Realtime Anomaly Detection with the CMS Level-1 Global Trigger Test Crate

Sioni Summers (CERN)
for the CMS Collaboration

EuCAIFCon 2024

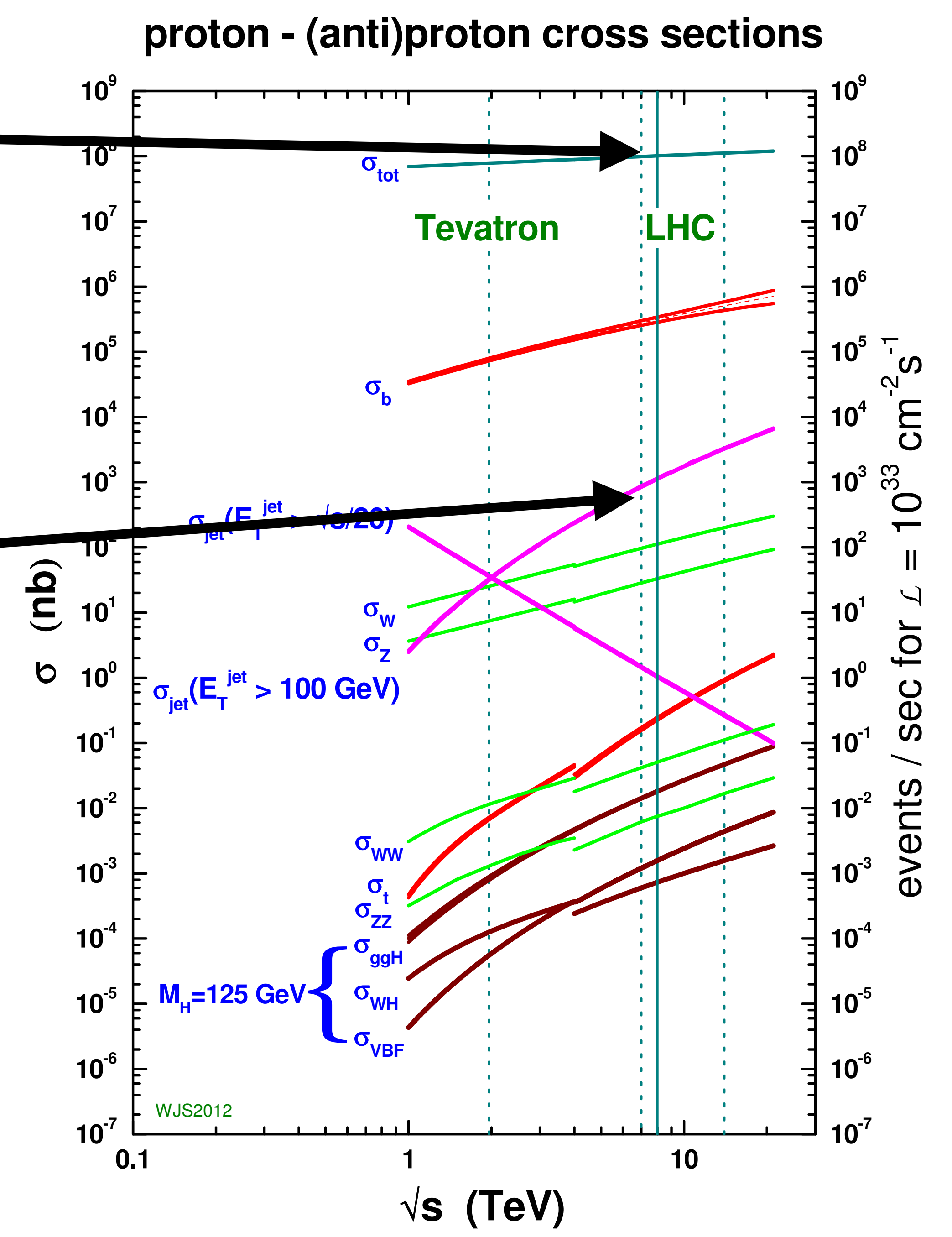
Trigger

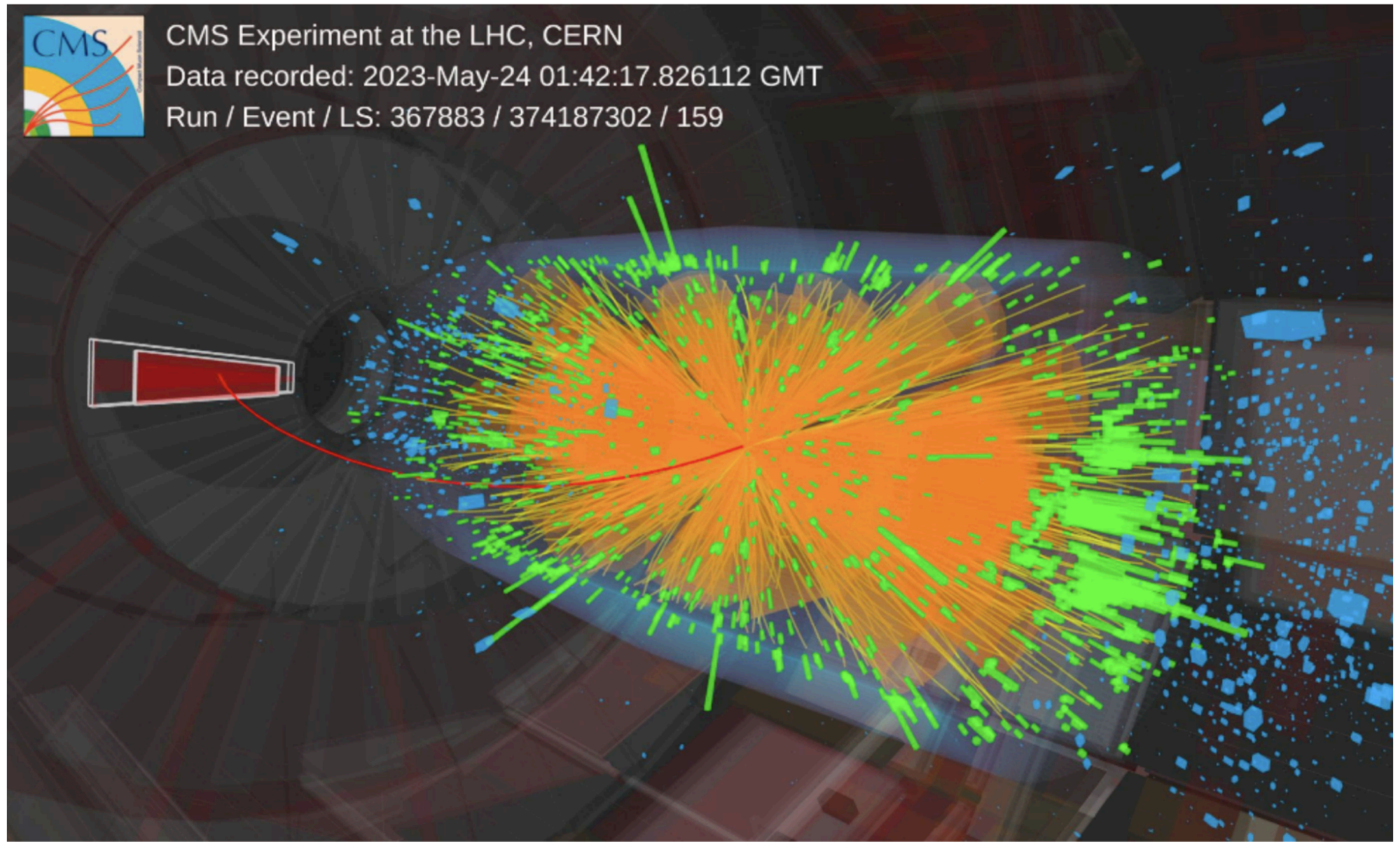
- At the CMS experiment at the CERN Large Hadron Collider
- Searching for rare phenomena → large number of collision events
- Granular detector for detailed analysis → high data rates
- Trigger is realtime data reduction, with 5 μs latency, in FPGA processors
  - Fast processing, decide which events to keep or discard

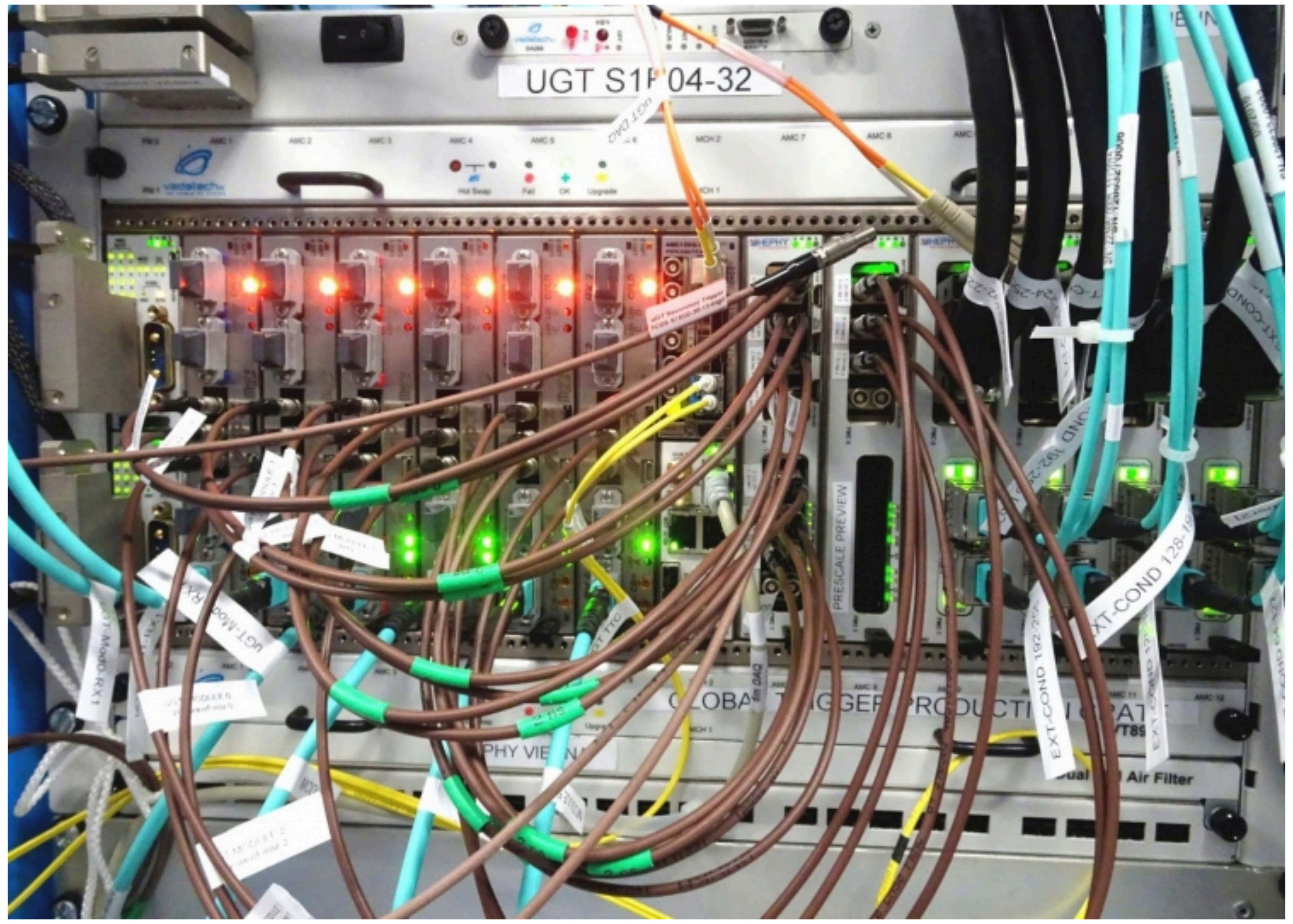
Total

Discard > 99%

To keep

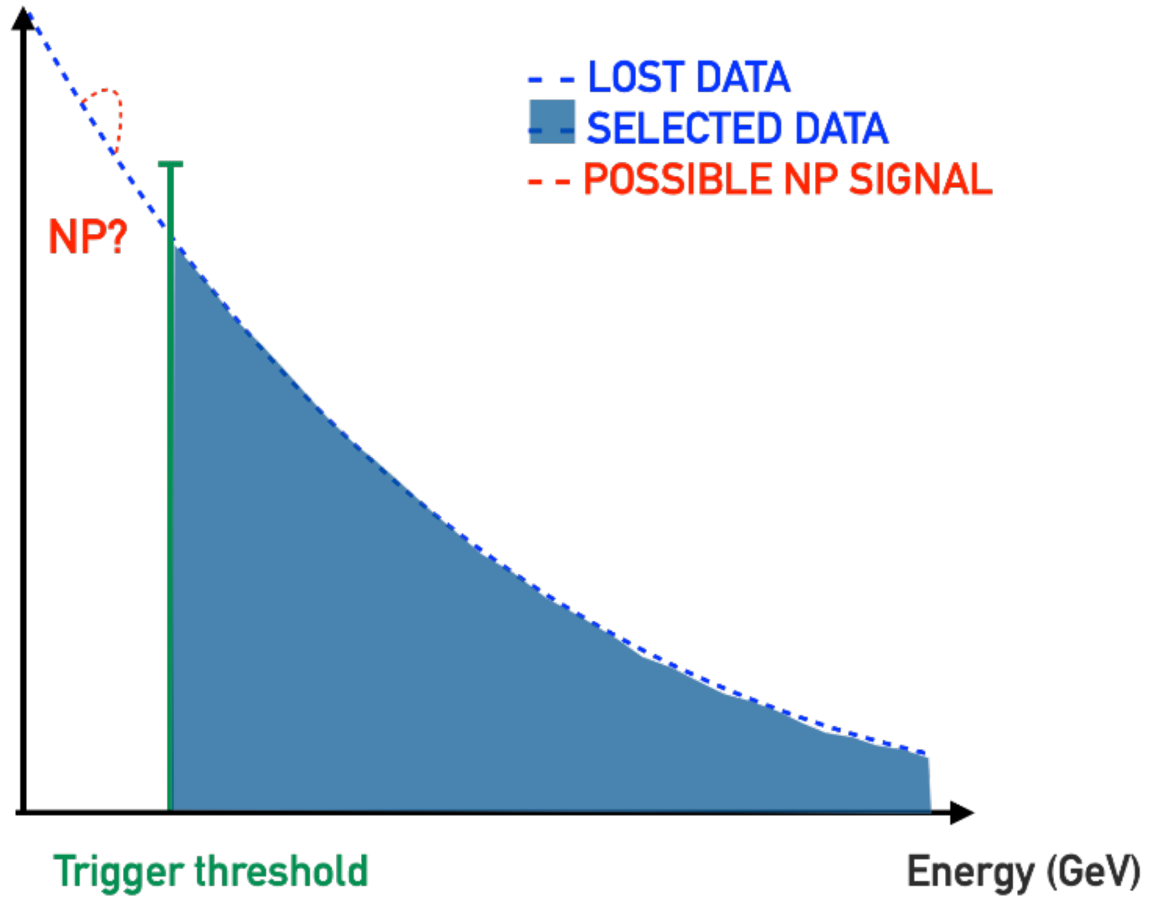




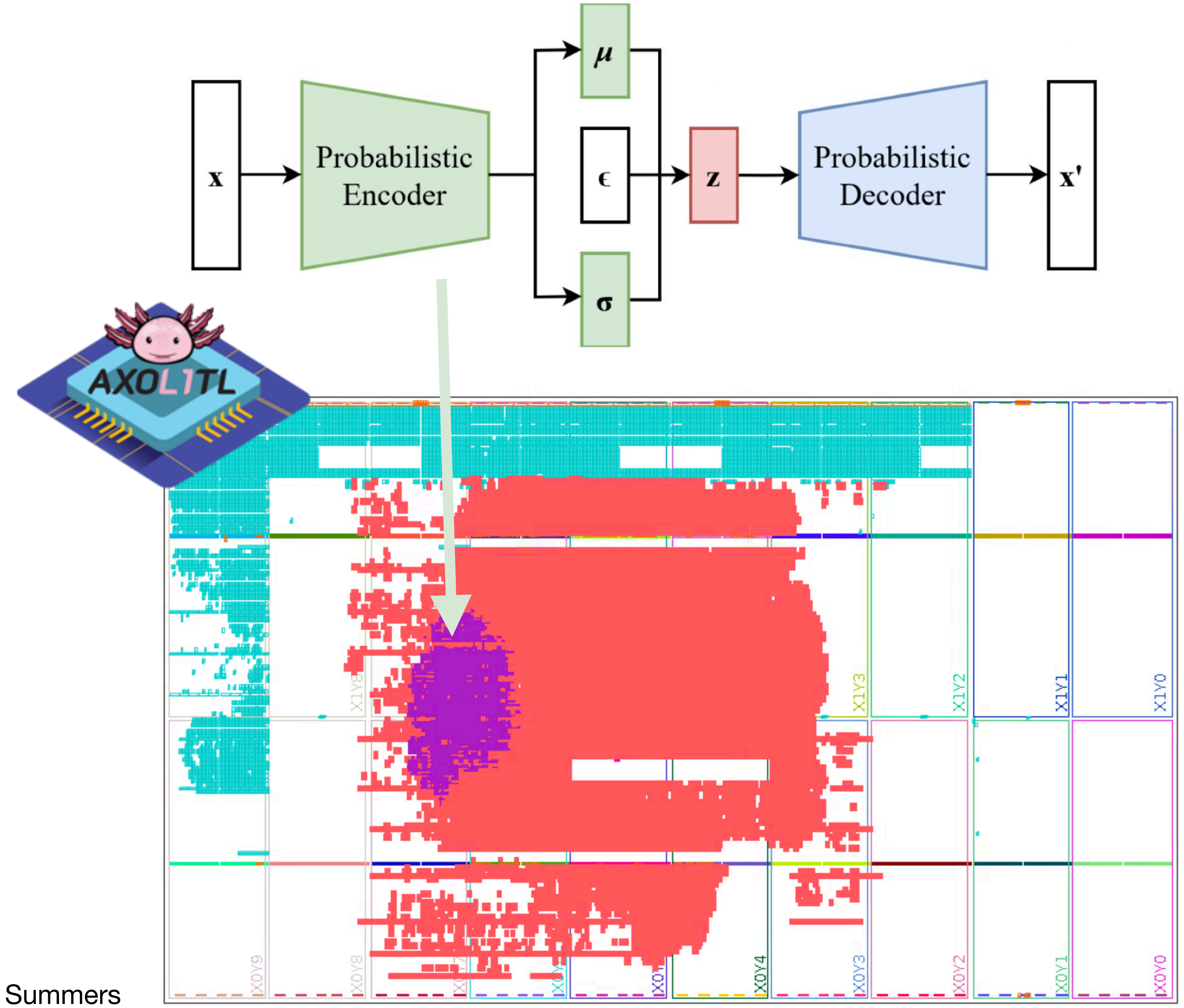


Keep / Discard

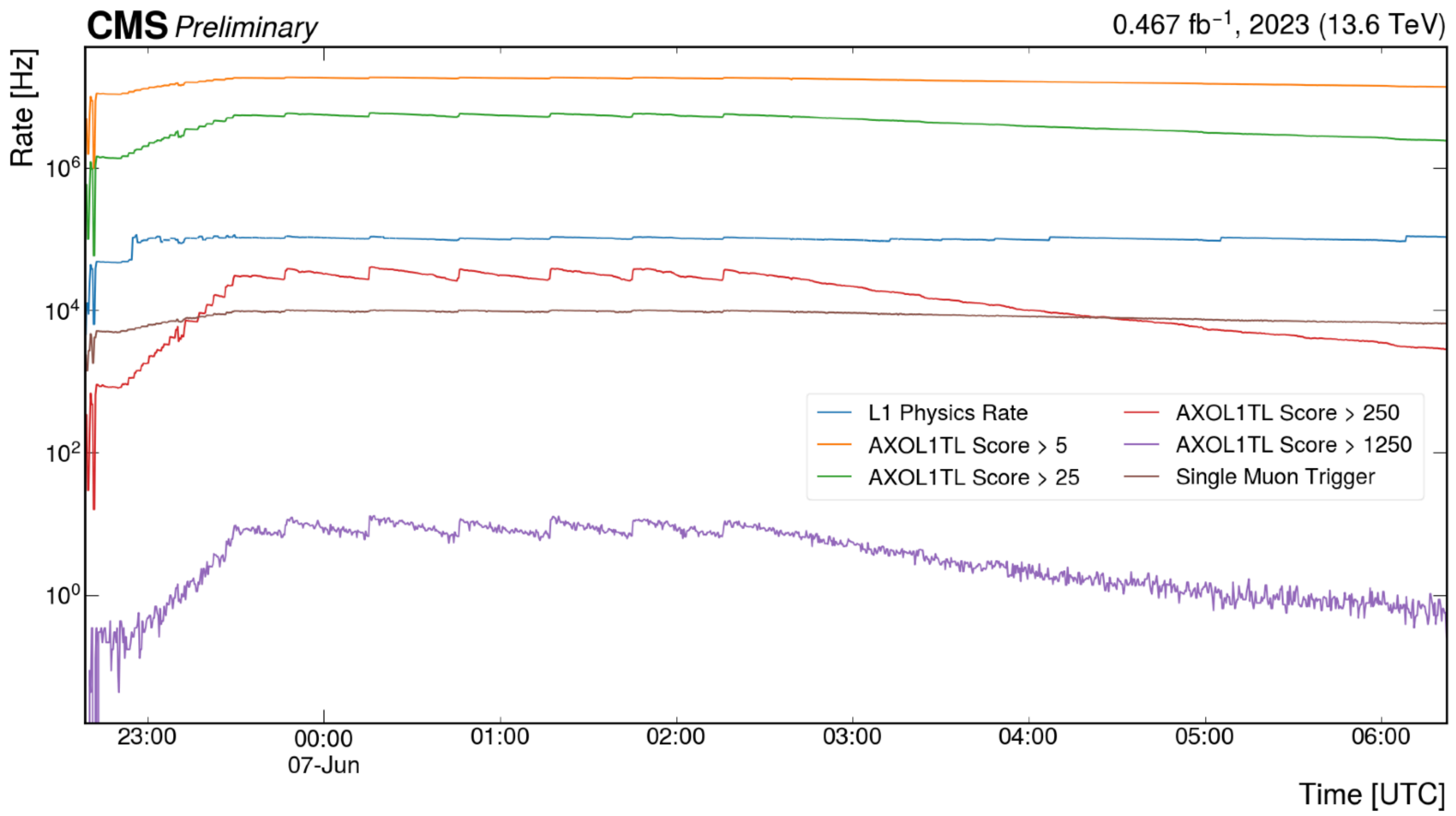
Anomaly Detection

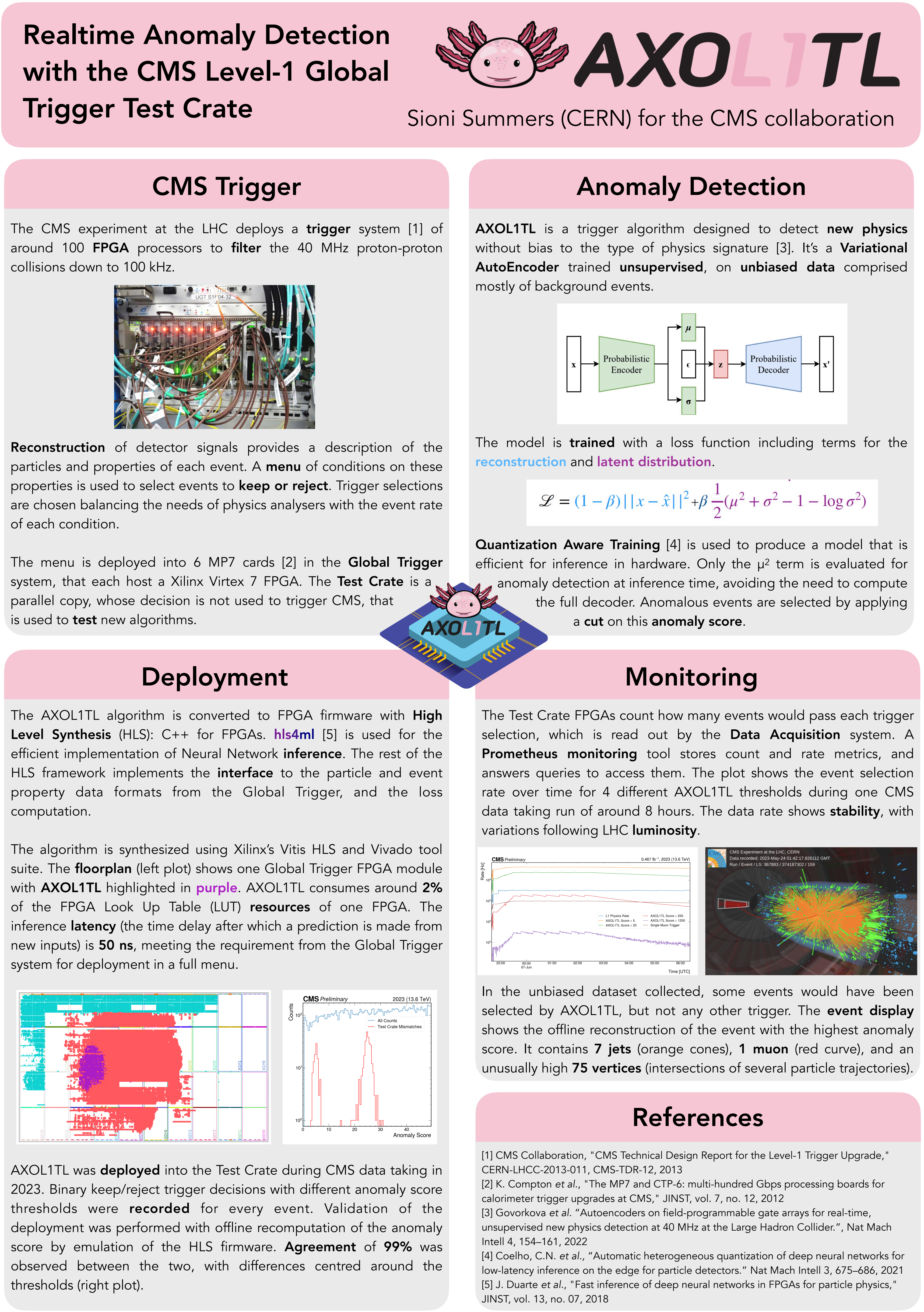


- Normal trigger selections compare the event particles to a table of rules
  - e.g. “is there a muon with transverse momentum above 22 GeV?”
- Could these selections reject the New Physics we’d like to see?
  - Especially low mass new particles
- AXOL1TL is an ML approach to Anomaly Detection searching generically for New Physics
- Tiny Variational AutoEncoder trained on unbiased data with Quantization Aware Training
- **hls4ml** converts the Neural Network to FPGA logic with 50 ns prediction latency
- To learn what this all means, visit the poster!



- Anomaly Detection has been deployed in the Global Trigger Test Crate in 2023
  - Run in “safe mode” alongside normal trigger
- Used to test performance and validate integration
- Check rate stability of selections and look at offline data





Realtime Anomaly Detection with the CMS Level-1 Global Trigger Test Crate

AXOL1TL

Sioni Summers (CERN) for the CMS collaboration

CMS Trigger

The CMS experiment at the LHC deploys a **trigger** system [1] of around 100 **FPGA** processors to **filter** the 40 MHz proton-proton collisions down to 100 kHz.

Reconstruction of detector signals provides a description of the particles and properties of each event. A **menu** of conditions on these properties is used to select events to **keep or reject**. Trigger selections are chosen balancing the needs of physics analysers with the event rate of each condition.

The menu is deployed into 6 MP7 cards [2] in the **Global Trigger** system, that each host a Xilinx Virtex 7 FPGA. The **Test Crate** is a parallel copy, whose decision is not used to trigger CMS, that is used to **test** new algorithms.

Anomaly Detection

AXOL1TL is a trigger algorithm designed to detect **new physics** without bias to the type of physics signature [3]. It's a **Variational AutoEncoder** trained **unsupervised**, on **unbiased data** comprised mostly of background events.

The model is **trained** with a loss function including terms for the reconstruction and latent distribution.

$$\mathscr{L} = (1-\beta)||x-\hat{x}||^2 + \beta\frac{1}{2}(\mu^2+\sigma^2-1-\log\sigma^2)$$

Quantization Aware Training [4] is used to produce a model that is efficient for inference in hardware. Only the μ^2 term is evaluated for anomaly detection at inference time, avoiding the need to compute the full decoder. Anomalous events are selected by applying a **cut** on this **anomaly score**.

Deployment

The AXOL1TL algorithm is converted to FPGA firmware with **High Level Synthesis** (HLS): C++ for FPGAs. **hls4ml** [5] is used for the efficient implementation of Neural Network **inference**. The rest of the HLS framework implements the **interface** to the particle and event property data formats from the Global Trigger, and the loss computation.

The algorithm is synthesized using Xilinx's Vitis HLS and Vivado tool suite. The **floorplan** (left plot) shows one Global Trigger FPGA module with **AXOL1TL** highlighted in purple. AXOL1TL consumes around **2%** of the FPGA Look Up Table (LUT) **resources** of one FPGA. The inference **latency** (the time delay after which a prediction is made from new inputs) is **50 ns**, meeting the requirement from the Global Trigger system for deployment in a full menu.

AXOL1TL was **deployed** into the Test Crate during CMS data taking in 2023. Binary keep/reject trigger decisions with different anomaly score thresholds were **recorded** for every event. Validation of the deployment was performed with offline recomputation of the anomaly score by emulation of the HLS firmware. **Agreement** of **99%** was observed between the two, with differences centred around the thresholds (right plot).

Monitoring

The Test Crate FPGAs count how many events would pass each trigger selection, which is read out by the **Data Acquisition** system. A **Prometheus monitoring** tool stores count and rate metrics, and answers queries to access them. The plot shows the event selection rate over time for 4 different AXOL1TL thresholds during one CMS data taking run of around 8 hours. The data rate shows **stability**, with variations following LHC **luminosity**.

In the unbiased dataset collected, some events would have been selected by AXOL1TL, but not any other trigger. The **event display** shows the offline reconstruction of the event with the highest anomaly score. It contains **7 jets** (orange cones), **1 muon** (red curve), and an unusually high **75 vertices** (intersections of several particle trajectories).

References

[1] CMS Collaboration, "CMS Technical Design Report for the Level-1 Trigger Upgrade," CERN-LHCC-2013-011, CMS-TDR-12, 2013
[2] K. Compton *et al.*, "The MP7 and CTP-6: multi-hundred Gbps processing boards for calorimeter trigger upgrades at CMS," JINST, vol. 7, no. 12, 2012
[3] 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, 154–161, 2022
[4] Coelho, C.N. *et al.*, “Automatic heterogeneous quantization of deep neural networks for low-latency inference on the edge for particle detectors.” Nat Mach Intell 3, 675–686, 2021
[5] J. Duarte *et al.*, "Fast inference of deep neural networks in FPGAs for particle physics," JINST, vol. 13, no. 07, 2018

Baler: A tool for machine learning based data compression

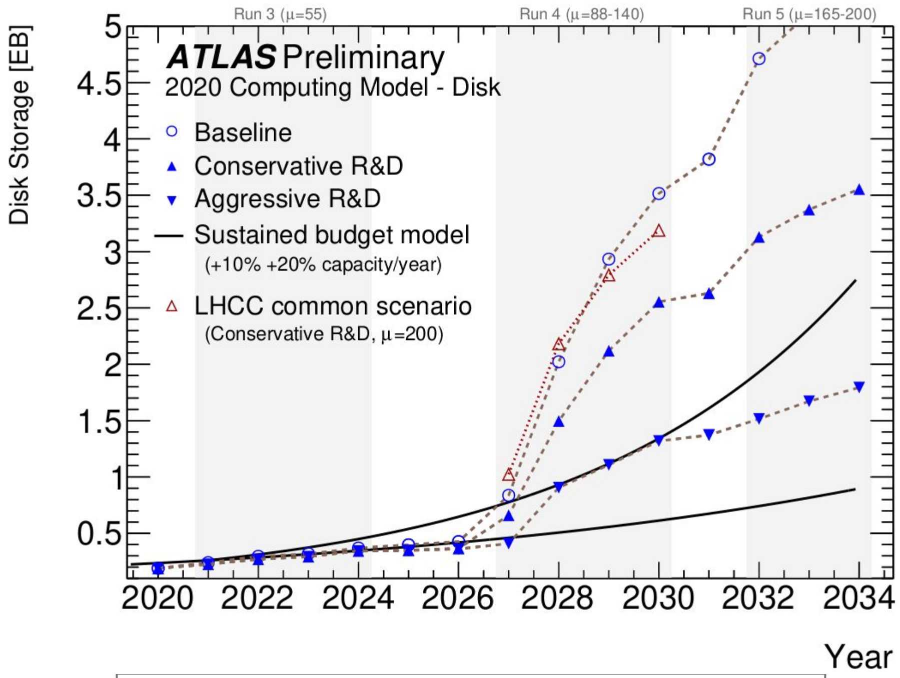
Alexander Ekman for the Baler collaboration

Problem: More data than storage

KUNGL. FYSIOGRAFISKA SÄLLSKAPET I LUND
AKADEMI FÖR NATURVETENSKAP, MEDICIN OCH TEKNIK

LUND UNIVERSITY

- Collection and generation of data is overwhelming processing and storage capacity in science and industry
- High demand for greater compression than traditional lossless and lossy methods



ATLAS HL-LHC Computing Conceptual Design Report
Calafiura, P ; Catmore, J ; Costanzo, D ; Di Girolamo, A
http://cds.cern.ch/record/2729668/

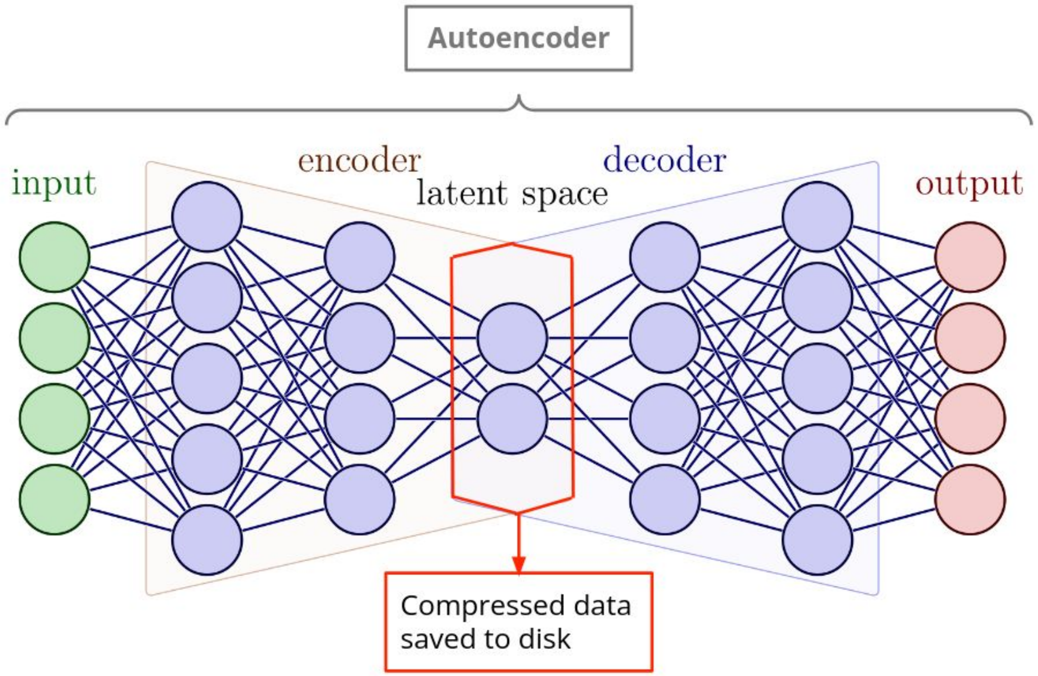
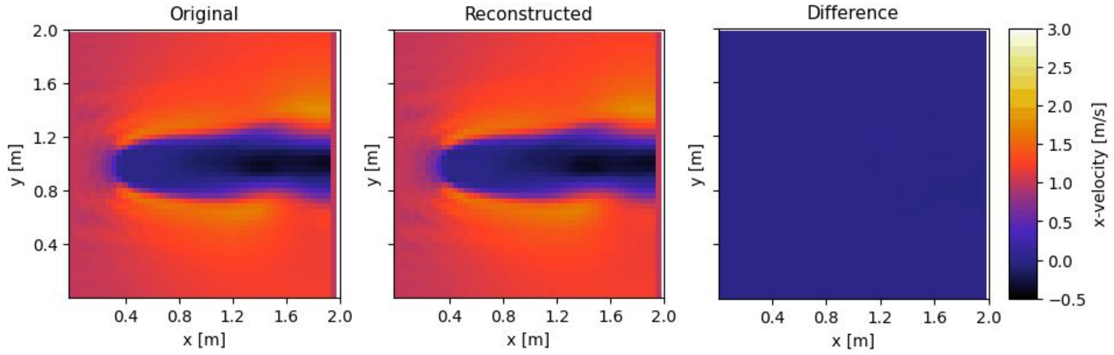


Figure modified from:
https://tikz.net/neural_networks/

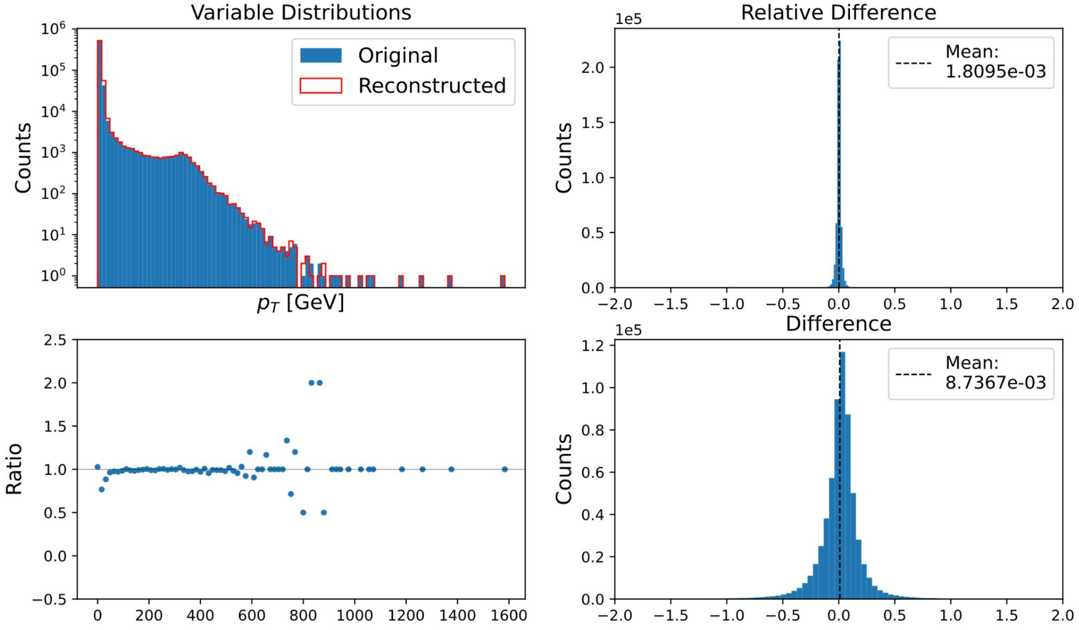
Our solution: “Baler”

- Multidisciplinary tool to investigate the viability of this compression method
  - https://github.com/baler-collaboration/baler
- Simple to install as a pip package
  - `pip install baler-compressor`
- Promising performance for varying scientific fields

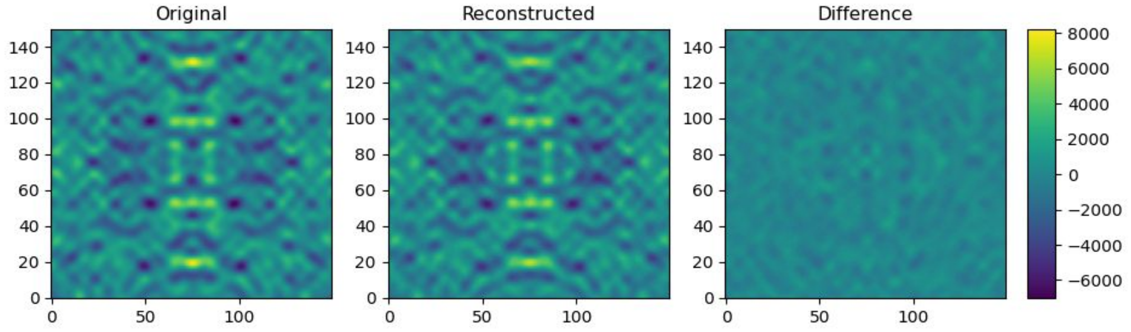
Computational Fluid dynamics



High Energy Particle Physics

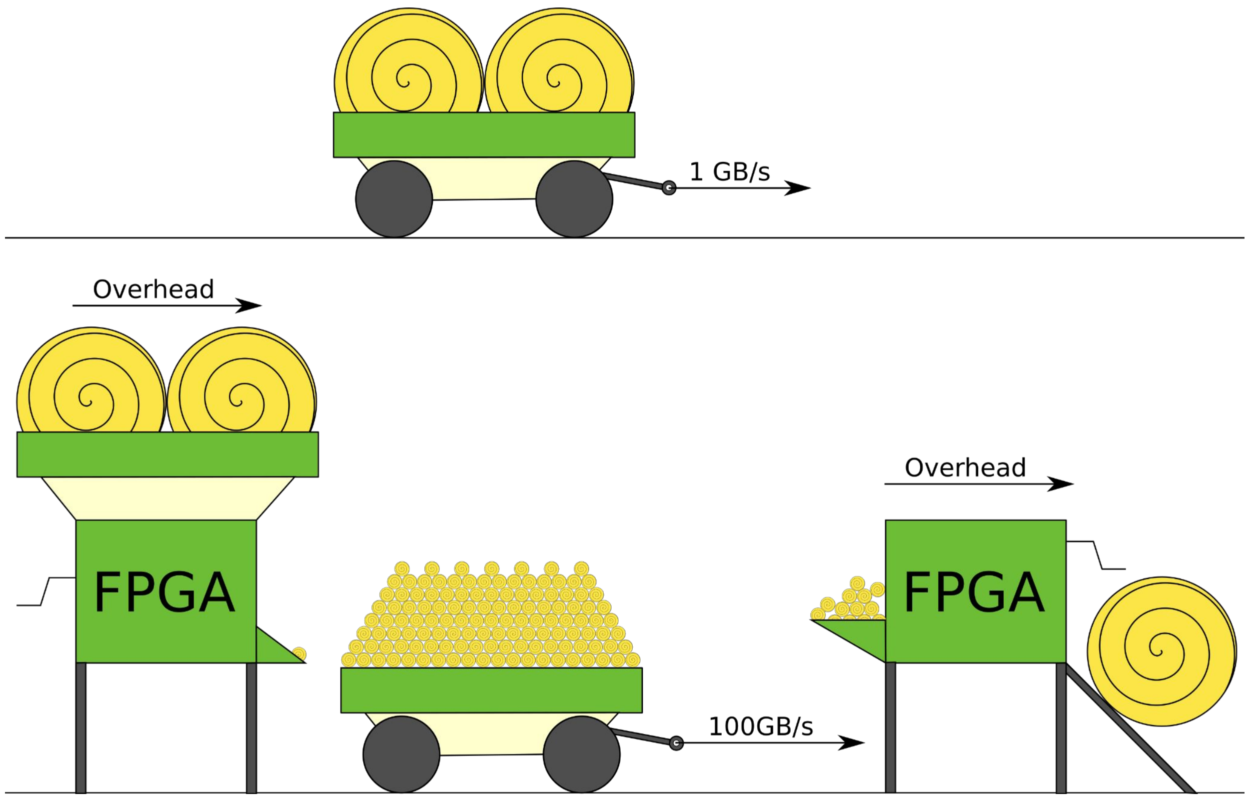


X-spectroscopy



Future outlook

- We found a small demand for lossy compression of scientific data in final storage
- New focus on “online” compression and bandwidth compression using FPGA technology
- Draw inspiration from progress in machine learning based image and video compression



Embedded Neural Networks on FPGAs for Real-Time Computation of the Energy Deposited in the ATLAS Liquid Argon Calorimeter

Raphaël BERTRAND

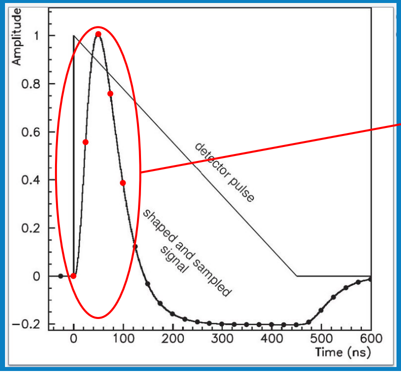
Aix Marseille Univ, CNRS/IN2P3, CPPM, Marseille, France

CENTRE DE PHYSIQUE DES PARTICULES DE MARSEILLE
CPPM

cnrs
IN2P3

ATLAS EXPERIMENT

Introduction



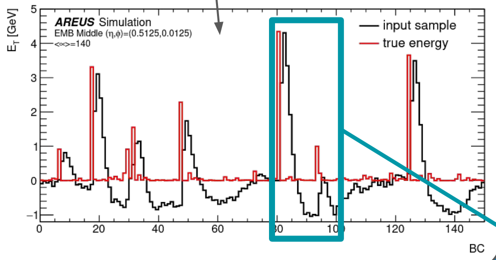
$$E_t = \sum_{i=1}^{5} c_i * x_{t+i}$$

Energy deposits in LAr calorimeter cell→electric pulse

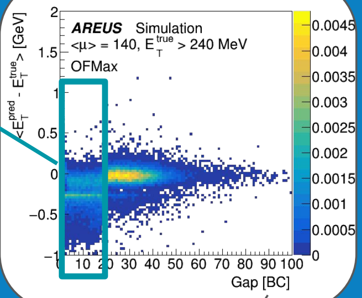
- **shaped**
- **sampled**
- **digitized** at **40 MHz**

The pulse **spans** about **625 ns**.

Simulated pulse chain with additional injected pulses at higher energy



Region where pulses overlap



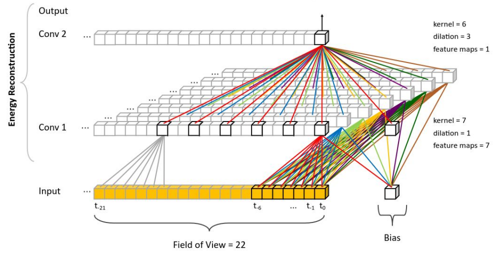
Bunch Crossing : 25 ns

- HL-LHC **high pileup → OF performance degradation**
- Neural networks are investigated as an alternative solution to the OF algorithm.

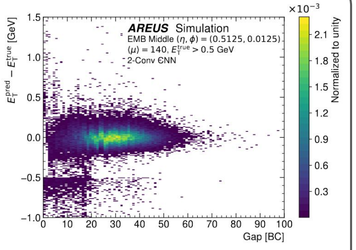
Neural network structure

CNN with ~100 parameters

Energy prediction for every BC

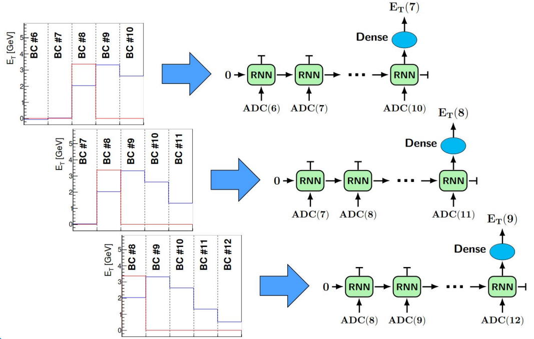


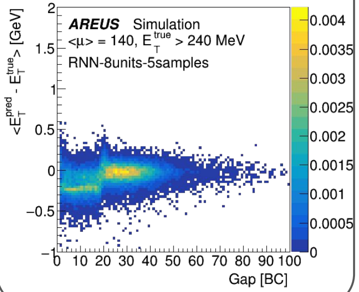
Continuous input from one detector cell



2-Conv CNN

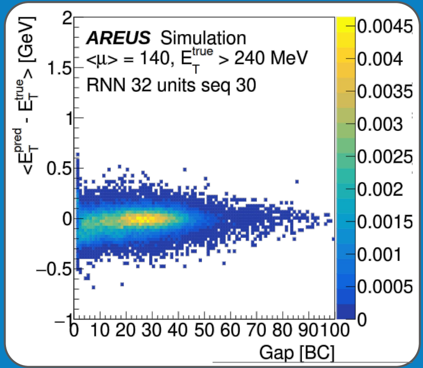
RNN 8 units 5 input samples





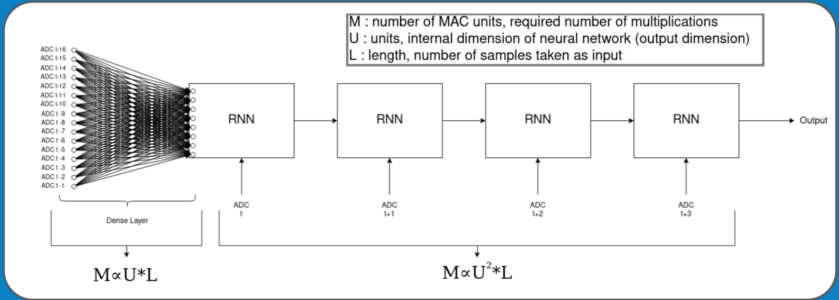
- **CNNs** and **RNNs** are designed to **compute deposited energy**.
- **NNs can correct the degradation** of the energy resolution.

Optimization of the RNN



RNN with 8 units and 5 samples as input can be upgraded :

- Increase nb of units
  ↳ better resolution overall
- Increase nb of input samples
  ↳better resolution with overlapped pulses



Increasing the number of units and input samples go with more computations and it can't be implemented.

- Dense Layer as input of the 1st RNN cell for input samples before the energy deposit.
  - RNN cells to compute the amplitude on the peak
  - Dense to correct for the pileup

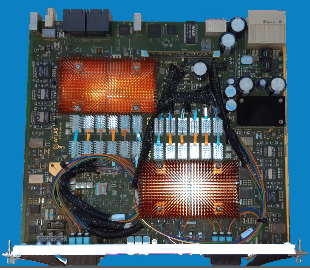
Firmware implementation

LASP demonstrator built with Stratix-10

- prototype with Agilex 7 ongoing

Each FPGA needs to reconstruct the energy for 384 channels :

- Impossible to implement 384 NNs on the FPGA
  - Need multiplexing
  - Need higher frequency



LASP board demonstrator

RNN and CNN Implemented on Stratix-10

- CNN implemented on Agilex, RNN still in progress
- CNN directly implemented in VHDL, RNN Implemented first in HLS for fast prototyping and then optimized in VHDL
- Fits LAr requirements for both

FPGA	Network	Multiplex.	Detector cells	f_{max}	ALMs	DSPs
Stratix-10	RNN (HLS)	10	370	393 MHz	90 %	100 %
	RNN (VHDL)	14	392	561 MHz	18 %	66 %
	CNN (100 param.)	12	396	415 MHz	8 %	28 %
Agilex	CNN (100 param.)	12	396	539 MHz	4 %	13 %
	CNN (400 param.)	12	396	510 MHz	19 %	50 %