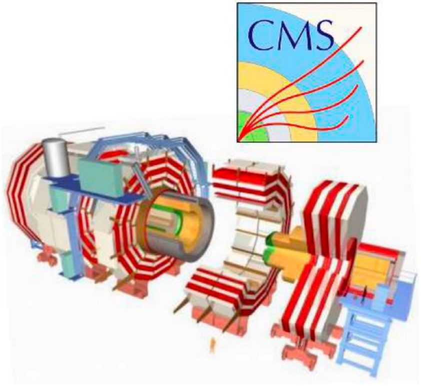
Tracker Data Quality certification in CMS with new Machine Learning tools



Atanu Pathak

On behalf of the CMS Collaboration

Purdue University Northwest

42nd International Conference on High Energy Physics (ICHEP 2024)
18 - 24 July, 2024
Prague, Czech Republic

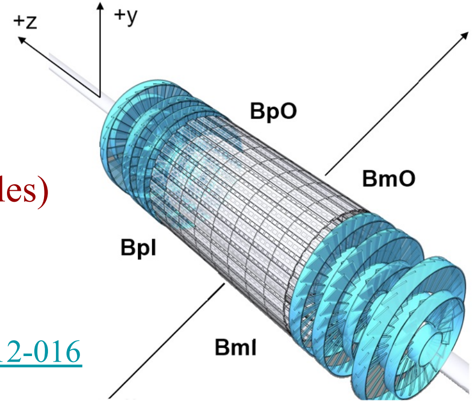
CMS Tracker: Pixel

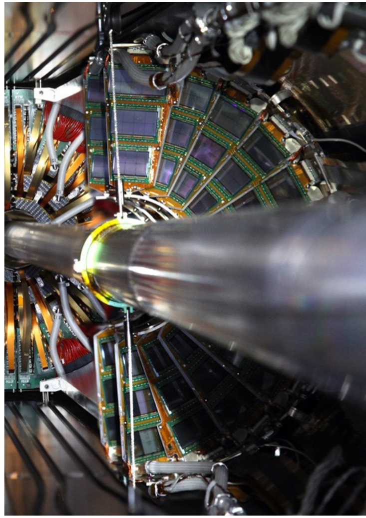
CMS Pixel detector (Phase 1 upgrade, installed in 2017): 4 hit coverage up to |η| < 3

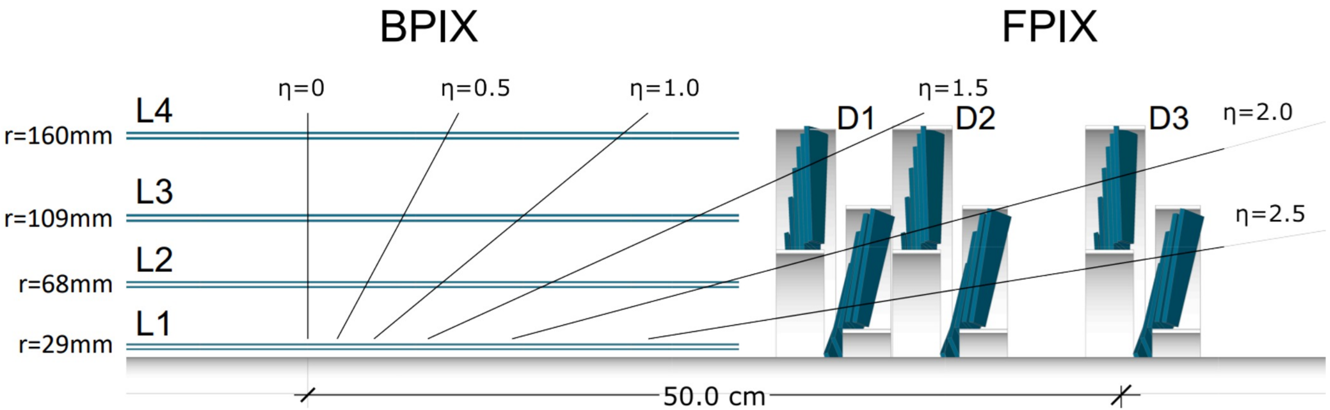
- **Barrel (BPIX):** 4 layers (1184 modules)
  - 4 Half Shells: BmI, BmO, BpI, BpO
- **Forward (FPIX):** 3 disks with 2 rings on each end (672 modules)
  - Half Cylinders: HCmI_1, HCmI_2, …

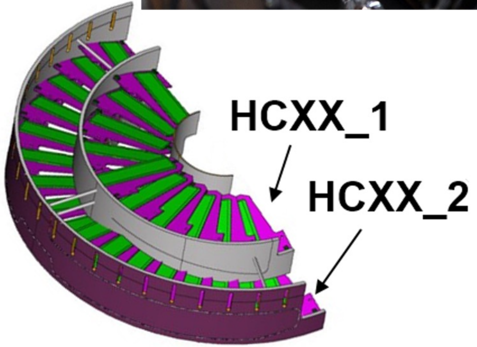
*p/m = plus/minus
*I/O= Inner/Outer w.r.t LHC ring center

CERN-LHCC-2012-016





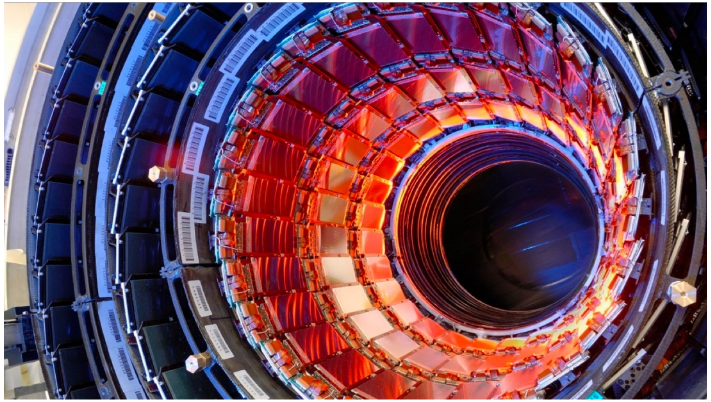


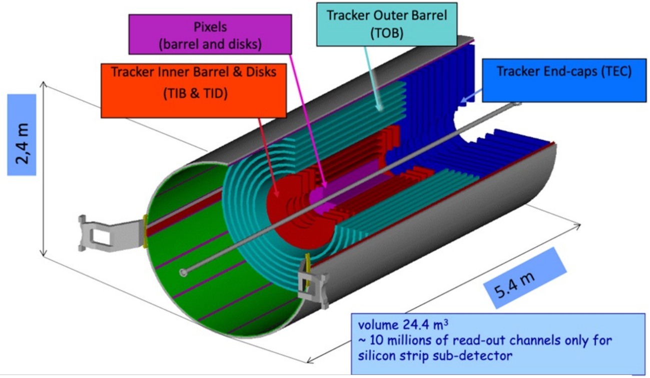


XX can be mO/I or pO/I

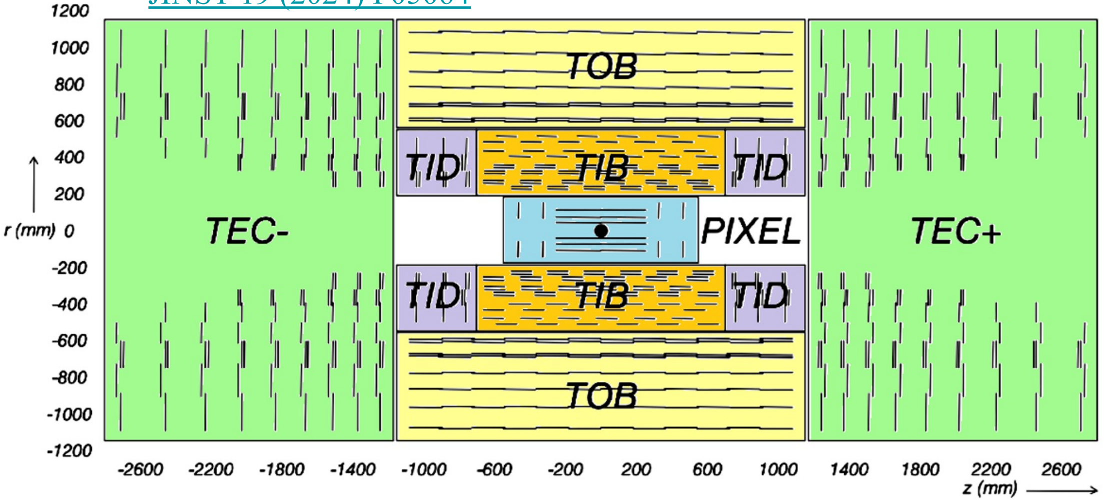
CMS Tracker: Strip

- CMS Strip detector (since 2007): 9.3 million strips (15148 modules)
  - 5 m long, 2.5 m diameter
- 10 layers in the Barrel region:
  - 4 inner barrel layers (TIB)
  - 6 outer barrel layers (TOB).
- 12+12 layers in the Endcap region:
  - 3 inner disks (TID plus/minus)
  - 9 end cap disks (TEC plus/minus).





JINST 19 (2024) P05064

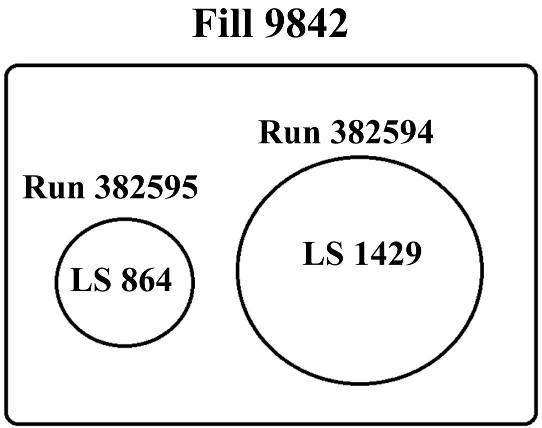


Why it is crucial to monitor Tracker Data

- CMS aims to collect as many collision events as possible to extract the best physics results
  - Any issue needs to be identified and solved in real time.
  - No Tracks mean No physics.
    - Essential to monitor Tracker conditions and performance **24x7.**

- Tracker data can be from **Cosmics** or **LHC collisions** (pp, Heavy ions).

- **LHC collisions** are organized as follows:
  - LHC fills
    - Fills have fixed beam configuration (bunches)
    - A fill may include several Runs
  - Runs
    - New run each time the data taking is stopped
    - Runs are composed of Lumi Sections
  - Lumi Sections (LS)
    - Roughly 23 seconds of datataking



- Data certification (DC) is done by evaluating each run (eventually excluding bad LS). CMS is storing ~7 kHz of data, so marking as bad 1 LS implies losing 23.31 x 7 kHz ~160k events/LS.

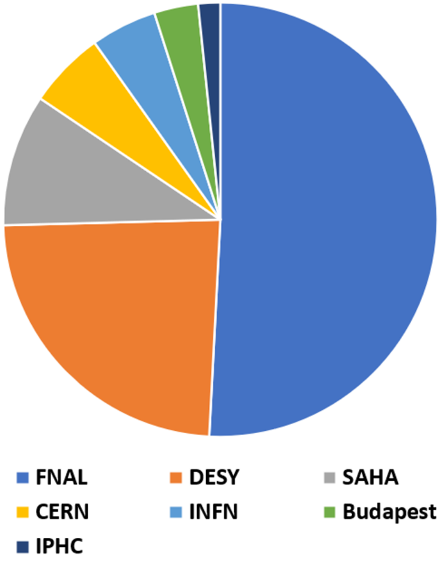
CMS Data Quality in numbers : 2023

Shifts were done from remote centers: spread across different parts of the worlds with different time zones

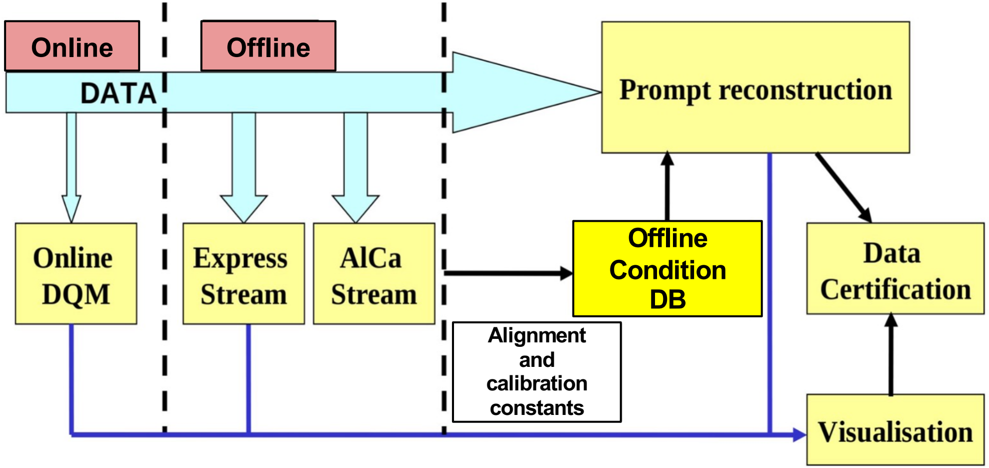
- In 2023, we certified about
  - 300 Proton-Proton collisions runs, 100 Heavy Ions collisions runs
  - 1000 Cosmics runs
  - Of these runs, 10% was marked Bad (mostly from non-stable collisions, and short runs with very low statistics)

- In order to ensure the data quality, several monitoring tools are used by experts
  - 1D and 2D distributions are monitored for each run separately:
    - 16000 Pixel plots
    - 2500 Strip plots
    - 7000 Tracking plots
  - Only a selection of the most important plots are checked (about 200) regularly.

- About 70 persons contributed to the data quality monitoring and data certification (DQM/DC) last year
  - It was total 34 weeks of operations, and we actively had
    - 46 shifters
    - 18 shift leaders
    - 6 experts



CMS DQM System and DC processes



- As soon as data are recorded, quality checks start
  - This is essential to guarantee data quality
  - Finding issues early allows quick reaction and solution

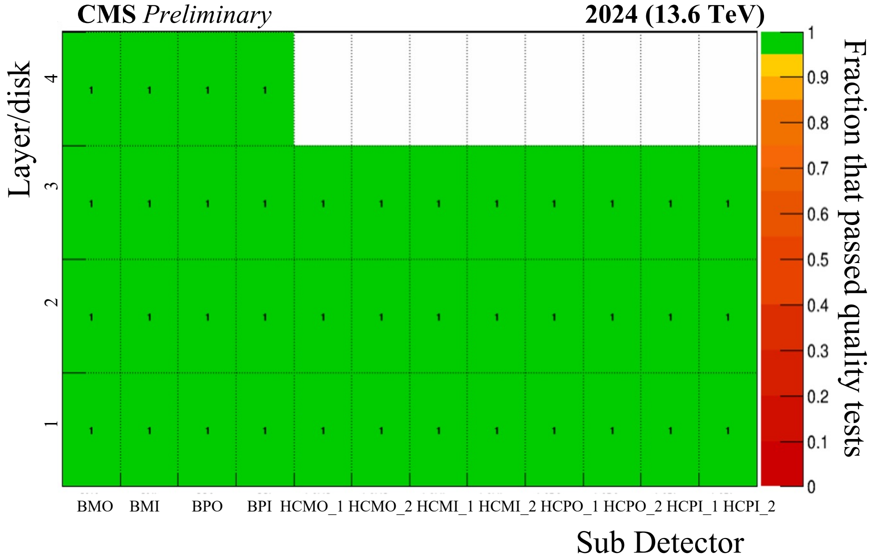
- **Online DQM**
  - A small fraction of the data are displayed "on-the-fly" for real-time monitoring purposes.

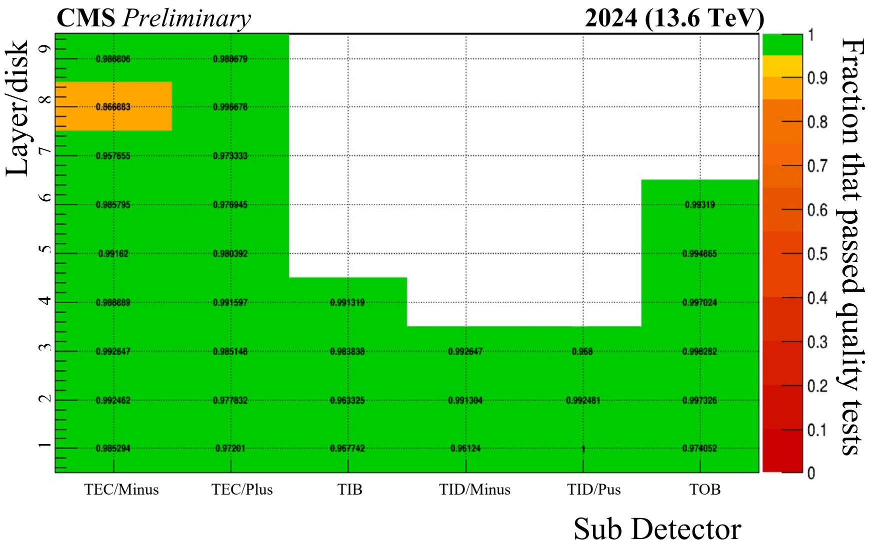
- **Offline DQM**
  - **Express stream**: Fraction of data, and it is fundamental to establish the "conditions" of the detector
  - **Prompt reconstruction**: Full offline processing of data, starts about 48h after data are collected, final data certification to ensure good quality for physics analysis.

CMS DQM tools : Summary Maps

- **Summary map** helps the experts in identifying any inefficient region of the detector during operations.
  - Fraction of modules that passed the Quality Tests for the different parts of the detector.

- Summary map of Pixel (left) and Strip (right) for run 380360 (May 4th 2024)
  - For this particular run, Strip TEC minus 8 box is yellow (86.7%):
    - Few modules were not being read due to a transient issue from two front-end driver modules.

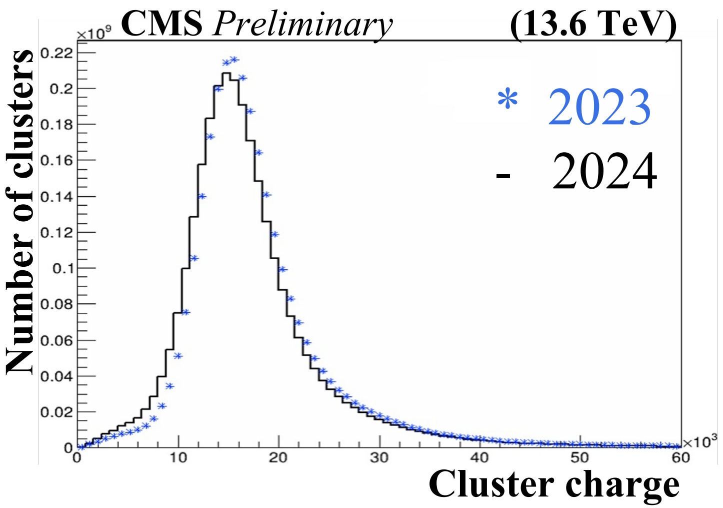


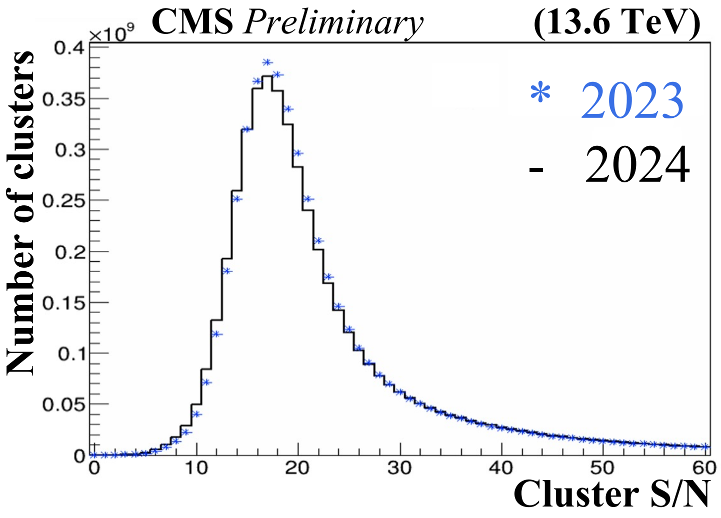


2024 vs 2023 Tracker Data Quality

- **1D distributions** helps to identify detector issue, by looking for change in shape
  - Distribution is typically compared to good reference run

- Comparison of 2023 (blue) vs 2024 (black): despite the harsher condition and the integrated radiation damage, **detector performance is very stable!**
  - Pixel (left): on-track cluster charge distribution, 4 layers of BPIX
  - Strip (right): on-track cluster Signal-to-Noise (S/N) ratio distribution, 4 layers of TIB

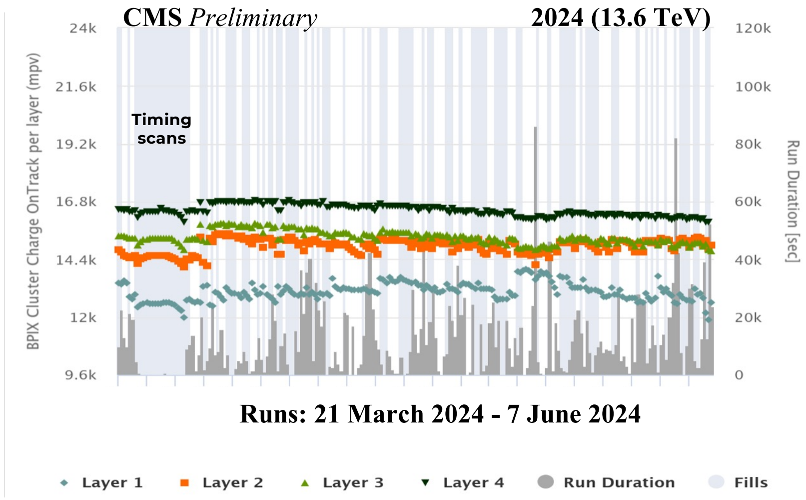


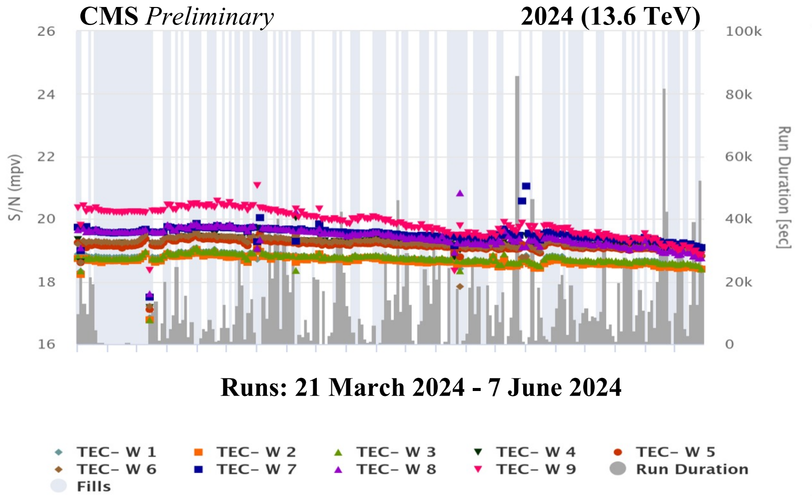


CMS DQM tools : HDQM

- **Historic DQM (HDQM)** helps to monitor the tracker performance over an extended time scale.
  - Outliers (problematic runs or simply low statistics)
  - Worsening trends (i.e. radiation damage)

- HDQM trend for pp collisions runs from 21 March to 7 June 2024
  - Pixel (left): Most Probable Value (MPV) of the on-track cluster charge distribution for each of the 4 BPIX layers
  - Strip (right): MPV of the on-track cluster signal-to-noise distribution for each of the 9 TEC minus wheels



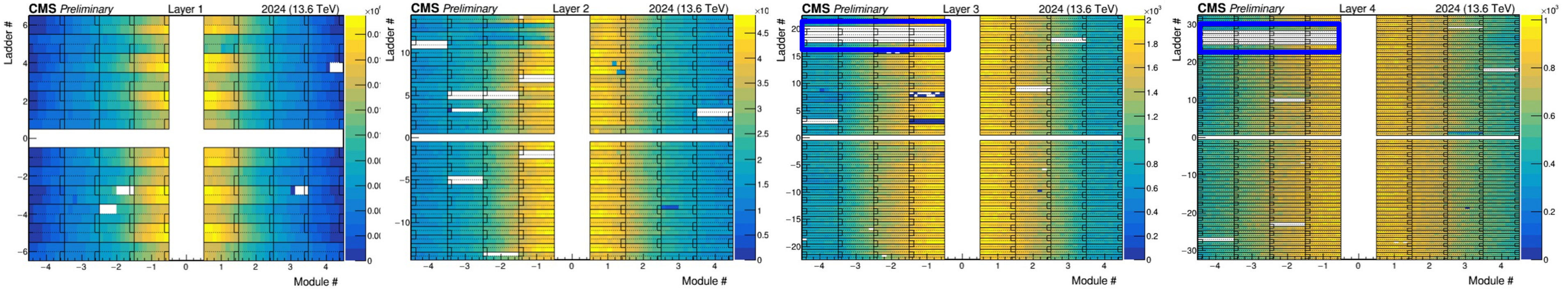


2024 Data Quality : Pixel Bad Components

- **2D distributions** helps to identify region of the detector with issue:
  - Single module: not significative impact on tracking performance
  - Large portion of the detector: severe impact on the tracking performance

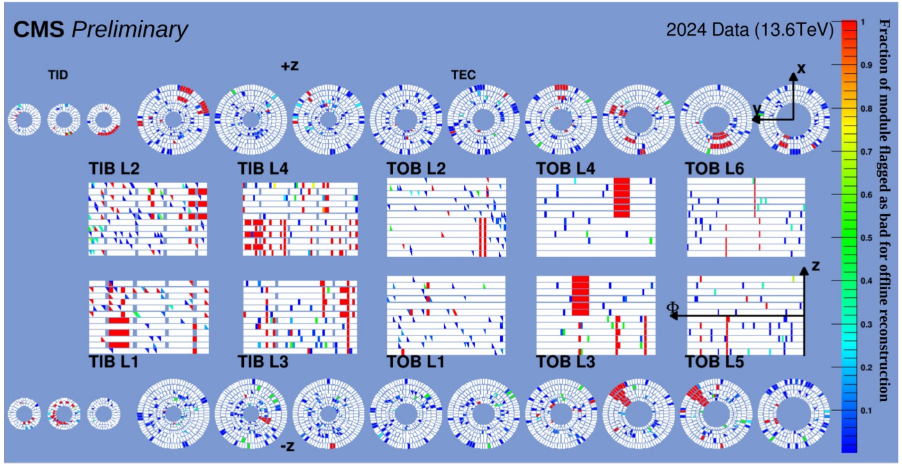
- 2D distributions of the on-track cluster occupancy: 4 BPIX layers (run 381053, May 22nd, 2024).

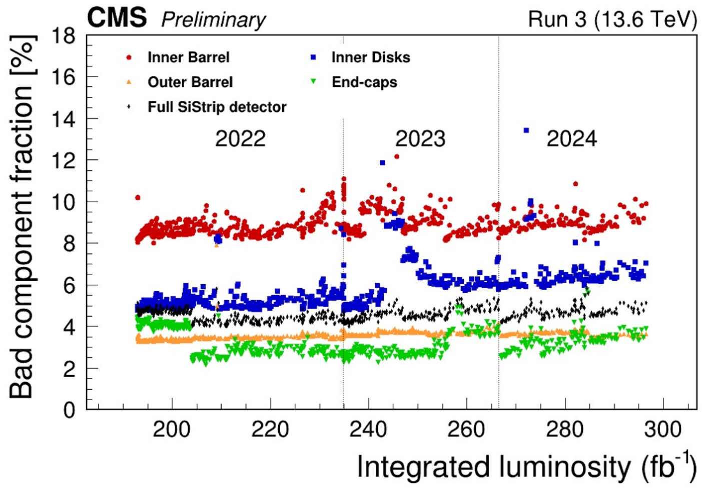
- The fraction of “bad components” is carefully monitored:
  - During 2024, the fraction of the non-functional readout chips (ROCs) in the Pixel detector is 3.2%
    - 3.9% BPIX
    - 2.2% FPIX
  - From June 2023 for BPIX L3 and L4, Quartz controlled PLL circuit does not lock to LHC clock.



2024 Data Quality : Strips Bad Components

- **Tracker maps** helps to monitor of single-module and multi-module performance (geometrical structure).
- Strip Tracker map (left): bad components in a run from June 03, 2024.
  - RED : completely masked (i.e not used) in the Prompt reconstruction.
  - Blue (faulty strips) and Green (optical fibres) : excluded from Prompt reconstruction
- Bad components trend (right) as a function of integrated delivered luminosity during Run 3
  - Increase in trend due to some issues in data-taking/powering that can either be promptly recovered/require more significant interventions.

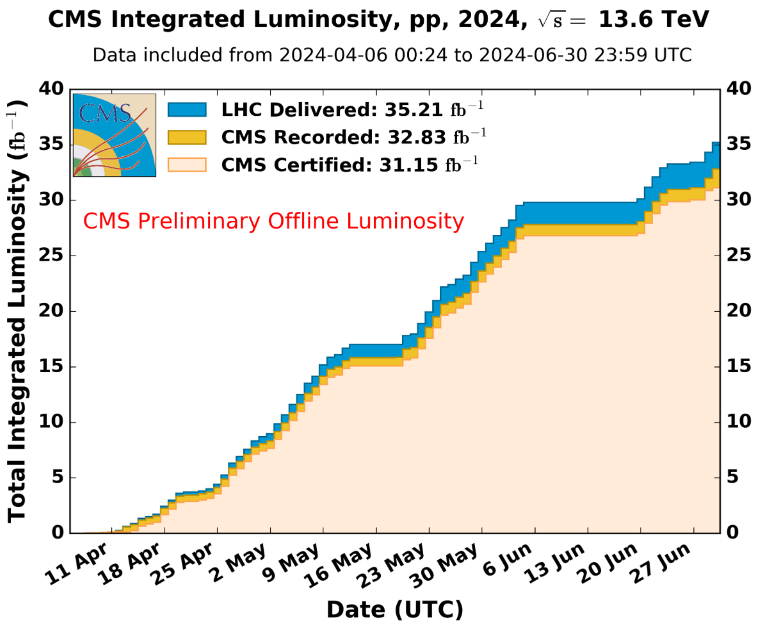




2024 CMS Data certification

- Proton-Proton (pp) collision runs at 13.6 TeV center-of-mass energy from 6 April to 30 June 2024
  - LHC delivered: 35.21 fb-1
  - CMS recorded: 32.83 fb-1
  - CMS certified: 31.15 fb-1
- Data taking efficiency
  - Detector issues that prevent taking data
  - Deadtime
  - Data taking efficiency: 93.2%
- Data certification efficiency
  - Recorded data with bad quality
  - Issue in detector components that degrades performance
  - Data certification efficiency: 97.3%
- Significant issue in Pixel during 2024
  - Technical issue in the Pixel CO2 cooling system
  - Switched off the Pixel detector for 2 days
  - About 0.4 fb^{-1} data recorded (but Bad)

https://twiki.cern.ch/twiki/bin/view/CMSPublic/DataQuality#Run_3_Data_Quality_Information



Machine Learning (ML) for DQM/DC

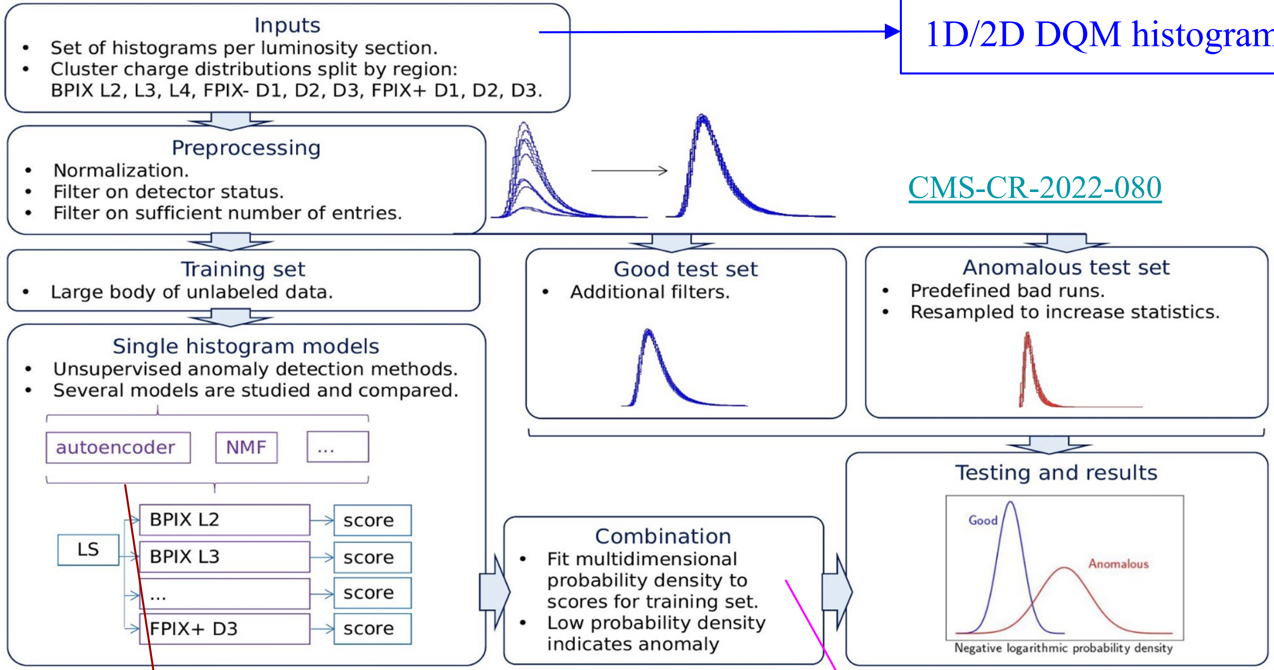
The current approach for DQM/DC at CMS

- Challenges to overcome:
  - **Personpower**
  - **Human error and Human driven decision process**
  - **Time granularity**
  - **Changing beam and detector(s) conditions**
  - **Anomalies may be unexpected**

- We are exploring the use of ML for anomaly detection, automating part of the process that includes
  - Tools to facilitate standardize certification tasks
  - Automate the evaluation of DQM histograms
  - Set proper Alarms/flags for threshold to monitor anomaly
  - Provide outputs robust against changing conditions and low statistics
  - Enable scaling to larger number of histograms

ML based DQM/DC: Case Study in Pixel

1D/2D DQM histograms bin contents, run number/LS number



CMS-CR-2022-080

Strategies for anomaly detection:

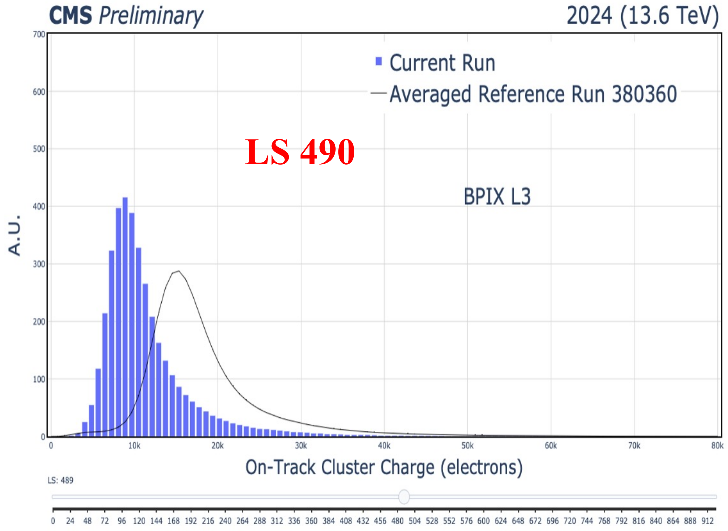
- “Learn” how to reconstruct an histogram (or set of them)
- “Reconstruct” the input histogram (inference)
- Use Mean Squared Error (MSE) to measure deviation (anomaly) between the reconstructed and the input histogram
- Flag anomalous LSs

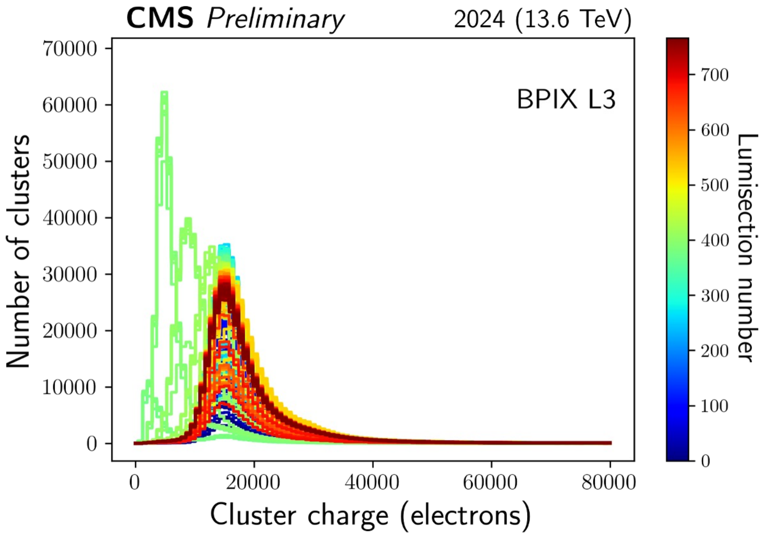
Training datasets: unsupervised approach

MSE : ML performance benchmark.

Example of anomaly: Pixel Cluster Charge

- During 2024, there were no runs with anomalous 1D shapes for On-track Cluster charge distributions
- Only anomalies were due to high voltage (HV) bias scans for a run 378981
  - Develop tools to check 1D distributions (vs reference run) per LS (html) (left)
- Develop ML models to check shape → find anomalous LS:
  - high voltage bias → charge collection efficiency reduced → shifted towards lower peak positions. (right)



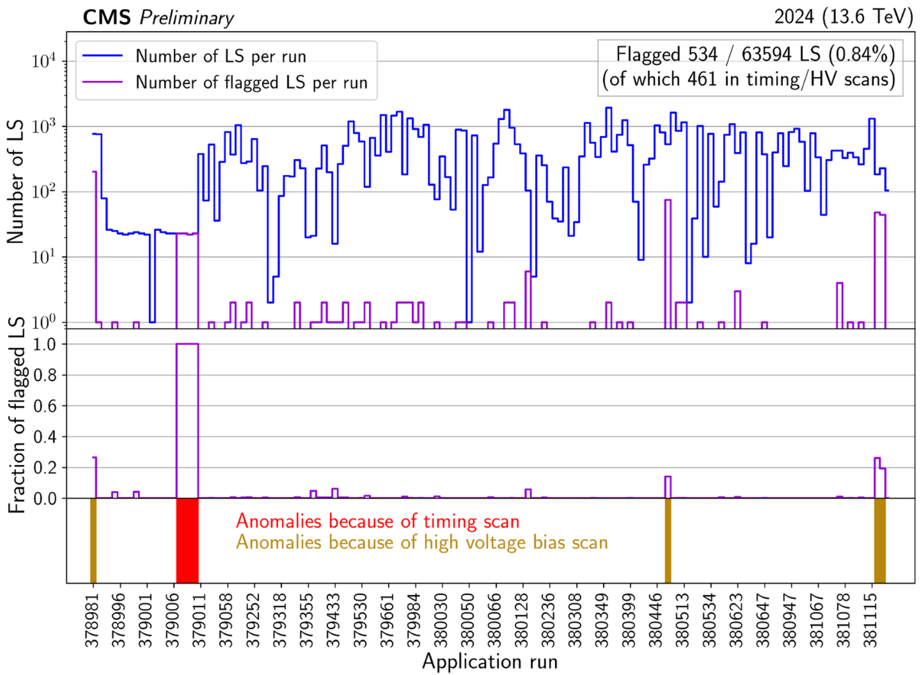


Anomaly detection (1D) with AutoEncoder ML model

Performance check of the anomaly detection method with autoencoders.

CERN-CMS-DP-2021-034

- Autoencoders are trained on the cluster charge distributions for all available LS in the ongoing 2024 data taking.



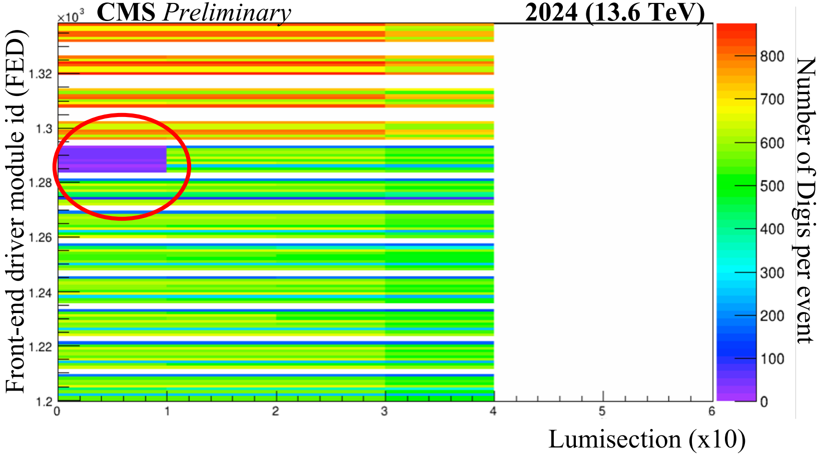
- The reconstruction quality is quantified by the MSE between a monitoring element and its autoencoder reconstruction.
- Anomalous LSs: High MSE value.
- A threshold is designed that maximizes the flagging of known anomalies while minimizing the false alarm rate.

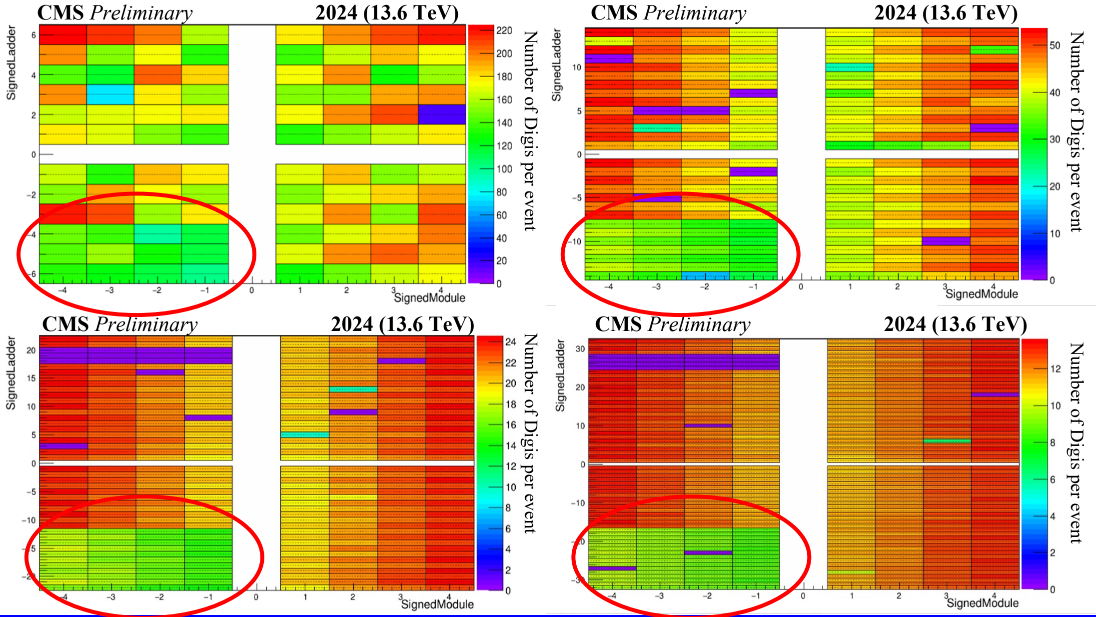
Upper panel: total number of LS in each run, as well as the number of LS flagged as anomalous by the autoencoder.
Middle panel: fraction of flagged LS in each run
Lower panel: known anomalies and their origin.

ML model is good at identifying the anomalies

Example of anomaly: 2D Digis maps for Pixel Barrel Layers

- For a particular run 380238, some FEDs of the Pixel detector were turned off due to LV trips up to 10 LS.
- 2D maps (right) show the average number of Digis (Hits) for each module in the 4 BPIX layers:
  - There is a large region with lower occupancy for all four layers, but from this it is very hard to identify when modules get fully recovered.





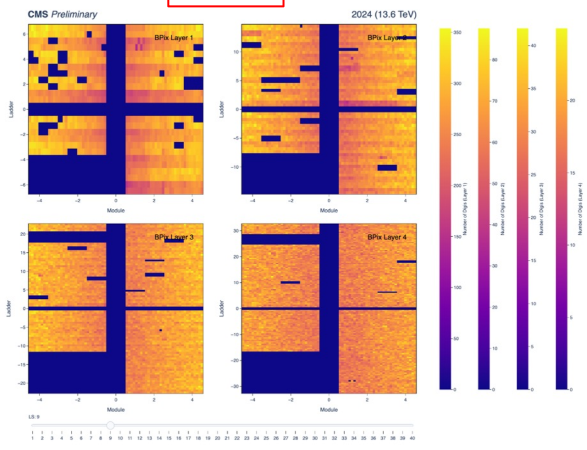
CERN-CMS-NOTE-2020-005

Example of anomaly: 2D Digis maps for Pixel Barrel Layers per LS

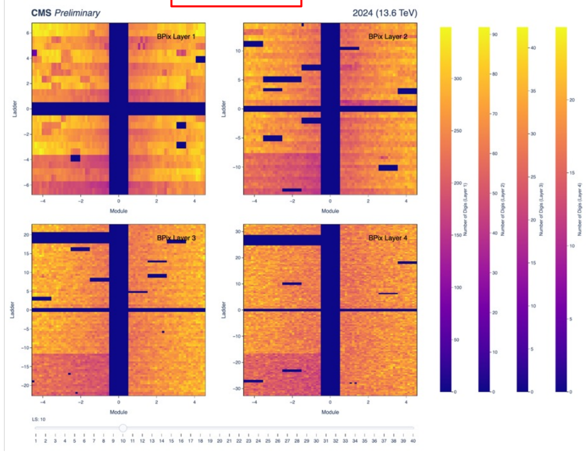
- Tools were developed to inspect 2D distributions in each LS: will allow to recover/exclude some good/bad LS

- 2D maps of the average number of Digis (Hits) for LS 9, 10 and 11, run 380238.
  - Each bin represents a Read Out Chip (ROC, 16 ROCs per each module).
  - A large portion turned off for all 4 BPIX layers up to LS 9. Then area is partly recovered in LS 10 and fully recovered after LS 11.

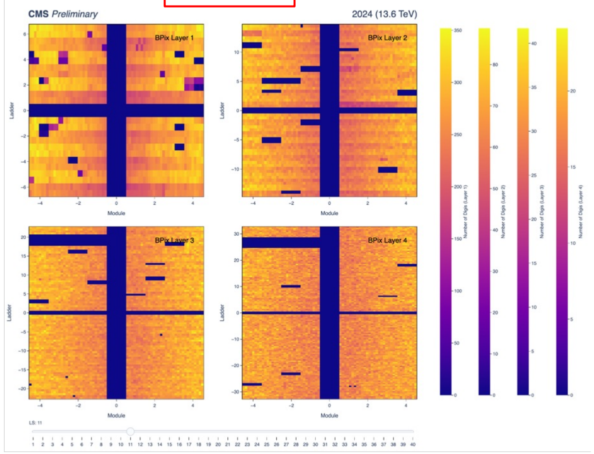
LS 9



LS 10



LS 11



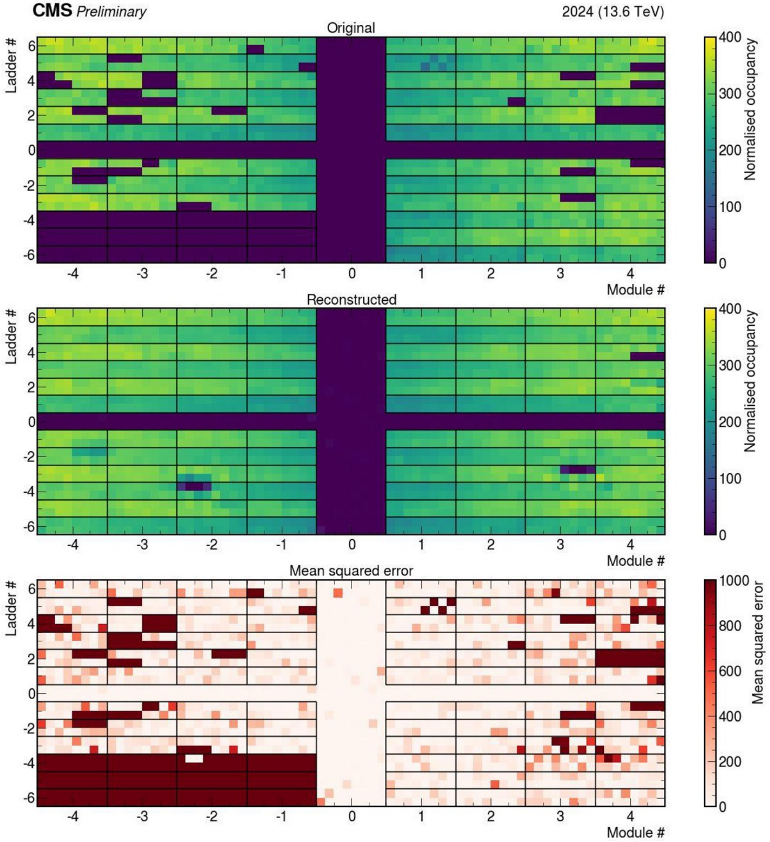
Pixel 2D Histograms ResNet AutoEncoder

- ML tools were trained to 2D map of the digi occupancy. Autoencoder based on residual networks, trained on 2024 data

- 2D map of the digi occupancy in BPIX layer 1
  - **Top:** original histogram
  - **Center:** reconstruction by the ML model
  - **Bottom:** Mean Squared Error (MSE)

- **LS 9**
  - Large region turned off
  - Correctly identified by ML model: Large area in red MSE

LS 9



Pixel 2D Histograms ResNet AutoEncoder

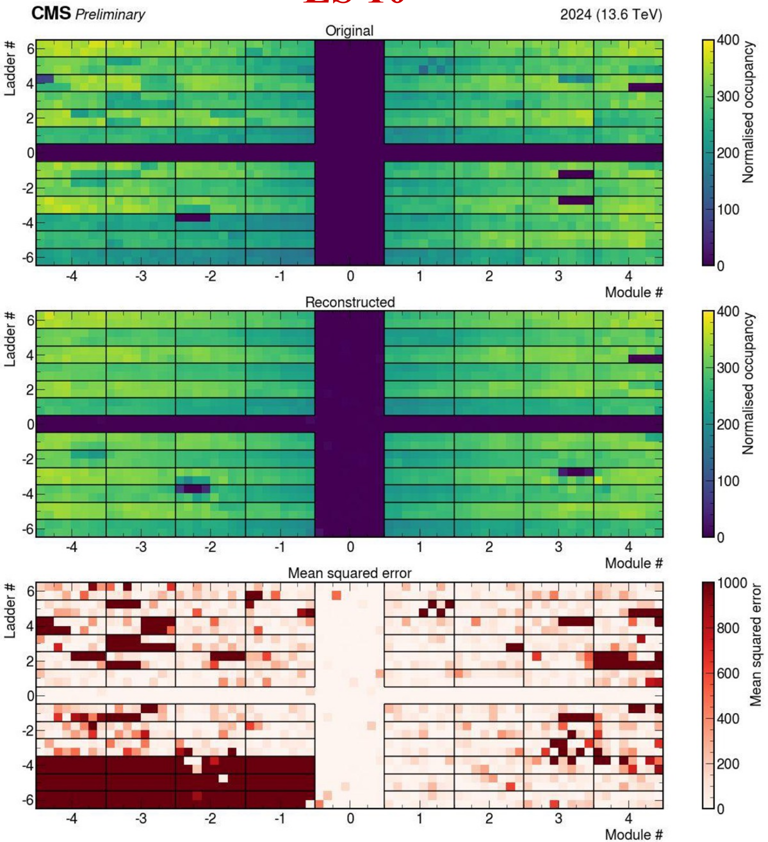
- ML tools were trained to 2D map of the digi occupancy. Autoencoder based on residual networks, trained on 2024 data

- 2D map of the digi occupancy in BPIX layer 1
  - **Top:** original histogram
  - **Center:** reconstruction by the ML model
  - **Bottom:** Mean Squared Error (MSE)

- **LS 9**
  - Large region turned off
  - Correctly identified by ML model: Large area in red MSE

- **LS 10**
  - Region partly recovered
  - Still identified by ML model as an anomalous LS

LS 10



Pixel 2D Histograms ResNet AutoEncoder

- ML tools were trained to 2D map of the digi occupancy. Autoencoder based on residual networks, trained on 2024 data

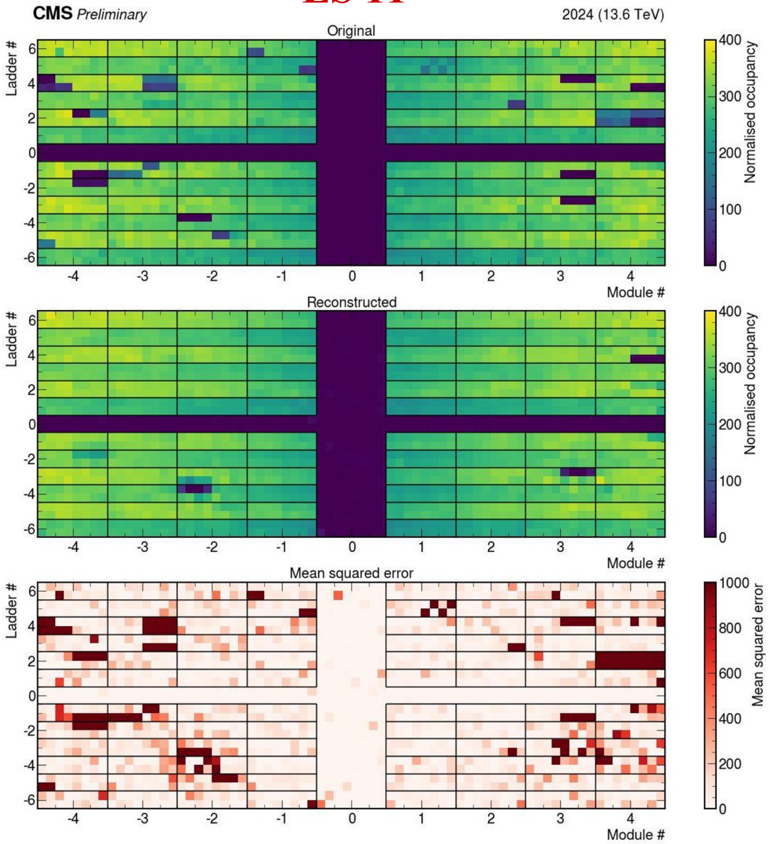
- 2D map of the digi occupancy in BPIX layer 1
  - **Top:** original histogram
  - **Center:** reconstruction by the ML model
  - **Bottom:** Mean Squared Error (MSE)

- **LS 9**
  - Large region turned off
  - Correctly identified by ML model: Large area in red MSE

- **LS 10**
  - Region partly recovered
  - Still identified by ML model as an anomalous LS

- **LS 11**
  - Fully recovered
  - Well reconstructed by ML model

LS 11



Summary

- We need to monitor **24x7 the Tracker conditions and performance during the data taking**:
  - Any issue need to be understood **very urgently**
  - **Bad tracker data → BAD data quality for the whole CMS**.

- The current DQM/DC procedure can be improved, mostly that:
  - Anomalies are **not tracked** unless they are staying for **longer time** enough to be an issue for analysis.
  - PerLS DQM → **finer time granularity** → point anomalous behaviour **efficiently** and **effectively**

- The strategy and goal of the ML based DQM/DC procedures are
  - **not to replace human decision-making,**
  - but to address challenges that make DQM/DC such a **labour intensive process**

- Our current ongoing efforts are :
  - Data exploration and data cleaning
  - ML studies with **other 1D/2D inputs**
  - Extend from unsupervised to fully **supervised** approaches (from anomaly detection to classification)
  - Deploy more tools together with **DQM perLS** harvesting in the ongoing Run 3.

Thank you for your attention

Back UP

Machine Learning for DQM/DC : DIALS

- **Data Inspector for Anomalous Lumi-Sections (DIALS)** is an data exploration tool.
  - Ability to explore all available data:
    - Designed to be an access point perLS monitoring elements (MEs).
  - Ability to get trend plots of multiple quantities (average, standard deviation, max, min)
  - Ability to flag/list outliers (LSes with no entries, LSes passing/ failing cuts on trend plots or other quantities)
  - Produce "anomaly" object listing Run(s)/LS(es)/ME(s)/ AnomalyType (with option for metadata)
  - It is responsible for indexing, storing pre-processed data and serving it via a WEB UI and REST Api.

- On-Going work:
  - Automatic ML pipeline inference on newly stored data for fast-DC on top of ML-flags
  - Extra data sources to add to DIALS:
    - OMS (Fill Information, Number of Bunches, Luminosity, Trigger Rates, DCS information, etc)
    - RunRegistry (DC flags, "Quality" JSONs, etc.) CertHelper (Flags, Problem classification, etc.)