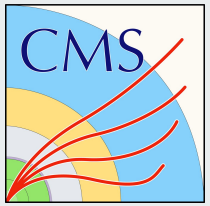
AutoDQM for Anomaly Detection in the CMS Detector

DPF-PHENO 2024

AutoDQM Team
autodqm-developers@cern.ch
https://github.com/AutoDQM

May 13-17 2024



Introduction

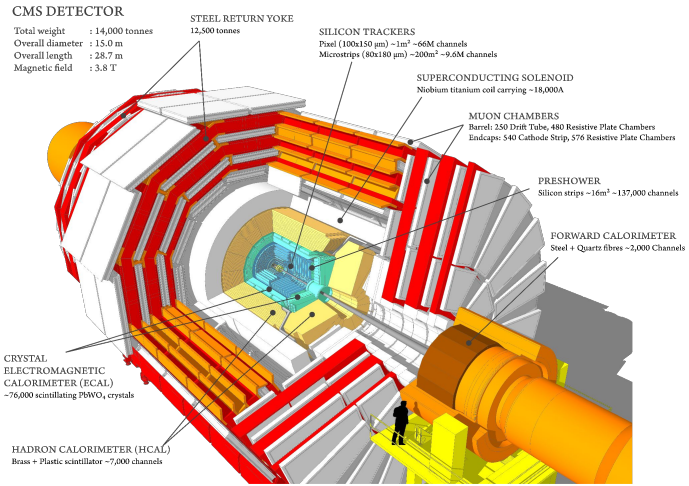
- Introduce DQM and AutoDQM
- Statistical tests + Machine Learning in AutoDQM
- AutoDQM performance studies

AutoDQM Tool

CMS Detector and Data Quality Monitoring

- The CMS detector collects a large amount of data
- Raw data are validated during data taking (Online DQM)
- Full-event PF reconstructions are validated few days later (Offline DQM)
- Trained DQM shifters validate runs by visually comparing selected histograms in a run with a previous reference run

→ Important task, but time consuming and labor-intensive.



AutoDQM Tool

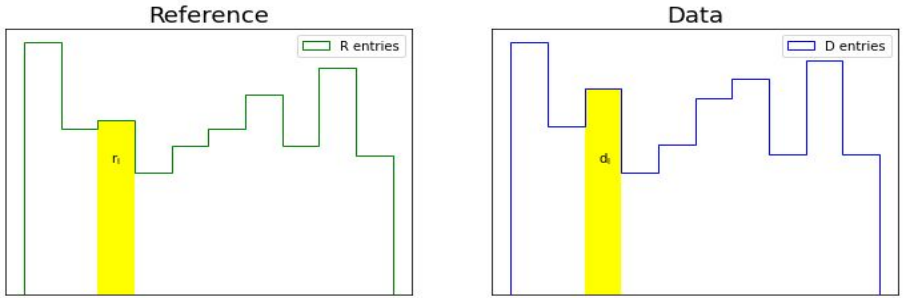
- AutoDQM is a web tool that semi-automates the process of comparing plots to references and looking for outliers, using statistical tests and machine learning algorithms

Tests in AutoDQM

- Beta-binomial statistical test
- Principal Component Analysis
- Autoencoder

Beta-binomial statistical test

- Computes the probability that, given a reference histogram with R total entries, and r_i entries in bin i, a data histogram with D total entries will show d_i entries in the same bin. This probability can then be converted into a pull value.
  - Good at catching single bin a anomalies
- Pull values can be used to calculate the $\chi 2$ score
  - Good at catching whole histogram anomalies



https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.betabinom.html

Machine learning (ML) for DQM anomaly detection

- Uses unsupervised ML algorithms
  - Many detector subsystems do not have enough bad data to effectively train supervised models
  - Past problems may not be representative of future issues
- ML algorithms learn to nearly reconstruct the histograms in the training set (learns general “shape” without the noise)
- Anomaly scores computed from the sum of square of errors (SSE) between the testing histogram and the reconstruction by ML algorithm

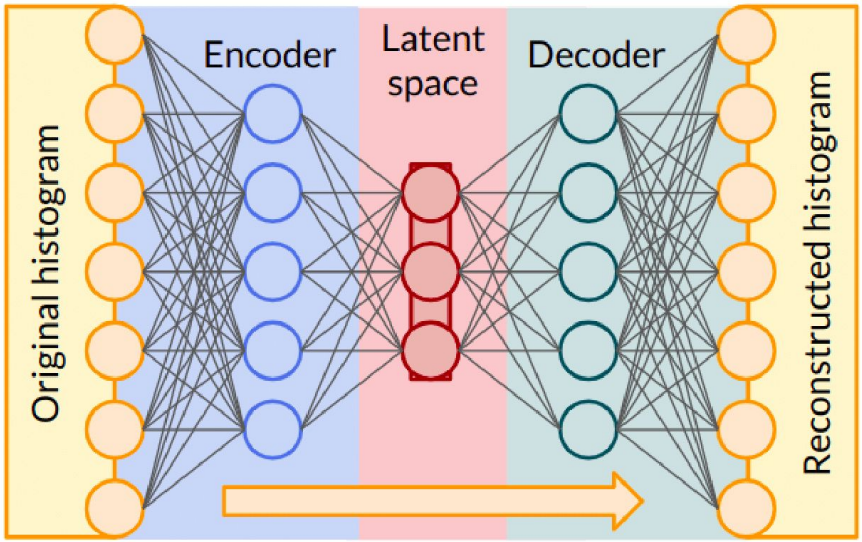
Principal Component Analysis (PCA)

- Dataset is transformed into new space, where the eigenvectors of this new space contain the variational information of the original dataset. These eigenvectors are called principal components
- First few principal components contains the general “shape” of the good histograms
- Good runs contain similar variational information, and can be reconstructed well. Bad runs may not contain or contain other information and will not be reconstructed well

https://doi.org/10.1038/s43586-022-00184-w

Autoencoder (AE)

- Type of neural network that can learn a representation of the data that can be used to reconstruct the input dataset
- Similar to PCA, they learn the general “shape” of the histograms. Good runs histograms will have good reconstructions, while Bad runs histograms may not.



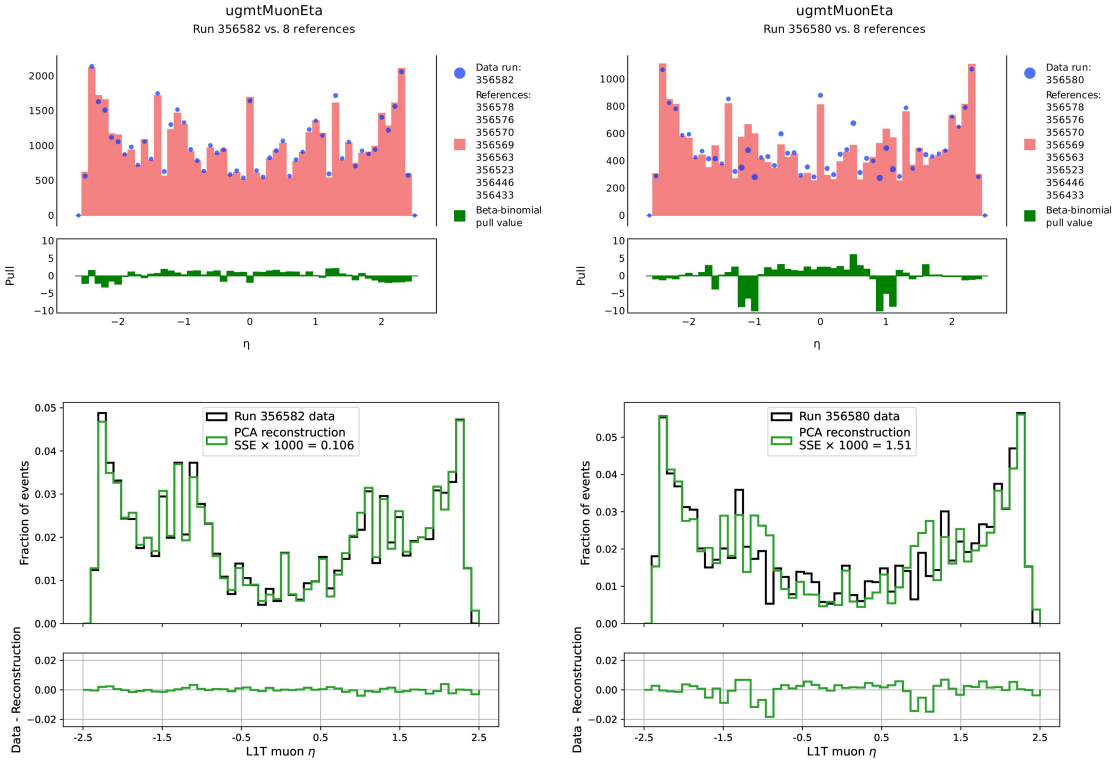
https://doi.org/10.1017/9781108955652.006

Performance Evaluation

Assessment strategy

- Used runs from 2022 that have been labeled as “good” or “bad” by the CMS Physics Performance and Data sets (PPD) group.
- 265 good and 43 bad runs, totalling 36 fb^{-1}
- Runs required to last at least 5 minutes and contain at least 3 pb^{-1} of collision data to ensure sufficient statistics
- Examined 62 histograms from the L1T online DQM set including inputs from ECAL, HCAL, and muon chambers.

Reconstruction Example

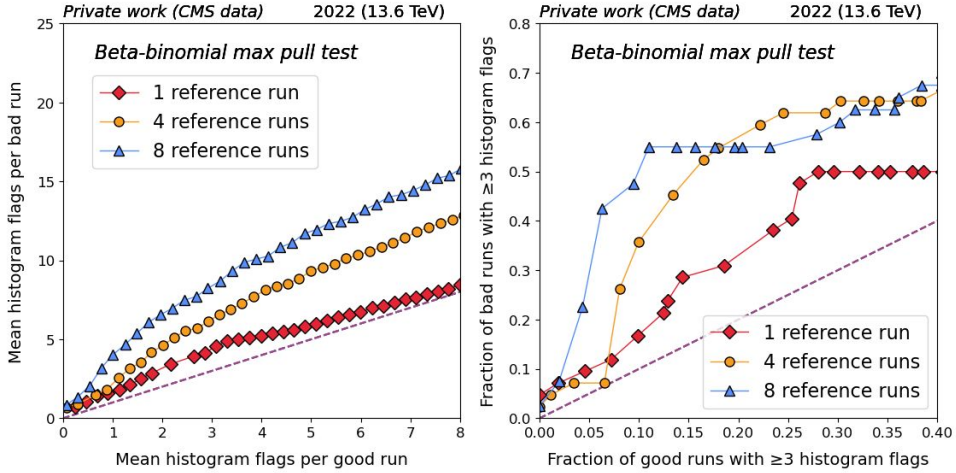


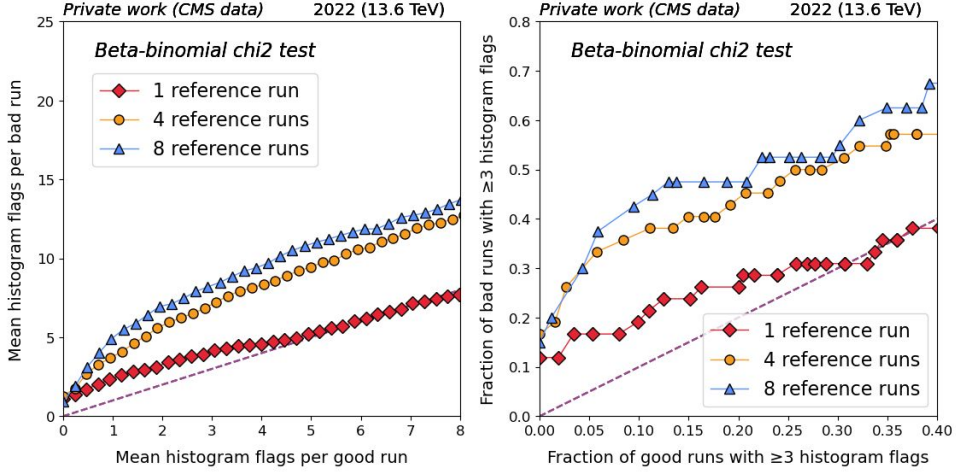
- Reconstruction of run 356582 (left) with expected distribution
- Run 356580 (right) show deficit at $0.9 < |\eta| < 1.2$
- Both PCA and beta-binomial show high anomaly scores for the deficit

Assessment metrics

- Beta-binomial: Max pull value and χ^2 are calculated for each histogram compared to 1,4,8 references
- PCA and AE : SSE scores are calculated for each histogram
- Using only scores from good runs, scores are ranked and thresholds are calculated such that the 1st threshold, only 1 histogram will score below the threshold and so on
- We then count how many histograms in each run fail each threshold.
- Using this information, we can form 2 ROC type curves.
    - Histogram Flags (HF) ROC shows the average number of histograms flagged per run
    - Run Flags (RF) ROC shows the fraction of runs with at least N histograms flagged

Results



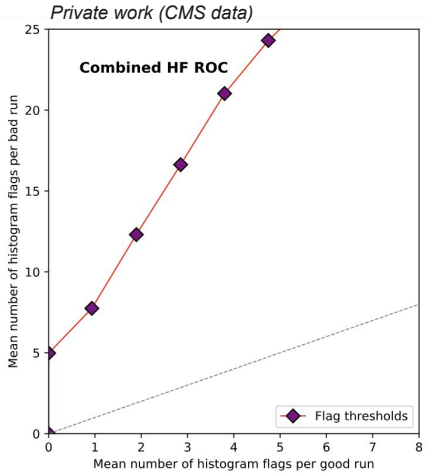


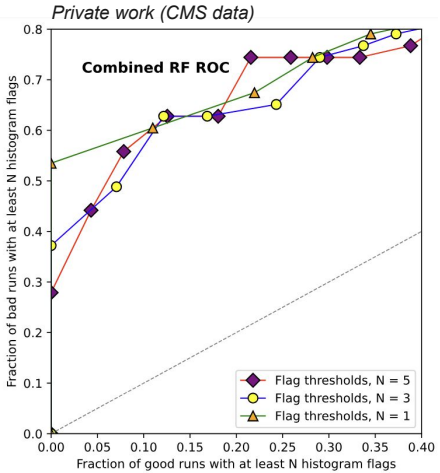
- Strong discrimination between good and bad runs
- At low number of good runs flagged, HF ROC shows 3-5 times more bad runs flagged
- RF ROC shows less than 15% of good runs flagged for 40-50% of bad runs flagged
- Performs better with more reference runs used
    - Expected as distributions can differ depending on pileup. Including multiple pileup conditions in references can improve performance

NOTE: we do not expect AutoDQM to identify 100% of bad runs in this set, as issues may not affect the L1T system, or issues may not show up in the DQM histograms

Results

- Best performance is obtained when all tests are applied simultaneously
- HF ROC shows 5-6 times more bad runs flagged compared to good runs
- RF shows 60% of bad runs have at least 3 flags compared to < 15% of good runs





Conclusion

- CMS Data Quality Monitoring task is time consuming and labor intensive
- We created the AutoDQM that aims to assist with DQM
- AutoDQM compares plots to known good references and calculate an anomaly score using statistical tests and ML methods
- We conducted a performance study on each test method, showing good discrimination between Good and Bad runs
- Best performance is acquired from using all the tests together