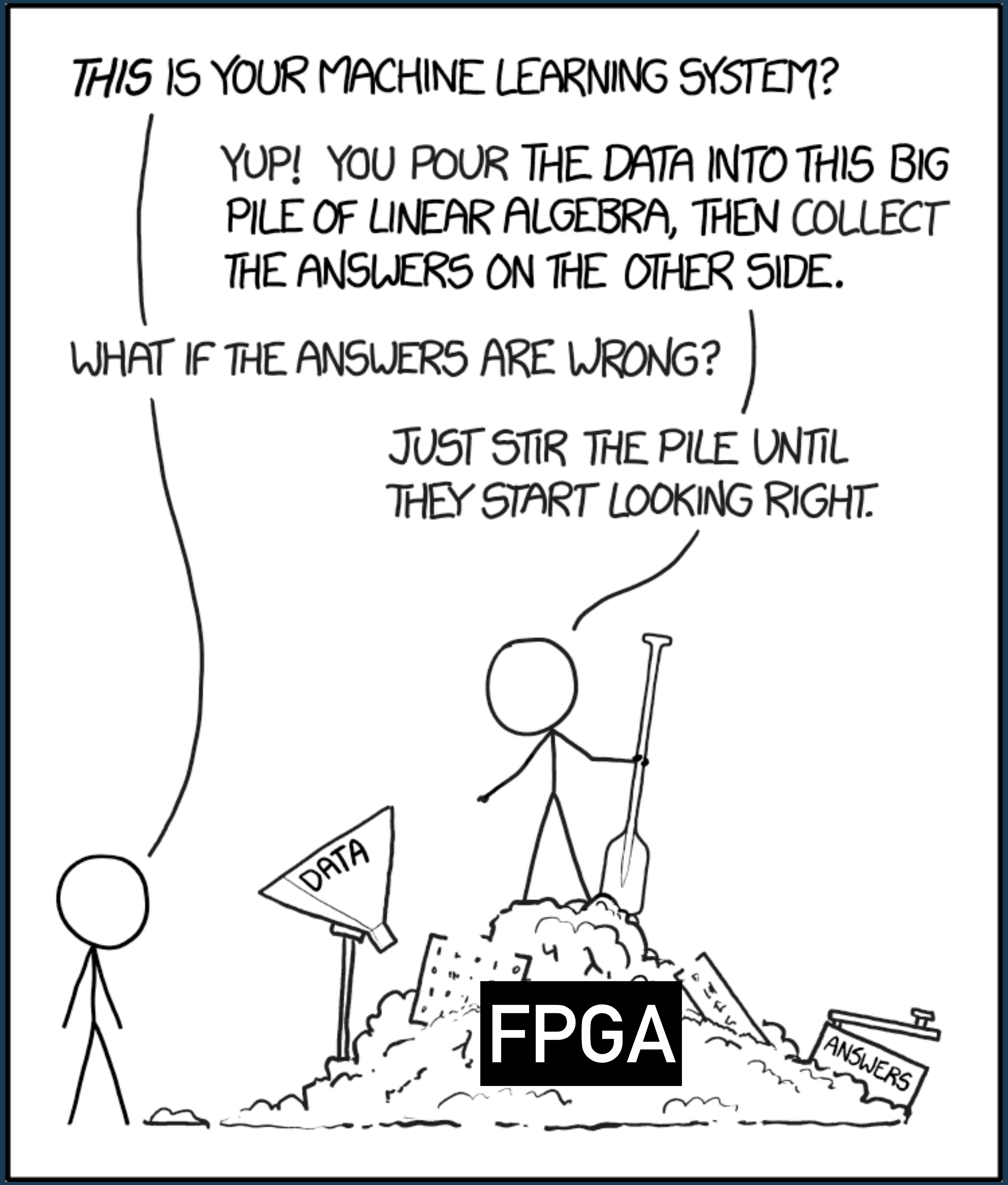


Image by Sioni Summers



xkcd "Machine Learning"

References

- CMS Collaboration. ”Anomaly Detection in the CMS Global Trigger Test Crate for Run 3”. CMS-DP-2023-079 , CERN-CMS-DP-2023-079 (2023): https://cds.cern.ch/record/2876546
- CMS Collaboration. "Level-1 Trigger Calorimeter Image Convolutional Anomaly Detection Algorithm”, CMS-DP-2023-086 ; CERN-CMS-DP-2023-086 https://cds.cern.ch/record/2879816
- J. Pearkes, “Realtime Anomaly Detection in the CMS Experiment Global Trigger Test Crate”, ML4Jets 2023, https://indico.cern.ch/event/1253794/timetable/?view=standard#57-realtime-anomaly-detection
- N. Zipper, “Testing a Neural Network for Anomaly Detection in the CMS Global Trigger test crate during Run 3”, TWEPP 2023 https://indico.cern.ch/event/1255624/contributions/5444028/
- C. Sun, “Realtime Anomaly Detection in the CMS Experiment Global Trigger Test Crate”, FastML 2023, https://indico.cern.ch/event/1283970/contributions/5554350/
- CMS Collaboration. “CMS Technical Design Report for the Level-1 Trigger Upgrade”, CERN-LHCC-2013-011 ; CMS-TDR-12 https://cds.cern.ch/record/1556311
- E. Govorkova, et al. "Autoencoders on field-programmable gate arrays for real-time, unsupervised new physics detection at 40 MHz at the Large Hadron Collider". Nat. Mach Intell. 4, 154 (2022). https://doi.org/10.1038/s42256-022-00441-3
- FastML Team. hls4ml (Version v0.7.1) [Computer software]. https://doi.org/10.5281/zenodo.1201549
- J. Duarte, et al. "Fast inference of deep neural networks in FPGAs for particle physics". JINST 13, P07027 (2018). https://doi.org/10.1088/1748-0221/13/07/P07027

References

- More results on Fast ML in CMS and beyond were shown at the Fast ML for Science Workshop series, e.g. https://indico.cern.ch/event/1283970/:
  - Fast ML inference in FPGAs for the Level-1 Scouting system at CMS
  - Realtime Anomaly Detection in the CMS Experiment Global Trigger Test Crate
  - Harnessing charged particle tracks in the Phase-2 CMS Level-1 Trigger with ultrafast Machine Learning
  - B-tagging and Tau reconstruction in the Level-1 Trigger with real-time Machine Learning
  - A Convolutional Neural Network for topological fast selection algorithms in FPGAs for the HL-LHC upgrade of the CMS experiment

Backup

AXOL1TL: LOSS TERM

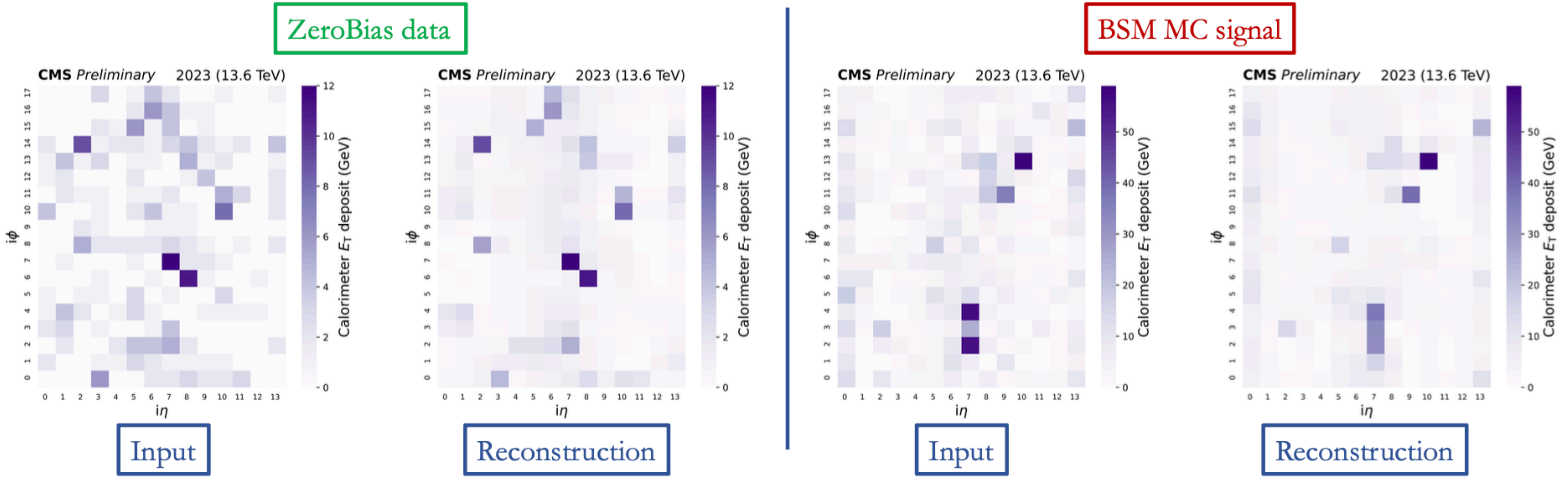
The AXOL1TL anomaly detection uses a Variational Autoencoder (VAE). A dense feed-forward neural network reads in (p_T, η, ϕ) hardware inputs of 19 L1 objects. The encoder network computes a latent space vector of Gaussian probability distributions, $N(\mu_8, \sigma_8)$. The decoder network reconstructs the original input from the latent space.

$$\text{Loss} = \underbrace{(1-\beta)\left\|x - \hat{x}\right\|^2}_{\text{Reconstruction term}} + \underbrace{\beta\frac{1}{2}(\mu^2 + \sigma^2 - 1 \;\; - \log\sigma^2)}_{\text{Full regularization term}}$$

Equation: VAE loss function. The reconstruction term is computed from the difference between the input (x) and output ($\hat{x}$) of the VAE. The second, full regularization term, is the Kullback–Leibler divergence (KL-divergence) between the latent space distribution and a standard normal distribution with mean μ and standard deviation σ. The parameter β can be tuned to balance the reconstruction performance with more efficient latent space encoding. At inference time, the loss is approximated by the mean-squared term $\Sigma\mu_i^2$ of the KL-divergence for latency considerations. This approximation has no impact on performance.

CICADA: Anomaly detection on Raw inputs

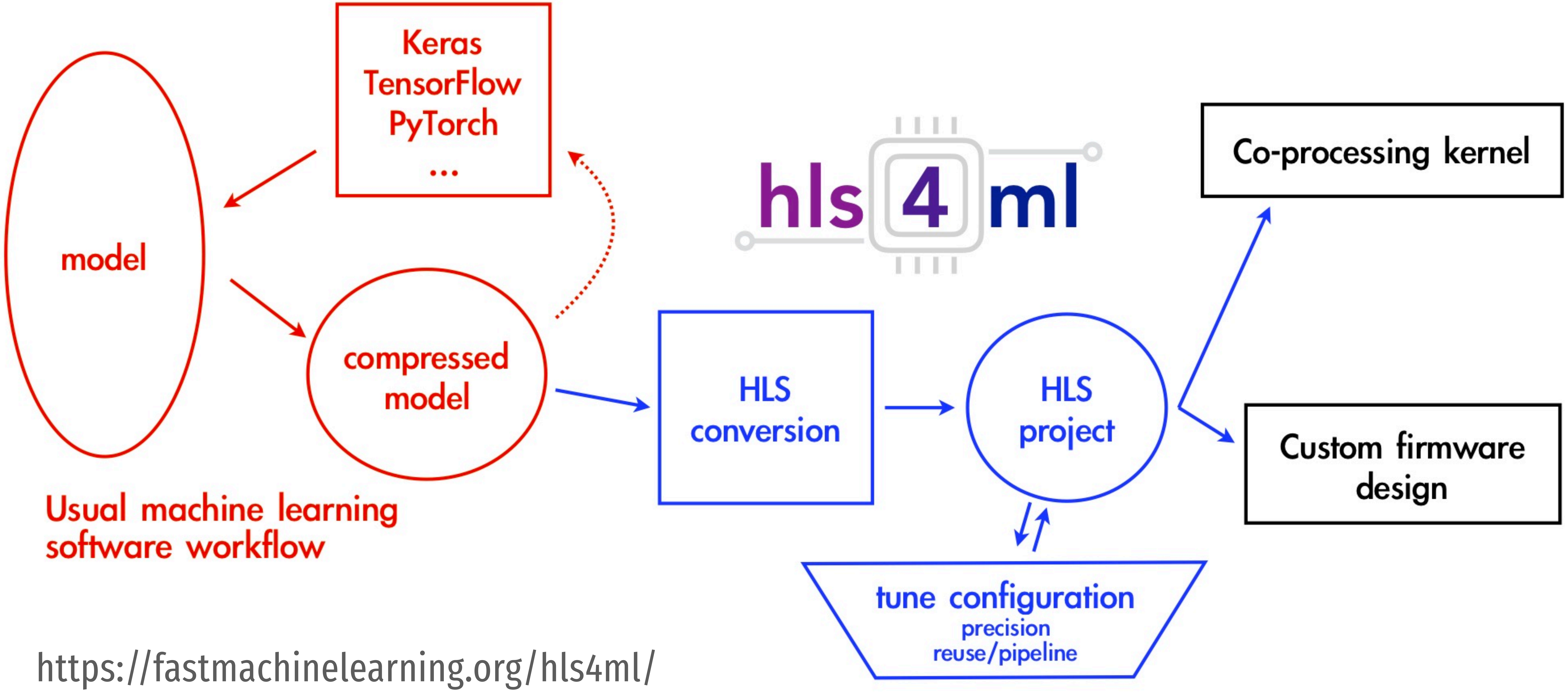
Artur Lobanov |
Machine Learning @ CMS L1 Trigger |
Workshop on Realtime ML, 10.4.24 |



Shown here is a comparison of the teacher model ability to reconstruct a Zero Bias (ZB) beam event (original: far left, reconstructed: center left) versus a signal sample, Soft Unclustered Energy Patterns (SUEP) on the right (original: center right, reconstructed: far right). In general, the teacher model is better able to reconstruct the Zero Bias beam event as evidenced by a far lower loss (0.81) compared to the SUEP loss (14.21). This example shows how the CICADA anomaly detection mechanism works to find anomalies. From [CMS DP-2023/086]

FROM PC TO FPGA

- hls4ml: package for translating NN to FPGA firmware



https://fastmachinelearning.org/hls4ml/