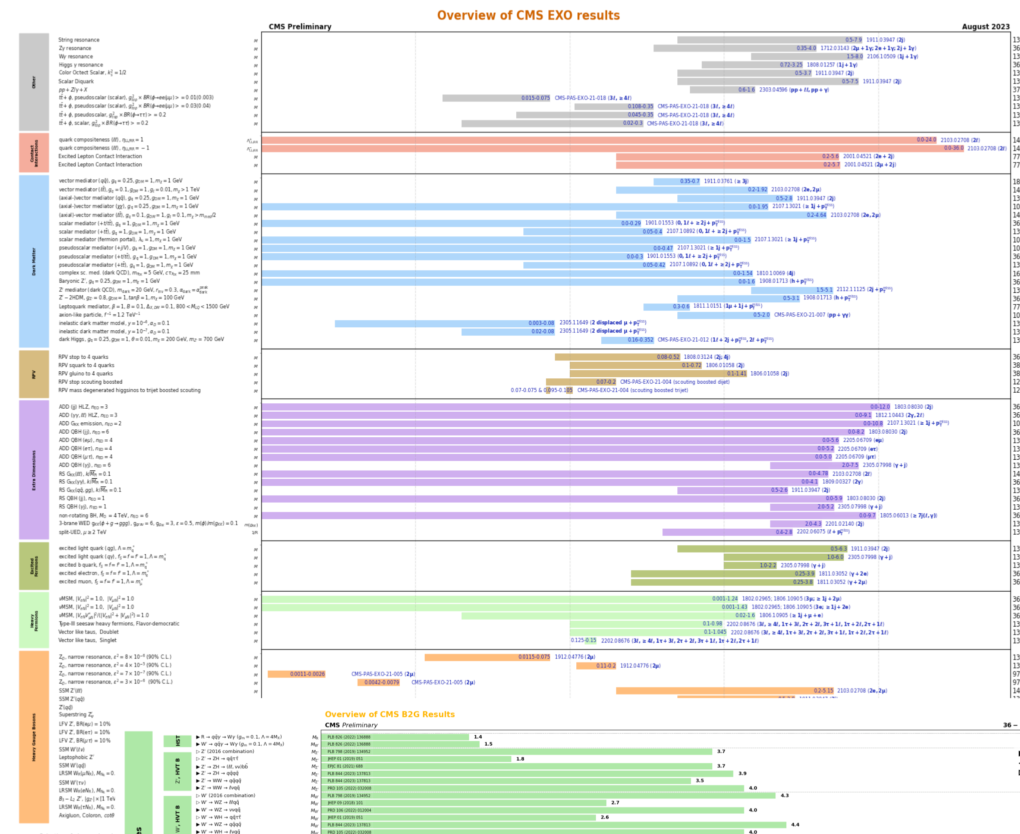
NexTGen
Next Generation Triggers

NGT Work Package 3.7: CMS L1T anomaly detection & data compression

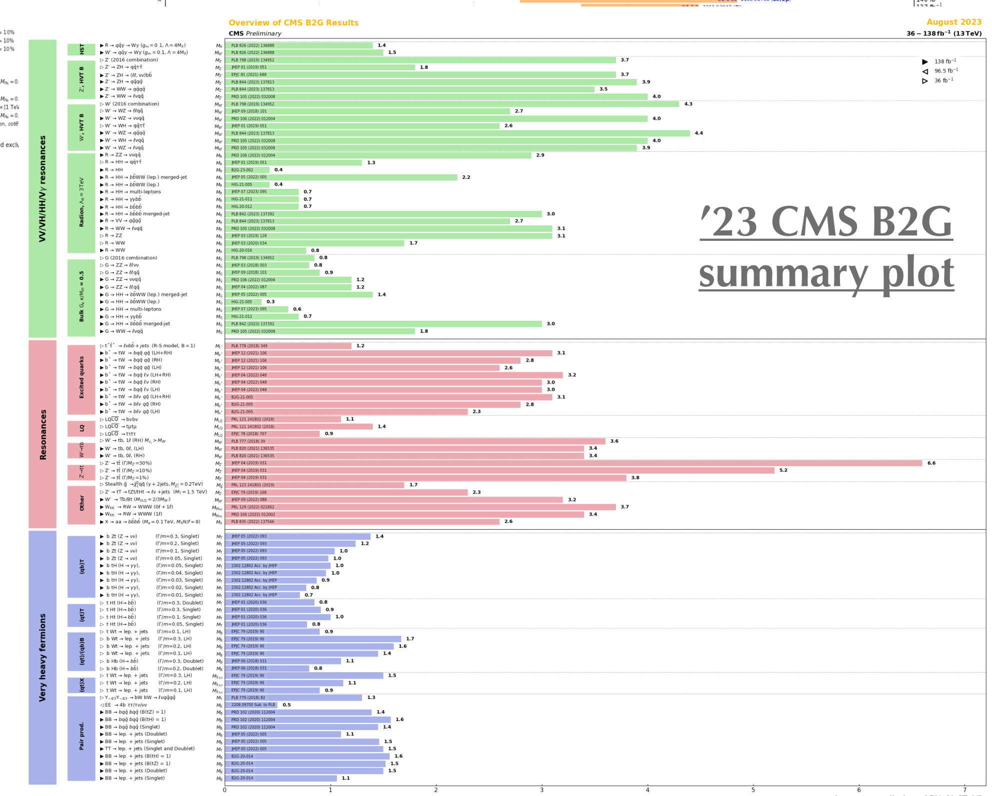
Jennifer Ngadiuba (Fermilab)
on behalf of the NGT WP 3.7 team

Anomaly detection @ LHC

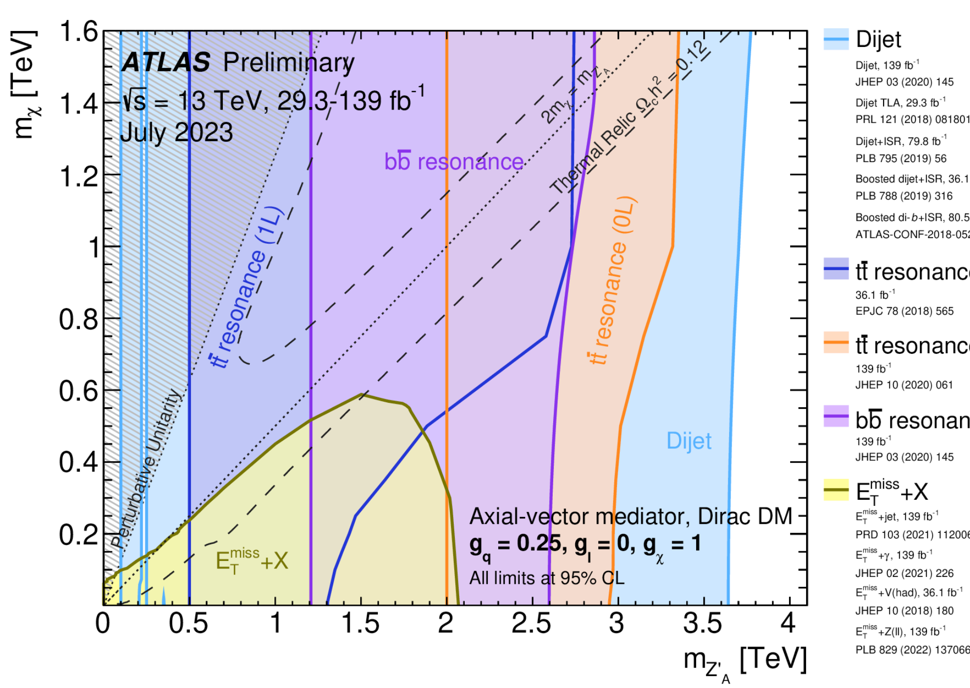
'23 CMS EXO summary plot



- **Goal: generalize new physics searches to a large variety of BSM models at once**
  - and even to the ones we have not thought about it yet !



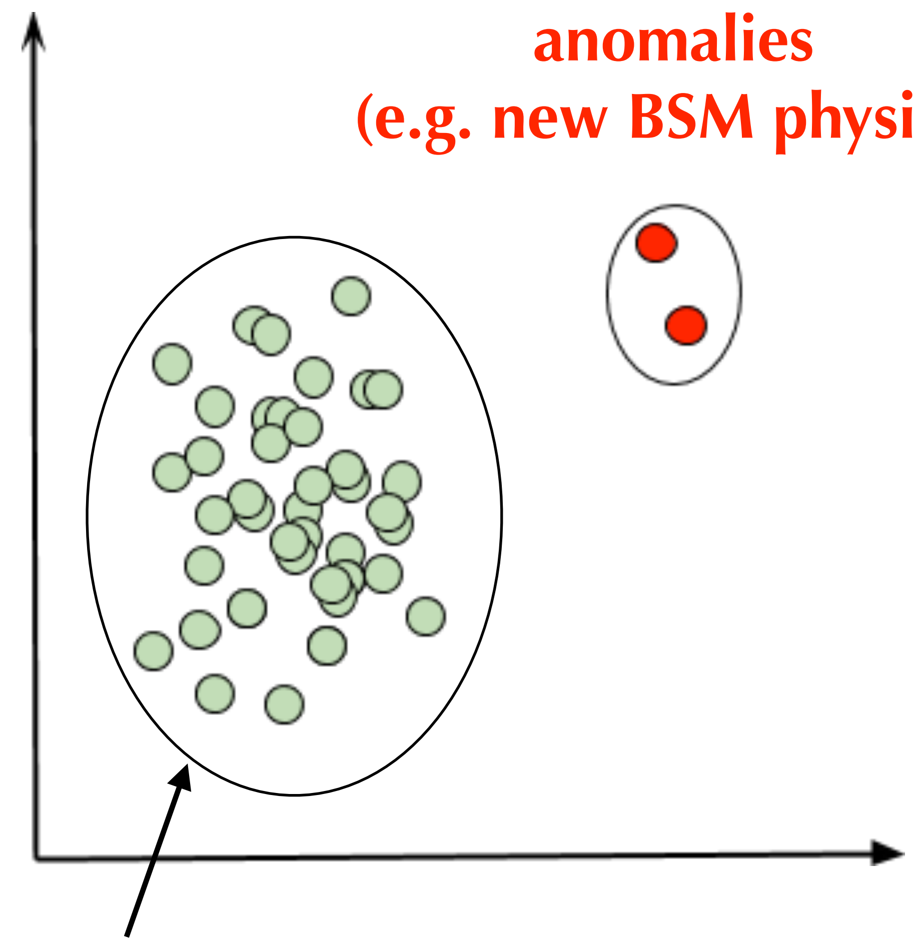
'23 CMS B2G summary plot



'23 ATLAS Dark Matter summary

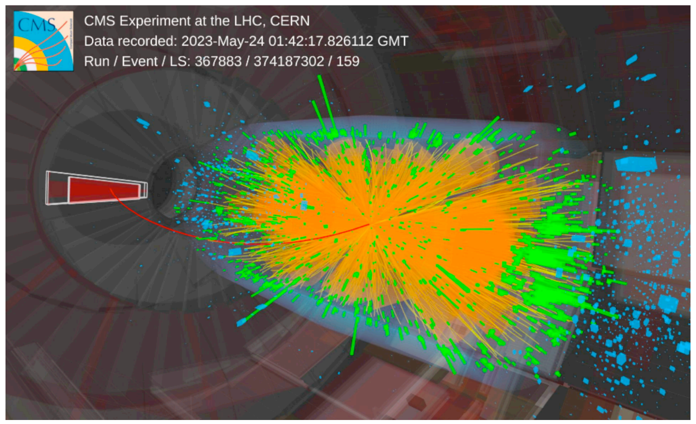
Anomaly detection @ LHC

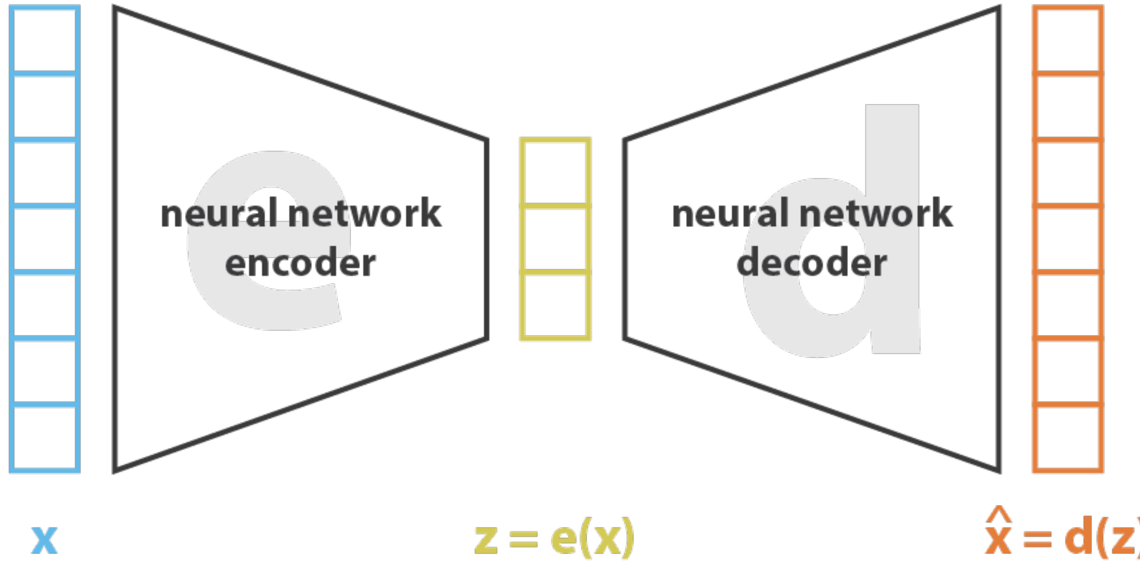
- **Goal: generalize new physics searches to a large variety of BSM models at once**
  - and even to the ones we have not thought about it yet !
- Identifying **rare events** in data sets which deviate significantly from the majority of the data and do not conform to “normal” behaviour
- **Normal behaviour can be learnt through a NN, for example with AUTOENCODERS**



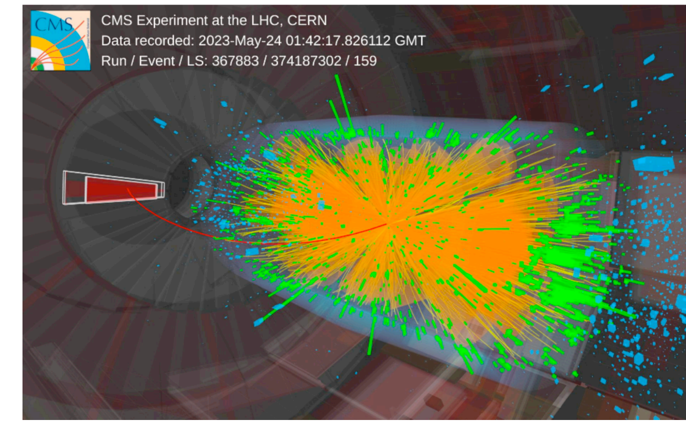
Autoencoders in a nutshell

Input data





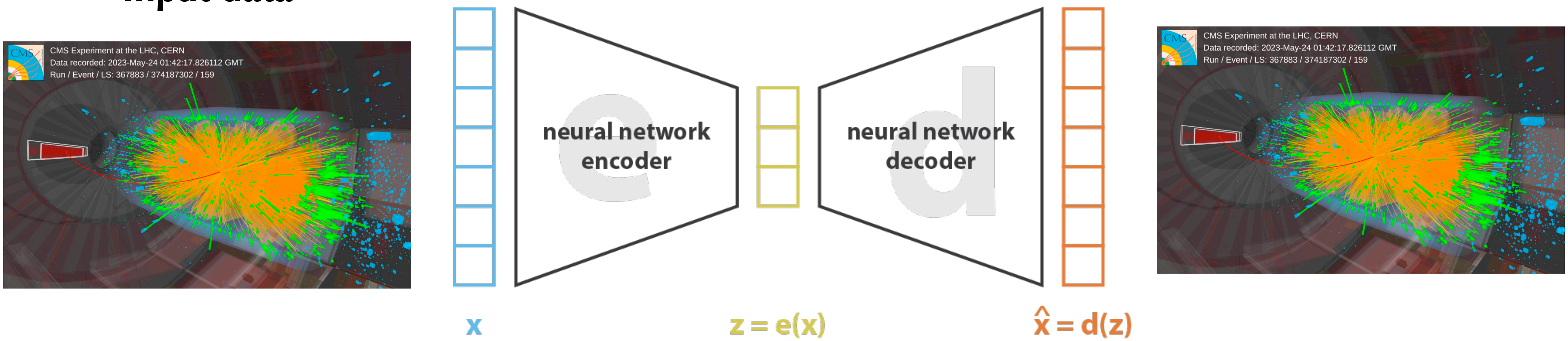
Reconstructed data



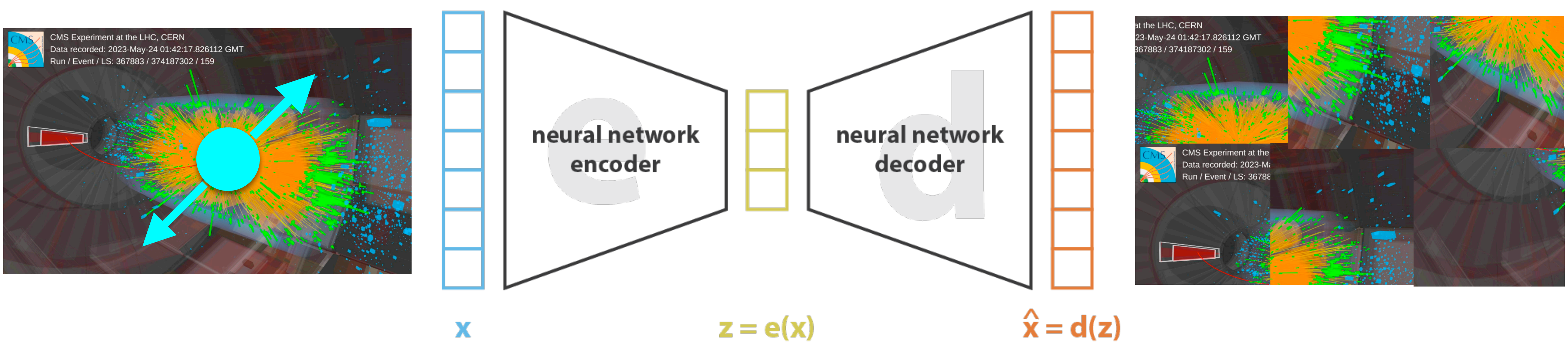
$$\mathcal{L}_{reco} = ||x - \hat{x}||^2 = MSE(input, output)$$

Autoencoders in a nutshell

Input data **Reconstructed data**

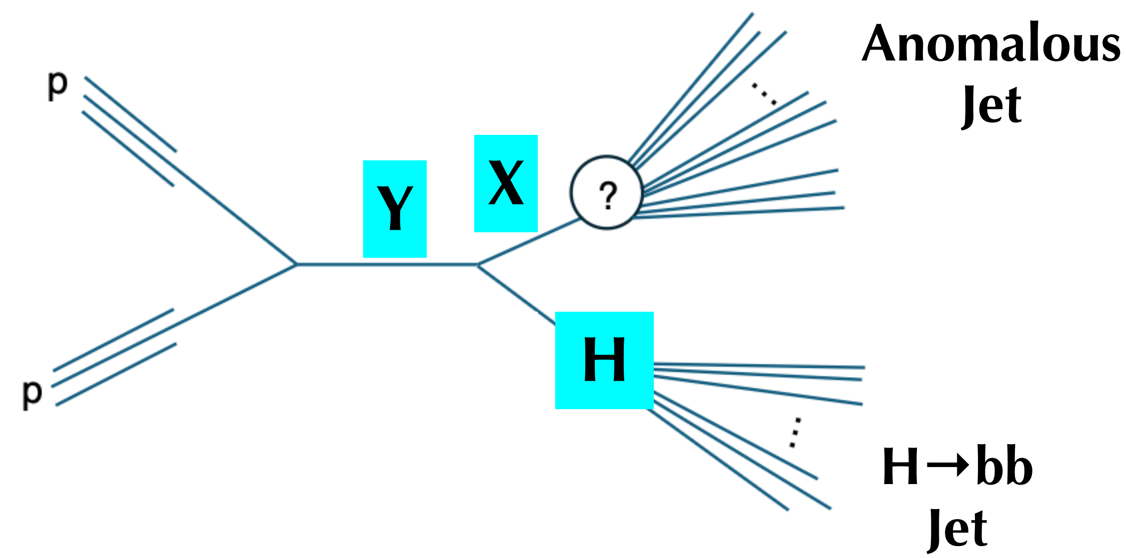


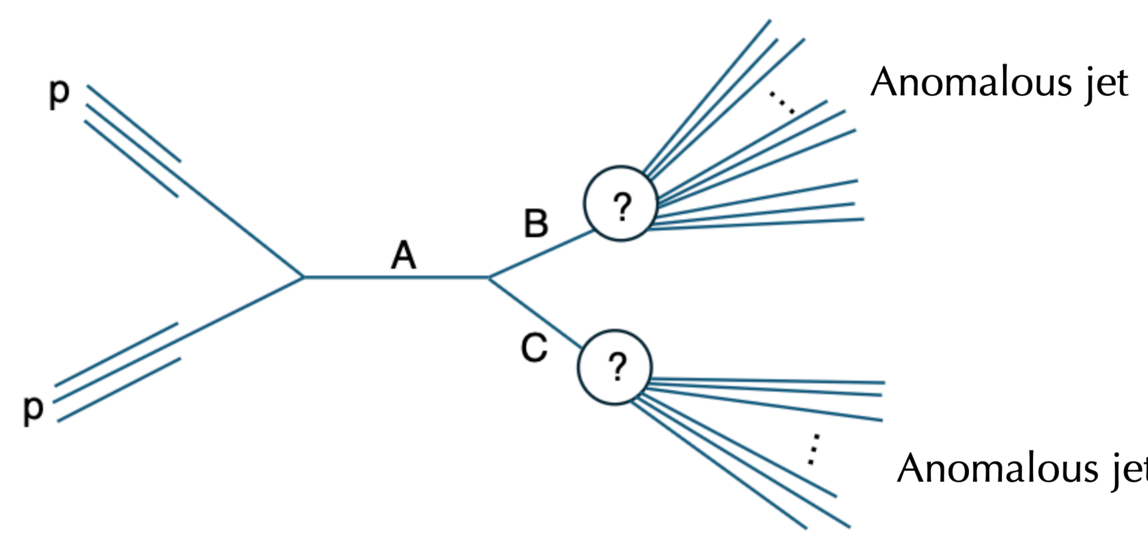
$$\mathcal{L}_{reco} = ||x - \hat{x}||^2 = MSE(input, output)$$

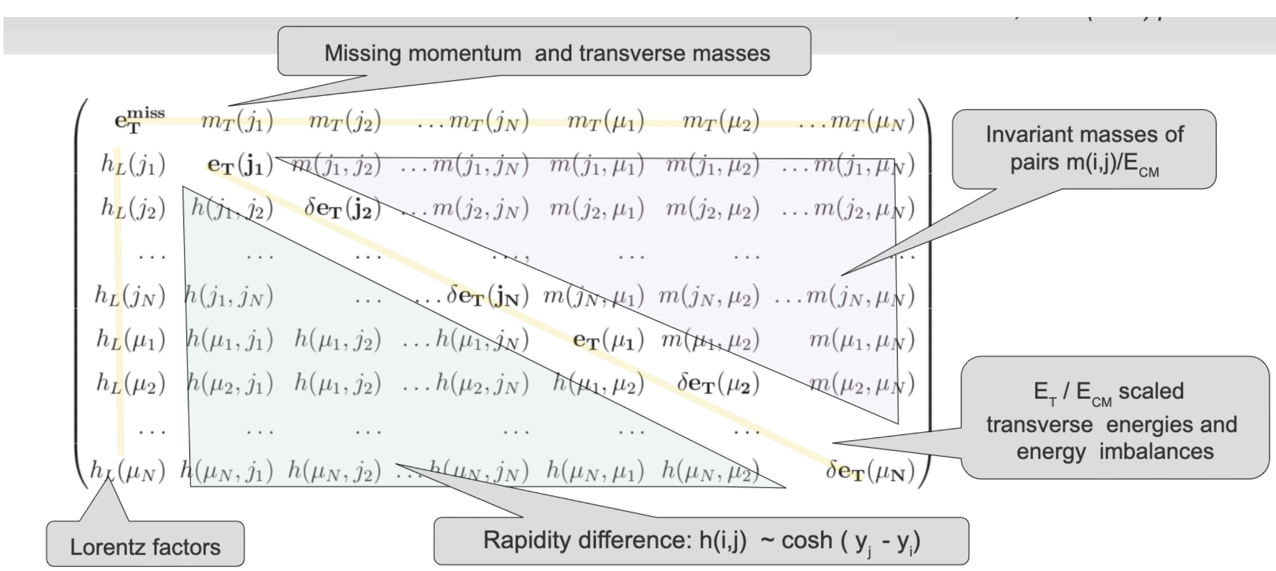


Anomaly detection @ LHC: results

- **Two ATLAS searches using autoencoders:**
  - two boosted jets [PRD 108 (2023) 052009]
  - dijet, lepton + jet(s), and photon + jet(s) [PRL 132 (2024) 081801]
- **One CMS search in boosted dijet final state** [CMS-PAS-EXO-22-026]:
  - several AD methods designed and applied, not only autoencoders

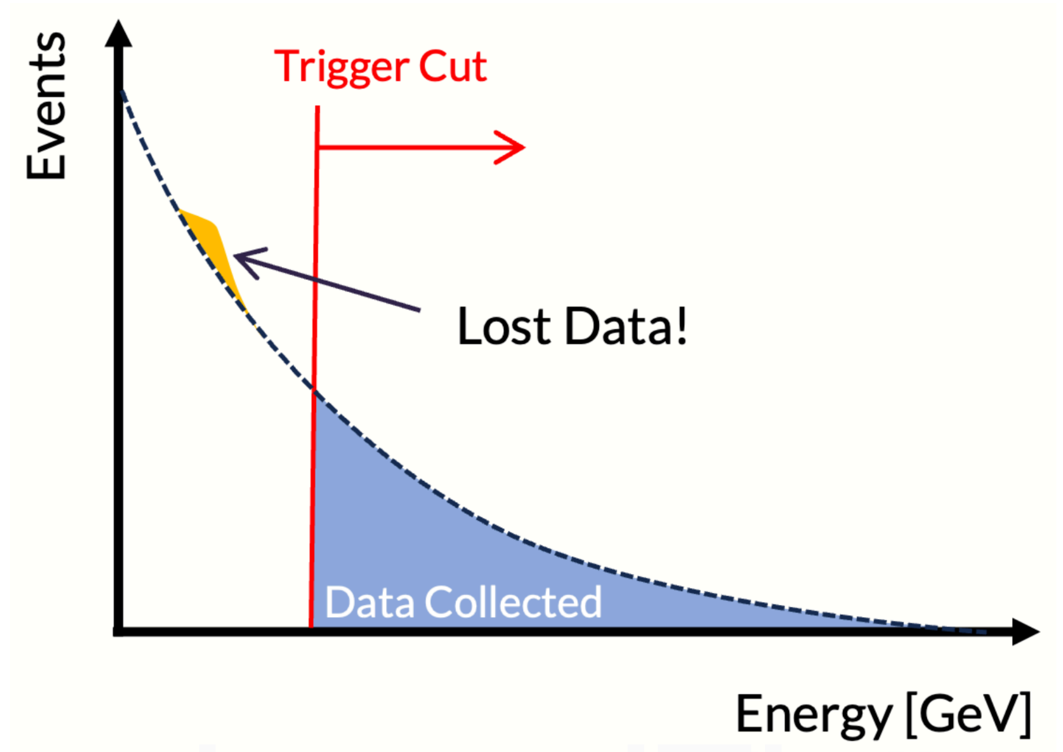




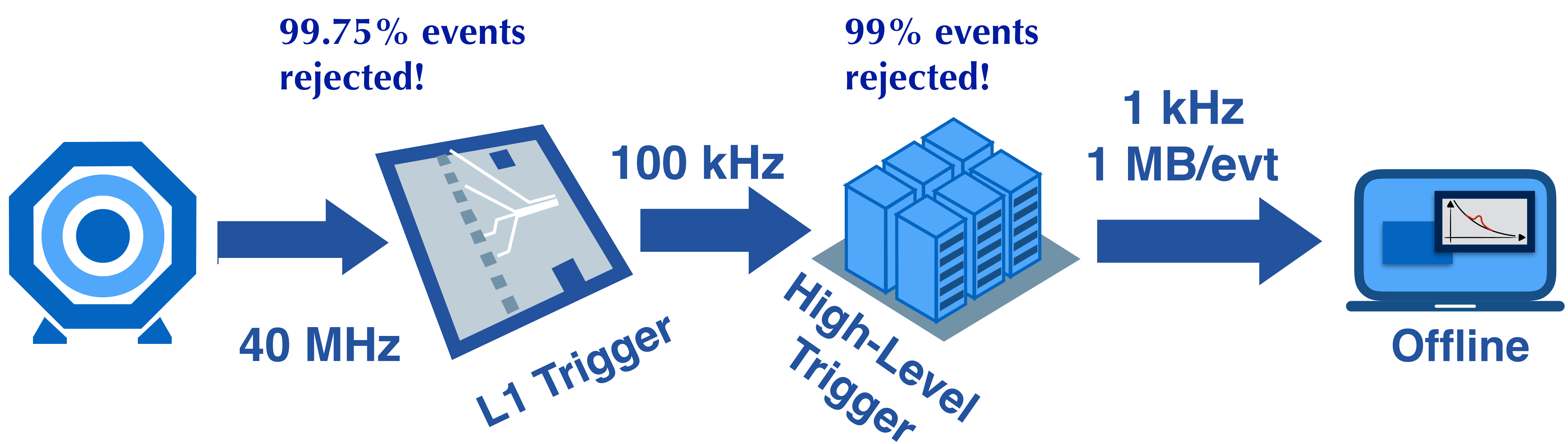


Data reduction @ LHC

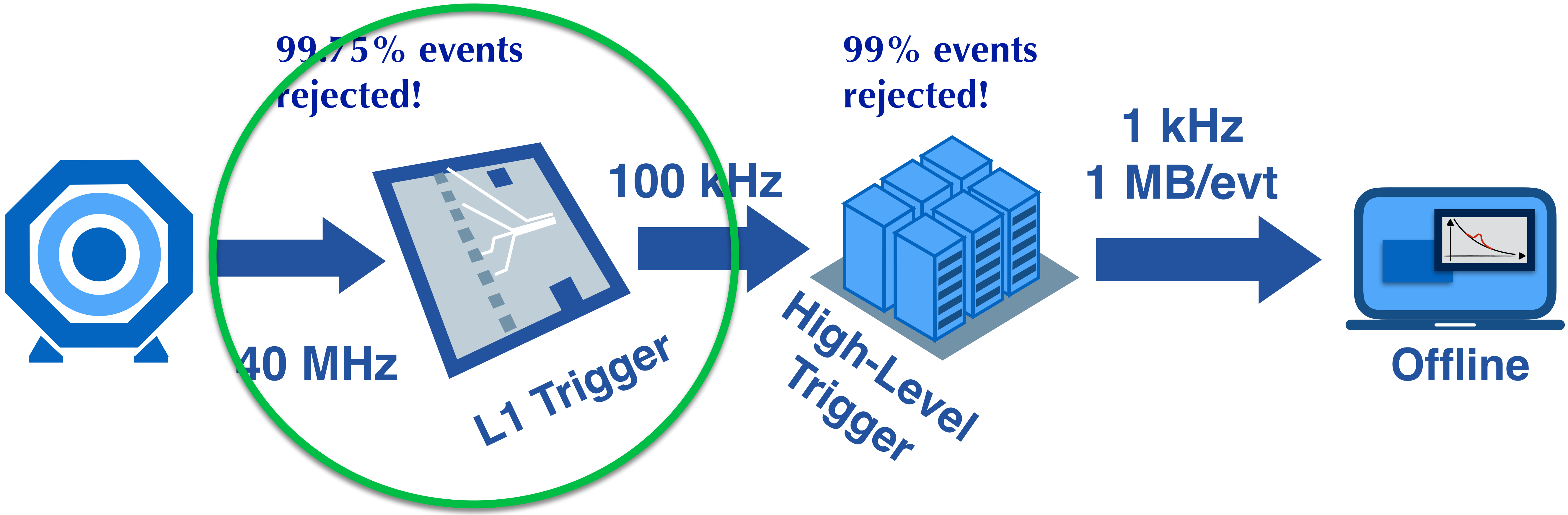
- The CMS L1T rejects 99.75% of the events
- Currently, we use simple heuristics to define trigger algorithms
  - Energy, charge, direction, momentum, etc.
- In this approach, we need to know what we're looking for to target it



  - **What if we are missing new physics because we did not design the right trigger?**



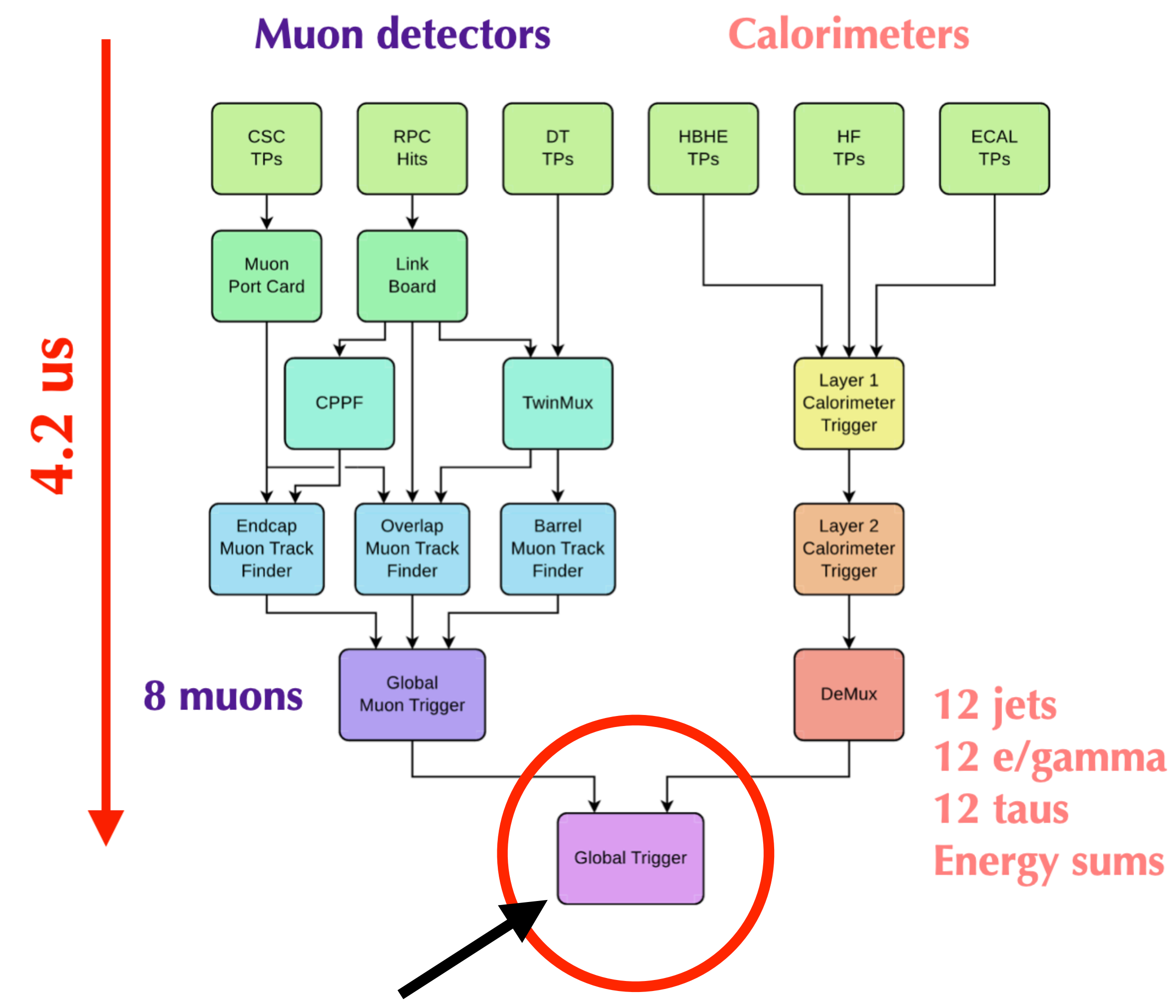
THE ANOMALY MIGHT BE DISCARDED BY THE TRIGGER



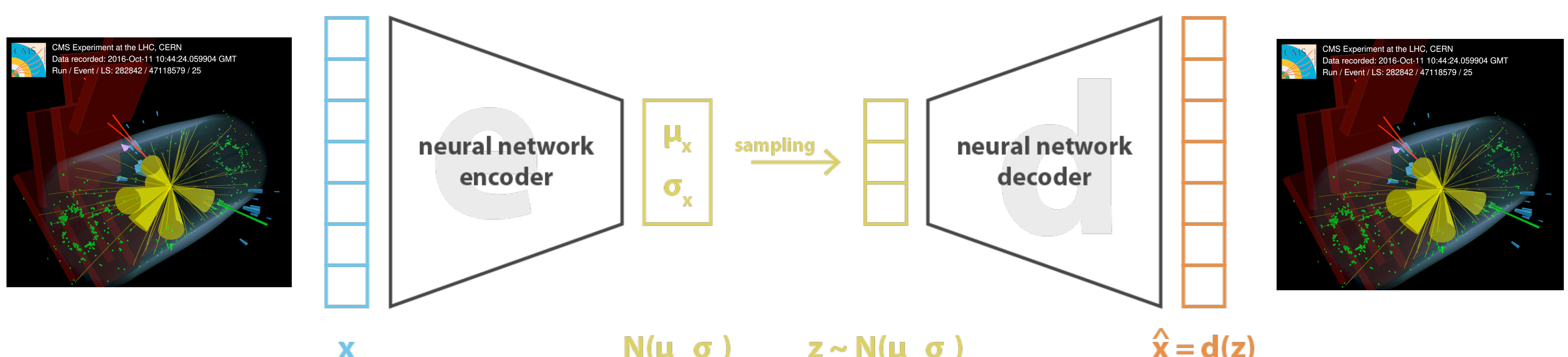
Correct the problem as early as possible in the data reduction workflow!

Ultra-fast anomaly detection @ CMS

- Train a variational autoencoder on **unbiased data collected by CMS in 2023 at 13.6 TeV** (~10.5 million)
  - ~ same inputs as Global Trigger (GT): **4-vector of muons, jets, MET, e/ɣ**
  - **learn to reconstruct the average collision event**, i.e. mostly soft hadronic collisions with large number of low energy jets
  - usually rejected by cut-based algo

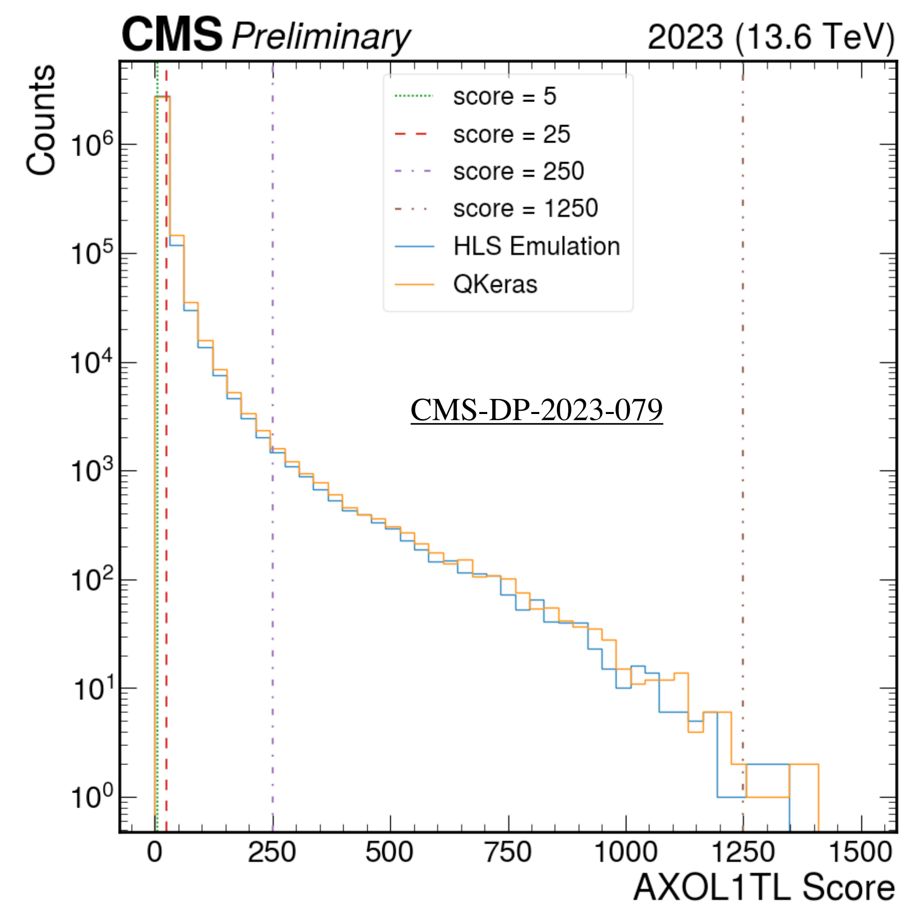


Strict latency constraint of 50 ns to run in the GT!

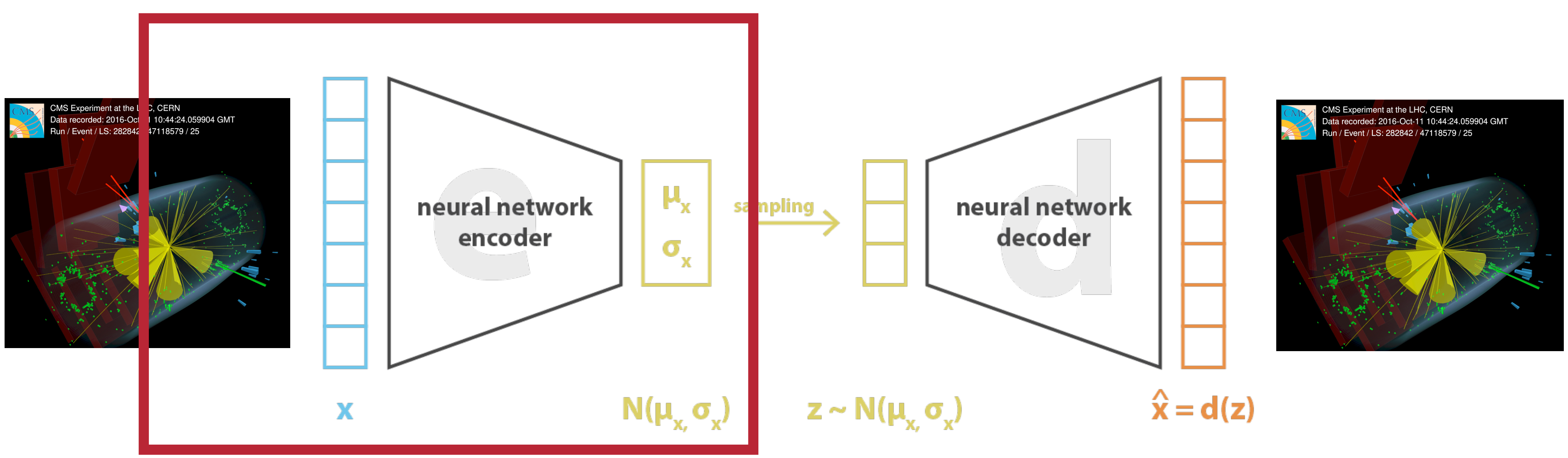


Ultra-fast anomaly detection @ CMS

- **Small, fully connected network architecture** (encoder: 32,16,8 nodes per layer)
- **TRICK:** define anomaly metric in the latent space (μ^2)
  → allows us to deploy only the encoder part
  → half model size and latency!
- **Quantization aware training with QKeras** to reduce FPGA resources utilization
- **hls4ml to translate NN into firmware**, then final integration with rest of trigger algorithms



- **Define different threasholds** on anomaly score based on allocated output rate

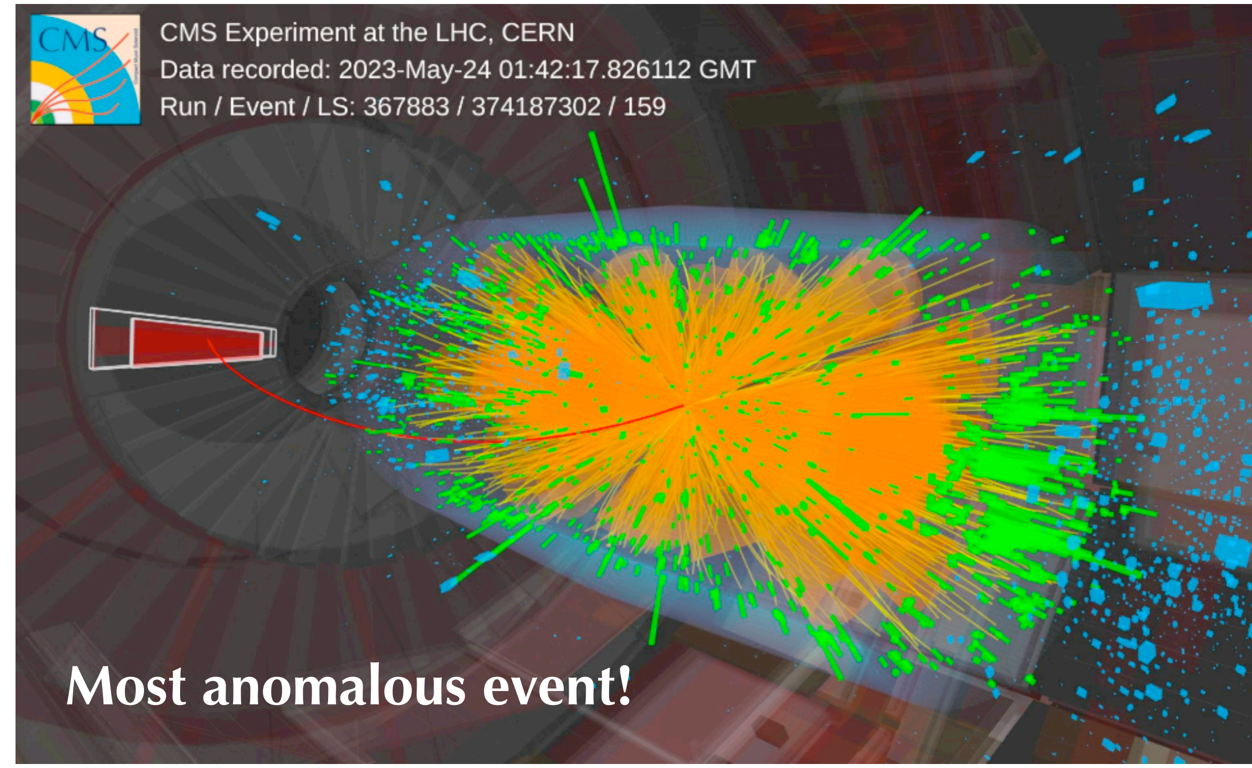


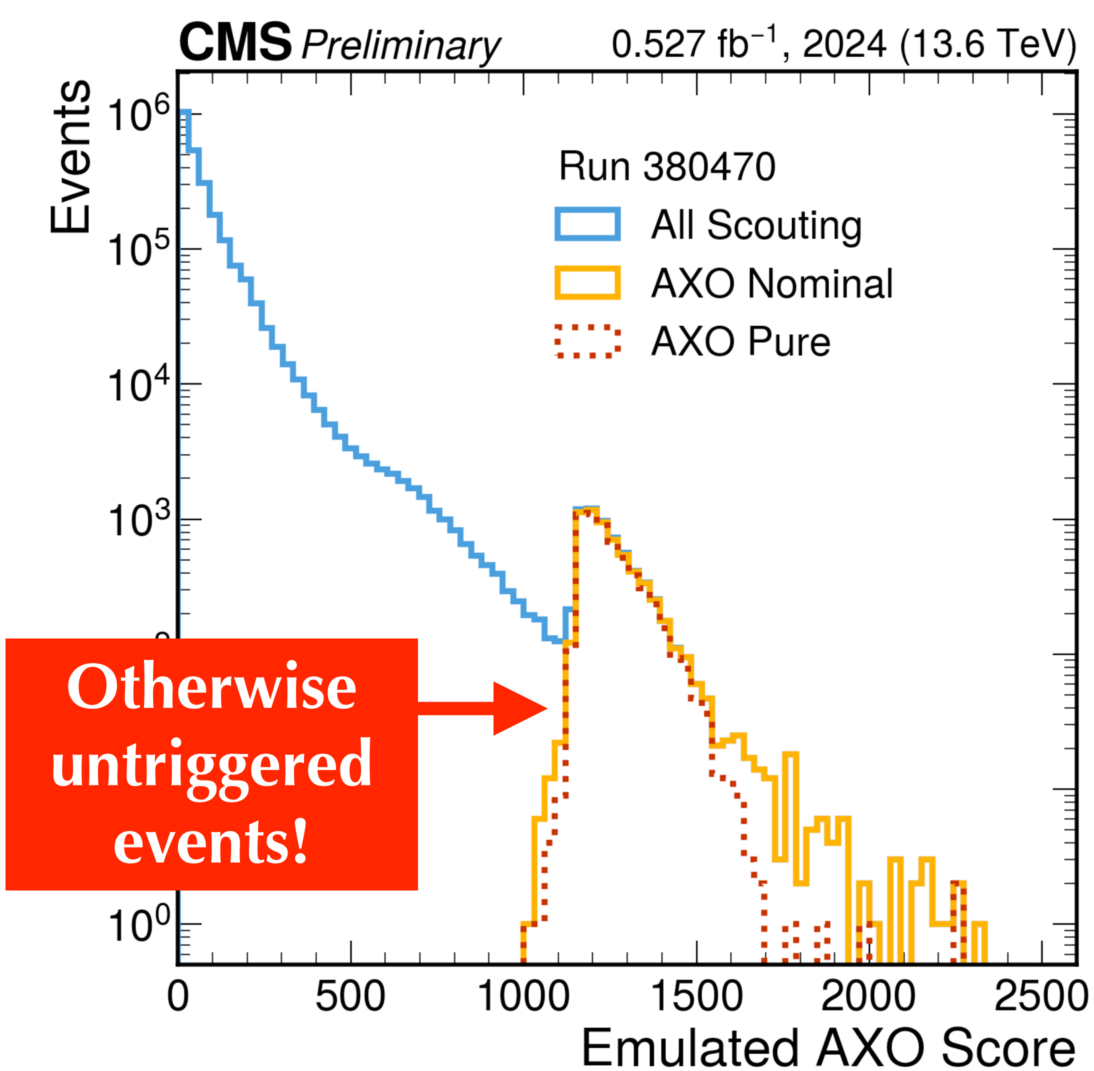
Ultra-fast anomaly detection @ CMS

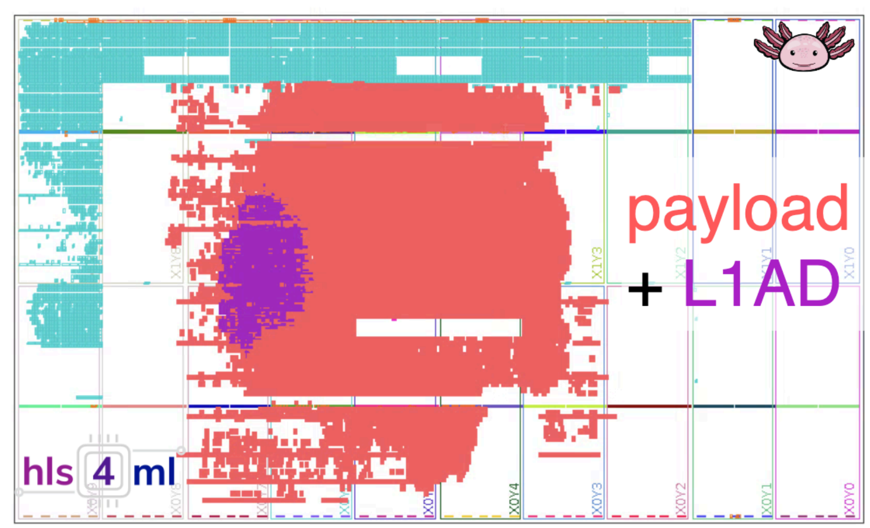
Anomaly eXtraction Online Level-1 Trigger aLgorithm

Online since Spring this year! (~ 100/fb)

CMS-DP-2023-079
CMS-DP-2024-059







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

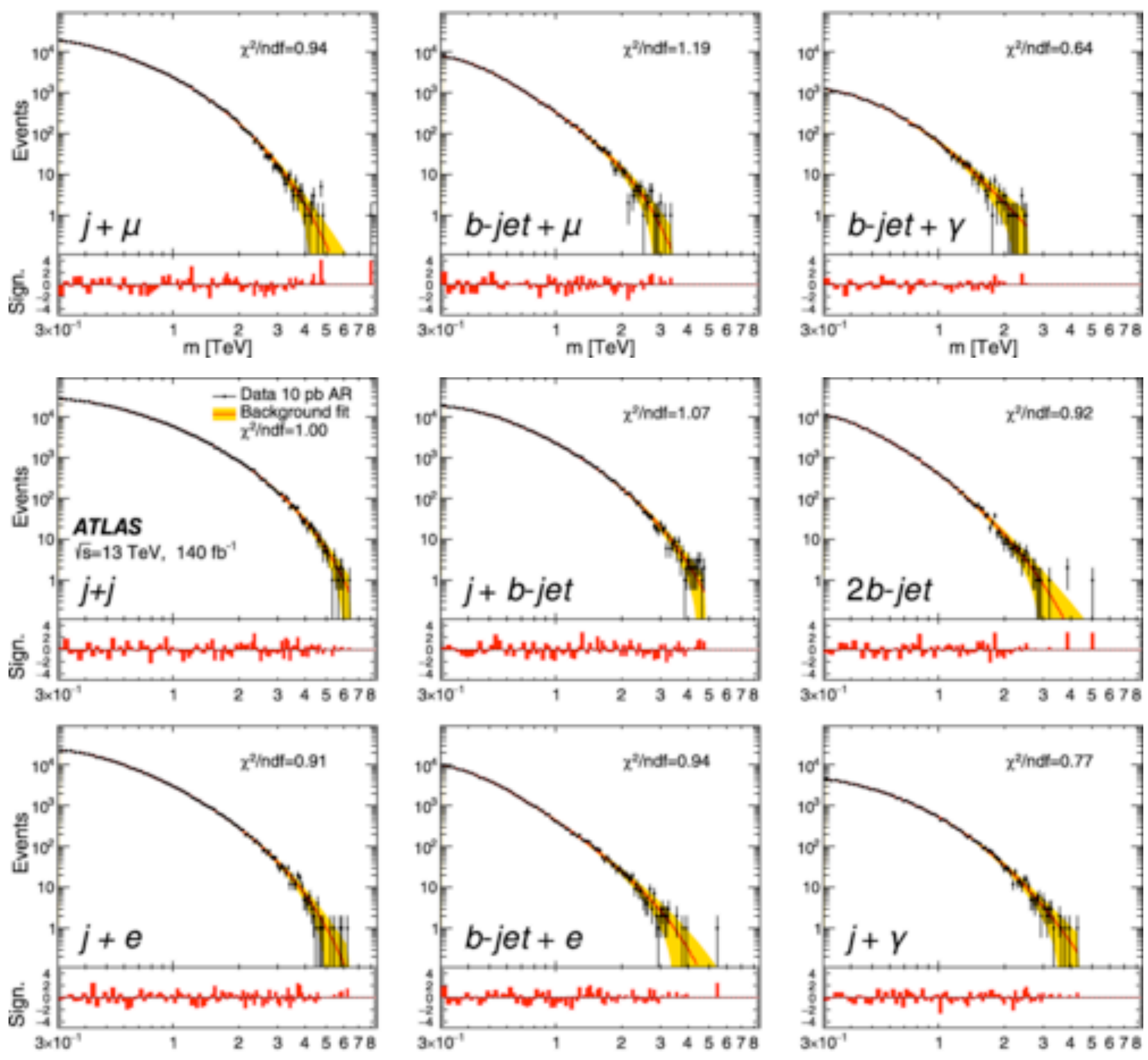
Boosting AXOL1TL with NGT

- AXOL1TL was designed and integrated over last ~ 3 years by CMS collaborators
- **Within NGT we aim at pushing this novel technology to its frontier!**
- The team is currently **advancing multiple aspects of the project** in synergy and collaboration with the original AXOL1TL team:
  - **Physics Analysis:** Investigating the collected anomalous event data for potential new physics signals [Sabrina Giorgetti, Phd student w/ Padova University + Jannicke Pearkes, Colorado Boulder Project Associate from Jan '25]
  - **Model Development:** Designing a more robust model based on representation learning techniques [Diptarko Choudhury, Technical student]
  - **Operational Automation:** Enhancing the efficiency and reliability of the trigger system's operations [Diptarko Choudhury, Technical student + Maciej Glowacki, CERN Fellow + Eric Moreno, Phd Student w/ MIT]
  - **Phase 2 Preparation:** Developing an upgraded model tailored to the Phase 2 trigger system, incorporating new inputs and architectures [Maciej Glowacki, CERN Fellow]

Sabrina Giorgetti
& Jannicke Pearkes

Physics analysis

- Characterization of trigger performance in 2024 data on SM candles (J/Ψ/Z peak, etc…)
- Designing first physics analysis with bump-hunt in many di-object invariant masses as in [PRL 132 (2024) 081801]

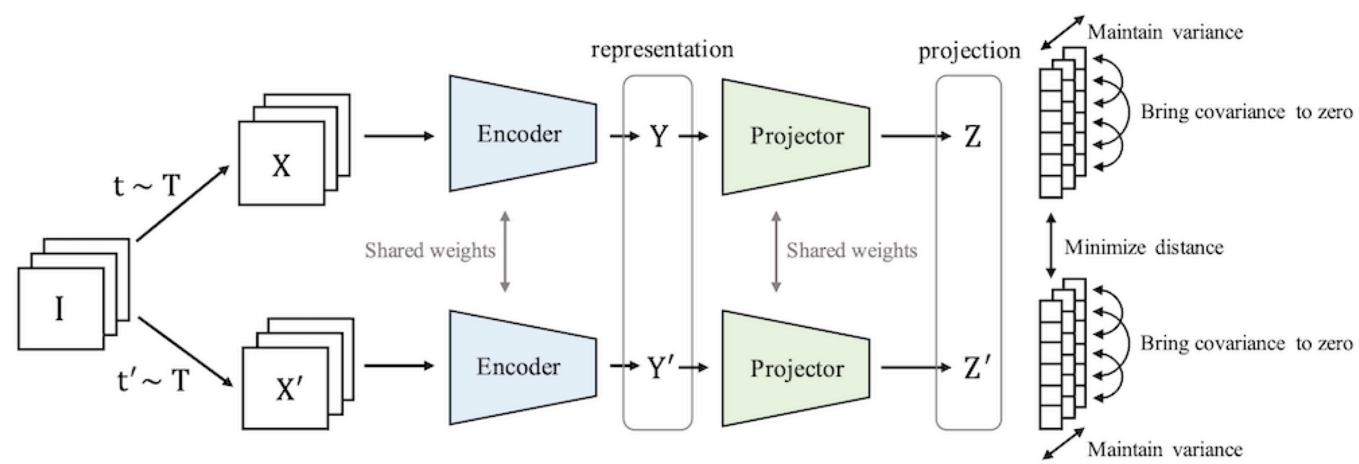


Model development

- Studying architectural improvement beyond VAE baseline → **Contrastive Learning** approach to improve embeddings (latent space) expressiveness

- **Contrastive learning** is a self-supervised learning (SSL) technique that aims to learn representations by comparing similar and dissimilar samples (called "augmentations")

- We are considering VicReg: *Variance-Invariance-Covariance Regularization for SSL*



  - **Invariance:** the two augmented views should produce similar embeddings

  - **High Variance:** each dimension of embeddings should contain meaningful information and not collapse to a constant value

  - **Low Covariance:** embedding dimensions should not have redundant information and should be independent

$$L_{\text{VicReg}} = \alpha \left(\frac{1}{N} \sum_{i=1}^{N} \| z_1^{(i)} - z_2^{(i)} \|^2 \right) + \beta \left(\frac{1}{d} \sum_{j=1}^{d} \max(0, \gamma - \sigma(z_j))^2 \right) + \gamma \left(\frac{1}{d} \sum_{i \neq j} \mathrm{Cov}(z_i, z_j)^2 \right)$$

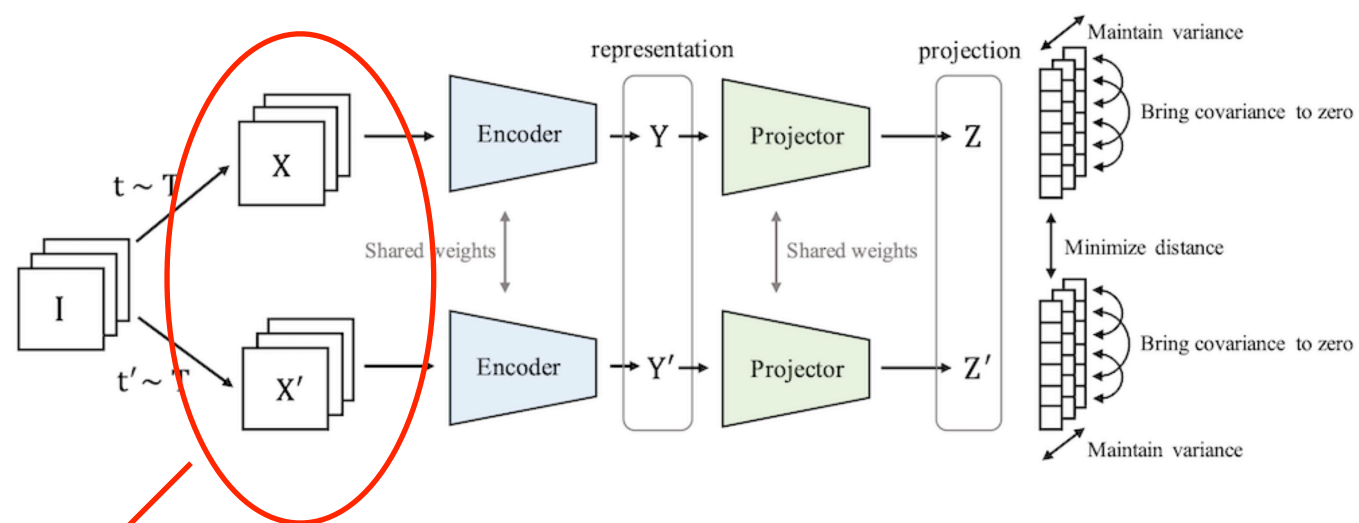
invariance **variance** **covariance**

Model development

- Studying architectural improvement beyond VAE baseline → **Contrastive Learning** approach to improve embeddings (latent space) expressiveness

- **Contrastive learning** is a self-supervised learning (SSL) technique that aims to learn representations by comparing similar and dissimilar samples (called "augmentations")

- We are considering [VicReg]: *Variance-Invariance-Covariance Regularization for SSL*



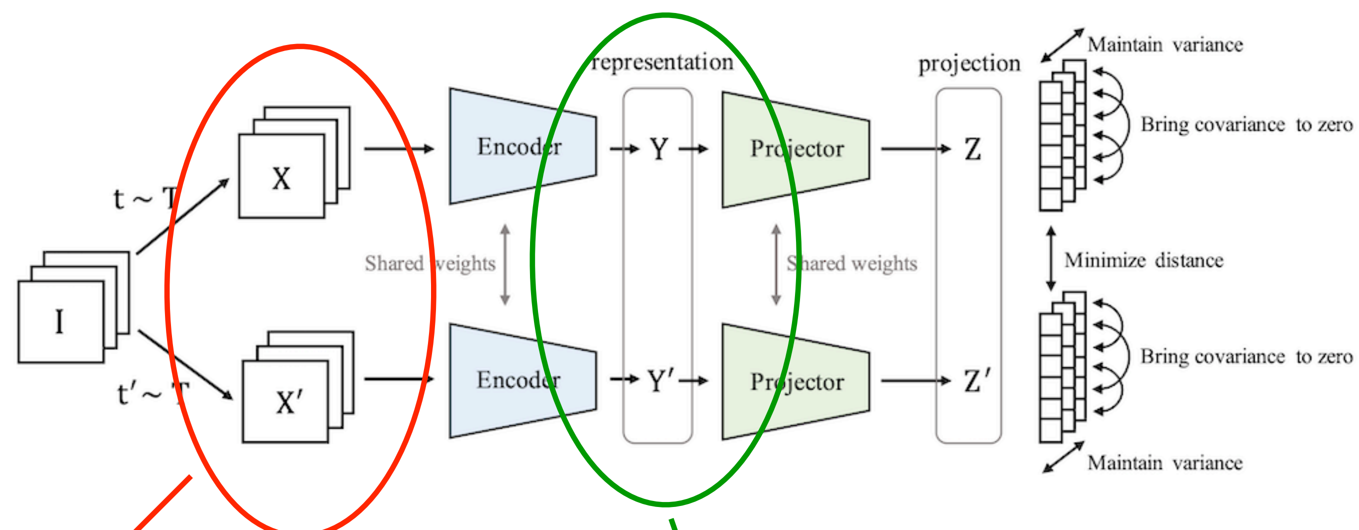
Augmentations considered:
gaussian smearing in within reconstruction resolutions and objects masking

Model development

- Studying architectural improvement beyond VAE baseline → **Contrastive Learning** approach to improve embeddings (latent space) expressiveness

- **Contrastive learning** is a self-supervised learning (SSL) technique that aims to learn representations by comparing similar and dissimilar samples (called "augmentations")

- We are considering VicReg: *Variance-Invariance-Covariance Regularization for SSL*

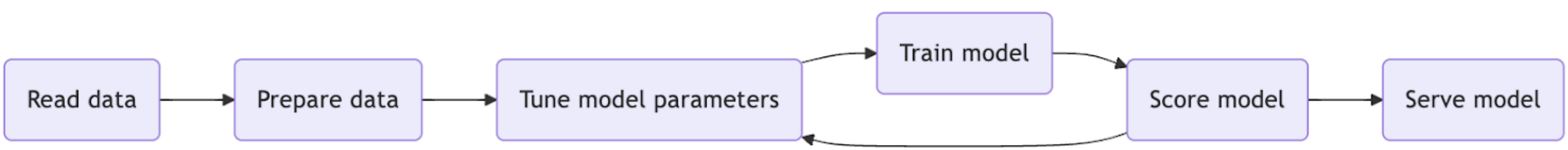


Augmentations considered:
gaussian smearing in within reconstruction resolutions and objects masking

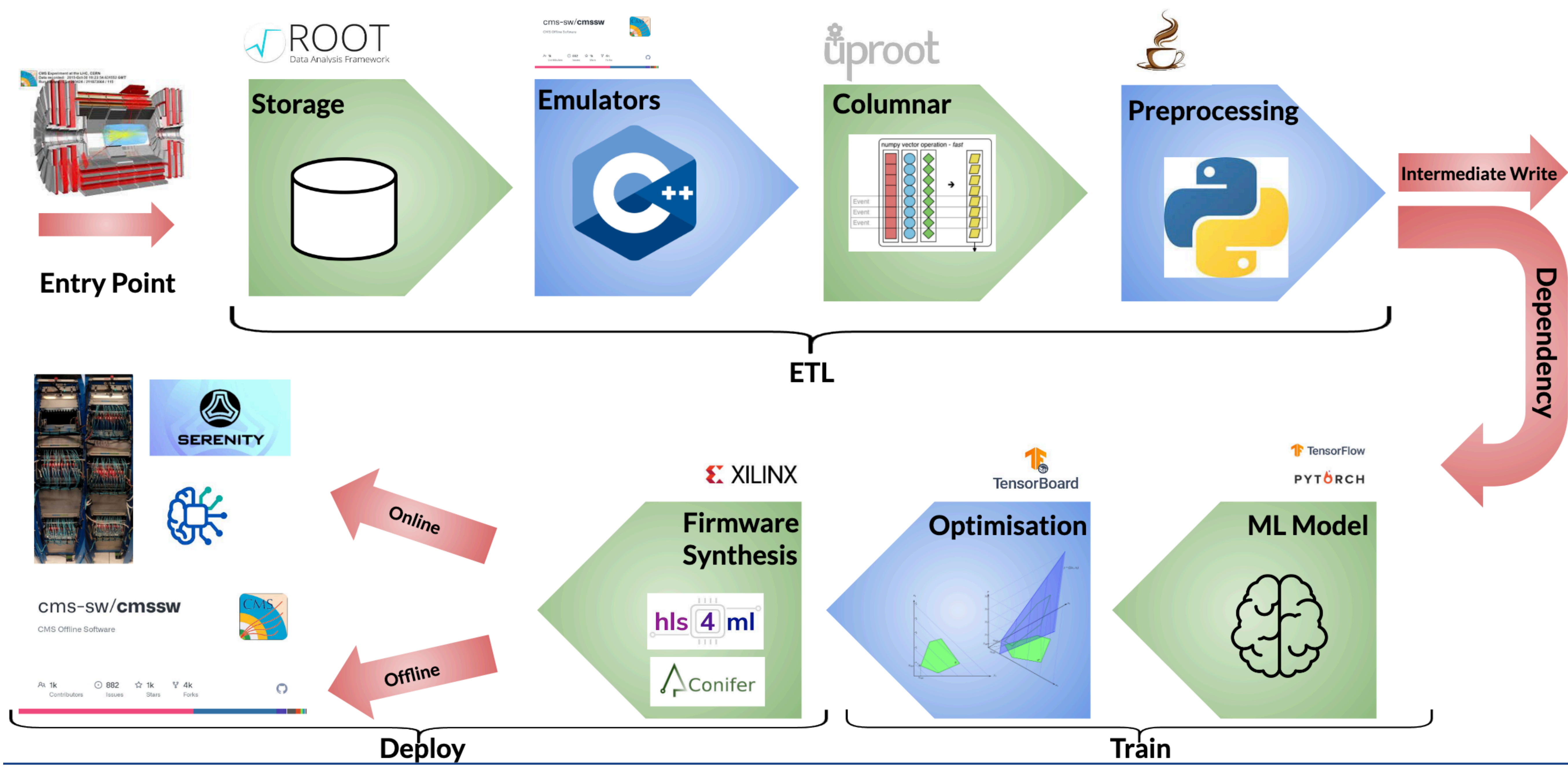
Different strategies being explored for AD downstream task:
VAE or multi-dimensional distances — TBD based on latency constraints and physics performance

Operational Automation: MLOps

- **A typical Machine Learning Lifecycle:**
  - Data integration from multiple sources
  - Data processing
  - Data loading and batching
  - Hyperparameter tuning, establish a Pareto front based on *some* metrics
  - Model deployment — Version *everything:* data, model, code



CMS L1T Workflow



How often?

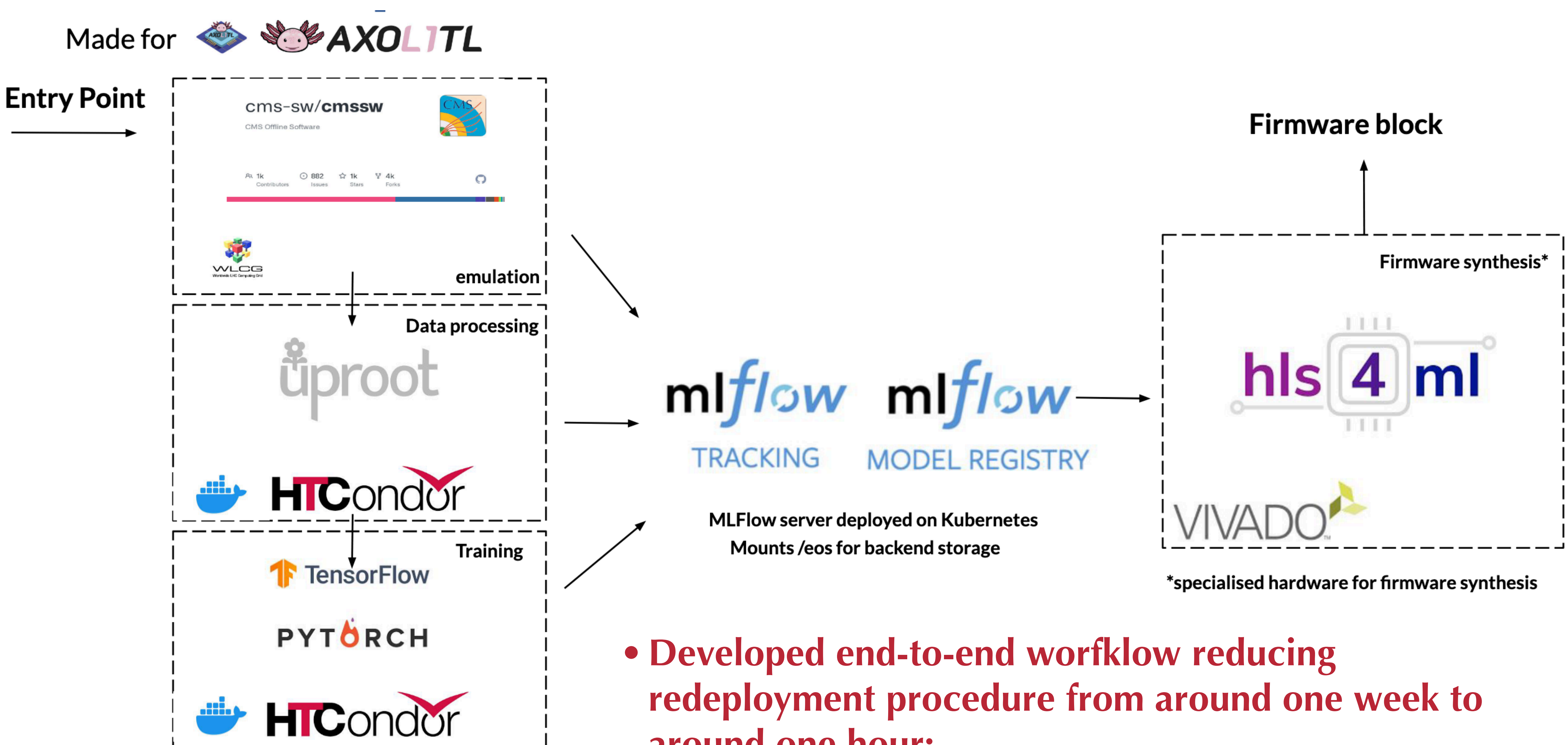
Necessary to automate and speed up the workflow!

Possible time scale of trigger ML retraining and redeployment

Current time scale of trigger ML retraining and redeployment

Seconds	Days	Months
Beam fluctuations	Beam conditions and detector variation	Large scale detector changes
		New physics goals
Built in trigger robustness	Subsystem calibration	Reconfigure and rebuild trigger

MLOPs initial implementation



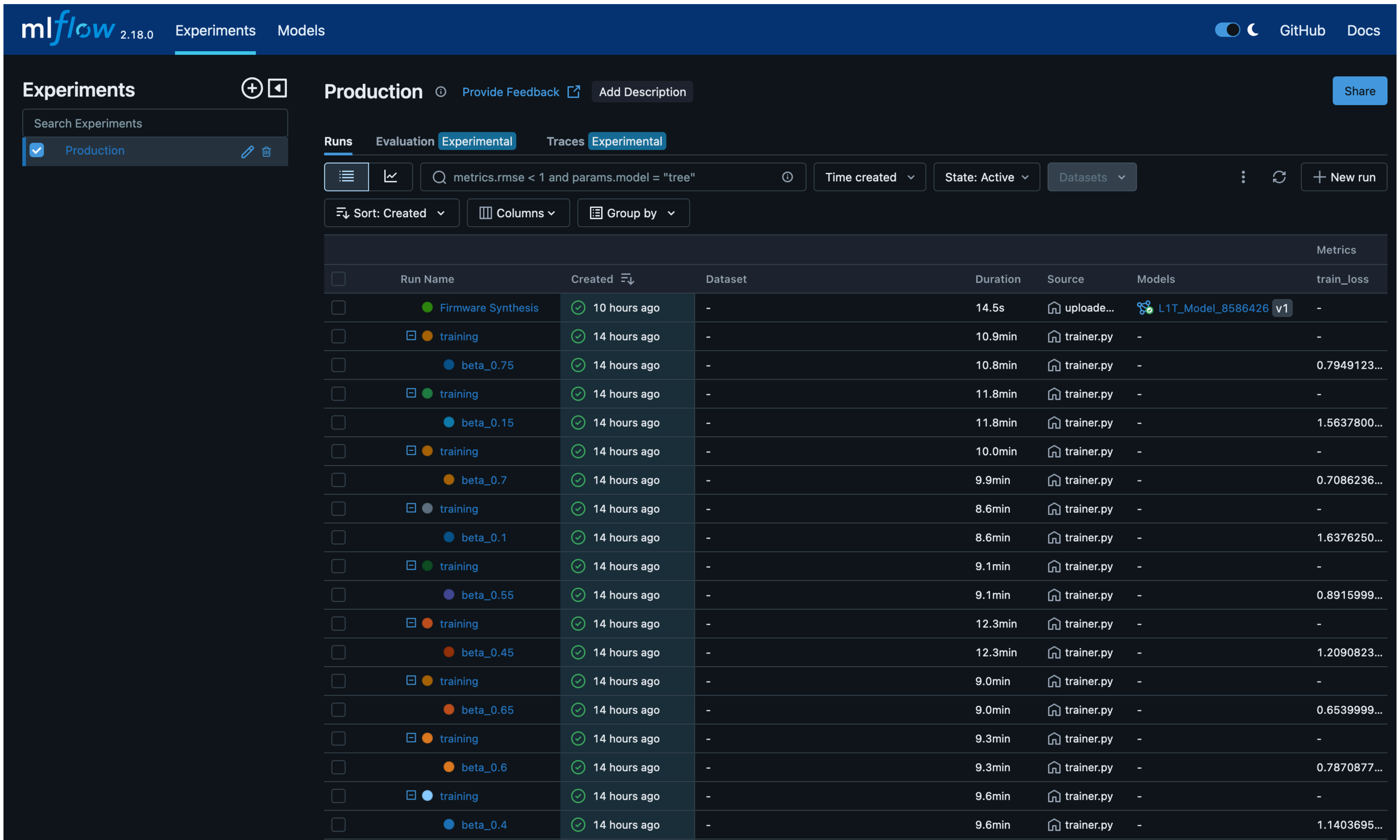
- **Developed end-to-end worfklow reducing redeployment procedure from around one week to around one hour:**
  - Producing data files for training of models
  - Training and evaluate models on produced data files
  - Producing firmware for trained models

- Code on CERN's GitLab instance with execution orchestrated using GitLab CI/CD

Diptarko Choudhury
& Maciej Glowacki

Example pipeline w/ MLFlow

- An MLFlow server was set up, used for logging ML training experiments and registering trained models

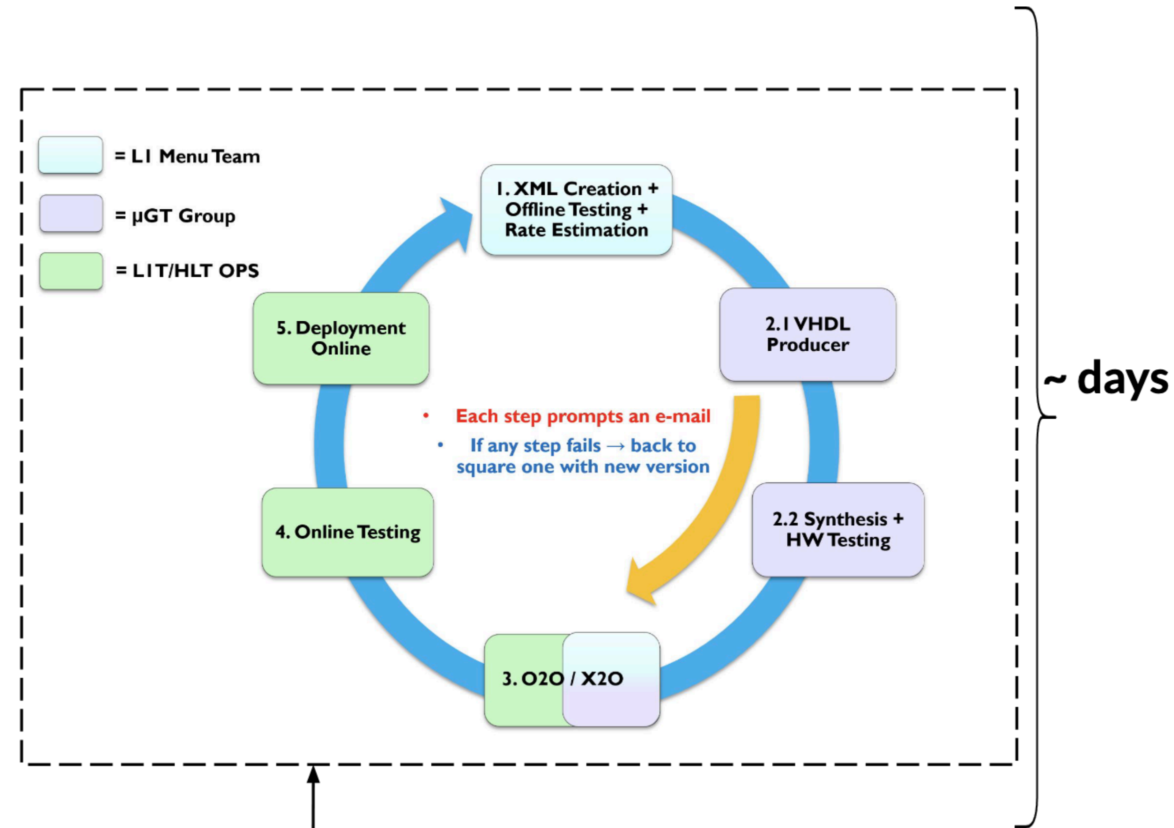


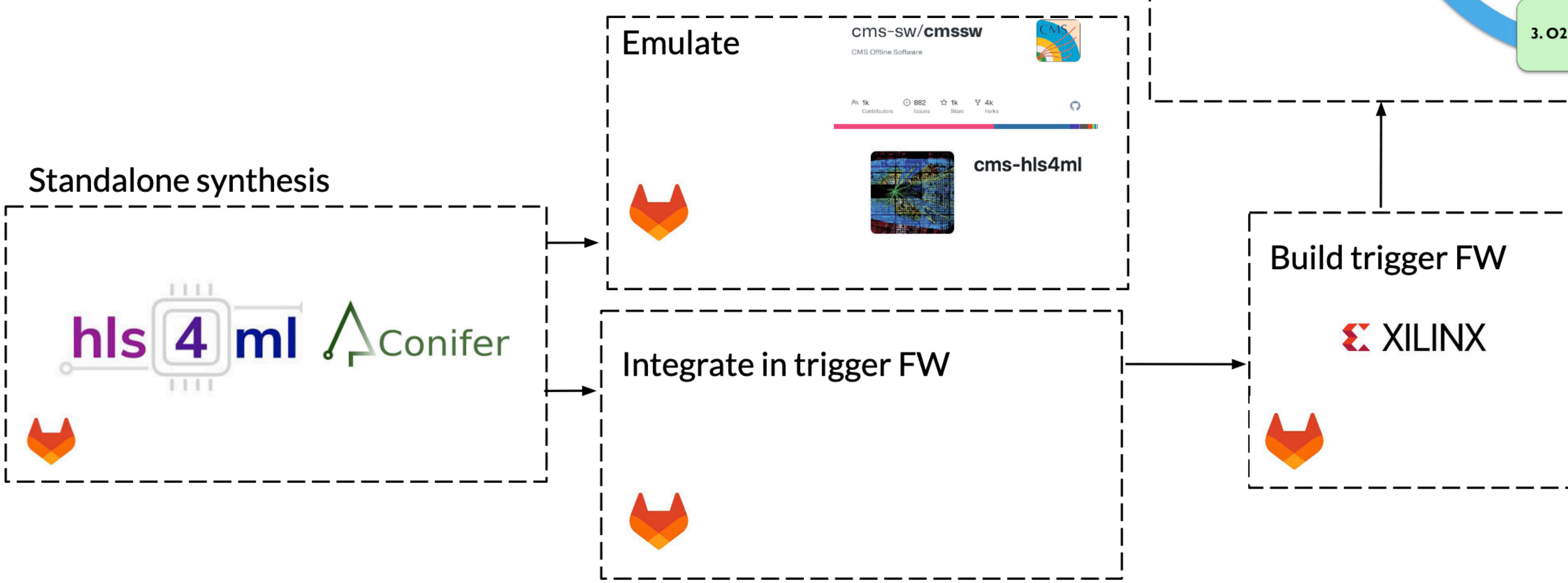
Diptarko Choudhury
& Maciej Glowacki

[link]

MLOPs initial implementation

- FW deployment into online (FPGA) and offline (CMSSW emulator) settings under development
- Interfacing with the GT protocol
- Currently to deploy a new model at P5 requires a CMSSW release
    - Tied to HLT CMSSW release (quarterly schedule)



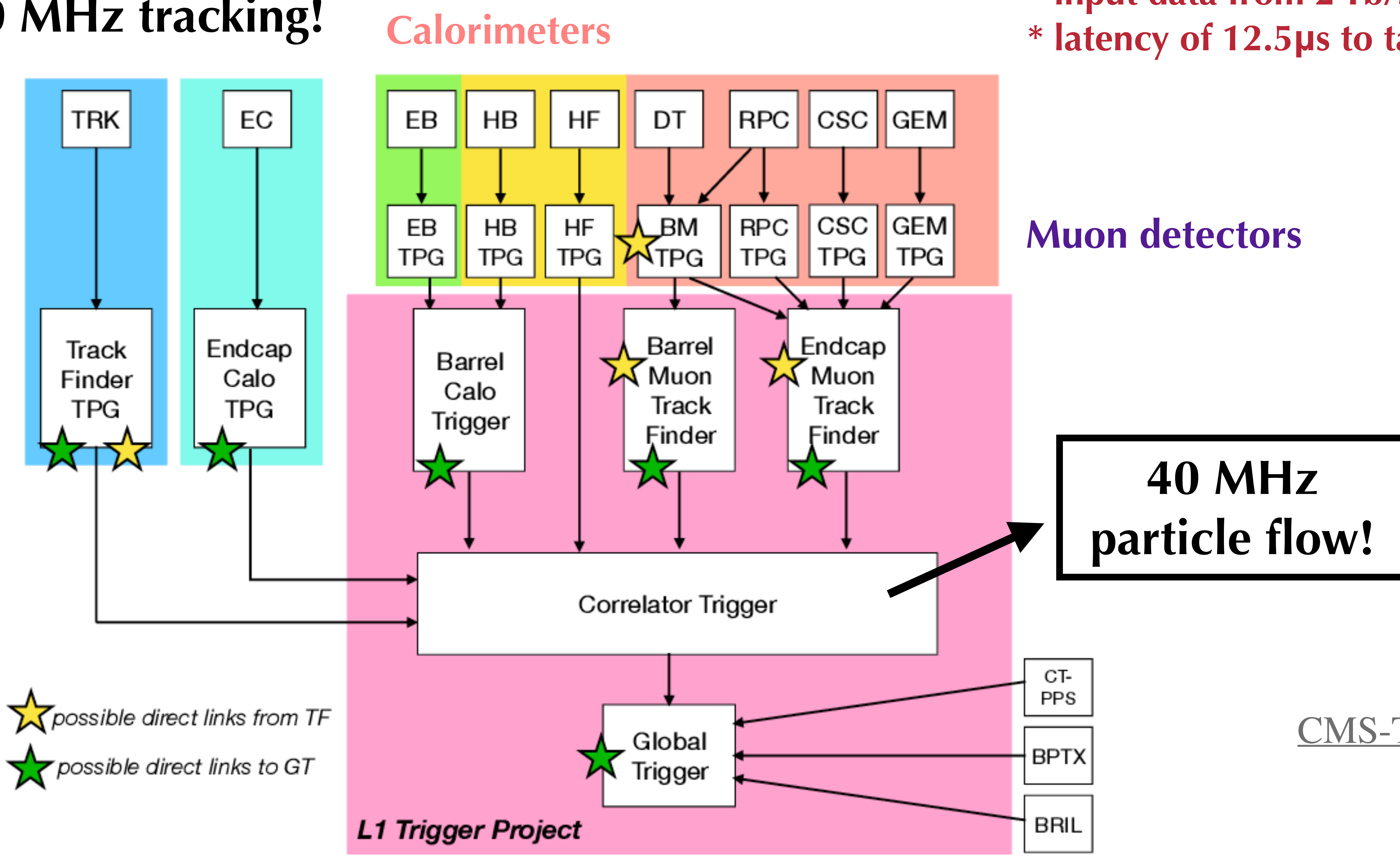


Anomaly detection @ Phase 2

At HL-LHC, up to 200 pile-up interactions: *CMS is upgrading the L1T and HLT to enable the same physics program we are doing now (at @60 PU)*

40 MHz tracking!

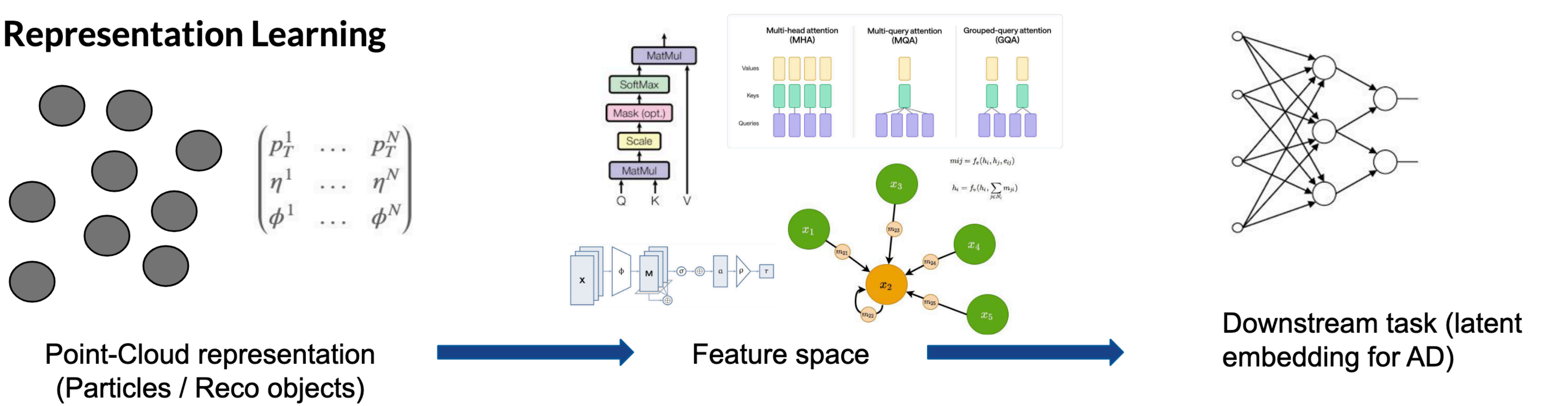
* input data from 2 Tb/s to 63 Tb/s
* latency of 12.5μs to take decision



CMS-TDR-021

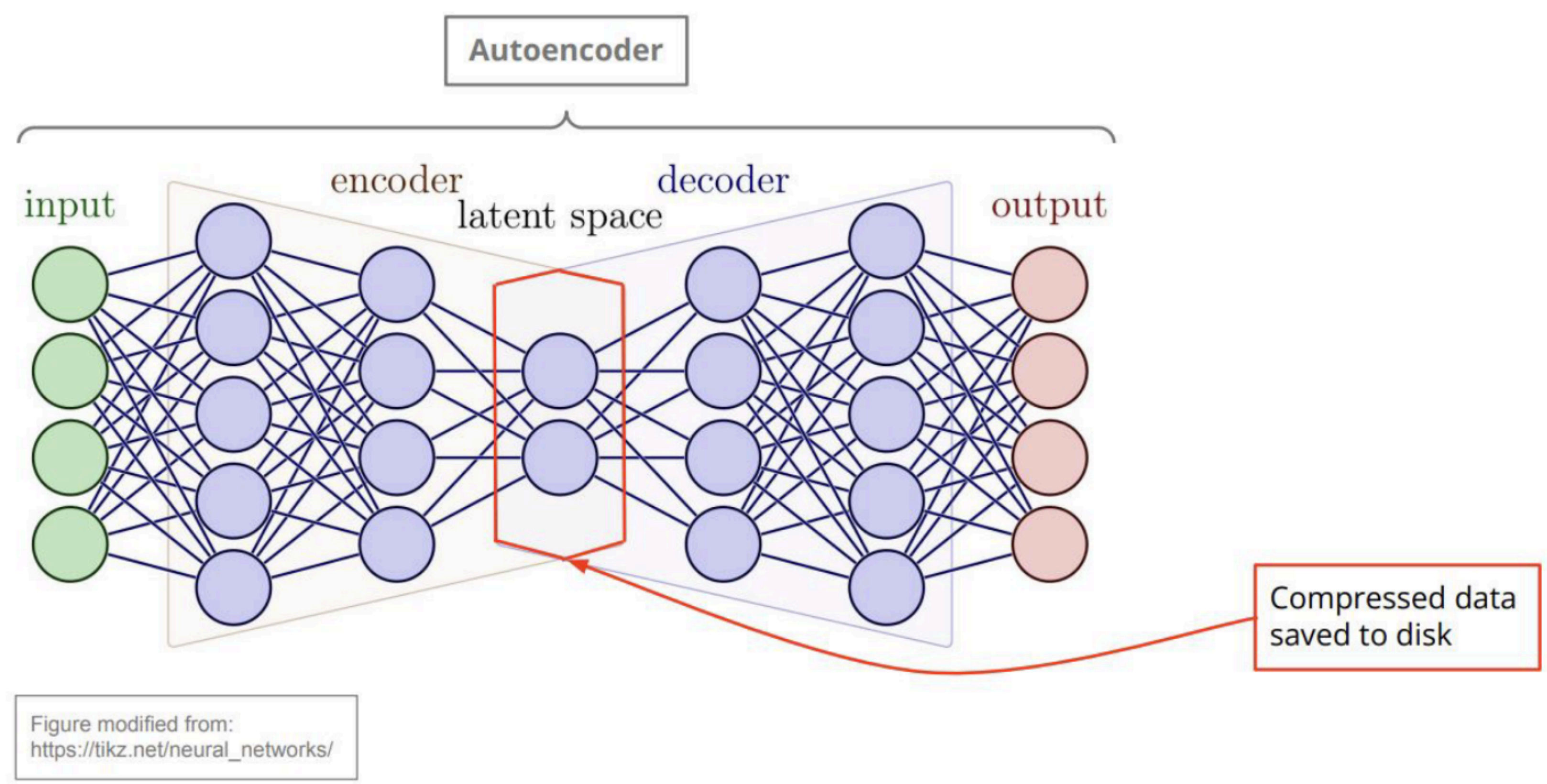
Kick starting: Anomaly detection @ Phase 2

- **First phase of R&D:** take improved reconstructed objects in the Global Trigger and reproduce baseline AXOL1TL with VAE to understand gain from better reconstruction
- **Second phase of R&D:** design novel point-cloud based AD algo that takes as input all reconstructed particles from L1 CT
  - inspiration from jet tagging work guaranteeing permutation invariance & equivariance through equivariant layers (DeepSets, GNN, Self-Attention)
  - multiple representation learning strategies to be explored: fully unsupervised, SSL, as well weakly supervised with noisy labels [e.g. Abhijith G. et all 2401.08777]



Kick starting: Phase 2 L1T Scouting

- CMS L1T data (trigger passthrough) reaching ~ Millions of PB
- Too much data to store
  - demand for efficient compression for downstream storage
- Use Machine Learning to obtain an expressive embedding for downstreams physics
- Scope for larger, sophisticated architectures due to relaxed constraints of the buffer
- Baseline idea to be explored makes use of autoencoders but imagine SOA representation learning approaches to be more expressive and still be implementable on hardware



Summary

- A first baseline CMS anomaly detection trigger was designed and integrated in the system by CMS collaborators in the past ~ 3 years
  - already collected ~ 100/fb this year
- NGT to push the frontier of this innovative technology to enhance the physics reach of CMS by allowing us to hire personnel fully dedicated to it
- The team is advancing on multiple fronts of the project and we expect major advancements in the next couple of years
  - 2025 & 2026 data taking and analysis
  - MLOps to aid current and future ML-based trigger algos
  - Phase 2 R&D
- Stay tuned!

See also public talks:

Noah's talk at FastML [Conference Talk]
Melissa's talk at CHEP [Conference Talk]
Jennifer's talk at ML4Jets [Conference Talk]