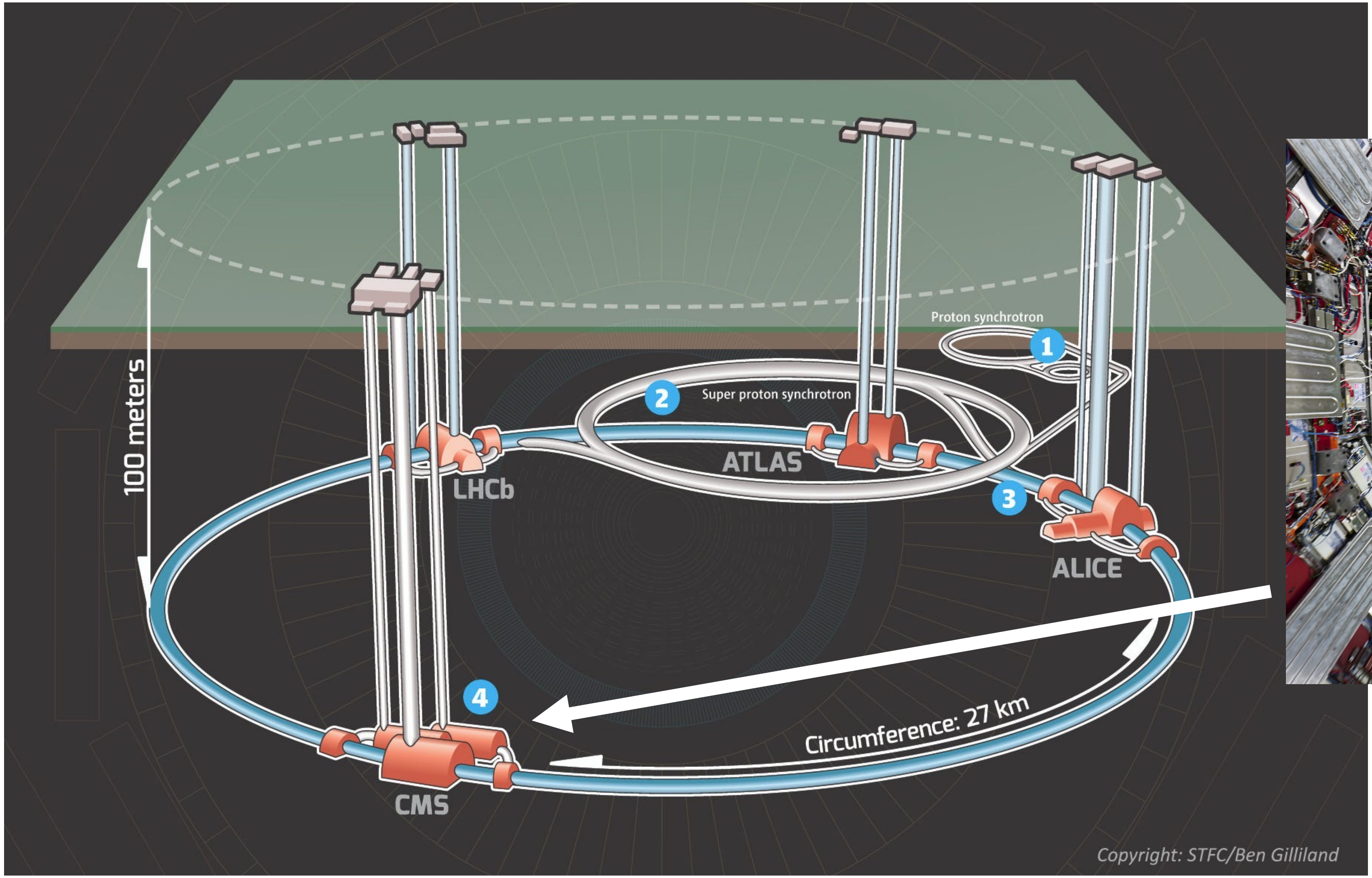
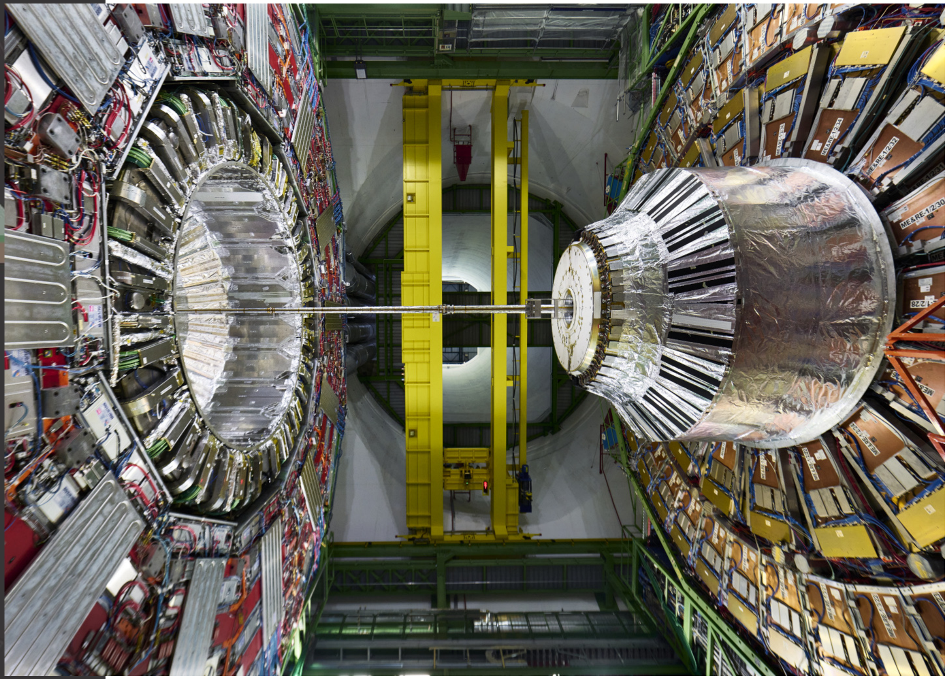
Real-time Anomaly Detection in the CMS Experiment

Noah Zipper on behalf of the CMS Collaboration

The Large Hadron Collider (LHC) @ CERN





https://home.cern/science/experiments/cms

The CMS Trigger

How can we deal with new collision data ~40 million times a second?

- We read in >60 TB/s from the detector!

The CMS Trigger

How can we deal with new collision data ~40 million times a second?

- We read in >60 TB/s from the detector!

The trigger cuts > 99.9% of incoming data, only picks interesting interactions

The CMS Trigger

How can we deal with new collision data ~40 million times a second?

- We read in >60 TB/s from the detector!

The trigger cuts > 99.9% of incoming data, only picks interesting interactions

We use a set of algorithms – the “trigger menu” – that looks at each event and decides to keep or toss data

The CMS Trigger

How can we deal with new collision data ~40 million times a second?

- We read in >60 TB/s from the detector!

The trigger cuts > 99.9% of incoming data, only picks interesting interactions

We use a set of algorithms – the "trigger menu" – that looks at each event and decides to keep or toss data

The trigger is broken up into two phases

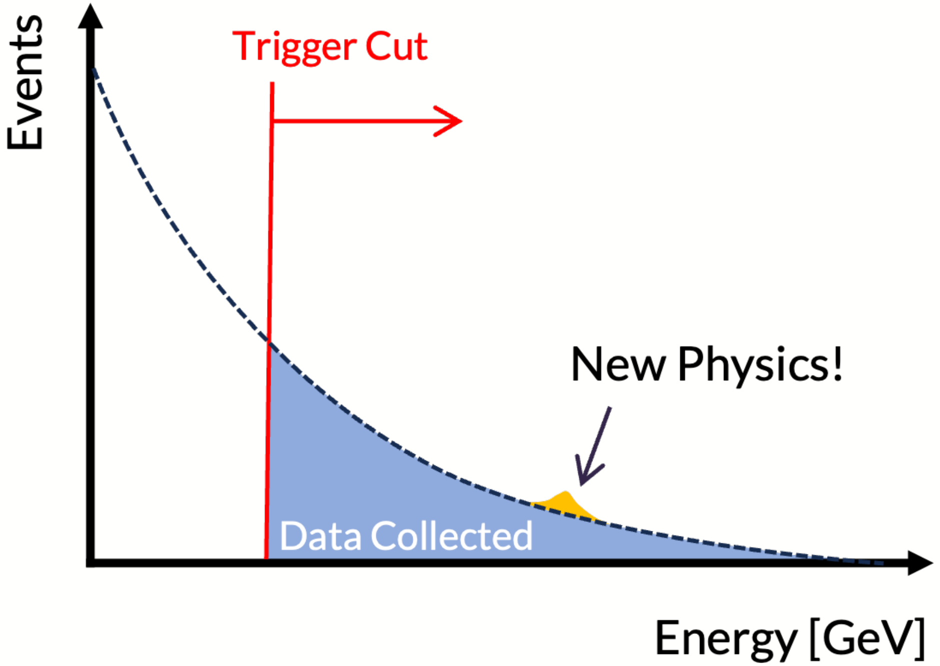
- Level-1 (L1T) – First step of real-time triggering, happens on hardware
  - Decisions in < 5 microseconds
- High-Level (HLT) – Data is passed from hardware to off-detector software
  - Decisions in < ½ second

Why Anomaly Detection?

Currently, we use simple heuristics to define trigger algorithms

- Energy, charge, direction, momentum, etc.

In this approach, we need to know what we're looking for to target it



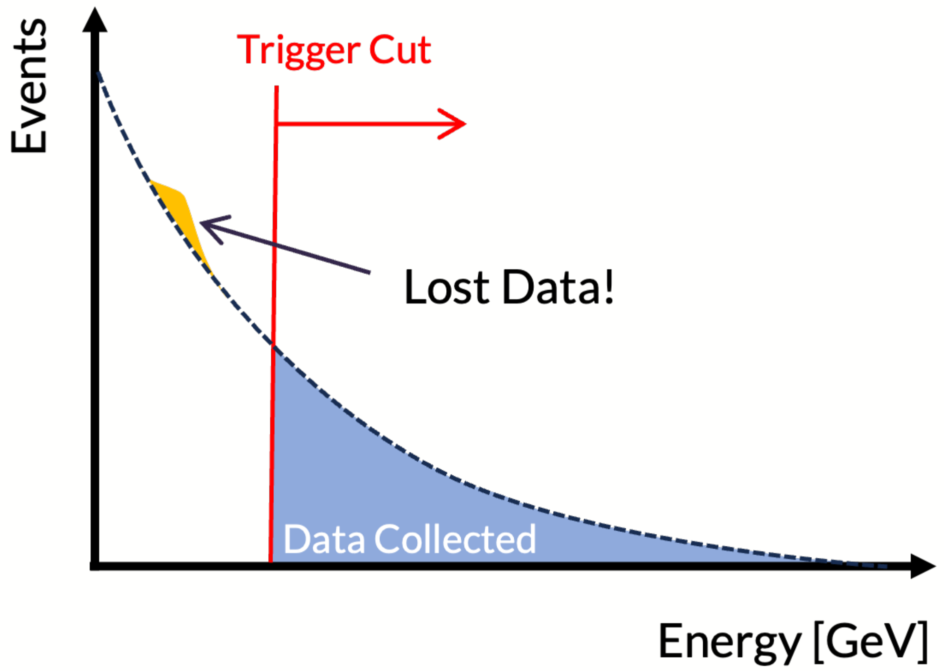
Why Anomaly Detection?

Currently, we use simple heuristics to define trigger algorithms

- Energy, charge, direction, momentum, etc.

In this approach, we need to know what we're looking for to target it

- **How do we stop rejecting data because we don't know what to look for?**



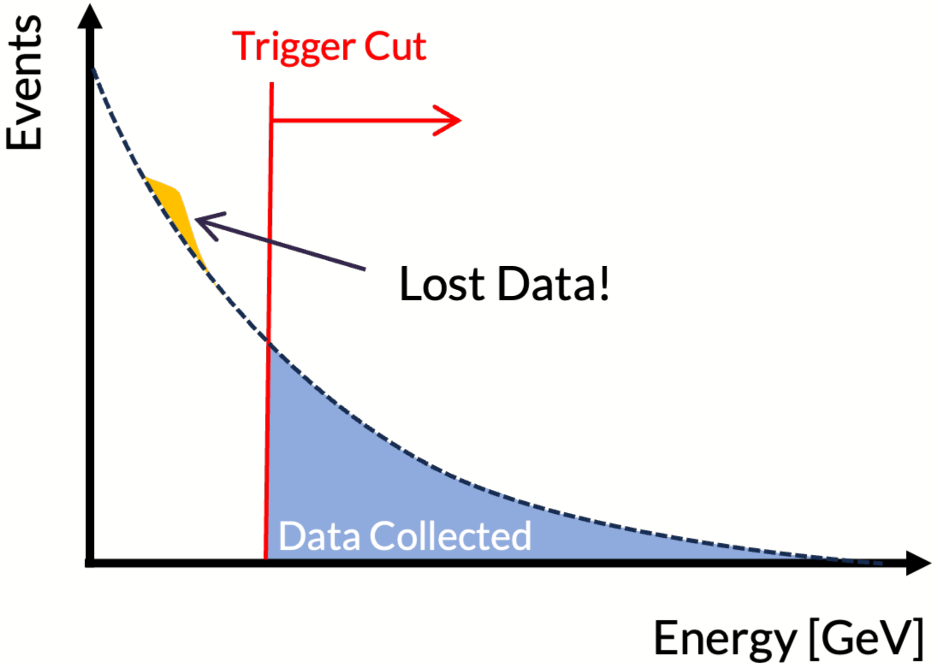
Why Anomaly Detection?

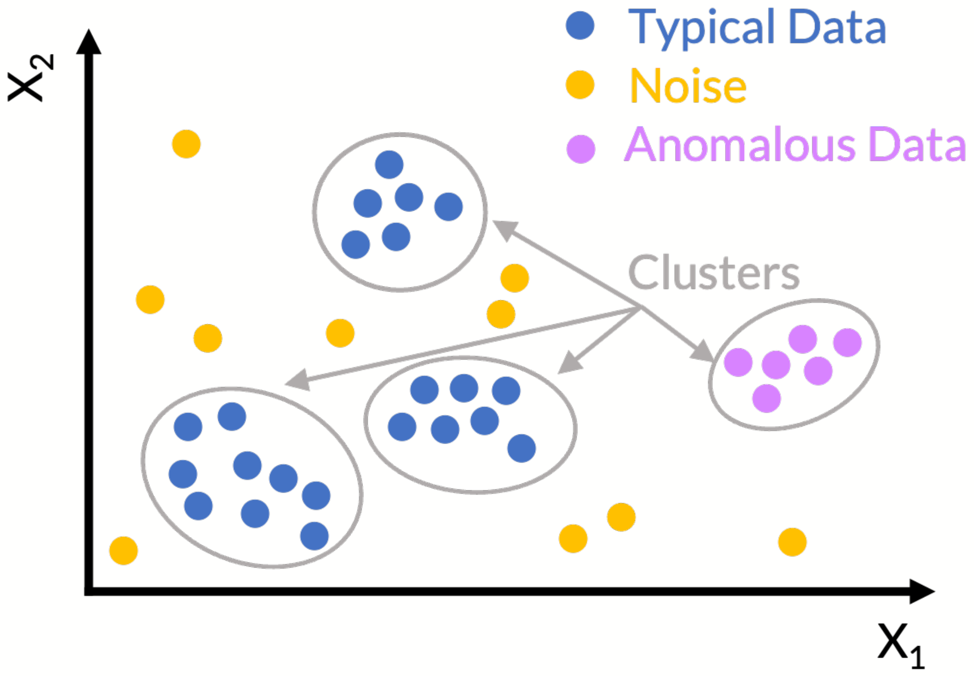
Currently, we use simple heuristics to define trigger algorithms

- Energy, charge, direction, momentum, etc.

In this approach, we need to know what we're looking for to target it

- **How do we stop rejecting data because we don't know what to look for?**





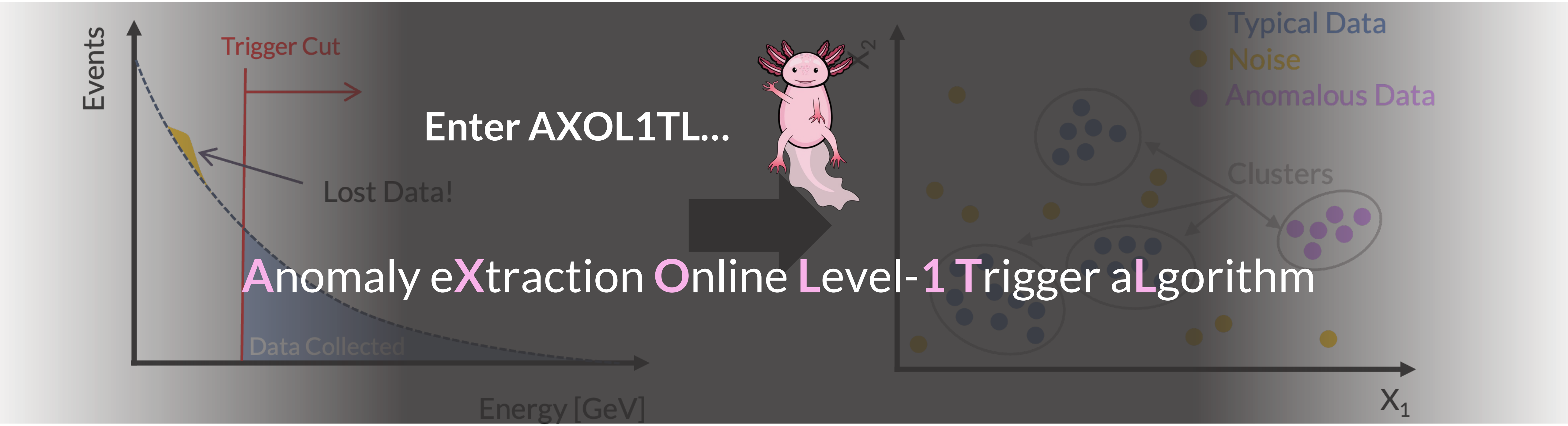
Why Anomaly Detection?

Currently, we use simple heuristics to define trigger algorithms

- Energy, charge, direction, momentum, etc.

In this approach, we need to know what we're looking for to target it

- **How do we stop rejecting data because we don't know what to look for?**



Enter AXOL1TL...

Anomaly eXtraction Online Level-1 Trigger aLgorithm

L1 Anomaly Detection @ LHC

“Zero Bias”

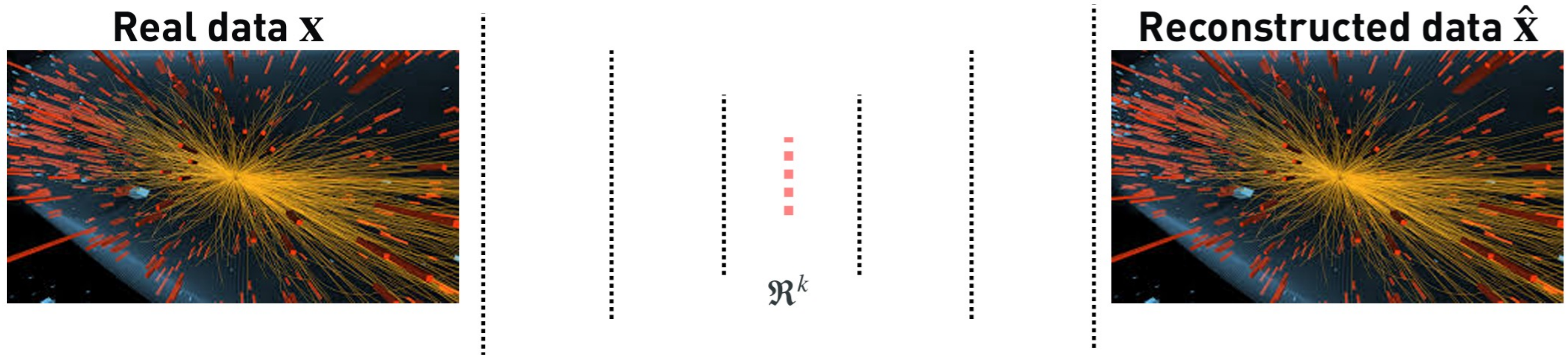
A dataset with no triggers, only turned on for small slices of time. Records events synched up with when collisions occur, saves everything.

AXO is an unsupervised Variational Autoencoder (VAE)

- Simple neural network(s), trained on real Zero Bias* data
- Basic L1 trigger objects as vector inputs
  - (p_T, η, ϕ) for 1 p_T^{miss}, 4 e/γ, 4 μ, and 10 jets

VAE uses encoder & decoder to compress and reconstruct the input data

- Squeeze data into a small dimension “latent space”
  - Forces efficient information encoding ➔ network “learns”
- Network gets good at encoding + decoding typical data examples



L1 Anomaly Detection @ LHC

"Zero Bias"

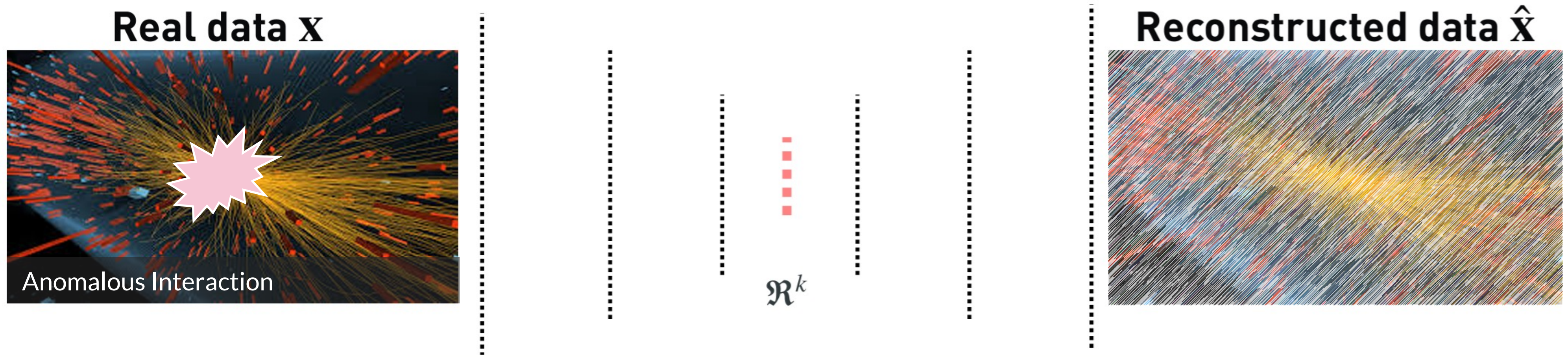
A dataset with no triggers, only turned on for small slices of time. Records events synched up with when collisions occur, saves everything.

AXO is an unsupervised Variational Autoencoder (VAE)

- Simple neural network(s), trained on real Zero Bias* data
- Basic L1 trigger objects as vector inputs
    - (p_T, η, ϕ) for 1 p_T^{miss}, 4 e/γ, 4 μ, and 10 jets

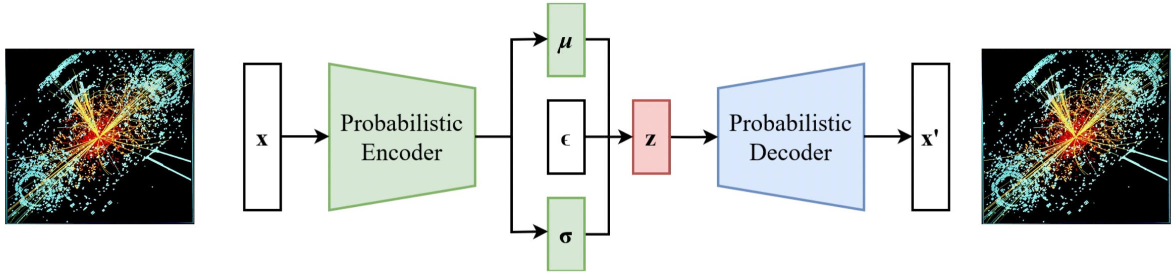
VAE uses encoder & decoder to compress and reconstruct the input data

- Squeeze data into a small dimension "latent space"
    - Forces efficient information encoding ➔ network "learns"
- Network gets good at encoding + decoding typical data examples
- Much worse for atypical examples



Model Design

Level-1 Trigger constraints informed design



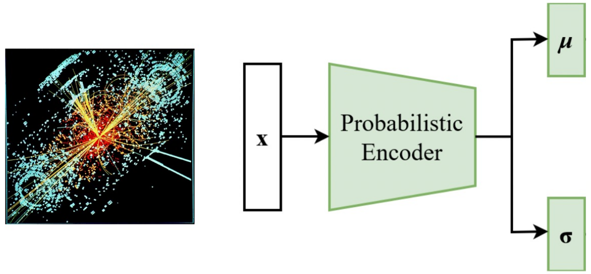
$$\text{Loss} = (1-\beta)||x-\hat{x}||^2 + \beta\frac{1}{2}(\mu^2+\sigma^2-1\ -\log\sigma^2)$$

Reconstruction term Full regularization term

- Standard optimization approaches for fast-ML
  - Pruning, truncation, quantization-aware training

Model Design

Level-1 Trigger constraints informed design



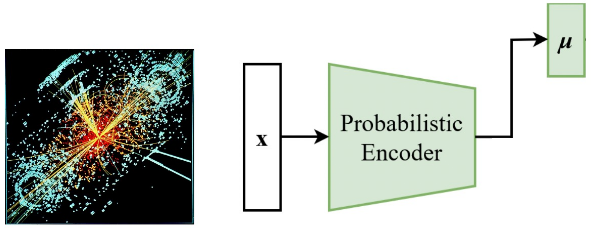
$$\text{Loss} = \cancel{(1-\beta)\|x-\hat{x}\|^2} + \beta\,\frac{1}{2}(\mu^2 + \sigma^2 - 1 \; - \log\sigma^2)$$

Reconstruction term   Full regularization term

- Standard optimization approaches for fast-ML
  - Pruning, truncation, quantization-aware training
- Remove decoder network
  - Significant latency & resource savings, minimal performance degradation

Model Design

Level-1 Trigger constraints informed design



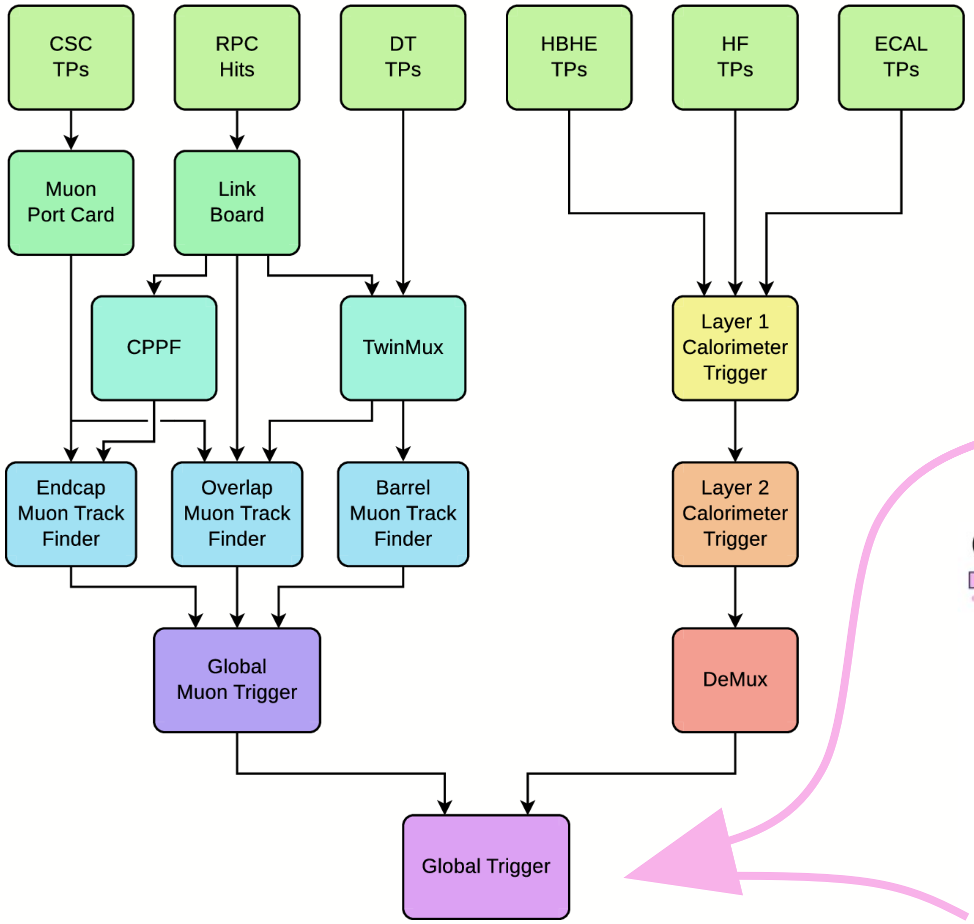
$$\text{Loss} = \cancel{(1-\beta)\|x-\hat{x}\|^2} + \beta\,\frac{1}{2}\left(\mu^2 + \cancel{\sigma^2 - 1 - \log\sigma^2}\right)$$

Reconstruction term | Full regularization term

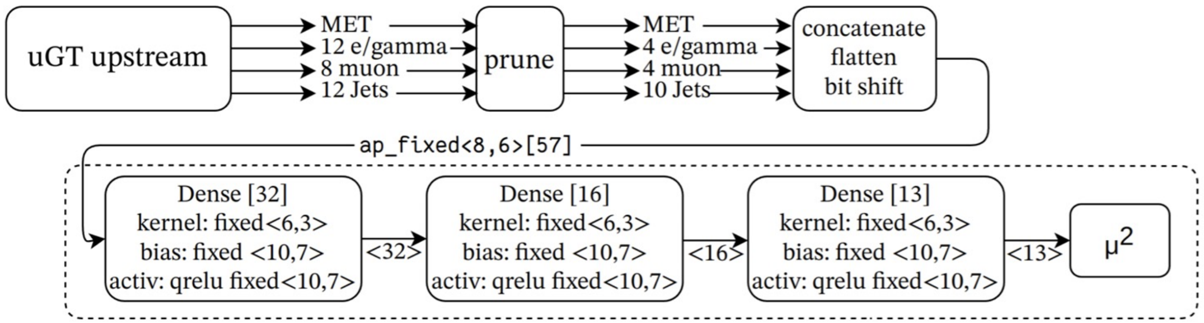
- Standard optimization approaches for fast-ML
  - Pruning, truncation, quantization-aware training
- Remove decoder network
  - Significant latency & resource savings, minimal performance degradation
- Remove latent σ term from loss calculation
  - Saves even more on timing, negligible performance degradation

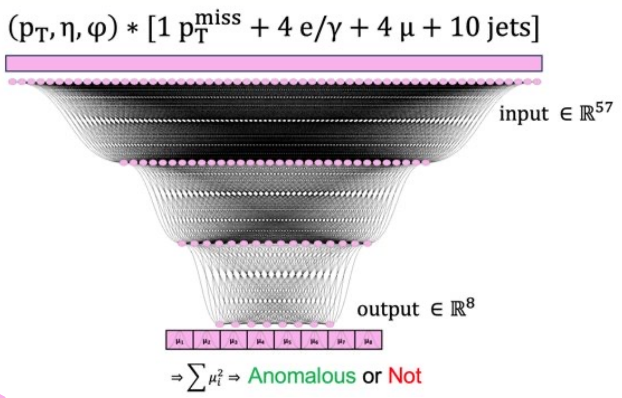
Integrating into the Trigger System

The CMS Level-1 Trigger System



AXO Algorithm

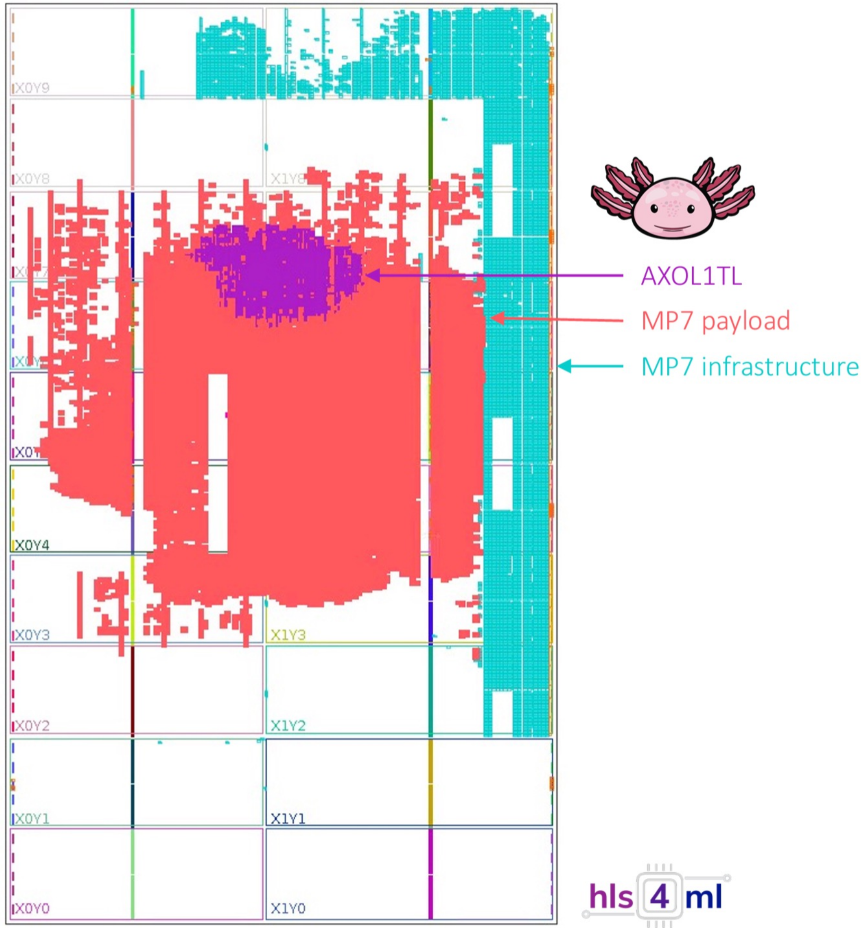




Integrating into the Trigger System

Algorithm must run on Field Programmable Gate Arrays (FPGAs)

- Firmware sits in MP7 Global Trigger board
    - Xilinx Virtex 7 chip
- AXO runs in < 50 nanoseconds
- Whole algorithm chain takes a few microseconds



Integrating into the Trigger System

Algorithm must run on Field Programmable Gate Arrays (FPGAs)

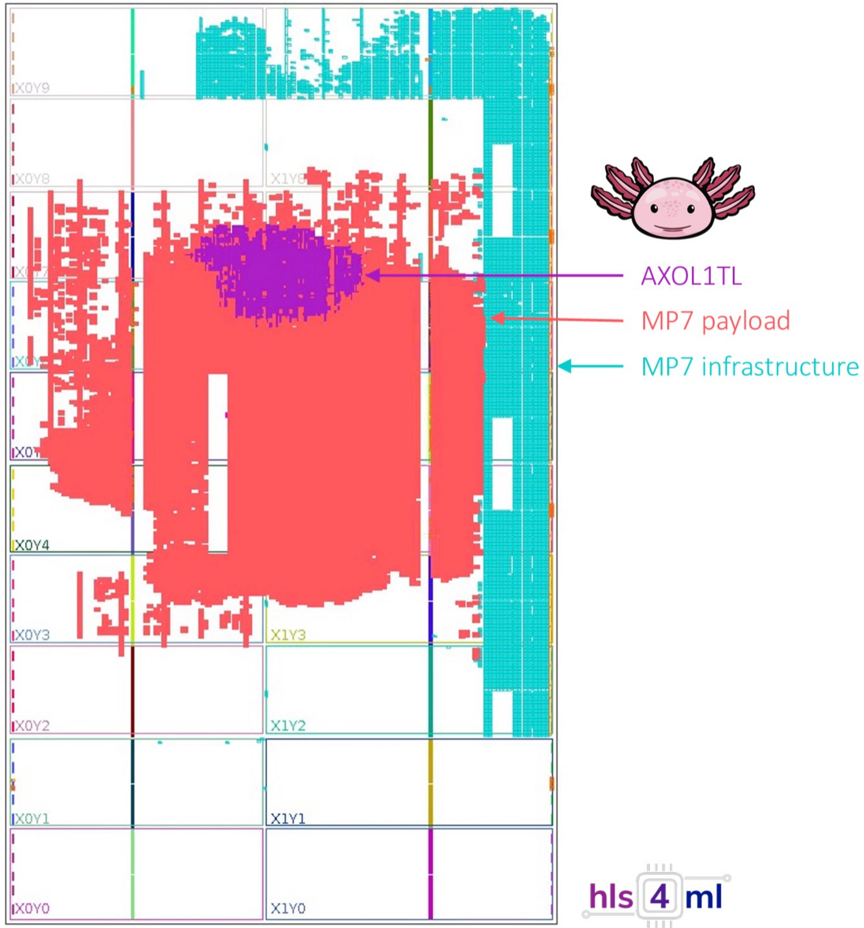
- Firmware sits in MP7 Global Trigger board
  - Xilinx Virtex 7 chip
- AXO runs in < 50 nanoseconds
- Whole algorithm chain takes a few microseconds

Build into existing global trigger firmware

- Test accuracy, timing, and resource usage in simulation

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

CERN-CMS-DP-2023-079 (2023). https://cds.cern.ch/record/2876546

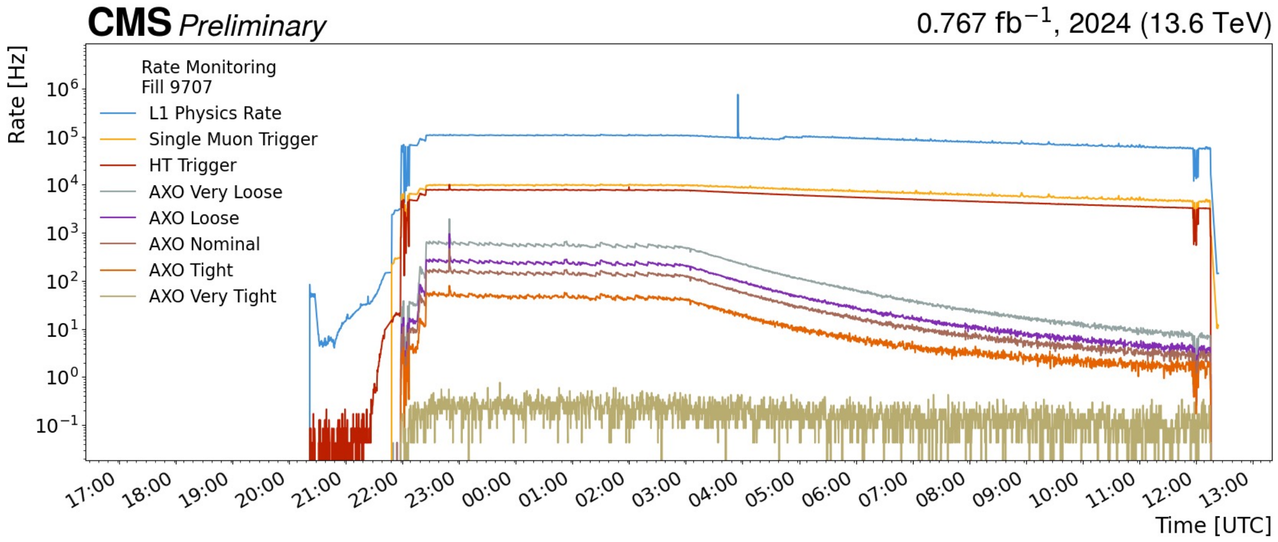


hls4ml

Performance and Validation

We validated stability in 2023

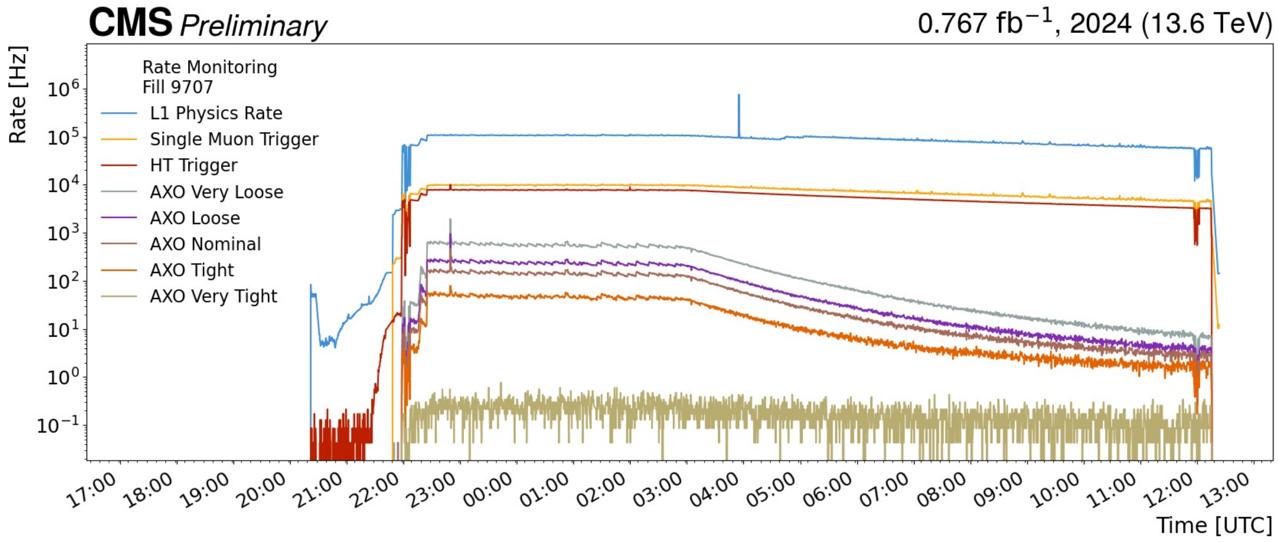
- Used “test crate” to monitor performance
- Trigger rates in data are stable and within expected ranges

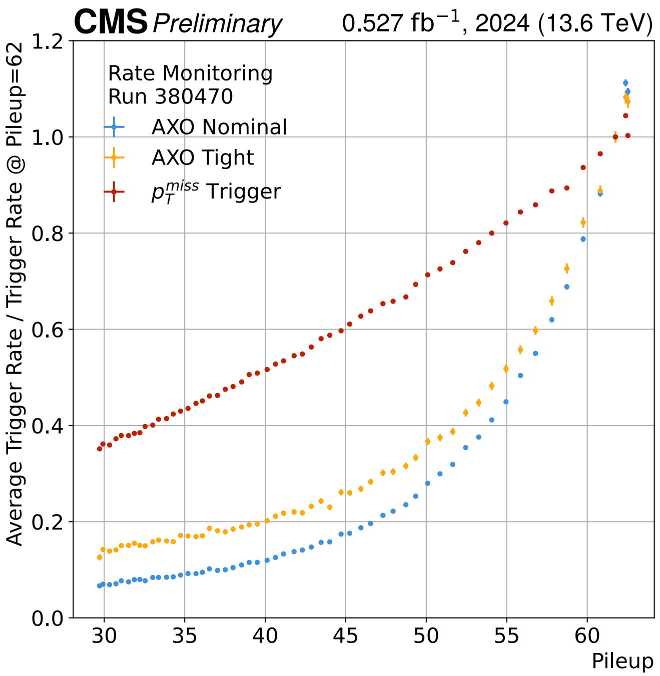


Performance and Validation

We validated stability in 2023

- Used "test crate" to monitor performance
- Trigger rates in data are stable and within expected ranges





CMS-CMS-DP-2024-059 (2024). https://cds.cern.ch/record/2904695.

Pileup* dependence

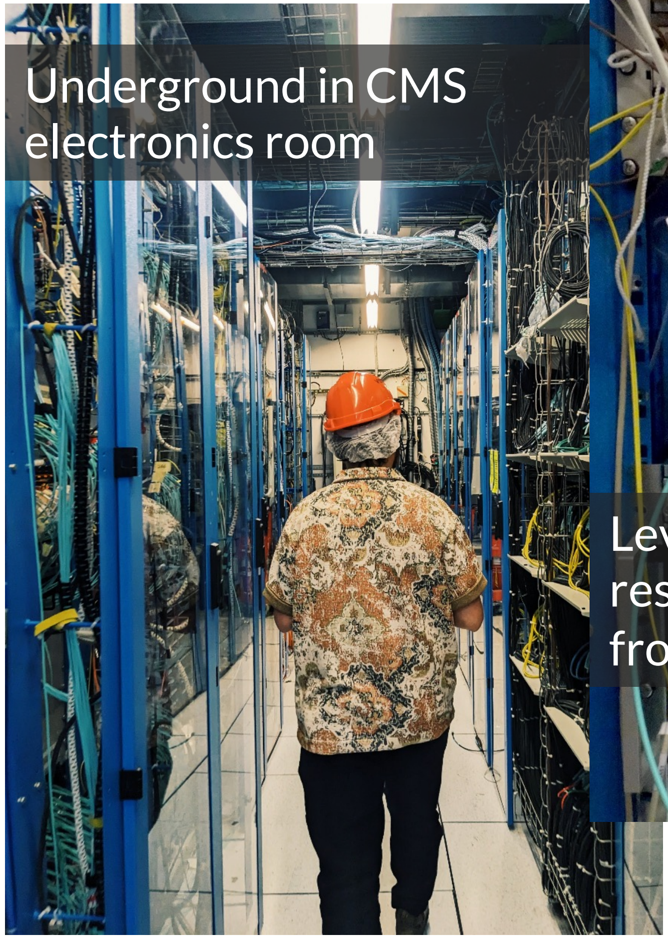
- Observed large, but anticipated correlation

"Pileup"

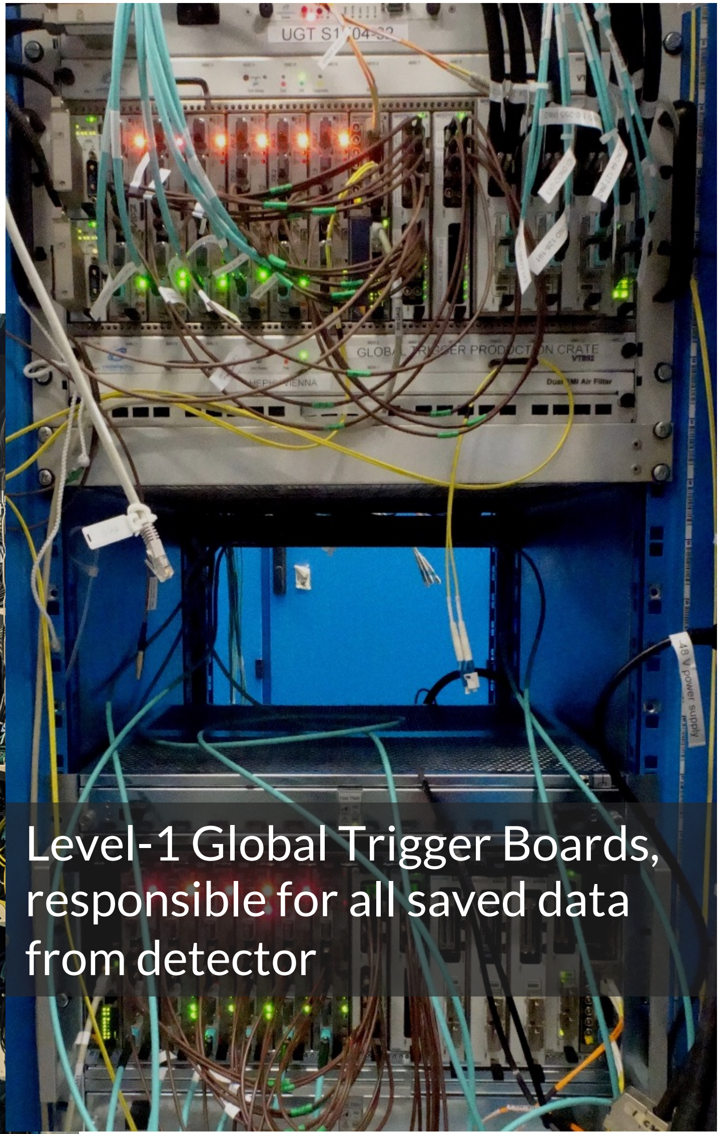
The number of concurrent interactions during a bunch collision. High pileup can spike trigger rate and lead to lost data.

Integrating into the Trigger System

Algorithm added into production system in May 2024, and taking data ever since 🎉



Underground in CMS electronics room



Level-1 Global Trigger Boards, responsible for all saved data from detector

First Results from Real 2024 Data!

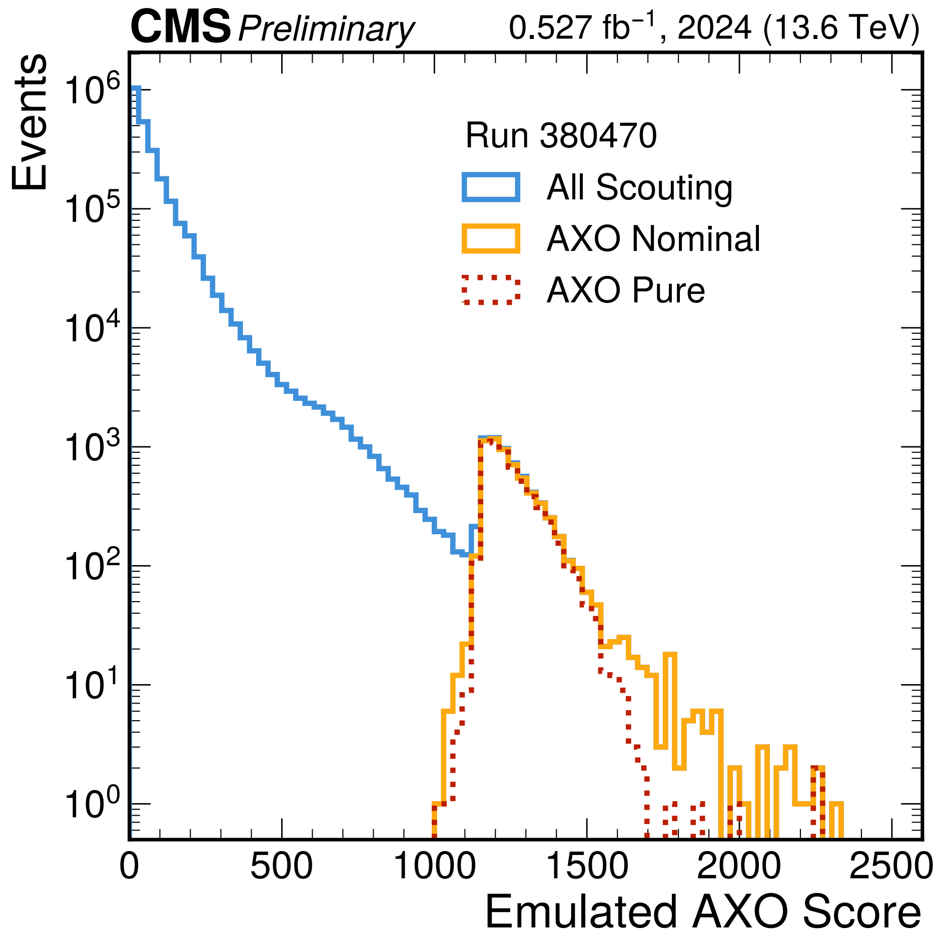
Still have lots of data to look through, but these are some first observations…

First Results from Real 2024 Data!

Still have lots of data to look through, but these are some first observations…

Anomaly score distributions

- We see a bump in “pure” events, where only AXO and no other L1 triggers select an interaction
- Correlation with other triggers at high scores

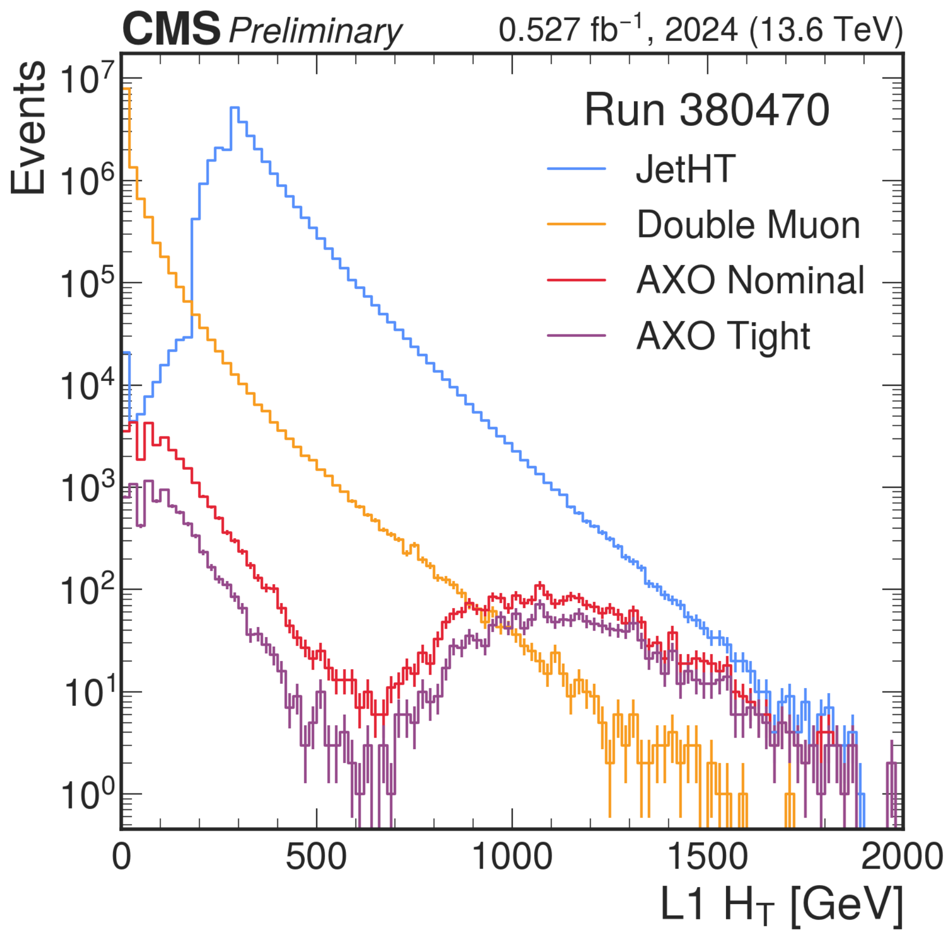


CMS-CMS-DP-2024-059 (2024). https://cds.cern.ch/record/2904695.

First Results from Real 2024 Data!

"H_T" or "Hadronic Energy Sum"

Quarks or gluons from collisions produce clusters of energy in the detector. We sum up all this energy in an event to get the H_T.



CMS-CMS-DP-2024-059 (2024). https://cds.cern.ch/record/2904695.

In some kinematic variables like H_T*, we see different shapes in AXO vs. other triggers

AXO decides certain known signals are too common

- Selects other, more anomalous, patterns
- We're still figuring out what the patterns are

First Results from Real 2024 Data!

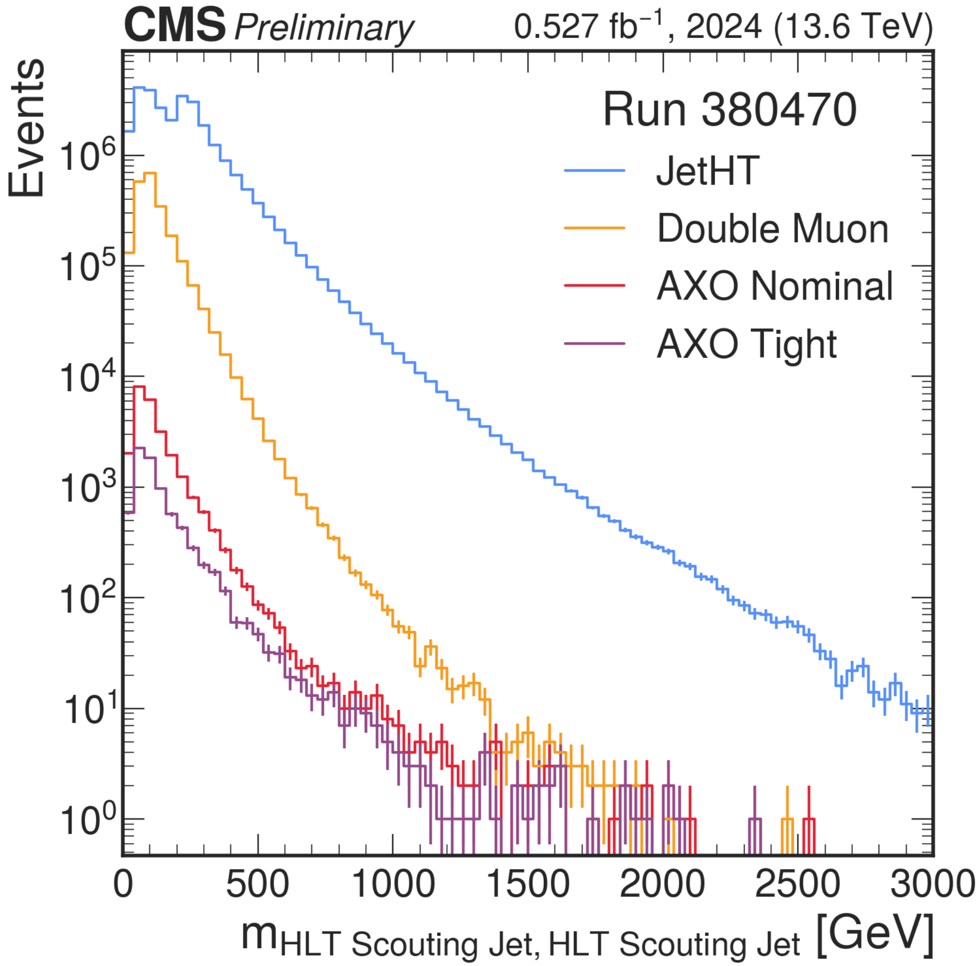
"H_T" or "Hadronic Energy Sum"

Quarks or gluons from collisions produce clusters of energy in the detector. We sum up all this energy in an event to get the H_T.

Invariant mass distributions

- Here, we combine objects to find a decaying particle mass
- Smooth and falling shape
- We can use this to search for new particles!

These shapes mean characterizing backgrounds to find signal is easier



CMS-CMS-DP-2024-059 (2024). https://cds.cern.ch/record/2904695.

First Results from Real 2024 Data!

"H_T" or "Hadronic Energy Sum"

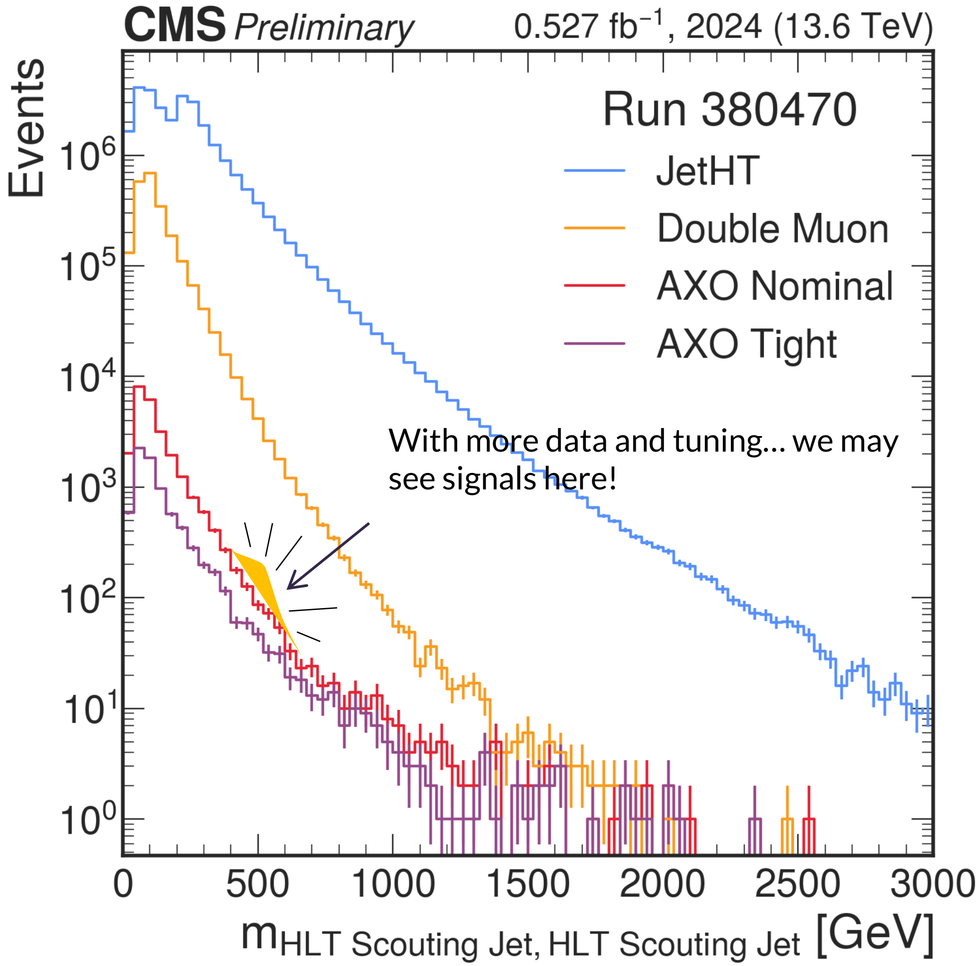
Quarks or gluons from collisions produce clusters of energy in the detector. We sum up all this energy in an event to get the H_T.

Invariant mass distributions

- Here, we combine objects to find a decaying particle mass
- Smooth and falling shape
- We can use this to search for new particles!

These shapes mean characterizing backgrounds to find signal is easier

- We can use this to search for new particles!



CMS-CMS-DP-2024-059 (2024). https://cds.cern.ch/record/2904695.

Next Steps

Dig more into the data, figure out what patterns AXO is finding

- Maybe something we haven't recorded before

Design analysis strategies with anomaly data

- Searching for mass resonances ("bump hunt")

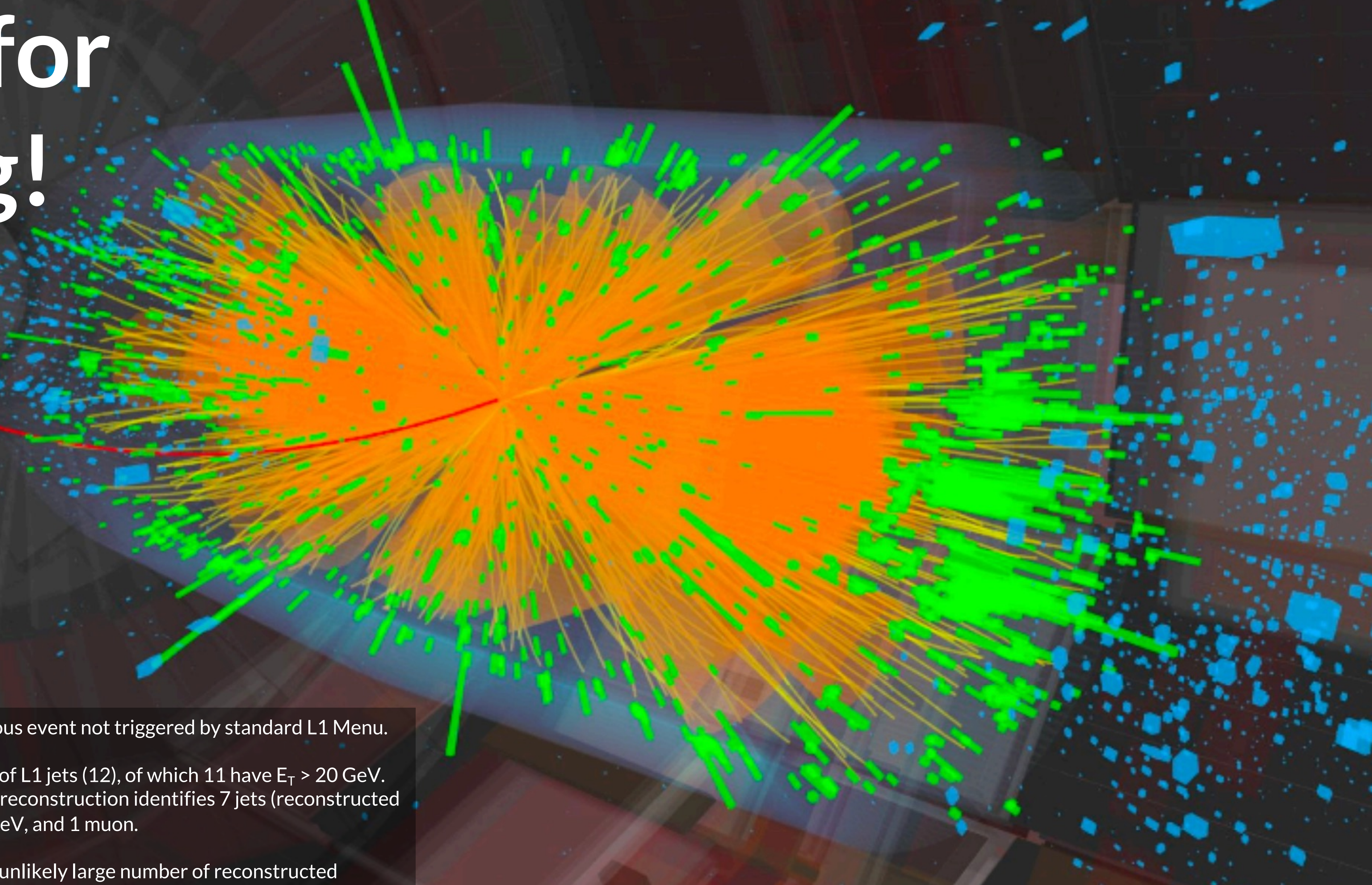
Update and upgrade algorithm

- AXO changes with changing detector conditions
    - Prepare for 2025!
- Improve performance with new kinds of ML models

References

- CMS Collaboration. “2024 Data Collected with AXOL1TL Anomaly Detection at the CMS Level-1 Trigger”. CMS-CMS-DP-2024-059 (2024). https://cds.cern.ch/record/2904695.
- CMS Collaboration. "CMS Technical Design Report for the Phase 1 Upgrade of the Hadron Calorimeter". CERN-LHCC-2012-015, CMS-TDR-10 (2012). https://cds.cern.ch/record/1481837.
- M. Jeitler, et al. "The level-1 global trigger for the CMS experiment at LHC". JINST 2, P01006 (2007). https://doi.org/10.1088/1748-0221/2/01/P01006.
- 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.
- Xilinx Virtex-7 FPGA. https://www.xilinx.com/products/silicon-devices/fpga/virtex-7.html.
- L1 Menu Repository. https://github.com/herbberg/l1menus.

Thanks for listening!



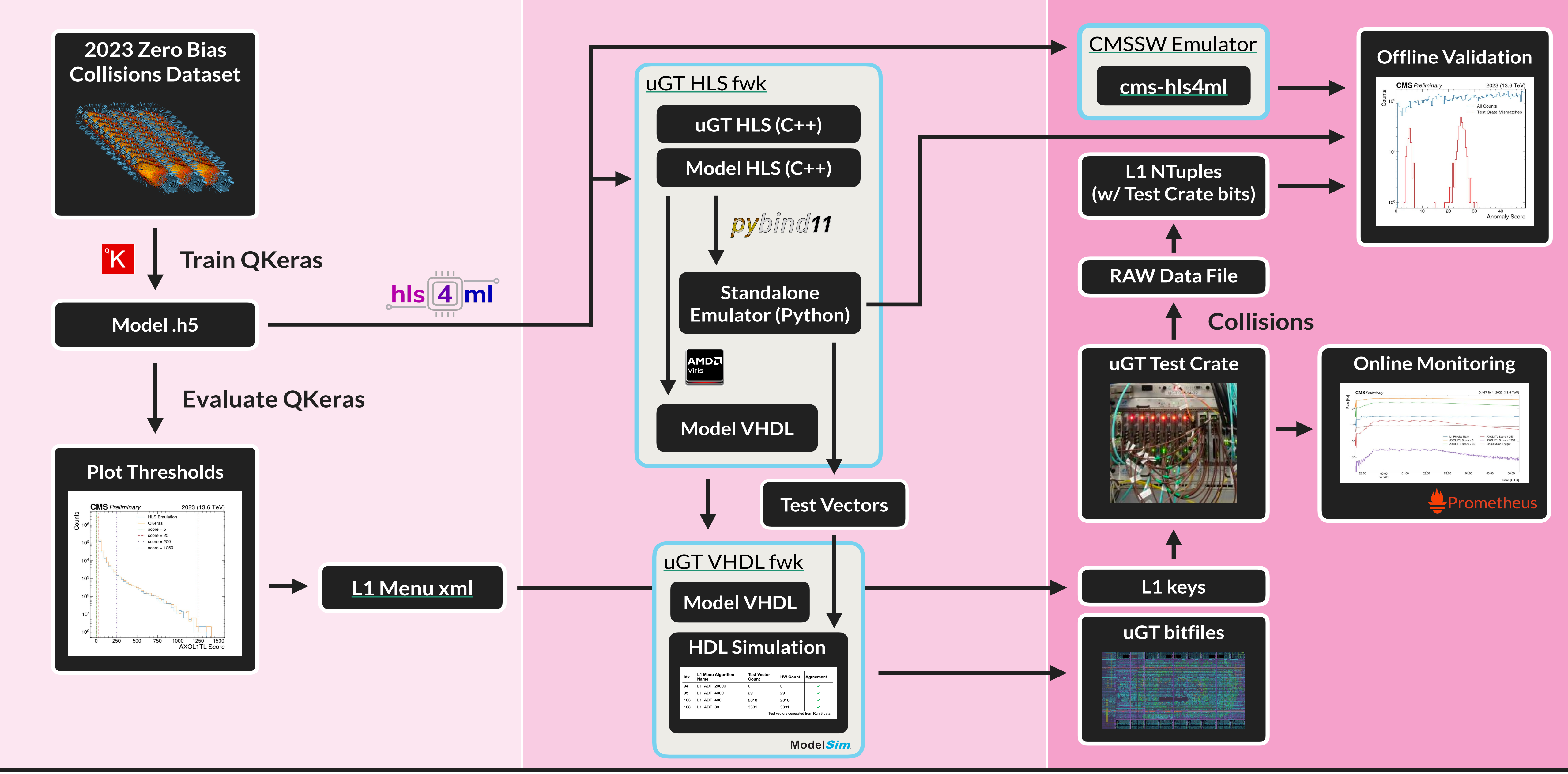
From 2023 ZeroBias dataset, an anomalous event not triggered by standard L1 Menu.

This event features the maximal number of L1 jets (12), of which 11 have E_T > 20 GeV. It also features a 3 GeV L1 muon. Offline reconstruction identifies 7 jets (reconstructed with the PUPPI algorithm) with p_T > 15 GeV, and 1 muon.

The event is also characterized by a very unlikely large number of reconstructed vertices (75), given the pile up profile of the data taken in Run 2 and Run 3.

CERN-CMS-DP-2023-079 (2023). https://cds.cern.ch/record/2876546

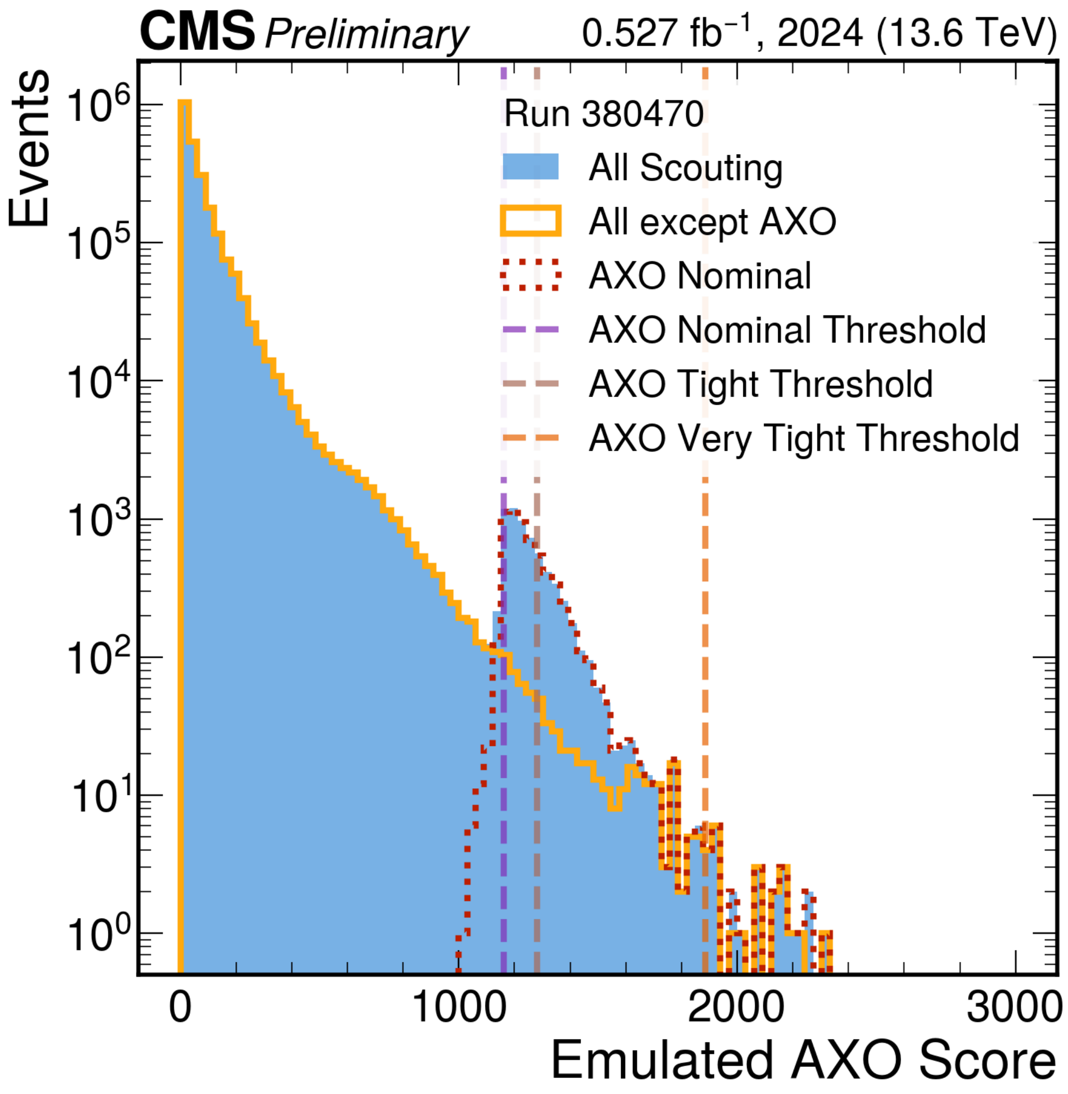
Backup



Development | Conversion | Implementation / Validation

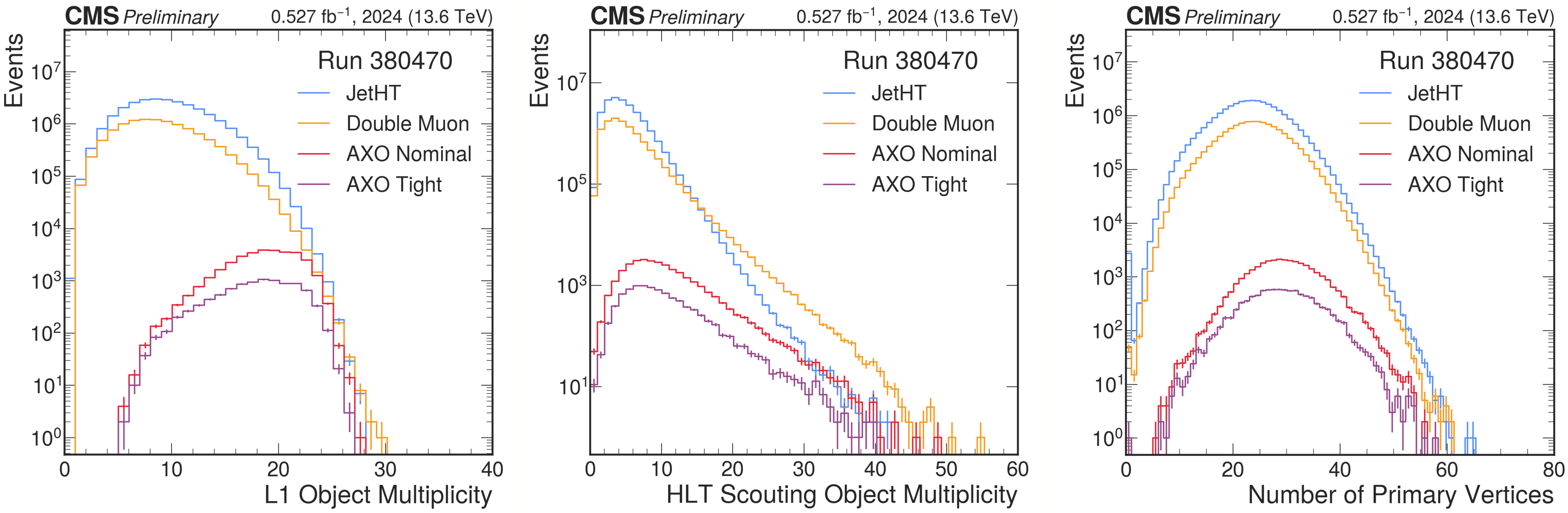
DP Note Plots

CMS-CMS-DP-2024-059 (2024). https://cds.cern.ch/record/2904695.



DP Note Plots

CMS-CMS-DP-2024-059 (2024). https://cds.cern.ch/record/2904695.



DP Note Plots

CMS-CMS-DP-2024-059 (2024). https://cds.cern.ch/record/2904695.