Finanziato dall'Unione europea
NextGenerationEU

Ministero dell'Università e della Ricerca

Italiadomani
PIANO NAZIONALE DI RIPRESA E RESILIENZA

ICSC
Centro Nazionale di Ricerca in HPC, Big Data and Quantum Computing

AISSAI Anomaly Detection Workshop

Anomaly detection for data quality monitoring of the CMS detector

Federica M. Simone on behalf of the CMS Collaboration
Bari Polytechnic & INFN Bari

de' remi facemmo ali
POLITECNICO DI BARI
1990

INFN
Istituto Nazionale di Fisica Nucleare

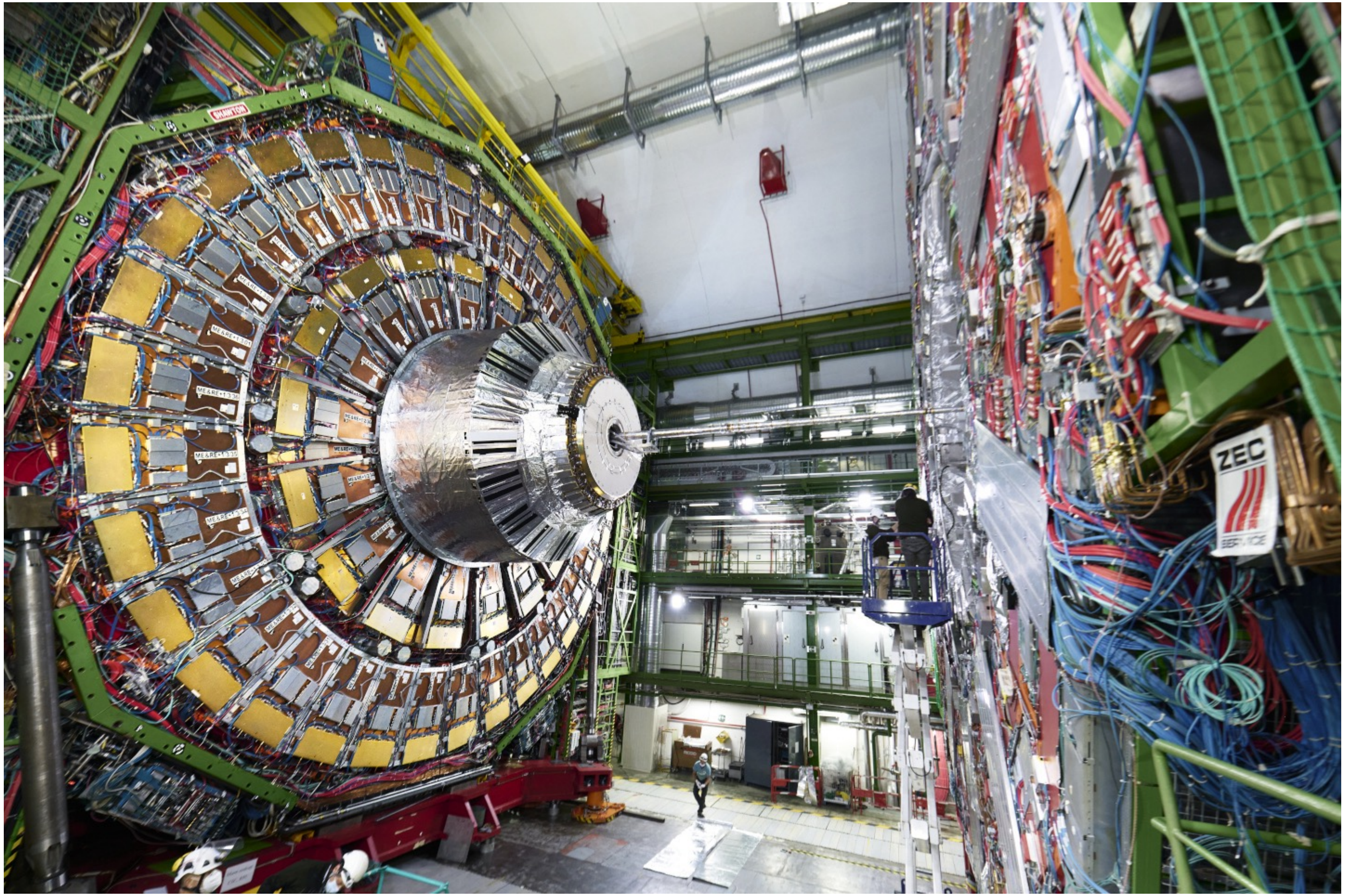
AISSAI Anomaly Detection Workshop

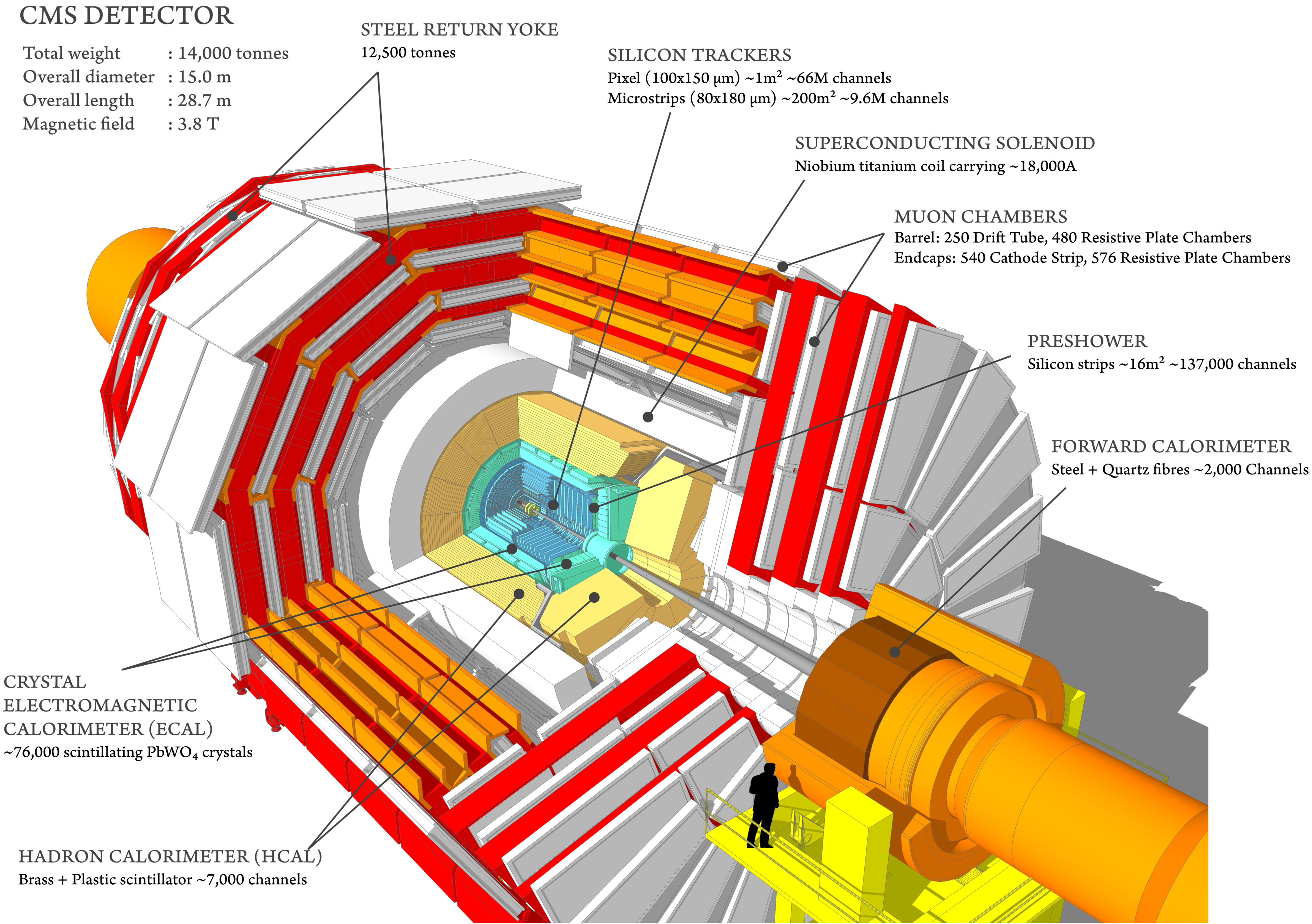
The Large Hadron Collider

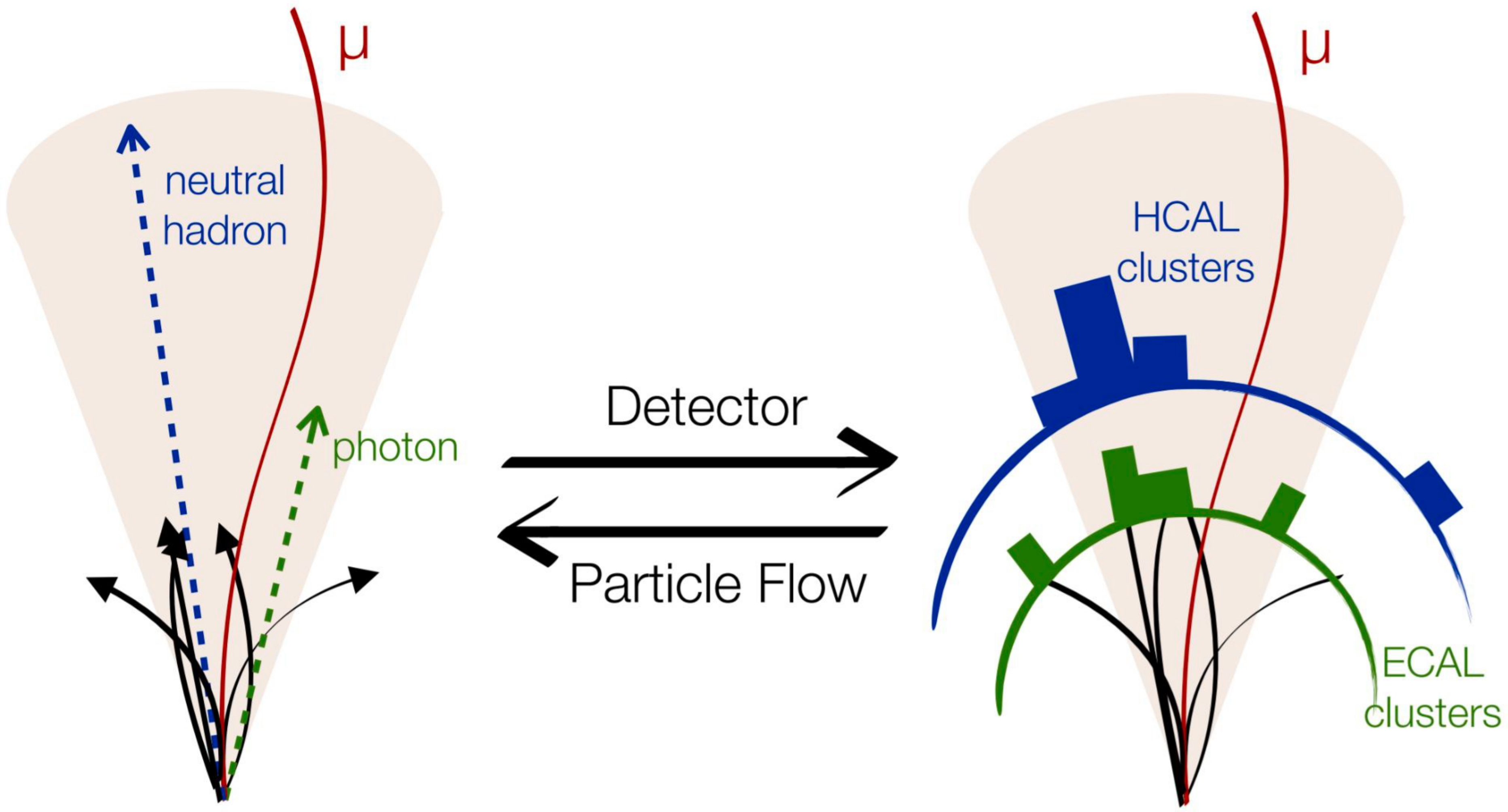
The Large Hadron Collider (LHC) is the largest and most powerful particle accelerator in the world, situated at the CERN near Geneva

The LHC accelerates two beams of protons that are made to collide at four points, around which the main experiments are located

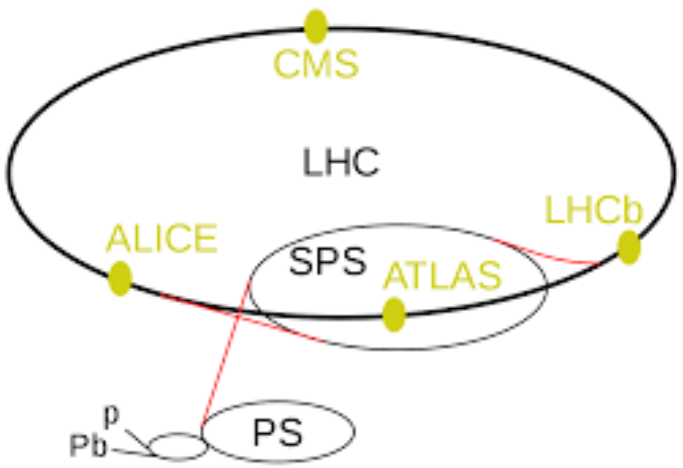
The CMS experiment

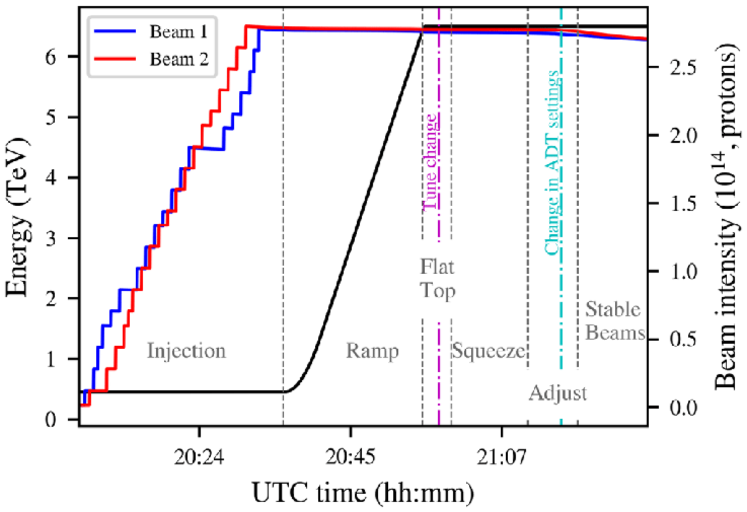


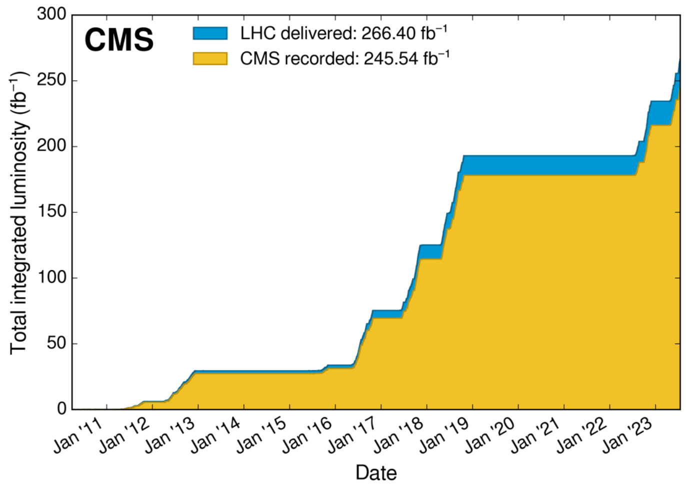

CMS DETECTOR
Total weight : 14,000 tonnes
Overall diameter : 15.0 m
Overall length : 28.7 m
Magnetic field : 3.8 T
STEEL RETURN YOKE
12,500 tonnes
SILICON TRACKERS
Pixel (100x150 μm) ~1m² ~66M channels
Microstrips (80x180 μm) ~200m² ~9.6M channels
SUPERCONDUCTING SOLENOID
Niobium titanium coil carrying ~18,000A
MUON CHAMBERS
Barrel: 250 Drift Tube, 480 Resistive Plate Chambers
Endcaps: 540 Cathode Strip, 576 Resistive Plate Chambers
PRESHOWER
Silicon strips ~16m² ~137,000 channels
FORWARD CALORIMETER
Steel + Quartz fibres ~2,000 Channels
CRYSTAL ELECTROMAGNETIC CALORIMETER (ECAL)
~76,000 scintillating $PbWO_4$ crystals
HADRON CALORIMETER (HCAL)
Brass + Plastic scintillator ~7,000 channels


μ
neutral
hadron
photon
Detector
Particle Flow
μ
HCAL
clusters
ECAL
clusters

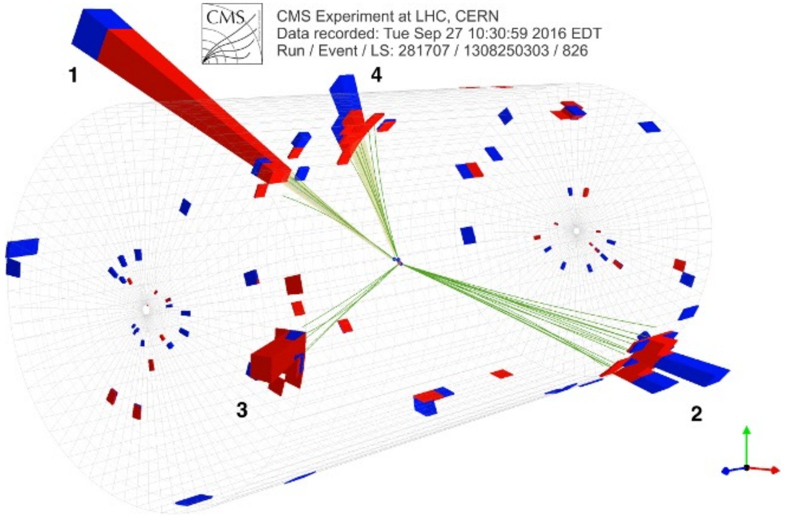
Data taking







Recorded event

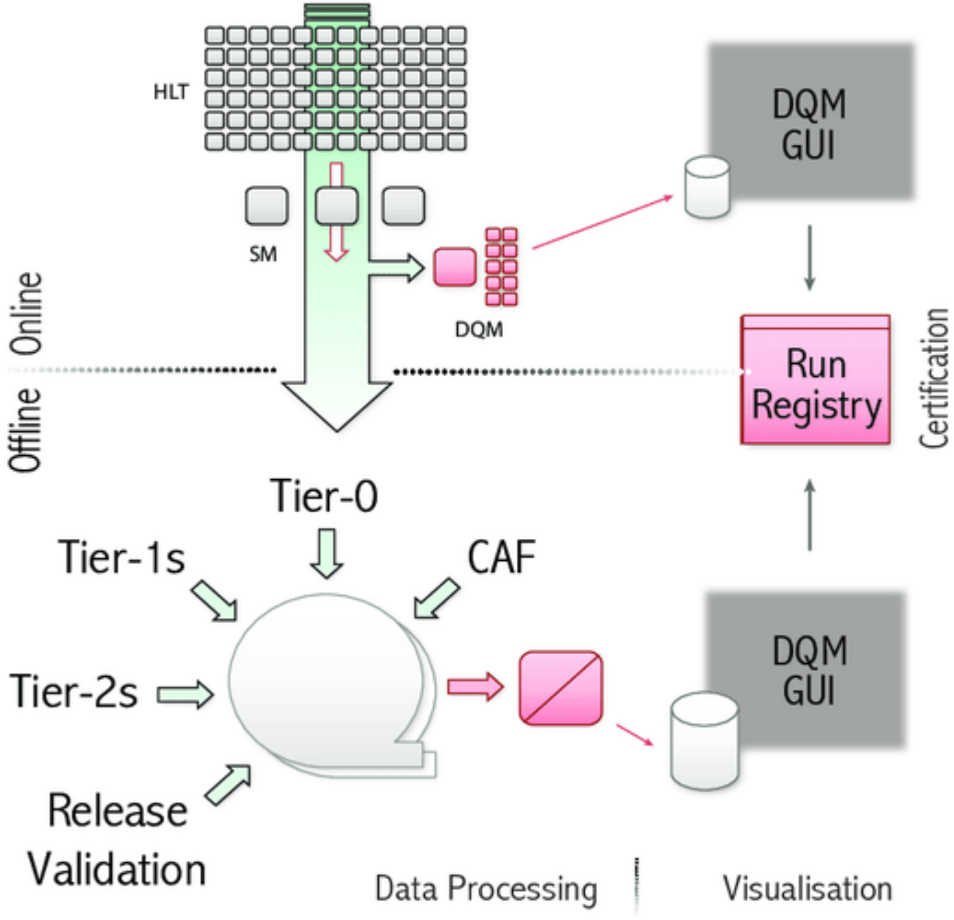


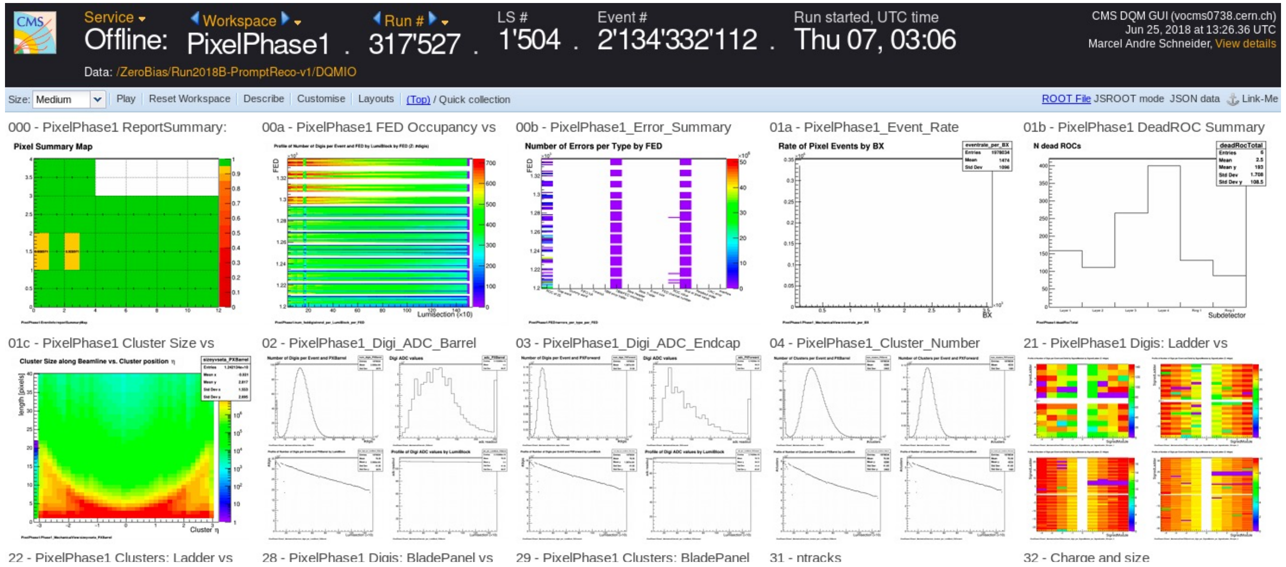
Lumisection (LS)

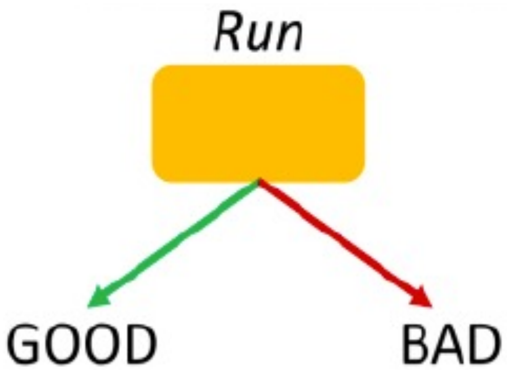
"Run" → thousands of LS

Data Quality Monitoring (DQM)

- **Online monitoring**: promptly raising alarms in case of detector malfunctioning
- **Offline Data Certification (DC):** identify high quality data usable for physics analysis
- **Offline monitoring and debugging**: providing inputs to experts to investigate spotted issues





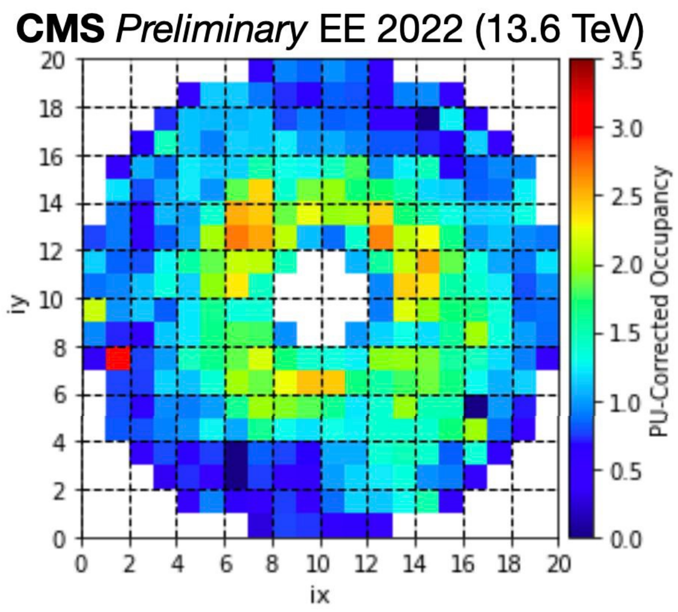


Monitoring Elements (ME)

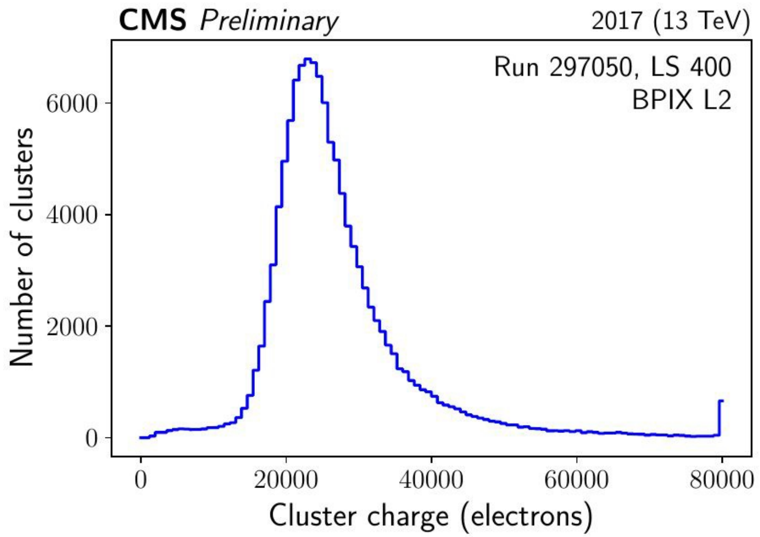
- Set of quantities which are typically inspected by experts with **per-run granularity**

- **Large variety of MEs**:
  - low level quantities e.g., hit occupancies in the detectors
  - high level quantities e.g., energy of reconstructed particles

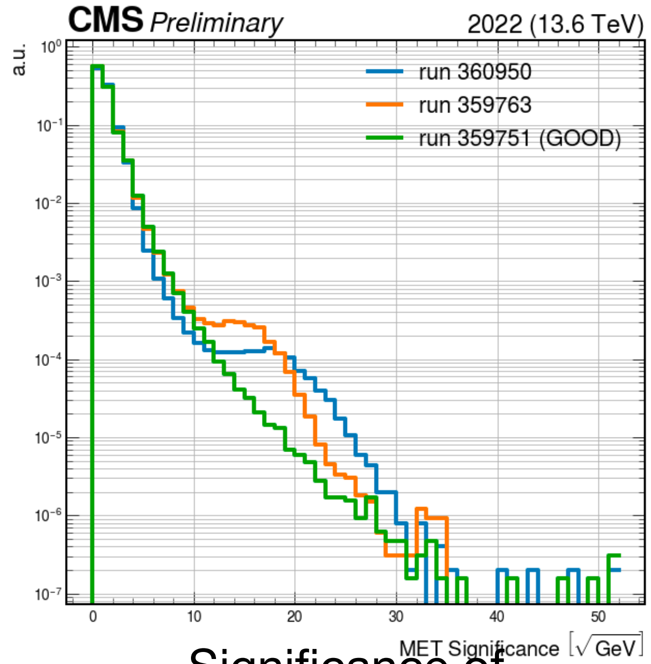
  → Specific for the different CMS subsystems and “physics objects”!



ECAL occupancy images



Pixel tracker cluster charge distribution

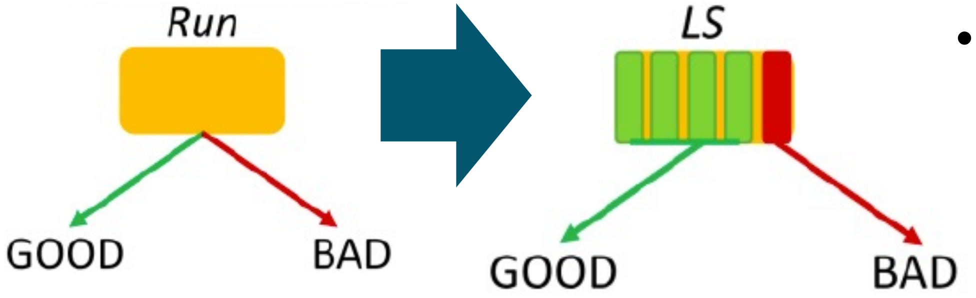


Significance of reconstructed MET

DQM challenges and limitations

- Online monitoring is a **highly time-sensitive** operational task

- Data Certification should **ensure high quality data**, while **limiting false positive rate** to fully exploit the luminosity delivered by LHC



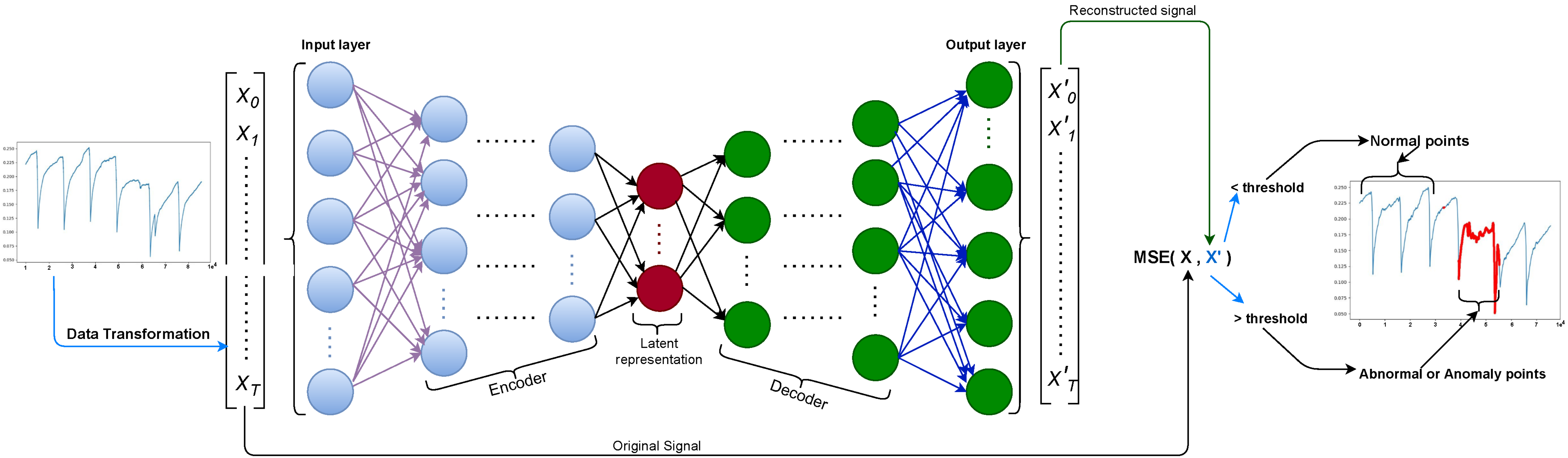
- Limited **time granularity** (run) can potentially hide transient issues only affecting few lumisections
  - Drawback: **per-LS approach** increases the number of MEs by a factor $\mathcal{O}$(10^3)
    → human inspection not feasible

- Impossible to foresee or simulate **all potential failure scenarios**

Autoencoders for Anomaly Detection

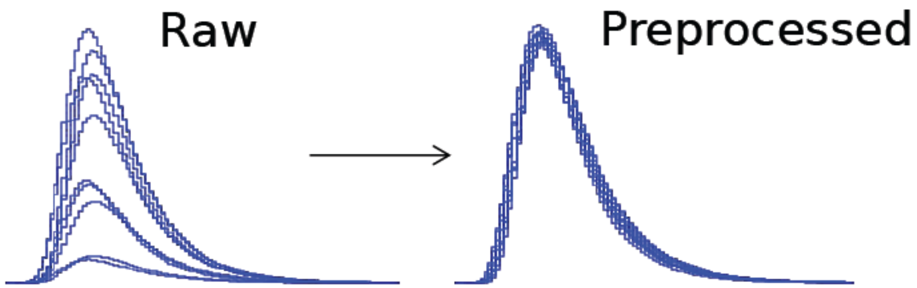
- High number of features
- Large class imbalance (most data is good)
- Non-exhaustive definition of failures

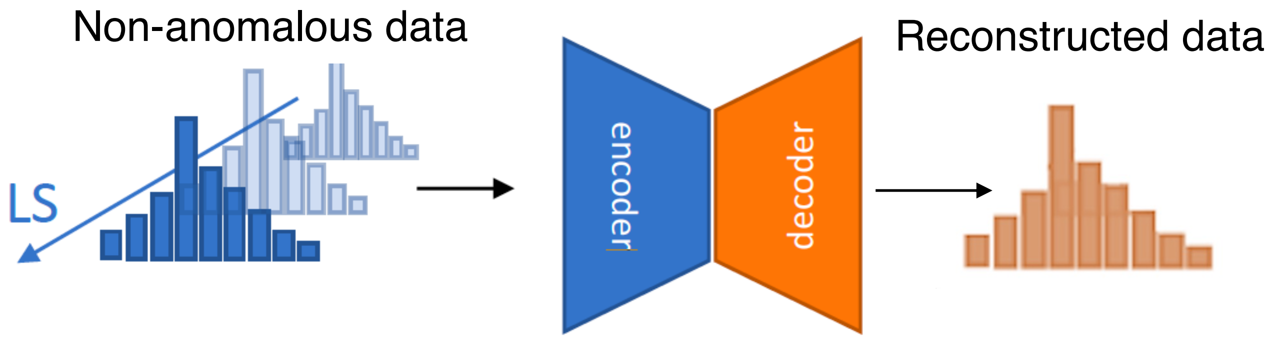
Unsupervised learning for anomaly detection



Encoding via dimensionality reduction

Machine Learning for DQM: general workflow





1 – Data preprocessing

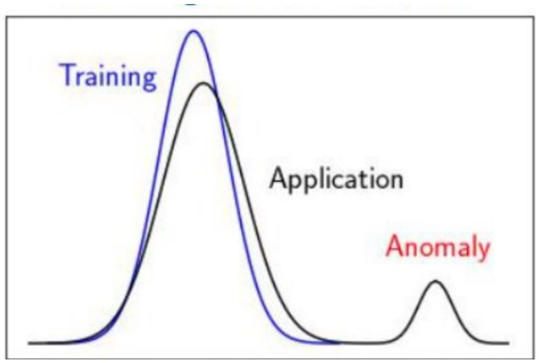
- Per-lumisection ME distributions
- Normalise to data taking conditions (e.g. pile-up)
- Filter over detector status, available statistics etc

2 – Training of NN

- Architecture depends on data dimensionality, sample size etc
- General idea: the model should learn an abstract representation of good data

3 – Testing

- Measure performance on labelled data
- Set metrics and thresholds

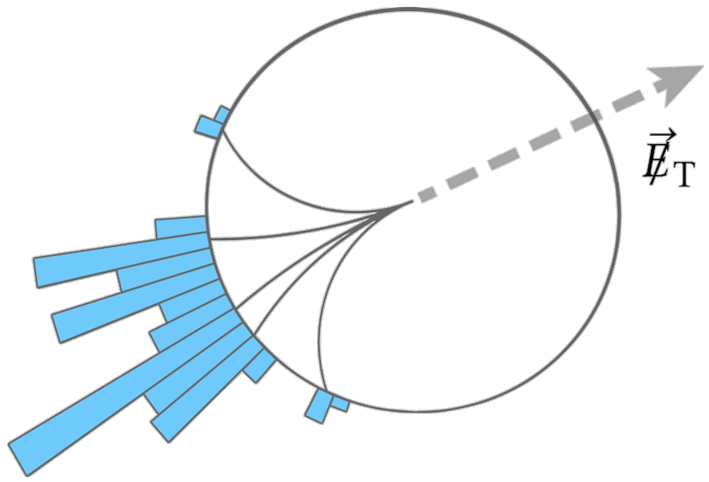


4 – Flag BAD/GOOD

- Flag data with LS granularity. Either reject anomalous data or further investigate

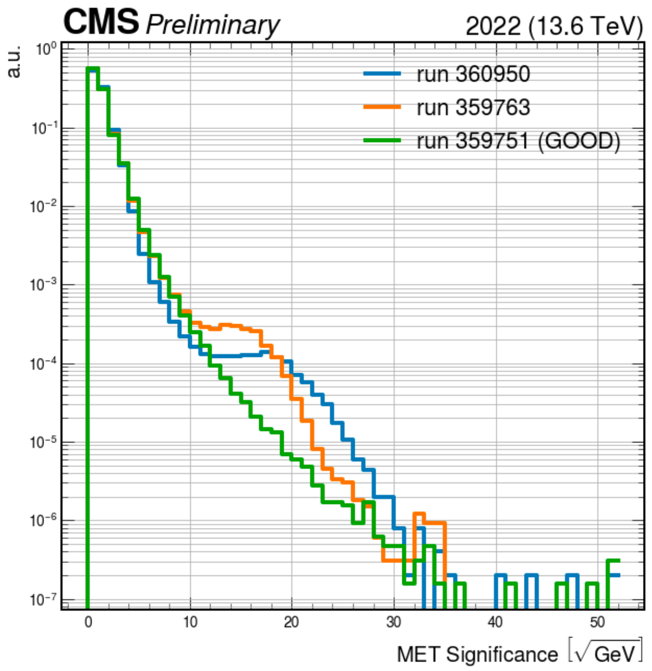
Run Number	dt	ecal	es
360131	GOOD	GOOD	GOOD
360130	GOOD	GOOD	GOOD
360129	GOOD	GOOD	GOOD
360128	GOOD	GOOD	GOOD

Specific applications: JetMET Data Certification



Given the variety of subdetector technologies and geometries, and the number of physics reconstructed objects, **CMS Data Certification is done separately for each sub-system** using a dedicated set of MEs

JetMET DC ensures quality of quantities related to reconstructed particle Jets and Missing Transverse Energy

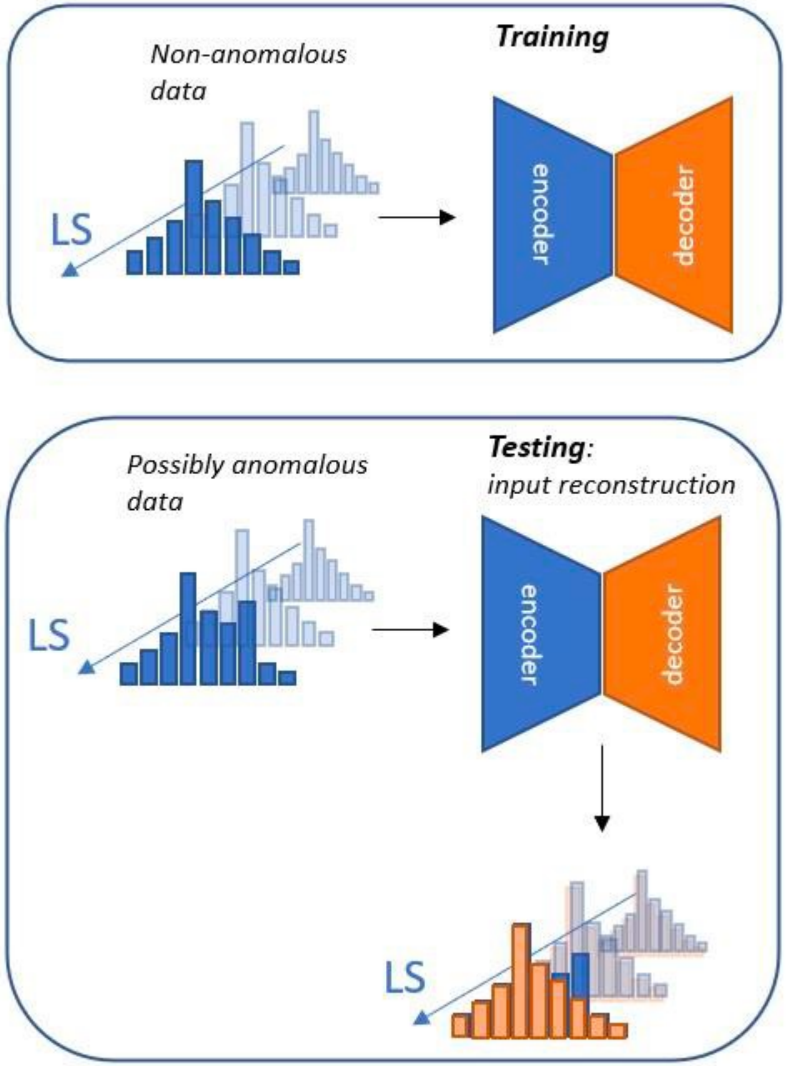


Anomaly detection approach born out of necessity:

- **bump in the MET** significance distribution causing **entire runs** to be flagged as BAD
- **per-LS inspection** would have shown that the issue was limited to a small number of LSs

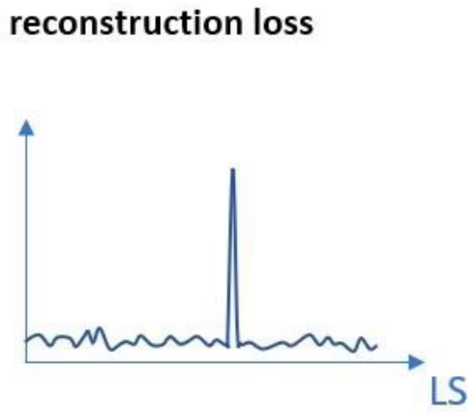
$$METSig \equiv \frac{MET}{\sqrt{SumET}} = \frac{MET}{\sqrt{\sum|\vec{p}_T|}}$$

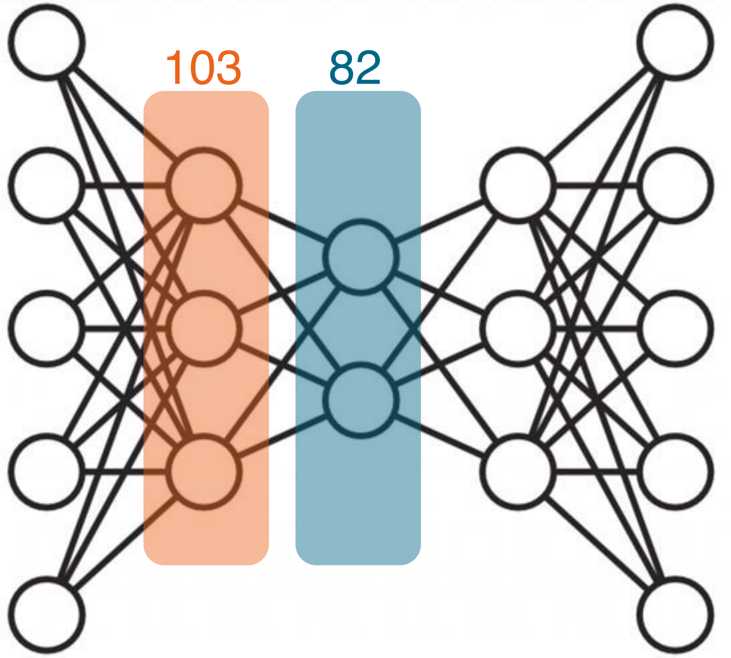
JetMET Data Certification: strategy



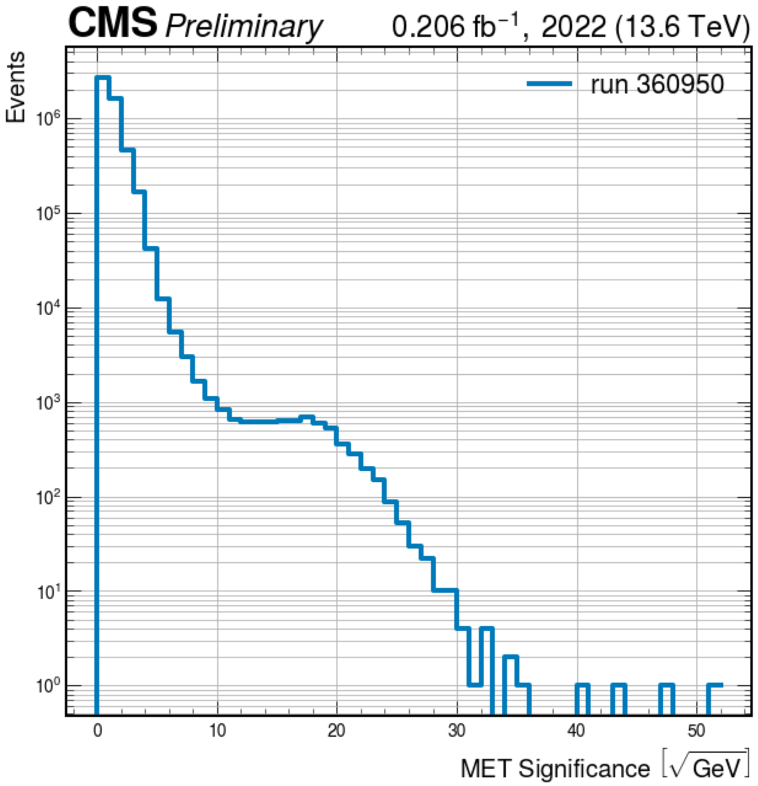
- **ME:** 1D Jet energy fractions, MET(sig) distributions
- **Autoencoder** model trained per ME
- **Training** on labeled GOOD runs, minimization of the reconstruction loss
- **NN parameters** set using Optuna (https://optuna.org)

$$metric = \frac{|mean(MSE_{after}) - mean(MSE_{before})|}{stddev(MSE_{before})}$$

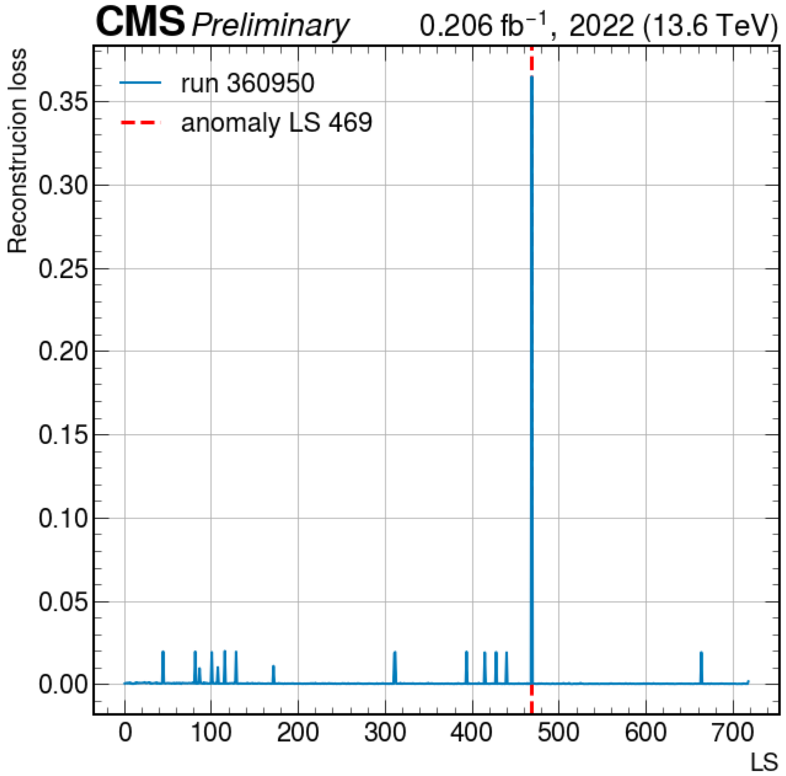




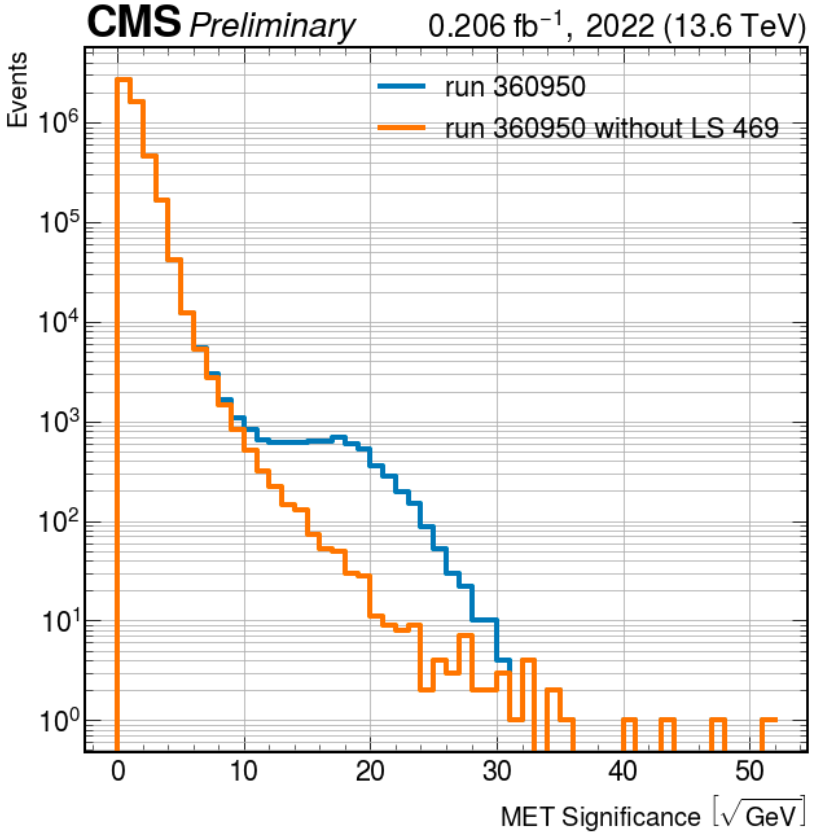
JetMET Data Certification: results



Run labelled as BAD due to anomalous MET sig shape



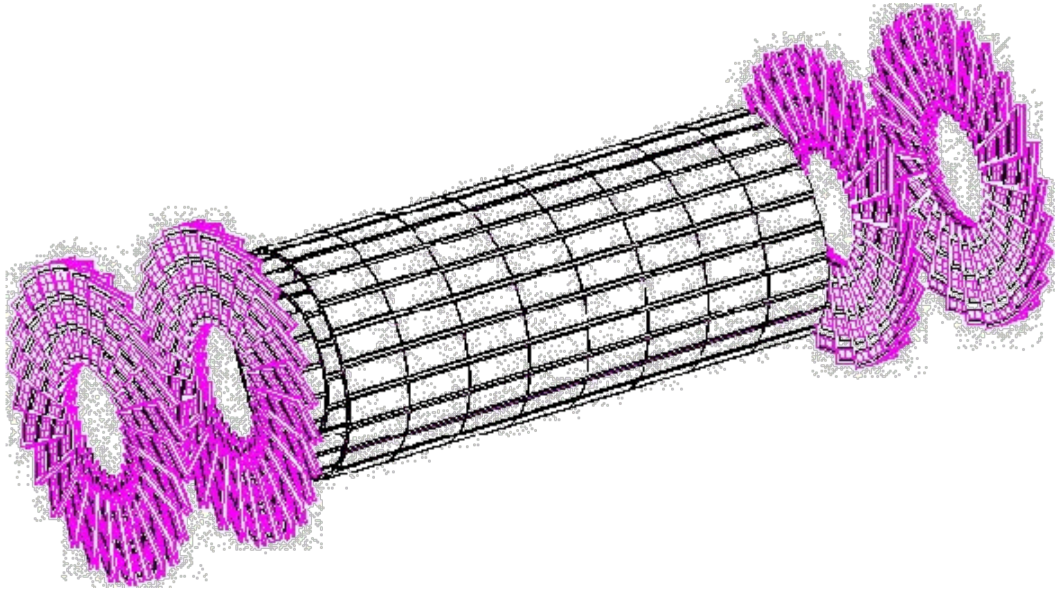
Per-LS reconstruction loss reveals the anomaly being limited to only one LS

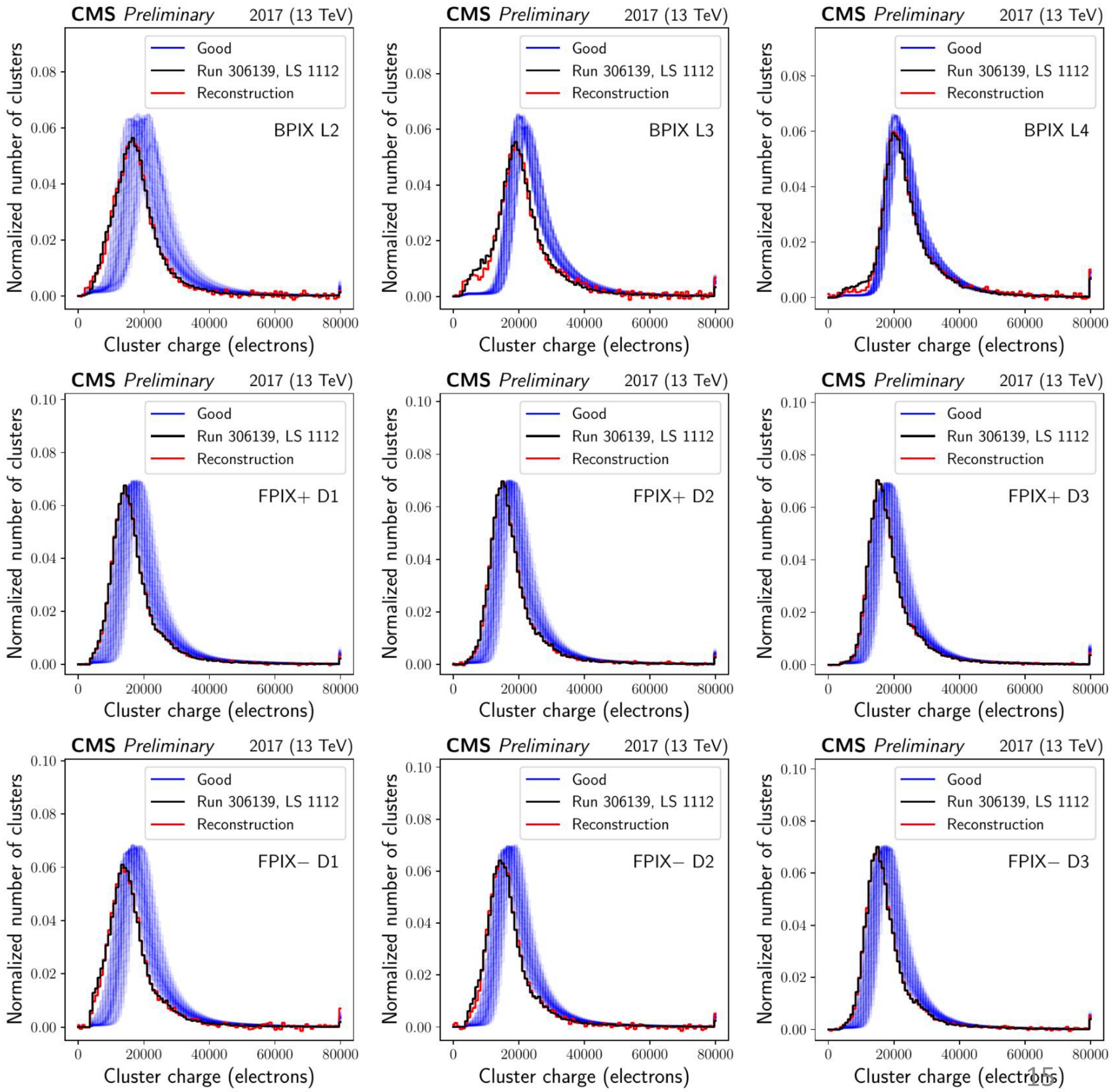


This system recovers quality of entire run after removing a small subset of anomalous data

Specific applications: Pixel tracker offline DQM

INPUT Monitoring Elements: hit occupancies and distributions of collected electric charge per cluster in the different layers and disks of the detector





Pixel tracker: studies of different unsupervised methods

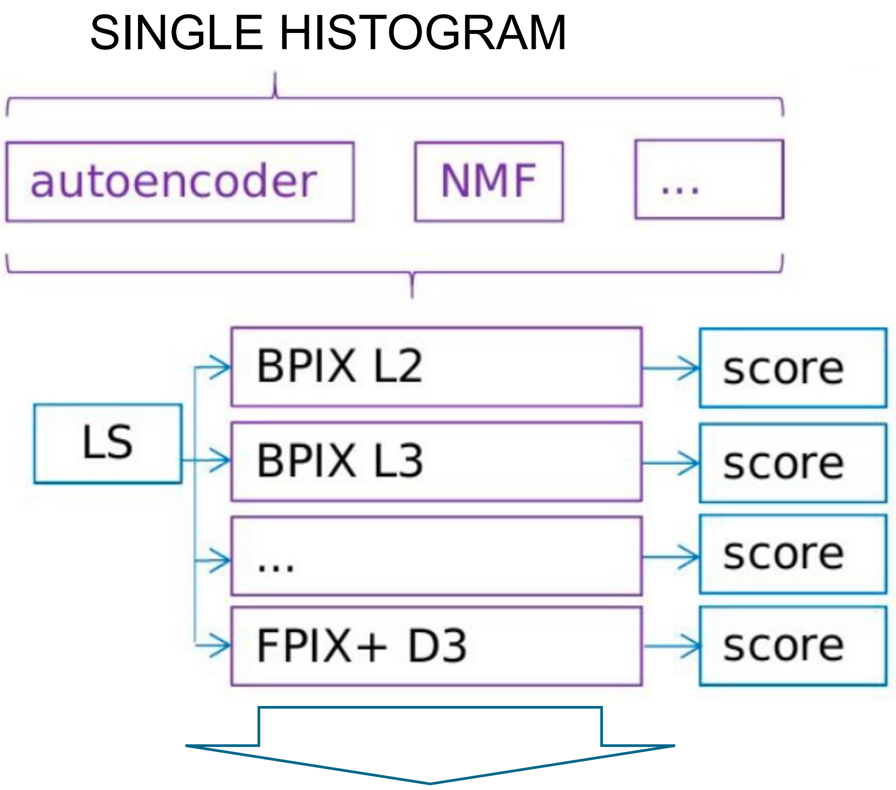
Moments: Comparison of the first and second order moments of a histogram to the distribution of those moments in the training set.

Landau fit: Mean-squared-error (MSE) between a histogram and a fitted Landau distribution.

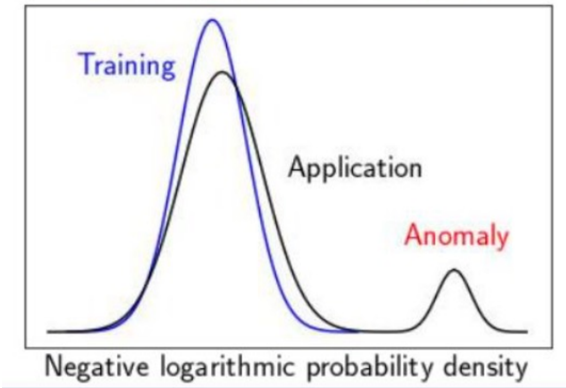
Templates: minimum MSE between a histogram and each of a set of well-chosen reference histograms.

NMF: MSE between a histogram and its nonnegative matrix factorization (NMF) reconstruction as an optimized linear combination of basis components extracted from the training set.

Autoencoder: MSE between a histogram and its autoencoder reconstruction.



Fit multi-dimensional probability density to scores of the training set



Pixel tracker: studies of different unsupervised methods

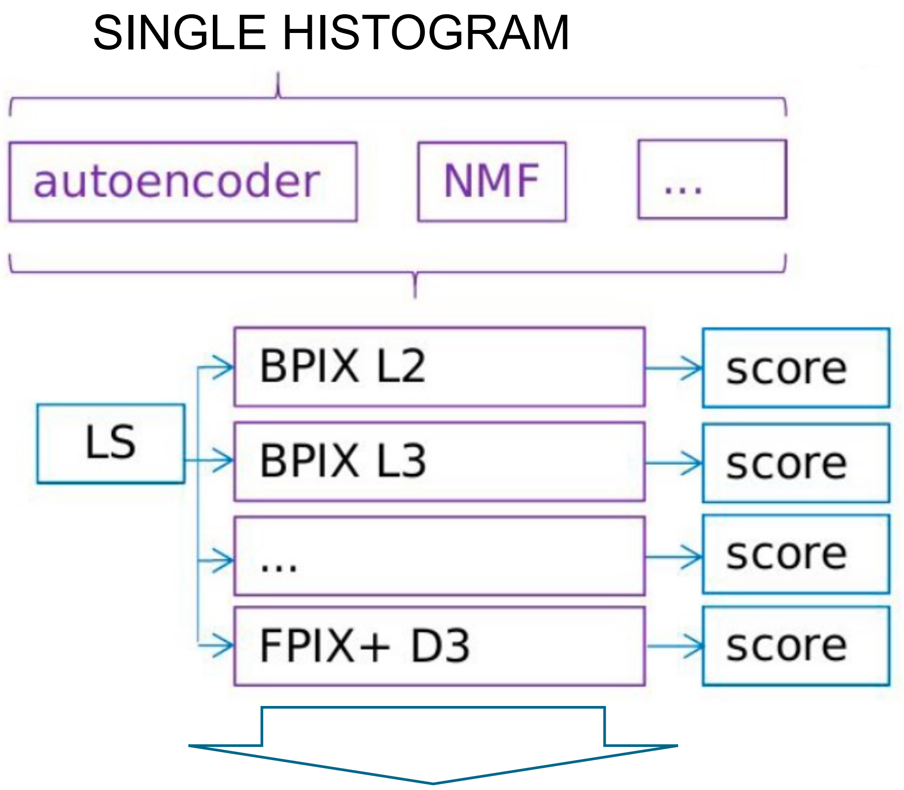
Moments: Comparison of the first and second order moments of a histogram to the distribution of those moments in the training set.

Landau fit: Mean-squared-error (MSE) between a histogram and a fitted Landau distribution.

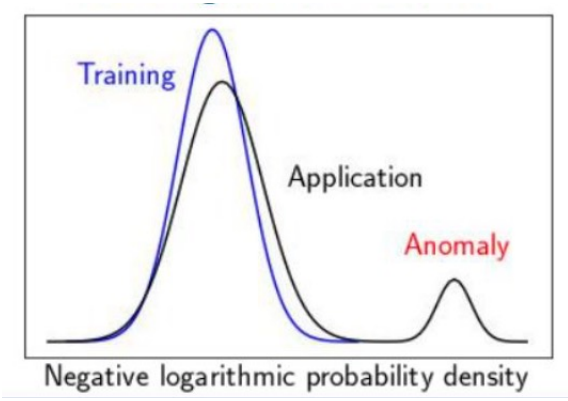
Templates: minimum MSE between a histogram and each of a set of well-chosen reference histograms.

NMF: MSE between a histogram and its nonnegative matrix factorization (NMF) reconstruction as an optimized linear combination of basis components extracted from the training set.

Autoencoder: MSE between a histogram and its autoencoder reconstruction.

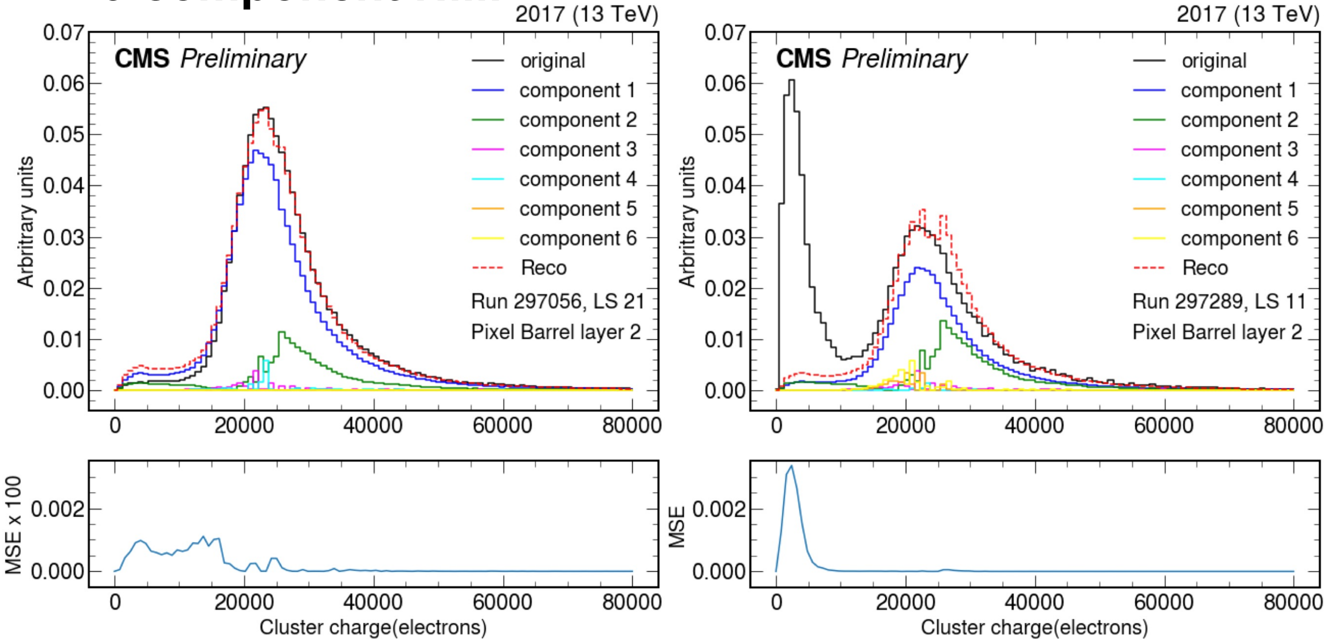


Fit multi-dimensional probability density to scores of the training set

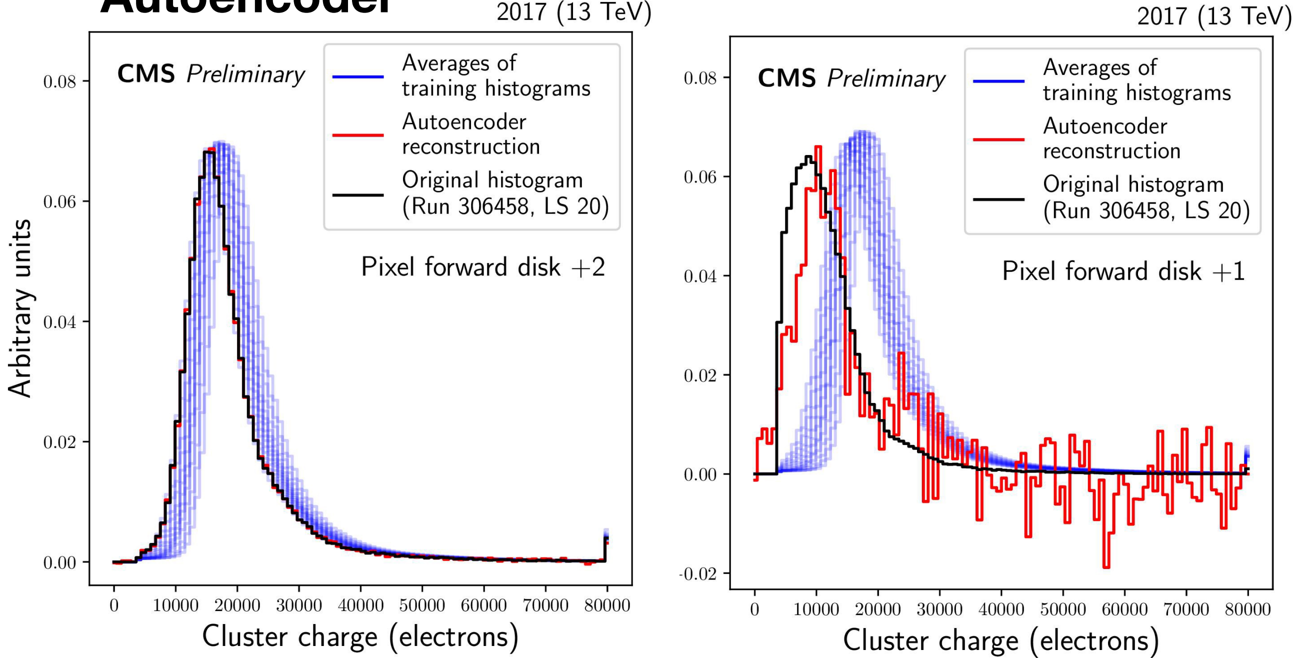


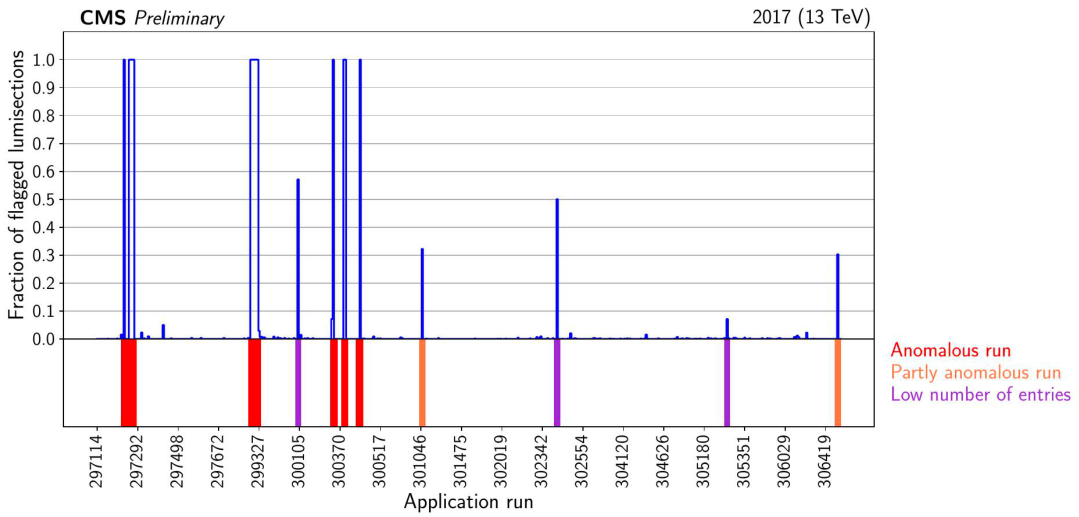
Pixel tracker: results

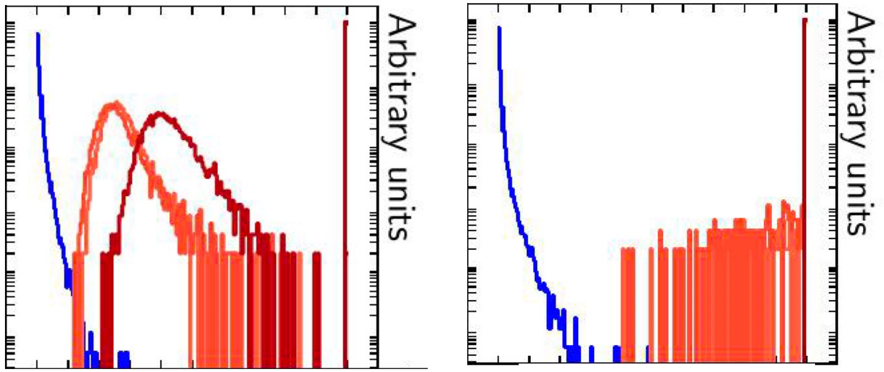
6-component NMF



Autoencoder

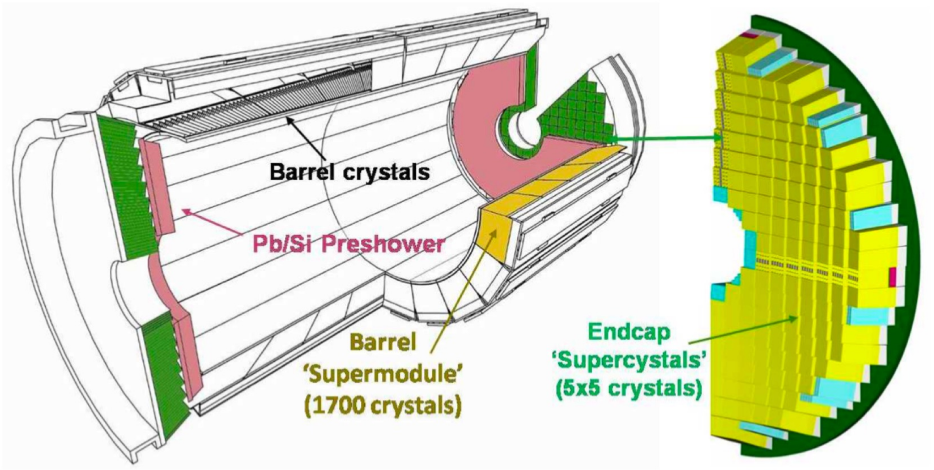






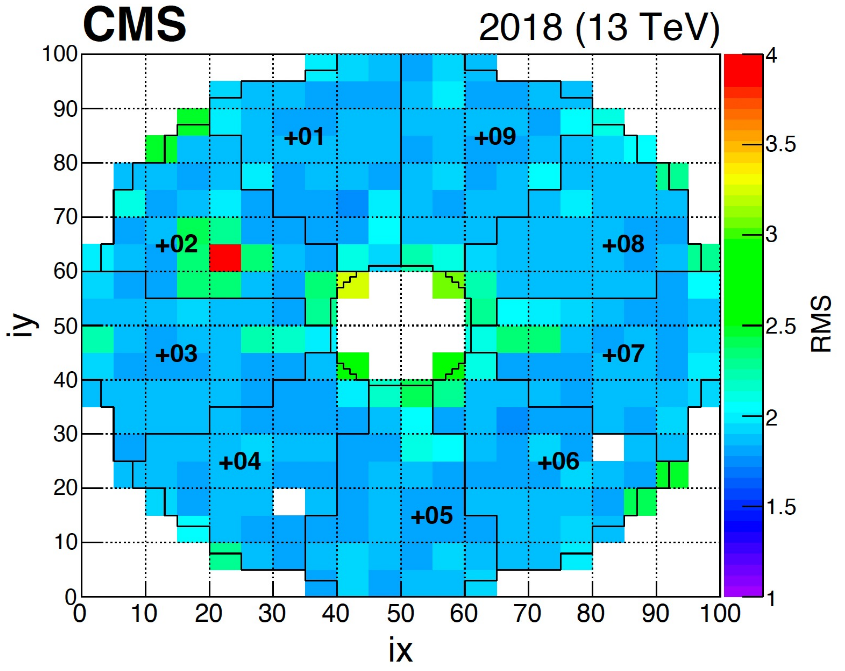
6-component NMF **Autoencoder**

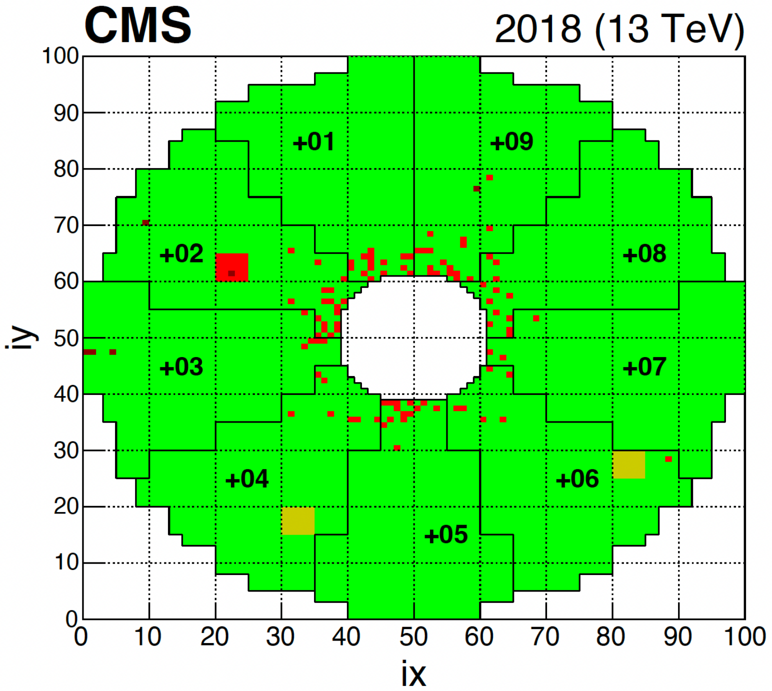
Specific applications: ECAL online DQM



Online ECAL DQM: real-time snapshot of a subset of the raw data by populating a set of histograms

- Histograms updated every LS over the run
- Continuously monitored by shifter
- **ME:** Occupancies (left) and quality plots (right) obtained by applying predefined thresholds

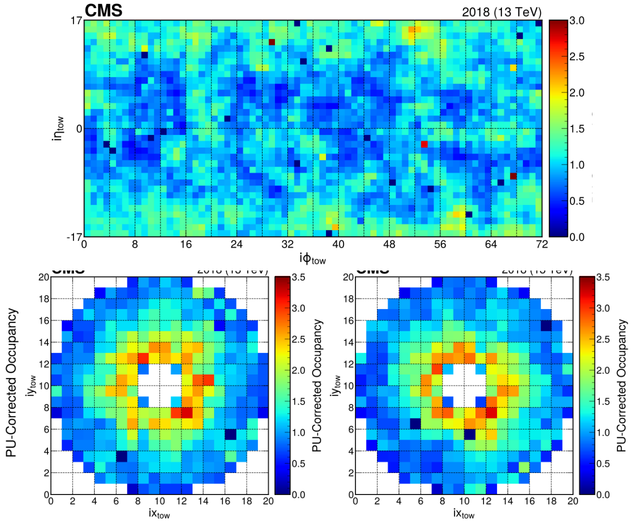


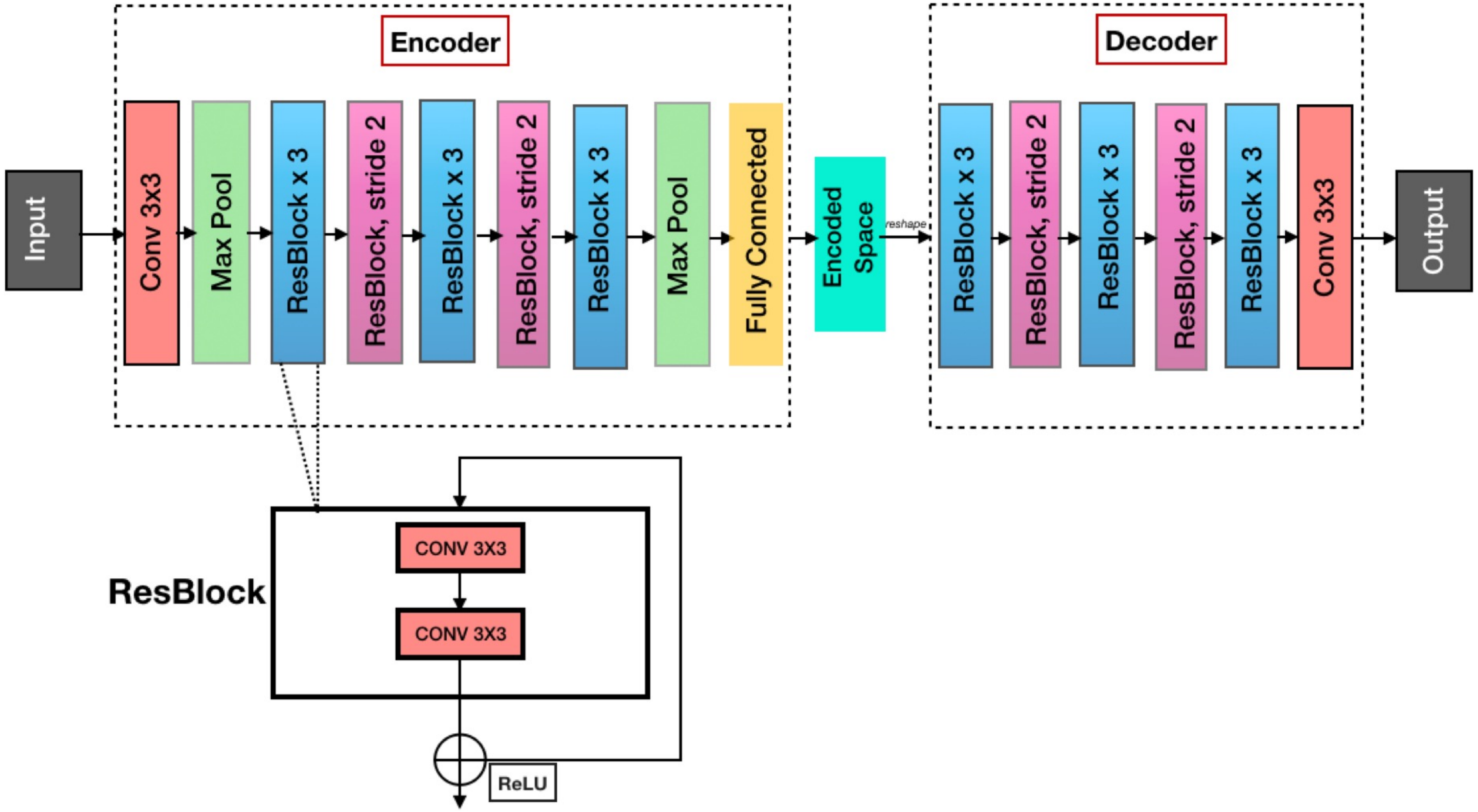


- green: “good”
- red: “bad”
- brown: “known problems”
- yellow: “no data”

ECAL online DQM: implementation

Architecture: Residual Neural Network (ResNet) CNN, separate NN models for detector regions
Input: Occupancy histograms as 2D images for each LS
Semi-supervised approach: training on certified good data

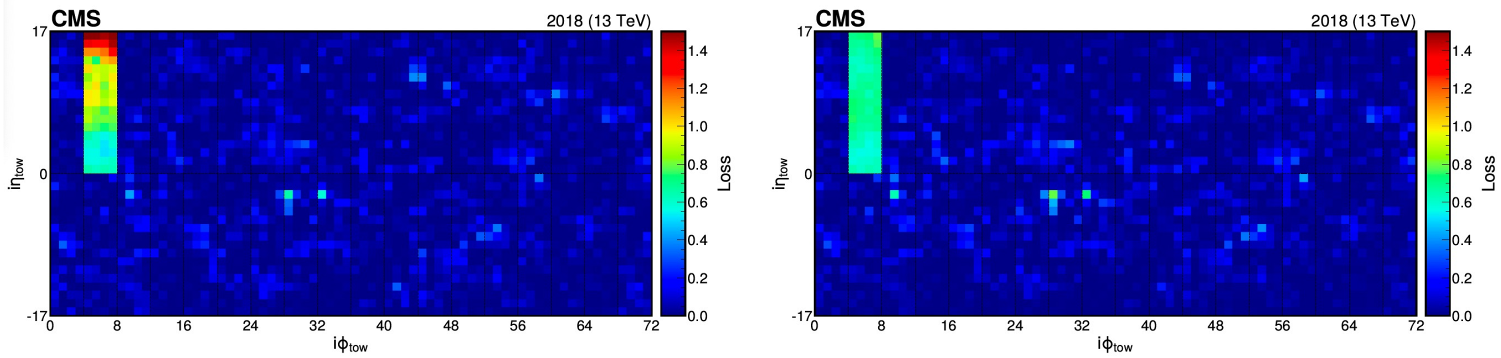




AE reconstructed images

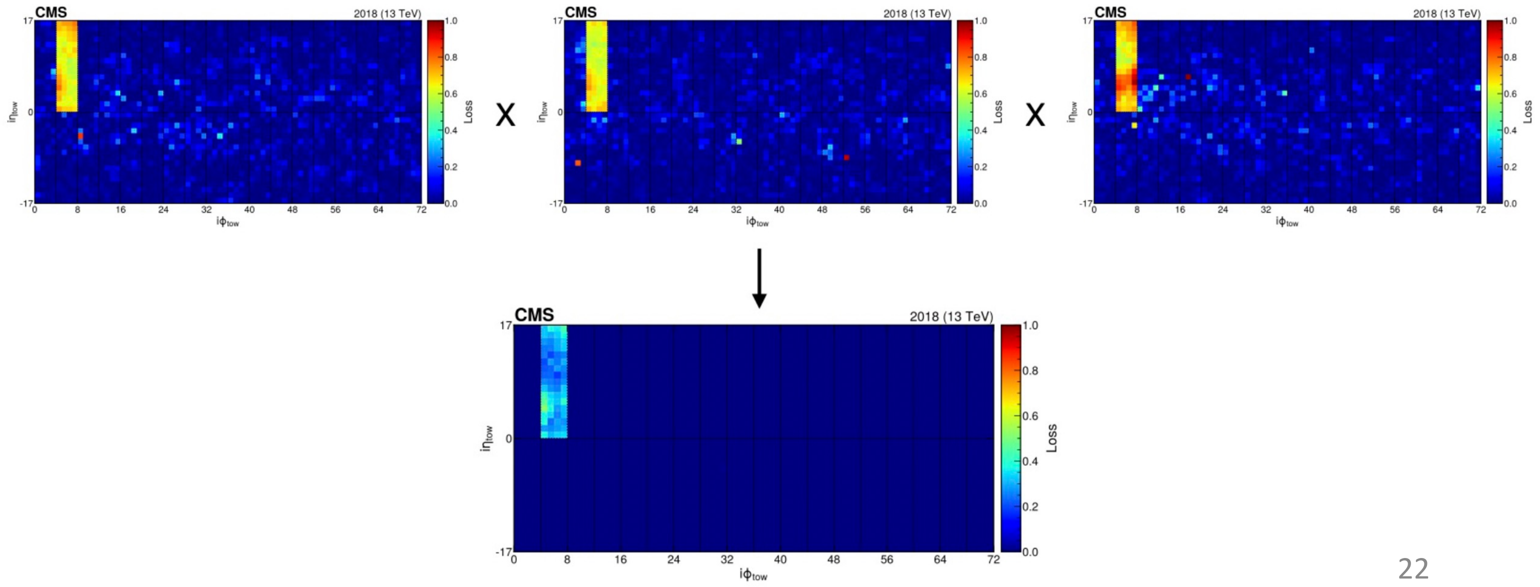
ECAL online DQM: spatial and time response correction

Spatial response correction: at a hadron collider, the higher the rapidity, the higher the number of produced particles → AE trained over the full rapidity range will return non-uniform loss vs rapidity → loss is corrected for the expected average occupancy



Loss map before (left) and after (right) spatial correction

Time response correction: anomalies will likely persist in consequent LSs, while random false positives will not → multiply loss maps from 3 consequent LSs

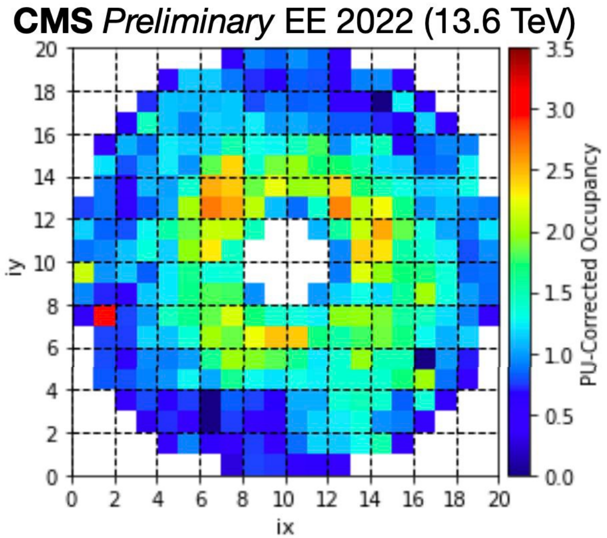


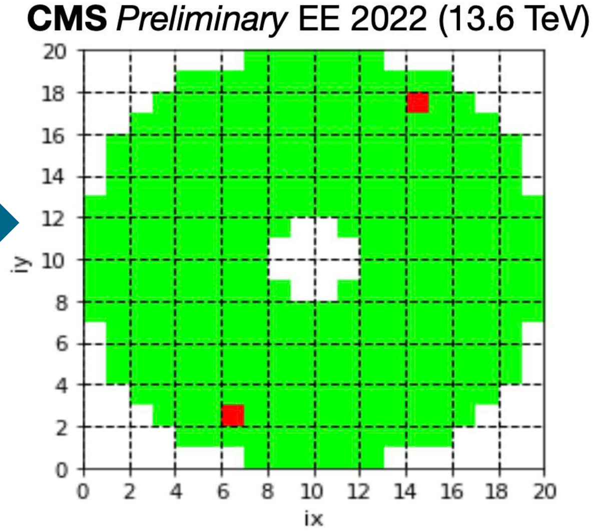
ECAL online DQM: results

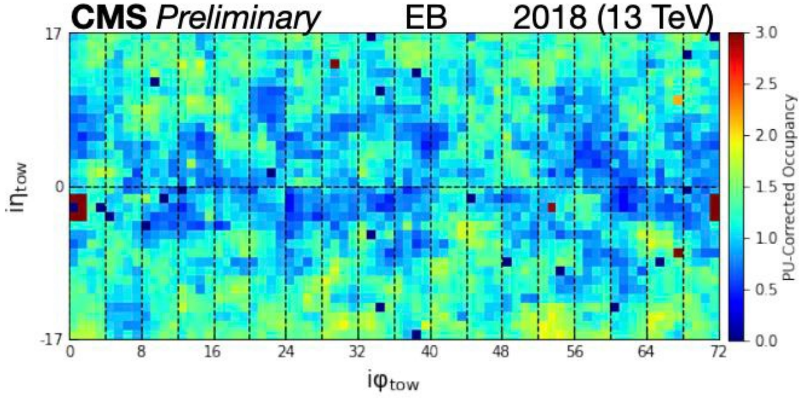
Validation on fake anomalies: different failure scenarios, loss thresholds set to reject 99% of anomalous data. False Discovery Rate (FDR) used as a metric

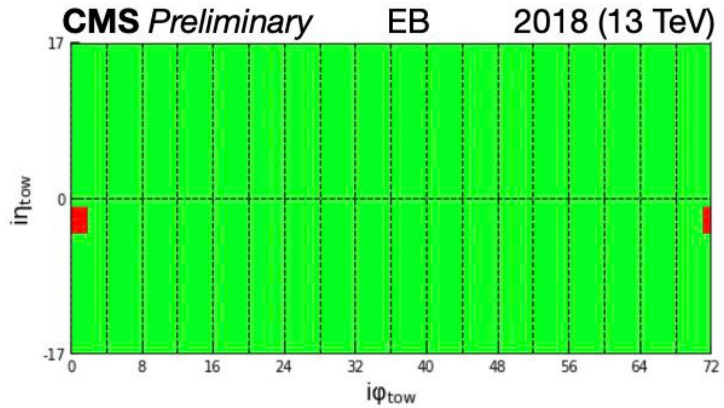
	FDR for 99% anomaly detection					
	Missing Sector		Zero Occupancy Tower		Hot Tower	
	EE+	EE−	EE+	EE−	EE+	EE−
AE no correction	29%	28%	86%	86%	$< 0.01\%$	$< 0.01\%$
AE after spatial correction	1.8%	2.2%	11%	14%	0.02%	0.04%
AE after spatial and time corrections	0.06%	0.18%	1.4%	4.4%	$< 0.01\%$	$< 0.01\%$

Testing on real unlabeled data using loss thresholds from validation: catches anomalies well with various shapes and sizes, also on recent data without retraining!

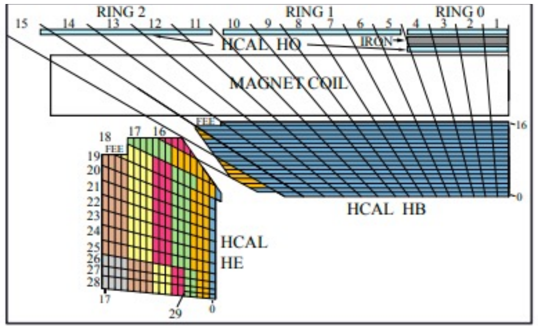


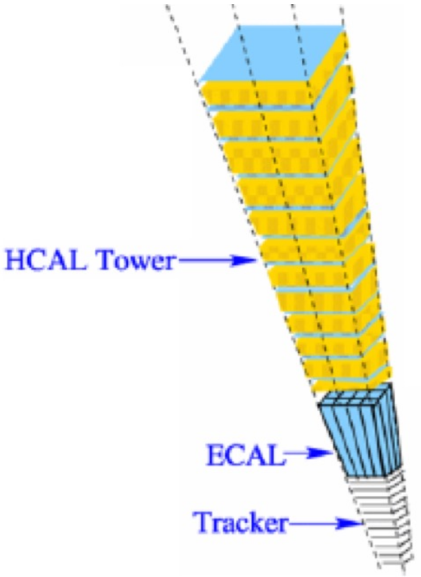






Specific applications: HCAL online DQM



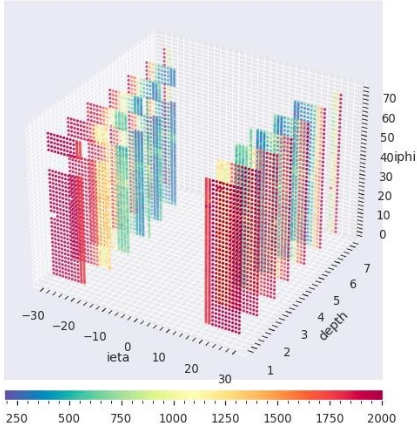


HCAL online DQM:

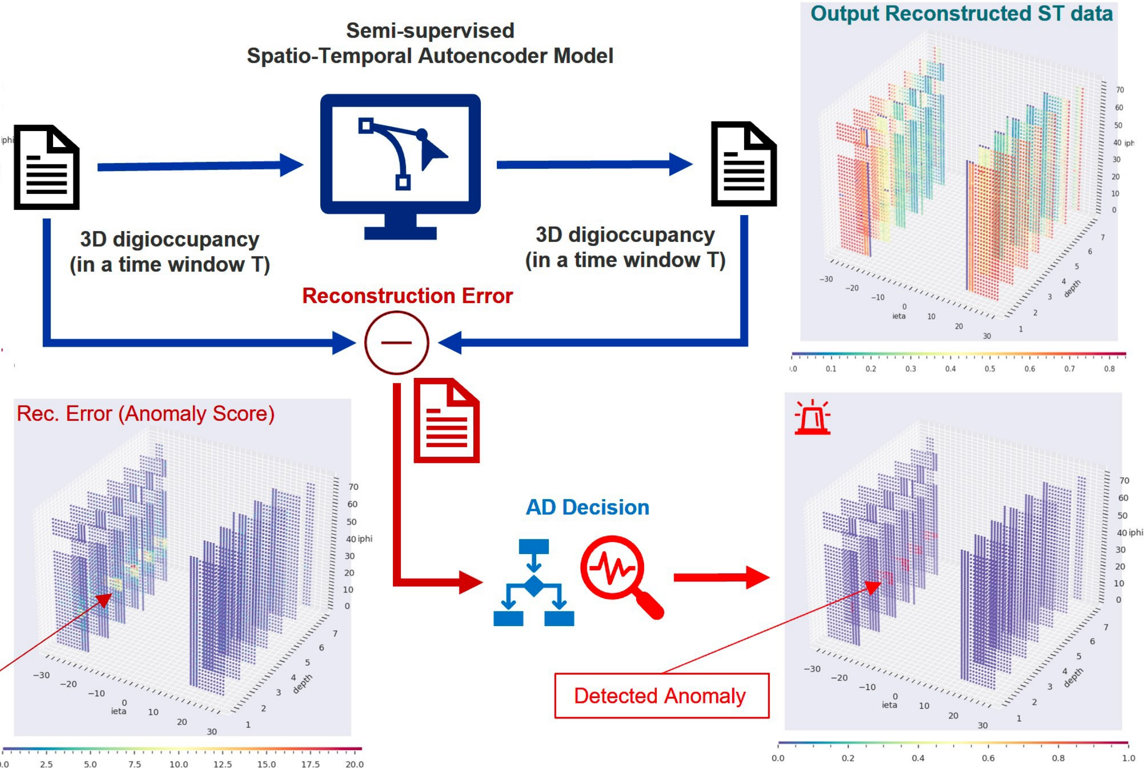
- a set of potential failures (communication issues, miscalibrations, hardware issues etc) can be spotted using **3-D occupancy maps**

- HCAL channels sharing services → **spatial correlations**

- **Temporal correlations** between failures (persistent issues over time, degrading channels) can be exploited for anomaly detection



→ semi-supervised spatio-temporal autoencoder model

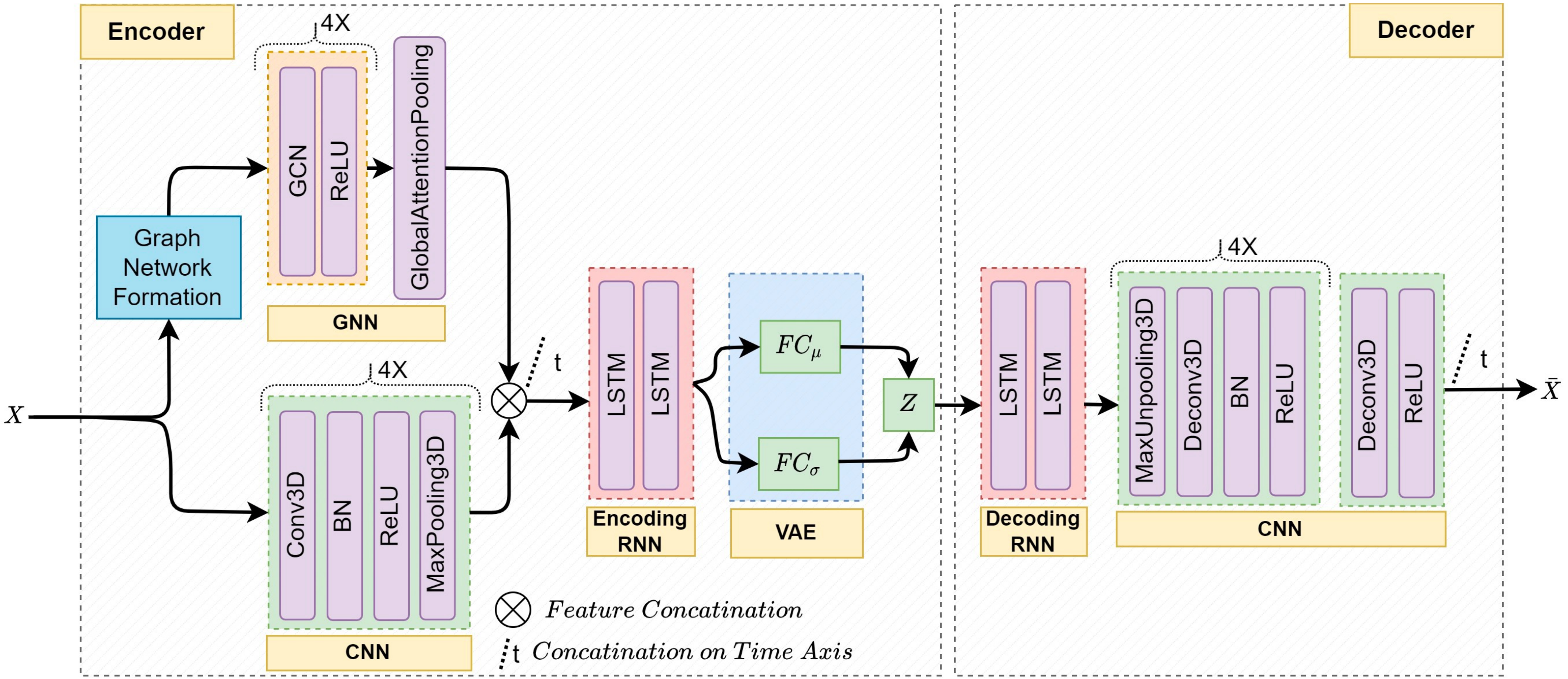


HCAL online DQM: model

Architecture: Graph Based ST AD model (GraphSTAD)

- CNN and GNN to capture Euclidean and non-Euclidean **spatial characteristics** of HCAL channels
- RNN captures the **temporal** behavior of the extracted features

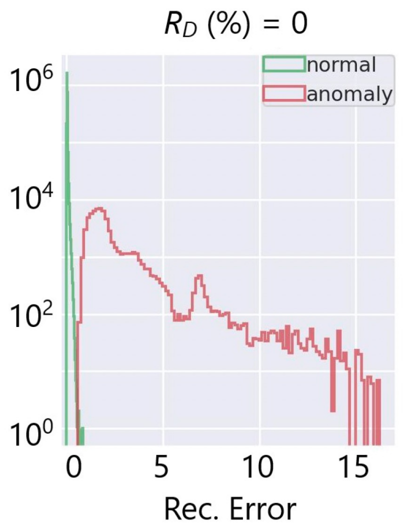
Training: 3D occupancy maps from certified good data



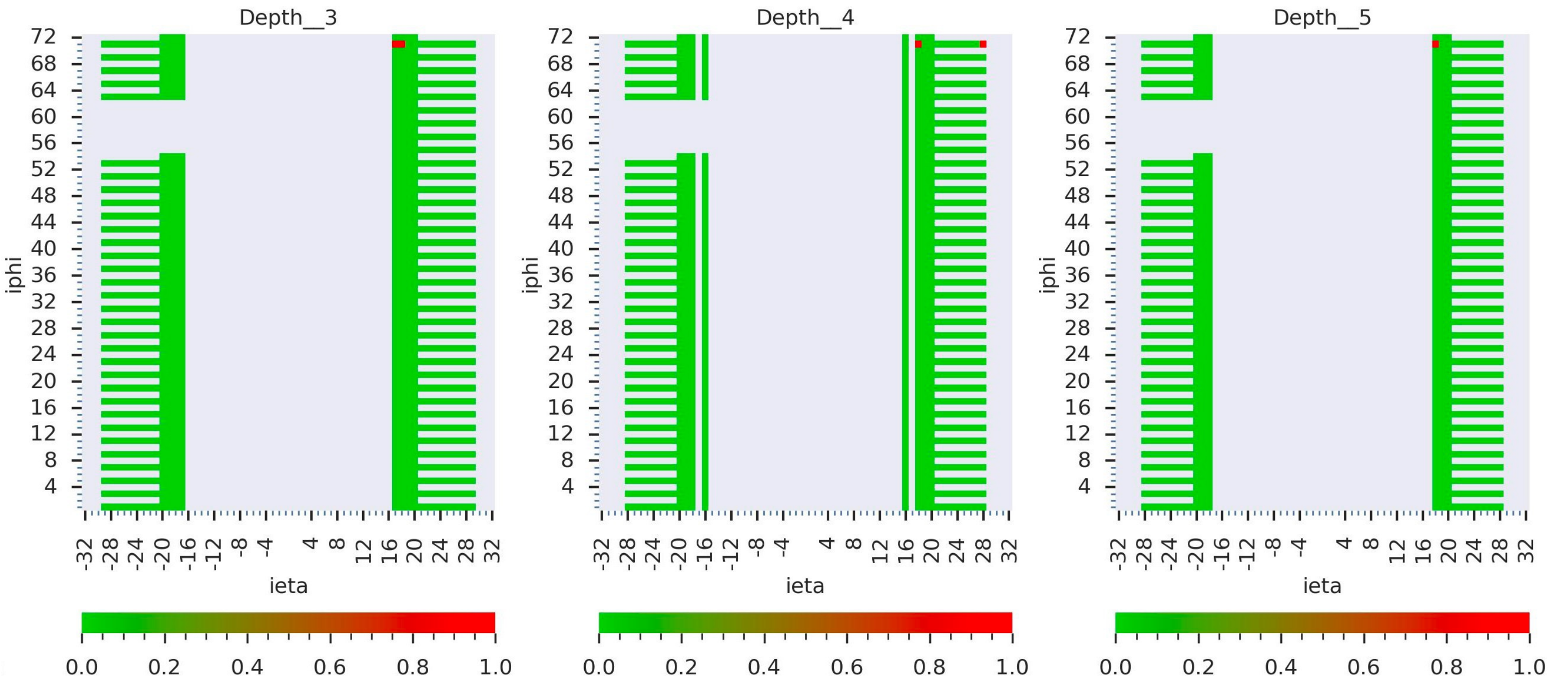
HCAL online DQM: results

Detection of degrading channels: simulated time-persistent degrading channel efficiently detected with low FPR

Health Rate	FPR (90%)	FPR (95%)	FPR (99%)
80%	1.636×10^{-3}	3.614×10^{-3}	2.988×10^{-2}
60%	1.329×10^{-4}	3.834×10^{-4}	1.550×10^{-3}
40%	8.405×10^{-6}	2.764×10^{-5}	2.242×10^{-4}
20%	2.263×10^{-6}	5.173×10^{-6}	2.505×10^{-5}
0%	9.699×10^{-7}	1.778×10^{-6}	6.142×10^{-6}



Detection of real anomalies in data: anomaly flag map spotting faulty HCAL channels during data taking



Summary

- Data Quality Monitoring is a crucial task in a large HEP experiment such as CMS
- Traditional approach based on visual inspection of a set of histograms (monitoring elements) either real-time (online monitoring) or offline for data certification
- Many advantages of ML approach in DQM operations: reduce human error, allow for finer time granularity monitoring (per lumisection), detect subtle anomalies
- Several developments toward an automated DQM for online or offline data quality monitoring within the different CMS subsystems show promising results
- Now working on common frameworks for a comprehensive anomaly detection system

References

- CMS Collaboration, “An AutoEncoder-based Anomaly Detection tool with a per-LS granularity”, CERN-CMS-DP-2023-010 (2023) https://cds.cern.ch/record/2854697
- CMS Collaboration, “Machine Learning Techniques for JetMET Data Certification of the CMS Detector”, CERN-CMS-DP-2023-032 (2923) https://cds.cern.ch/record/2860924
- CMS Collaboration, “Tracker DQM Machine Learning studies for data certification”, CERN-CMS-DP-2021-034 (2021) https://cds.cern.ch/record/2799472
- CMS ECAL Collaboration, “Autoencoder-based Anomaly Detection System for Online Data Quality Monitoring of the CMS Electromagnetic Calorimeter” (2023) https://doi.org/10.48550/arXiv.2309.10157
- CMS HCAL Collaboration, “Spatio-Temporal Anomaly Detection with Graph Networks for Data Quality Monitoring of the Hadron Calorimeter” (2023) https://doi.org/10.48550/arXiv.2311.04190