



# Anomaly Detection for Automated Data Quality Monitoring in the CMS Detector

Andrew Brinkerhoff<sup>1,11\*</sup>, Chosila Sutantawibul<sup>1</sup>, Indara Suarez<sup>2</sup>, Robert White<sup>3,9</sup>, Caio Daumann<sup>8</sup>, Jonathan Guiang<sup>10</sup>, Chad Freer<sup>4,5</sup>, Samuel May<sup>2</sup>, Bennett Marsh<sup>10</sup>, Darin Acosta<sup>7,11</sup>, Alex Aubuchon<sup>5</sup>, Emanuela Barberis<sup>5</sup>, Aaron Bundock<sup>9</sup>, Claudio Campagnari<sup>10</sup>, Evan Collins<sup>1</sup>, Preston Epps<sup>5</sup>, Johannes Erdmann<sup>8</sup>, Henning Flaecher<sup>9</sup>, Junshen Huang<sup>1</sup>, Vivan Nguyen<sup>5</sup>, Ryan Nie<sup>2</sup>, Sudarshan Paramesvaran<sup>9</sup>, John Rotter<sup>7,11</sup>, Kaitlin Salyer<sup>2</sup>, Siddhesh Sawant<sup>1</sup>, Tanvi Sheokand<sup>6</sup>, Darien Wood<sup>5</sup> and the CMS Muon Detector Collaboration

\*Correspondence:

Andrew\_Brinkerhoff@baylor.edu

<sup>1</sup> Baylor University, Waco, USA

<sup>2</sup> Boston University, Boston, USA

<sup>3</sup> INFN Sezione di Torino, Turin, Italy

<sup>4</sup> Massachusetts Institute of Technology, Cambridge, USA

<sup>5</sup> Northeastern University, Boston, USA

<sup>6</sup> Panjab University, Chandigarh, India

<sup>7</sup> Rice University, Houston, USA

<sup>8</sup> RWTH Aachen University III.

Physikalisches Institut A, Aachen, Germany

<sup>9</sup> University of Bristol, Bristol, England

<sup>10</sup> University of California Santa Barbara, Santa Barbara, USA

<sup>11</sup> University of Florida, Gainesville, USA

## Abstract

Successful operation of large particle detectors like the Compact Muon Solenoid (CMS) at the CERN Large Hadron Collider requires rapid, in-depth assessment of data quality. We introduce the “AutoDQM” system for Automated Data Quality Monitoring using advanced statistical techniques and unsupervised machine learning. Anomaly detection algorithms based on the beta-binomial probability function and principal component analysis are tested on the full set of proton-proton collision data collected by CMS in 2022. AutoDQM identifies anomalous “bad” data affected by significant detector malfunction at a rate 4 – 6 times higher than “good” data, demonstrating its effectiveness as a general data quality monitoring tool.

**Keywords:** Particle Physics, Anomaly Detection, PCA, Data Quality Monitoring

## Introduction

The Compact Muon Solenoid (CMS) experiment is a multipurpose particle detector designed to collect large amounts of data from high-energy proton-proton (pp) collisions at the CERN Large Hadron Collider (LHC) [1, 2]. The ATLAS and CMS experiments jointly discovered the Higgs boson using LHC collision data collected between 2010 and 2012, and are currently seeking evidence for new physics which could explain dark matter, dark energy, or the matter–antimatter asymmetry of the universe [3–5].

CMS identifies and measures electrons, muons,  $\tau$  leptons, photons, and hadrons using a global “particle-flow” (PF) algorithm, which combines information from various sub-detectors situated concentrically around the proton beam [6]. A multi-layer silicon tracker traces the helical path of charged particles emerging from the collision point as they are deflected by the CMS solenoid’s 3.8 T magnetic field [7]. Both charged and neutral particles then deposit energy into the electromagnetic calorimeter (ECAL), made from lead tungstate crystals, and the hadron calorimeter (HCAL), made from interleaved brass absorbers and plastic scintillators [8–10]. Beyond the calorimeters, layers of muon tracking detectors, including drift tubes, cathode strip chambers, resistive plate

© The Author(s) 2026. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

chambers, and gas electron multipliers are embedded in the CMS magnet's flux-return yoke [11]. While the LHC collision rate can exceed 30 MHz, at most around 100 kHz of events pass the initial selection by the CMS Level-1 Trigger (L1T), which uses calorimeter and muon detector inputs to perform preliminary event reconstruction on custom hardware boards in less than  $4\mu\text{s}$  [12]. The High-Level Trigger (HLT) then reconstructs events in more detail using CPUs and GPUs, and sends around 1 kHz of collision data to storage on hard disks [13].

One of the main challenges in operating CMS is to monitor the detector, trigger, and particle reconstruction continuously to ensure that the collected data satisfy the stringent criteria necessary for precise measurements and sensitive searches for new physics phenomena. Data-taking "runs" comprising a few minutes to several hours of collision data are processed in real time to create thousands of histograms which measure various aspects of detector performance. Trained "shifters" monitor these histograms and may intervene if the plots indicate anomalous behavior compared to previous runs. This data quality monitoring (DQM) is performed for each subdetector system to immediately identify issues "online" based on raw data, and also "offline" after a few days when full-detector PF reconstruction is available. Each year a few percent of the total data collected by CMS, corresponding to a dozen or more hours of beamtime, are designated as "bad" due to detector or reconstruction issues during data taking. Without DQM, this fraction would be considerably higher, as problems would go unnoticed for a longer period of time. Data quality monitoring is thus a time-consuming and labor-intensive but important task, so it is critical to develop robust tools that can help shifters quickly and reliably identify problems in any part of the highly complex CMS detector.

In this paper, we introduce the AutoDQM tool<sup>1</sup>, a web-based service that employs a generalized approach to automated DQM using statistical techniques and unsupervised machine learning (Sect. 2). Anomaly detection algorithms based on the beta-binomial probability function and principal component analysis (PCA) are discussed in Sects. 2.1 and 2.2. AutoDQM performance studies using L1T monitoring plots from the entire 2022 data set are presented in Sect. 3, along with examples of muon detector monitoring using AutoDQM. These results and plans for future developments are summarized in Sect. 4.

### The AutoDQM tool

Traditional DQM in CMS is performed using the online and offline DQM web-based GUIs, which contain hundreds of histograms for each CMS subdetector system [14]. DQM shifters for each subsystem examine selected histograms in a particular run and look for differences compared to histograms from previous reference runs. Such anomalies could indicate degraded detector, trigger, or reconstruction performance which would compromise the final physics analysis of the collected data. Of course, visual comparison of dozens or hundreds of histograms is fatiguing and error-prone. Basic Kolmogorov-Smirnov tests and approximate  $\chi^2$  comparisons between two histograms are available in the DQM GUIs, but are extremely unreliable at detecting true anomalies.

---

<sup>1</sup> AutoDQM source code publicly available at <https://github.com/AutoDQM/AutoDQM>

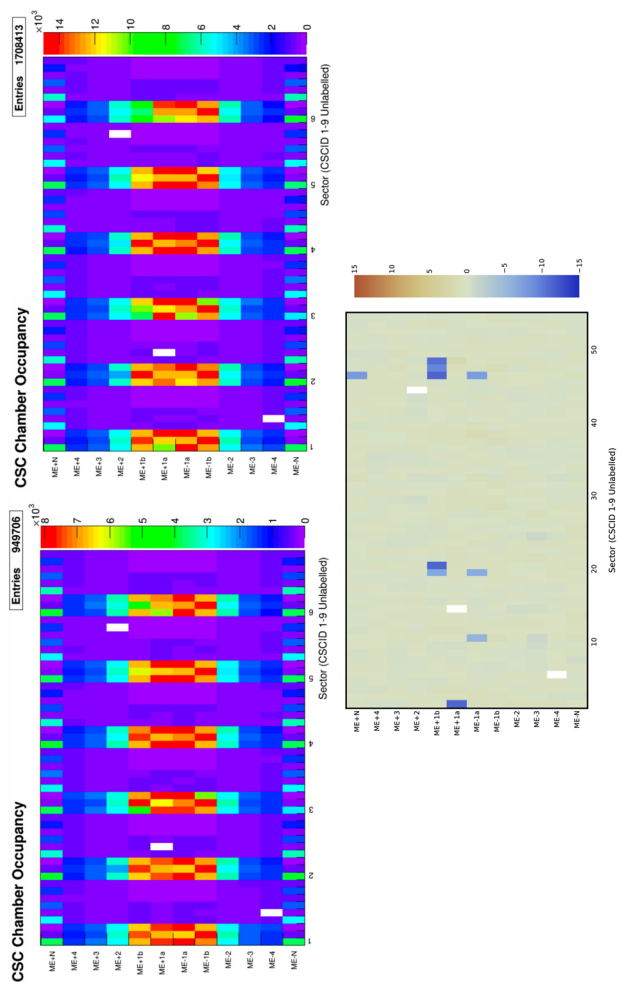
More advanced DQM techniques have long been considered [15, 16], and in two recent cases were deployed by CMS [17, 18], but for a very limited selection of plots. The AutoDQM anomaly detection tool is a web-based service which evaluates all forms of online and offline DQM histograms and assists shifters in rapid and effective data monitoring. AutoDQM uses  $\chi^2$  and single-bin pull value tests based on the beta-binomial probability function to look for anomalies in a given set of histograms compared to data from multiple previous “reference” runs. PCAs are also trained on larger sets of reference histograms to generate complex models of good data, which are then used to identify histograms which deviate from the expectation. AutoDQM graphically represents these statistical and machine learning (ML) tests to highlight anomalous regions within the histograms, allowing shifters and detector experts to quickly identify and locate issues as they arise.

### Statistical tests

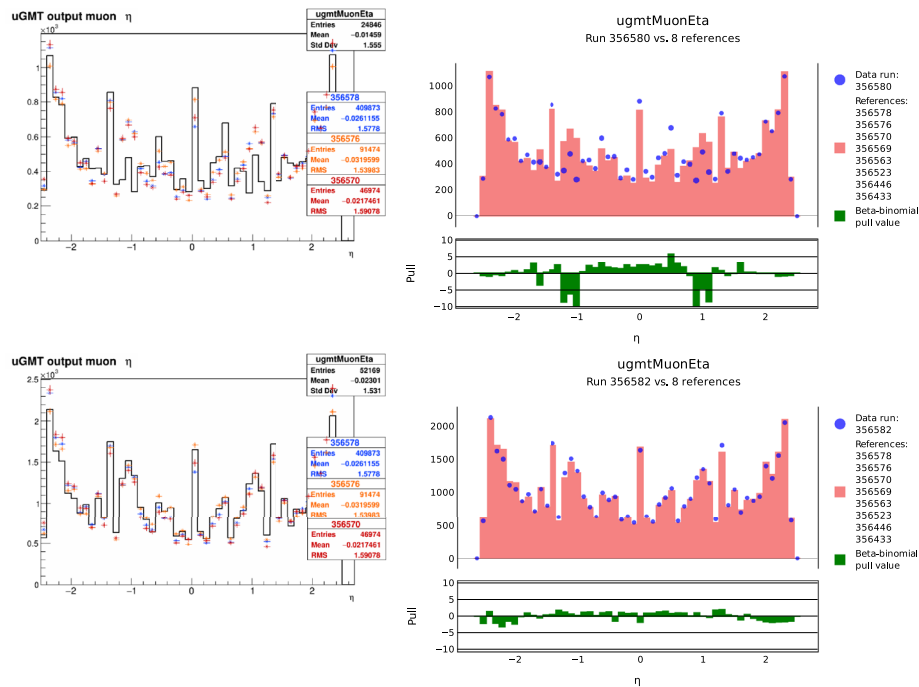
Most one-dimensional (1D) and two-dimensional (2D) histograms in the CMS DQM GUI have an integer number of entries in each bin. Depending on the type of histogram and the duration of the data-taking run, a specific histogram may contain millions of entries, or just a few; and these entries may be distributed evenly, or concentrated in a small number of bins. For a given data histogram, the number of entries  $d_i$  in each bin  $i$  may be treated as the frequency of a distinct outcome out of  $D$  trials, where  $D$  is the integral of the histogram. A reference histogram from a prior run with integral  $R$  and  $r_i$  entries in each bin can be used to compute the likelihood  $\mathcal{L}_i$  to observe  $d_i$  in each corresponding bin from the later data run, using the beta-binomial function:

$$\mathcal{L}_i = f(d_i|D, \alpha, \beta) = \binom{D}{d_i} \frac{B(d_i + \alpha, D - d_i + \beta)}{B(\alpha, \beta)} \quad (1)$$

where  $B$  is the beta function,  $\alpha = \alpha_0 + r_i$ , and  $\beta = \beta_0 + R - r_i$ . We set  $\alpha_0 = \beta_0 = 1$ , consistent with uniform priors, and use the `betabinom` probability mass function (pmf) implementation in `SciPy` to compute  $\mathcal{L}_i$  for all bins simultaneously using a `numpy` array representation of the histogram [19]. These  $\mathcal{L}_i$  values can be compared to the maximum likelihood  $\mathcal{L}_i^{max}$  for each bin, corresponding to  $d_i^{max} = D \times r_i/R$  (rounded up or down). The ratio  $\mathcal{L}_i/\mathcal{L}_i^{max}$  gives a relative likelihood  $\mathcal{L}_i^{rel}$ , which is converted to a pull value  $Z_i$  in units of standard deviations using the relation  $Z_i^2 = -2 \ln \mathcal{L}_i^{rel}$ . In order to ensure a minimum “tolerance” of  $\approx 1\%$  in the prediction, we scale both  $R$  and  $r_i$  by a factor of  $\tau = 1/\sqrt{1 + (10^{-4} r_i)^2}$  when computing  $\mathcal{L}_i^{rel}$  for each bin, such that  $\tau \times r_i \rightarrow 10^4$  as  $r_i \rightarrow \infty$ , yielding a minimum uncertainty of about  $1/\sqrt{10^4} = 1\%$ . When comparing a data run to multiple reference runs,  $Z_i$  is derived using the average of the  $\mathcal{L}_i^{rel}$  values, which are computed separately with respect to each reference run. With this approach, if the observed data matches the expectation from at least one of the reference histograms well, the pull values will not be very large. For example, if seven reference runs each give  $\mathcal{L}_i^{rel} \approx 0$  ( $Z_i \approx \infty$ ) for a given bin, but one reference run gives  $\mathcal{L}_i^{rel} = 0.33$  ( $Z_i \approx 1.5$ ), the final  $Z_i = \sqrt{-2 \ln (0.33/8)} \approx 2.5$ , which is larger than 1.5 but much smaller than  $\infty$ . This allows the statistical test to account for systematic variations in histogram shapes due to changing collision conditions between runs.



**Fig. 1** A 2D muon track “stub” occupancy histogram for cathode strip chambers (CSCs) in reference run 356,937 (left), data run 357,001 (right), and the AutoDQM heat map showing regions of statistically significant deficits in blue when comparing the data run to 8 prior “good” reference runs (bottom). These deficits in run 357,001 are almost invisible in the original DQM GUI histogram



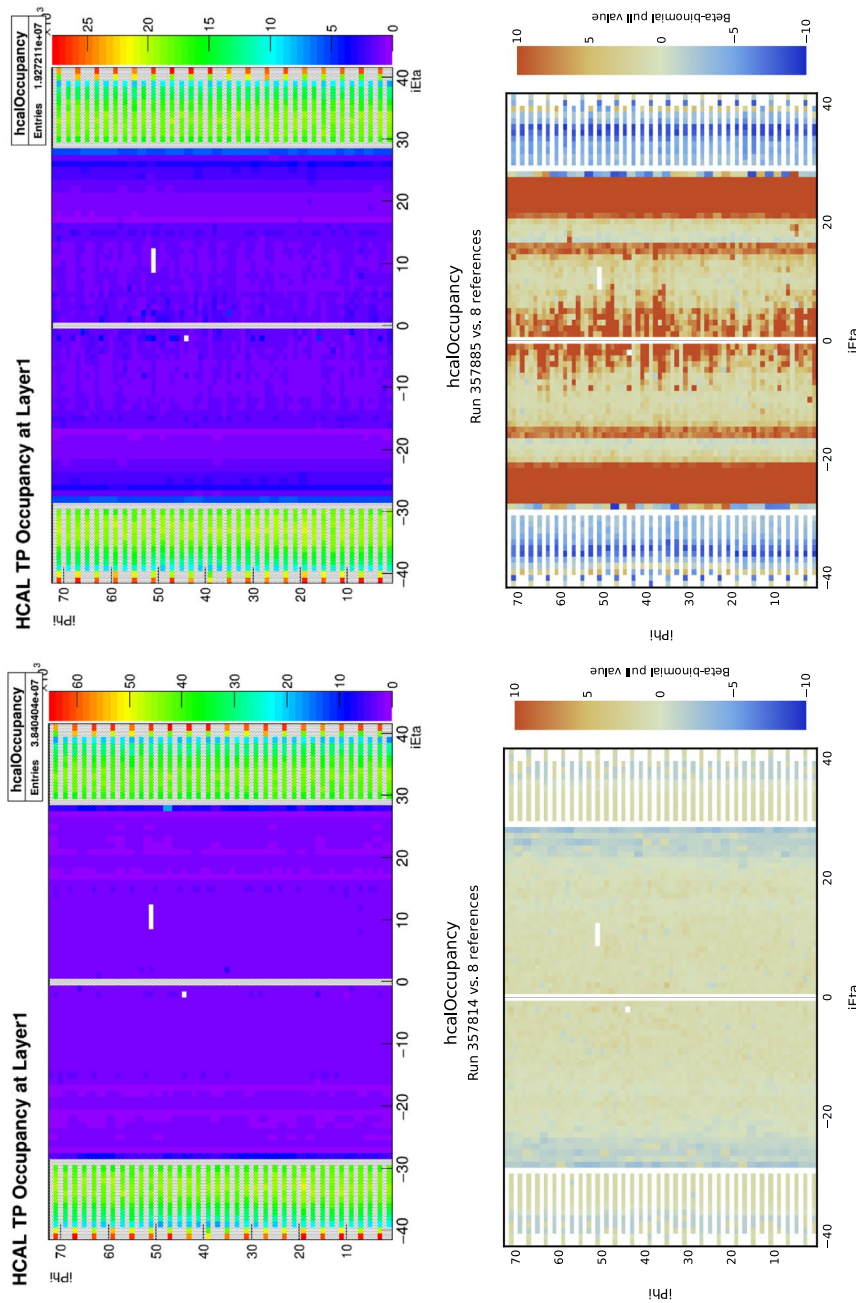
**Fig. 2** The standard DQM histogram for the pseudorapidity distribution  $\eta$  of reconstructed muon tracks in the LIT from data run 356,580 in black, overlaid with normalized distributions from 3 previous reference runs (upper left). AutoDQM flags the data as anomalous compared to 8 previous reference runs, and makes the local deficit more visible with the beta-binomial pull value histogram in green (upper right). The corresponding plots for run 356,582, where the muon detector issue was resolved, are shown below

In the AutoDQM GUI, the pull values for each bin in a 2D histogram are displayed as a heat map, allowing shifters to quickly identify the location of any significant excess (red) or deficit (blue), as shown in Figs. 1 and 3. For 1D histograms, the data distribution in blue overlays a per-bin probability-weighted average of the reference distributions in red, with the pull values in a separate panel in green (Fig. 2). Only histograms identified as anomalous are displayed immediately to shifters, focusing their attention on confirmed discrepancies. Shifters can view the full set of histograms by clicking a button in the GUI (Fig. 8).

The first statistical anomaly metric used by AutoDQM is  $\chi^2 = \sum Z_i^2/n$ , where  $n$  is the number of bins. The second anomaly metric is the modified maximum pull magnitude  $Z'_{max}$  out of all the bins, where the smallest single-bin relative likelihood  $\mathcal{L}'_{i min}$  is first adjusted for the look-elsewhere effect:  $\mathcal{L}'_{i min} = 1 - (1 - \mathcal{L}_{i min}^{rel})^n$ . The  $\chi^2$  and  $Z'_{max}$  thresholds to flag histograms as anomalous can be configured independently for each type of histogram.

### Machine learning for DQM anomaly detection

Machine learning algorithms can detect anomalies without the need for specific reference data, while also accounting for expected systematic variations in the histogram shapes. AutoDQM currently uses unsupervised ML algorithms based on Principal Component Analysis (PCAs), and we are developing algorithms based on normalized autoencoders and normalizing flows. This is preferred over the supervised approach,



**Fig. 3** The 2D  $\phi$  vs.  $\eta$  geometrical distributions of energy deposits in the HCAL detector transmitted to the LIT, from data runs 357,814 (upper left) and 357,885 (upper right). There is very little visible difference between these DQM histograms, which are monitored daily by LIT experts. The AutoDQM beta-binomial pull value “heat map” indicates no anomalous behavior for run 357,814 when comparing to 8 previous reference runs (lower left), but correctly flags run 357,885, which had an issue with HCAL timing (lower right)

in which algorithms are trained with explicitly labeled “good” and “bad” data, for two reasons. First, bad data are rare – many detector subsystems do not have enough bad data to effectively frame this as a supervised problem. Second, past problems resulting in bad data may not be representative of future issues. The unsupervised approach only requires a sufficient quantity of good data to train the algorithm, and is agnostic to the particular type of anomaly which could indicate bad data.

More formally, our unsupervised approach to DQM anomaly detection can be framed as follows: given a collection of histograms from good runs, we seek to learn a transformation  $\mathcal{T}$  from the input space (i.e. the entries in each bin) into a lower-dimensional “latent space”, such that the latent space can be used to approximately reconstruct the original histogram. The form of the transformation  $\mathcal{T}$  and the cost function used to optimize its parameters is described for PCAs in section 2.2.1.

Once a transformation  $\mathcal{T}$  has been learned, the anomaly score for data histogram  $d$  (normalized to area 1) can be calculated as the sum of the squares of the errors (SSE) between the original and reconstructed histograms:

$$\text{SSE} = \sum_{i=1}^n (d'_i - d_i)^2, \quad (2)$$

where  $d_i$  are the contents of bin  $i$  of the normalized data histogram, and  $d'_i$  are the bin contents obtained after applying the transformation  $\mathcal{T}$  and subsequently its inverse:  $d' = \mathcal{T}^{-1}(\mathcal{T}d)$ . Histograms that match the training data have a self-similar reconstruction under transformation  $\mathcal{T}^{-1}\mathcal{T}$ , leading to low SSE scores, while those with significant deviations compared to the training set do not. In this way, bad runs which behave differently than the good runs used in training can be identified by DQM histograms with high SSE scores, regardless of the nature of the anomaly.

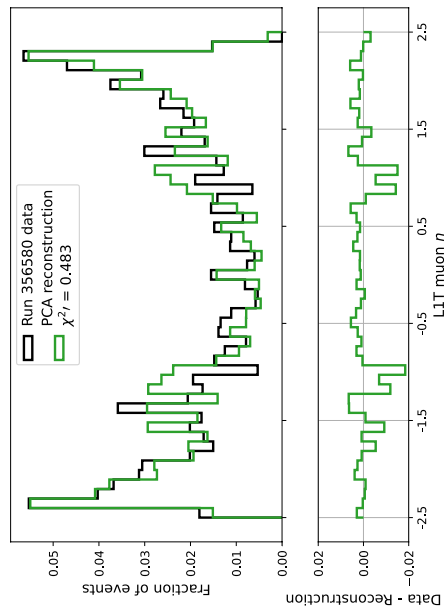
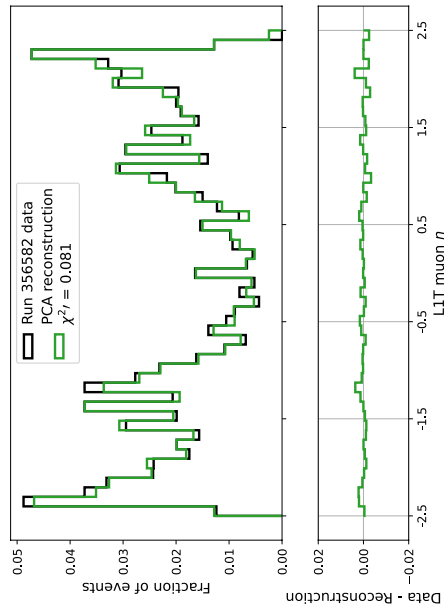
One shortcoming of the normalized SSE metric is its anti-correlation to the number of entries in a histogram, such that histograms in shorter runs consistently get higher anomaly scores (see Appendix Appendix A). To address this, we instead use a  $\chi^2$  metric, which is intrinsically less sensitive to statistical fluctuations than the SSE. We use the beta-binomial probability function (Eq. 1) with the original data histogram  $d$  (integral  $D$ , not normalized to 1), and take  $d'$  (integral  $D'$ , again not normalized) scaled by 100 as the “reference” histogram; so mathematically:

$$\alpha = \alpha_0 + d'_i \times 100 \text{ and } \beta = \beta_0 + D' \times 100, \text{ with } \alpha_0 = \beta_0 = 1 \quad (3)$$

The factor of 100 suppresses any statistical uncertainty in the transformed bin contents  $d'_i$ , leaving only the uncertainty in the original  $d_i$ . However, this  $\chi^2$  score is consistently *low* for histograms with fewer entries, so a modified version is used, scaling by the number of entries  $D$ :

$$\chi^{2'} = \chi^2 / D^{1/3}. \quad (4)$$

This scaling mitigates the  $\chi^2$  score dependence on  $D$ , as shown in Appendix Appendix A.



**Fig. 4** Normalized reconstructions of the  $\eta$  distribution of muon tracks from runs 356,580 (left) and 356,582 (right), using the PCA reconstruction. A deficit of tracks in  $0.9 < |\eta| < 1.2$  in run 356,580 is indicated by the PCA  $\chi^2/\nu$  score over 0.4, and the data-reconstruction difference in the lower panel

### Principal component analysis

Principal component analysis is an unsupervised ML approach frequently used in the context of dimensionality reduction [20]. AutoDQM uses the PCA implementation from `scikit-learn` [21]. The PCA takes a 1D or 2D histogram from a given run as input, and transforms it into 2 key components extracted from the training set of 216 good runs. 2D histograms are first flattened into 1D for both training and evaluation. We found that merging low-occupancy bins improves the PCA reconstruction by reducing the impact of statistical fluctuations. Merging proceeds iteratively until each bin contains at least 0.33% of the histogram entries, averaging the bin occupancy over the full training dataset. The reconstruction is produced from the latent space via re-transformation, with negative bins set to zero using a rectifier function as introduced in Ref. [22]. This avoids non-physical reconstructed histograms, and is applied in the last stage to avoid biasing the weights in the PCA re-transformation workflow.

The PCA reconstruction ought to closely agree with the input histogram for good histograms, but anomalous features will not be identified as principal components in the initial transformation, and thus will not appear in the same way in the reconstruction. By comparing the input and reconstructed histograms, the PCA can flag anomalous histograms based on high  $\chi^2$ ' scores (Eq. 4), typically  $> 0.4$ , as shown in Fig. 4.

## Performance evaluation

### Assessment strategy

When developing anomaly detection tools, it is often difficult to rigorously measure their performance identifying true anomalies in data. An unbiased assessment using real anomalies in real data requires an independent measure of “anomalousness” on the same set of data. Ideally, this independent measure would also focus on *important* anomalies, as many types of data variation have no bearing on whether the data is “good” – in the case of CMS, usable for later physics analysis. Previous anomaly detection studies for CMS DQM have either used histograms which were individually labeled as bad by visual inspection, or have generated artificial anomalies to mimic problematic detector behavior [15–18].

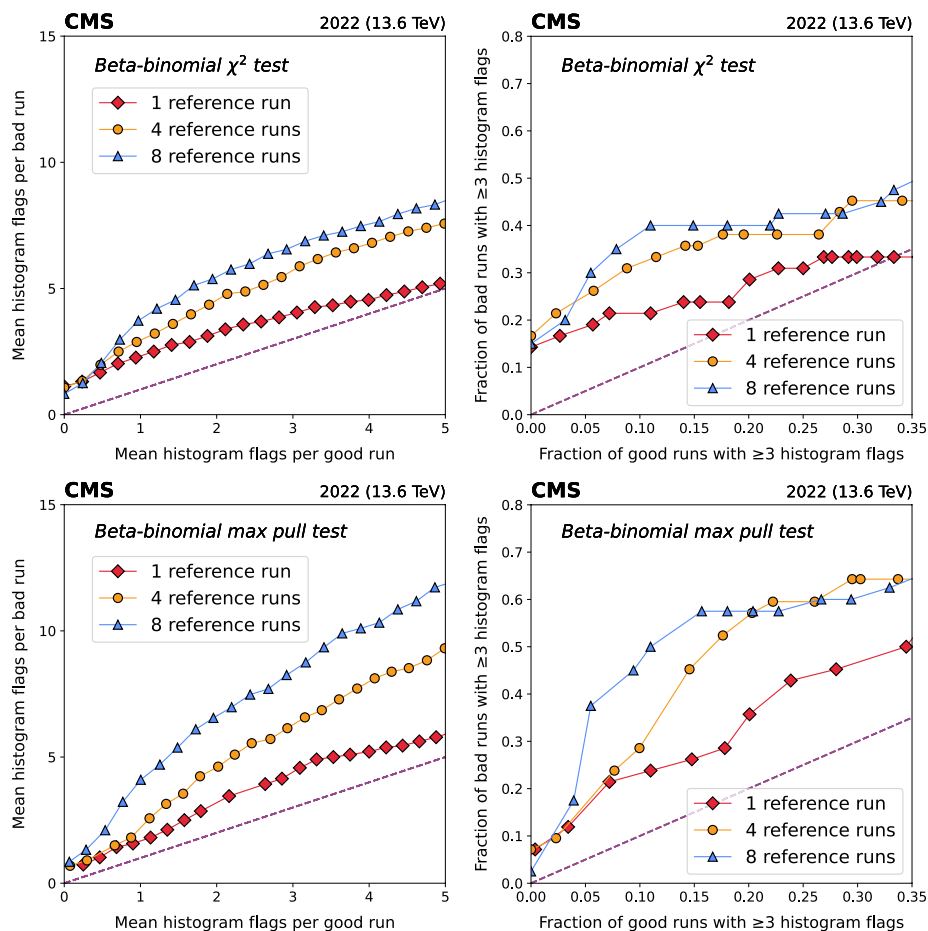
To measure the AutoDQM performance, we use a full year’s worth of data collection runs which were labeled as good or bad by the CMS Physics Performance and Datasets (PPD) group. The PPD team synthesizes information about each run from detector subsystem experts and analyzers who study reconstructed hadron jets,  $\tau$  leptons, photons, electrons, and muons. PPD then decides whether any detector or reconstruction issues observed are serious enough to exclude all or part of the data in a run from CMS physics analyses. The final determination is made without reference to the AutoDQM tool, and frequently uses information which is not available in the DQM GUI at all. Thus the PPD evaluation is both independent of AutoDQM, and reflects the seriousness of anomalous behavior in the CMS detector.

Our assessment data set includes 265 good and 43 bad runs collected in 2022, representing an integrated luminosity of  $36 \text{ fb}^{-1}$ . Selected runs are required to have lasted at least 5 min and contain at least  $3 \text{ pb}^{-1}$  of collision data, to ensure a sufficient number of entries in the DQM histograms. Good runs must have at least 90% of their data labeled good by PPD; bad runs must be over 50% bad, or contain over  $40 \text{ pb}^{-1}$  of bad data. For

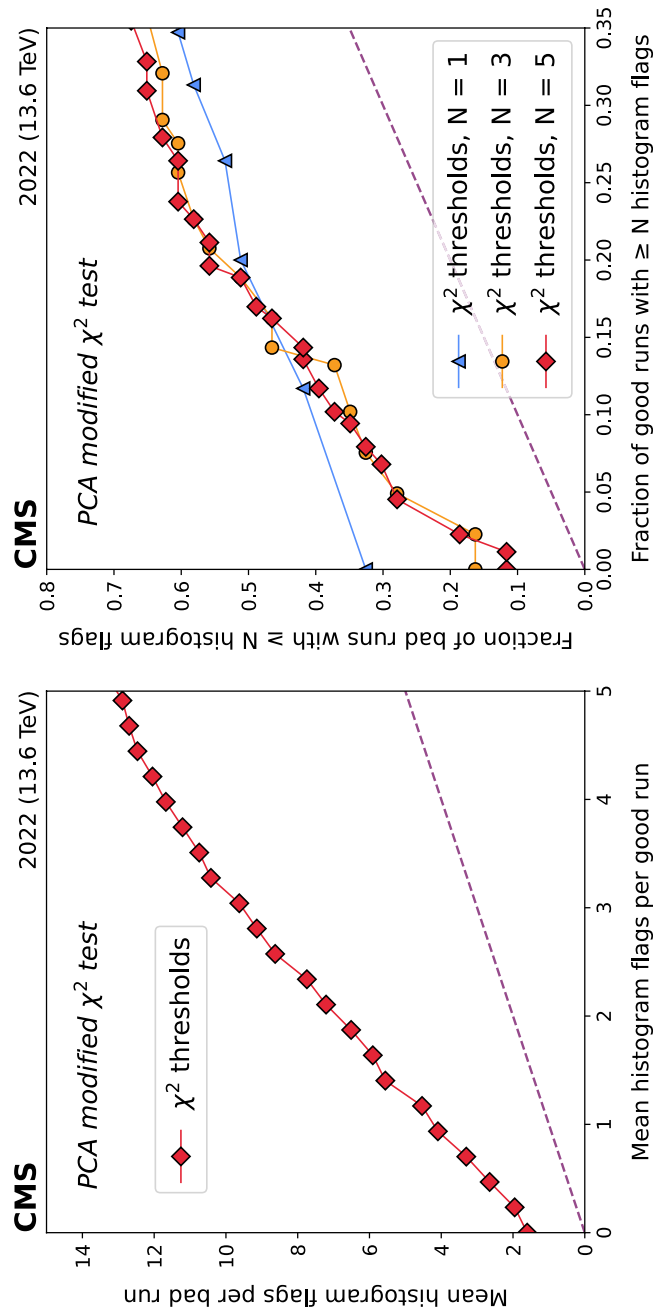
each run, we examine 62 histograms from the L1T online DQM, covering inputs from the ECAL, HCAL, and muon chambers. Typical histograms produced for L1T reconstructed hadron jets, electrons and photons,  $\tau$  leptons, and muons include 1D and 2D distributions of their  $\eta$  and/or  $\phi$  location, transverse momentum ( $p_T$ ), and isolation- and identification-related quantities. Because the L1T does not receive any information from the silicon tracker, we exclude runs that PPD labeled bad due to tracker issues.

**Assessment Metrics**

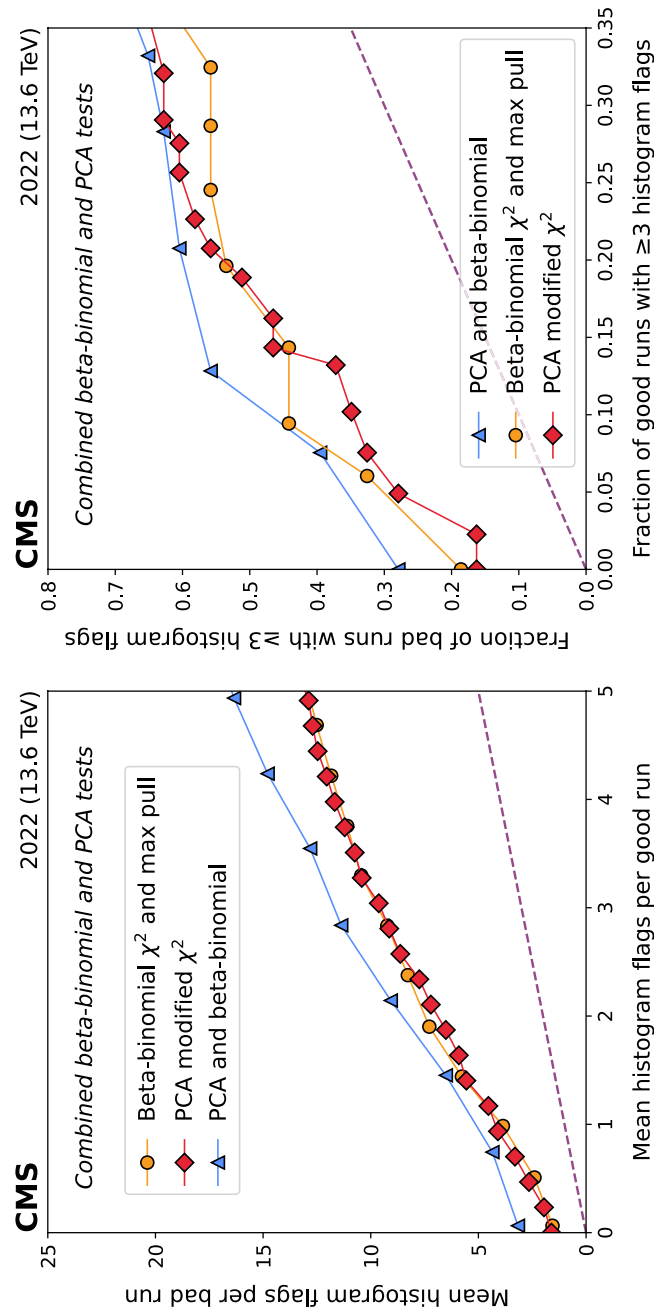
For each histogram in each run, we compute the  $\chi^2$  and  $Z'_{max}$  anomaly score values from the beta-binomial statistical test (Eq. 1), comparing to 1, 4, or 8 prior reference runs. Reference runs must be labeled as good by PPD, and must have lasted at least 30 min. This matches typical shifter procedures, where longer reference runs with no known issues are compared to the most recent data run. The PCA algorithms were trained on all 216 good runs from 2022 lasting at least 30 min, and the  $\chi^2$ ' from Eq. 4 is used as the anomaly score. The anomaly scores  $s$  for each histogram from the 265 good runs



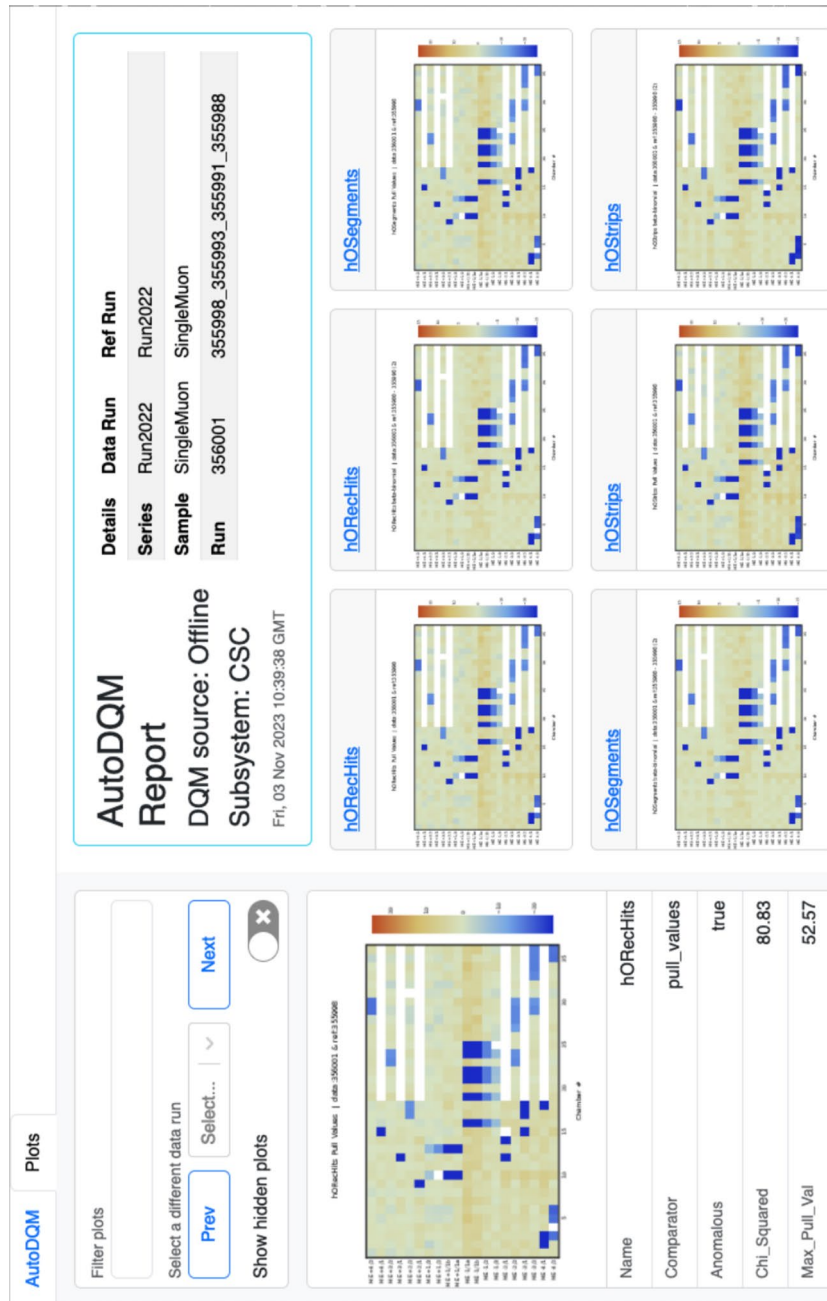
**Fig. 5** Performance of the beta-binomial  $\chi^2$  and maximum single-bin pull ( $Z'_{max}$ ) statistical tests (Sec. 2.1) on L1T DQM histograms from 308 runs containing 2022 data. ROC curves are constructed based on the mean number of histograms flagged per run (left), and the fraction of runs with at least 3 histograms flagged (right), comparing the data to 1, 4, or 8 prior reference runs



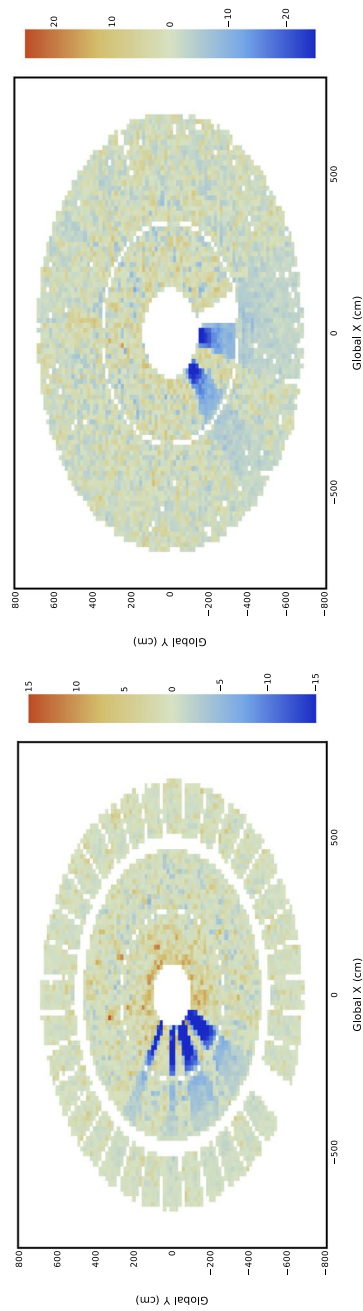
**Fig. 6** Performance of the PCA modified  $\chi^2$ -test (Sec. 2.2) on L1T DQM histograms from 308 runs containing 2022 data. ROC curves are constructed based on the mean number of histograms flagged per run (left), and the fraction of runs with at least 1, 3, or 5 histograms flagged (right)



**Fig. 7** Performance of the combined beta-binomial and PCA algorithms on LIT DQM histograms from 308 runs containing 2022 data. ROC curves are constructed based on the mean number of histograms flagged per run (left), and the fraction of runs with at least 3 histograms flagged (right)



**Fig. 8** AutoDQM GUI webpage for the cathode strip chambers (CSCs) in run 356001 in 2022, showing numerous chambers with anomalously low occupancy of reconstructed muon “hits” in blue. Each plot can be expanded by clicking, and histograms not flagged as anomalous can be viewed using the “Show hidden plots” toggle. The precise anomaly scores for each histogram are displayed in a panel on the left



**Fig. 9** AutoDQM GUI plots of the geometrical reconstructed muon “hit” distribution in the CSC detectors for run 356,001 in 2022, showing regions with anomalously low occupancy in blue. Regions which are consistently empty across multiple runs appear in white, allowing the shifter to distinguish between new and long-running issues

(including those lasting less than 30 min) are ranked, allowing us to set variable anomaly thresholds  $t_i = (s_i + s_{i+1})/2$  for each histogram type, based on the ranked values. At the lowest threshold for a given histogram type,  $t_0 = (s_0 + s_1)/2$ , that histogram will be flagged as anomalous in all but one good run. At  $t_1 = (s_1 + s_2)/2$ , this histogram will be flagged in all but two good runs, with the pattern continuing to higher thresholds. The maximum threshold for a given histogram type is set *above* the highest value seen in any good run:  $t_{max} = s_{max} + (s_{max} - s_{max-1})/2$ . For each good and bad run, at each threshold index, we count how many of the 62 different histograms are flagged as anomalous, i.e. for which  $s \geq t_i$ .

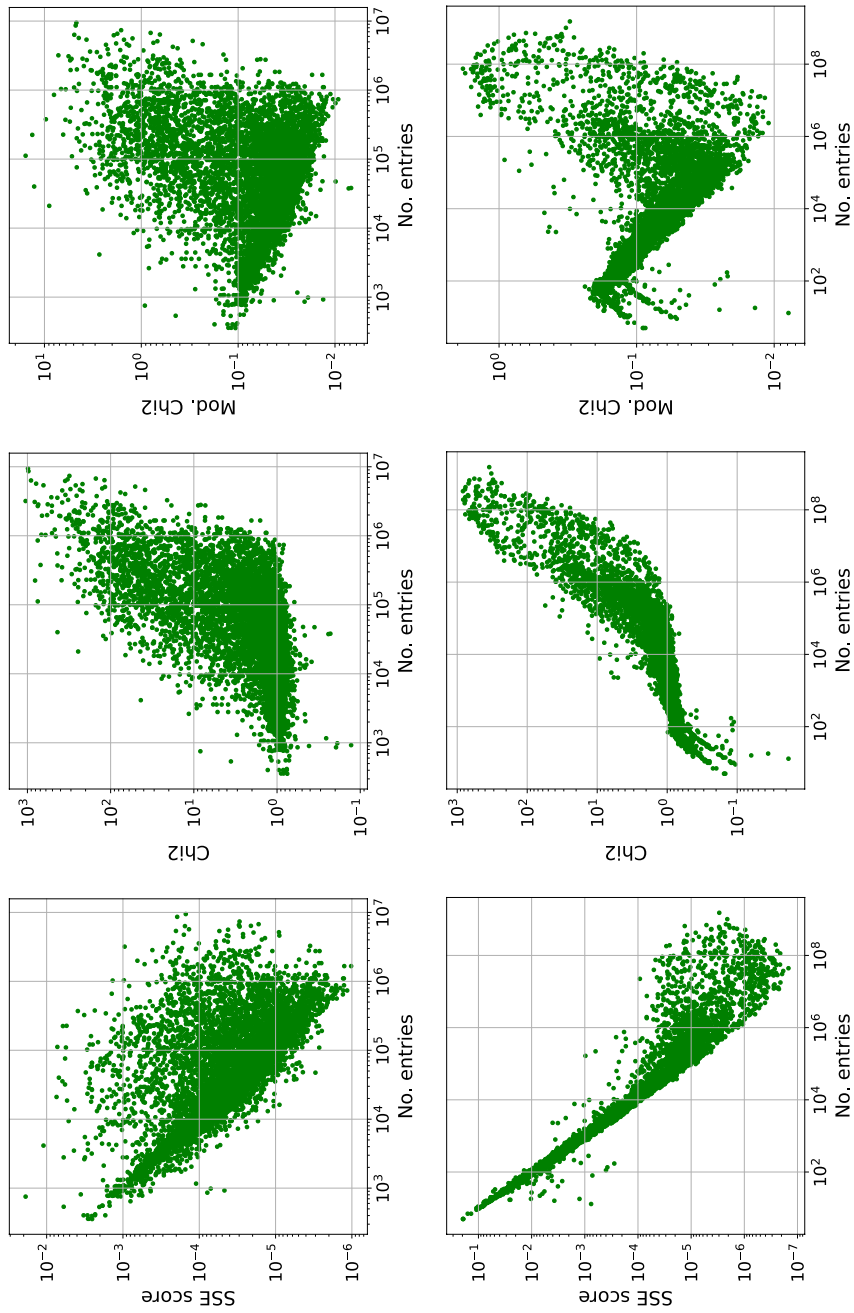
Ideally AutoDQM would flag a large number of histograms from bad runs, and only a few from good runs. It is also important to check what fraction of good and bad runs have a significant number of anomaly flags. If good runs consistently have many histograms flagged as anomalous, “alert fatigue” will set in, and shifters will start to ignore the AutoDQM output. Having a low number of flagged histograms is also important to allow shifters and experts to follow up on notable anomalies even in good runs, in case some intervention is needed.

We thus construct two types of receiver operating characteristic (ROC) curves based on the number of flagged histograms in each run. The histogram flags (HF) ROC shows the average number of *histograms* flagged as anomalous by a given test for a variety of thresholds in good and bad runs, as shown on the left-hand side of Figs. 5 – 6. The run flags (RF) ROC is based on the fraction of *runs* with at least  $N$  histograms (out of the 62) flagged as anomalous, and can be seen on the right-hand side of Figs. 5 – 6. We evaluate the RF ROC for  $N = 1, 3, \text{ and } 5$ , to test both “tight” and “loose” anomaly thresholds.

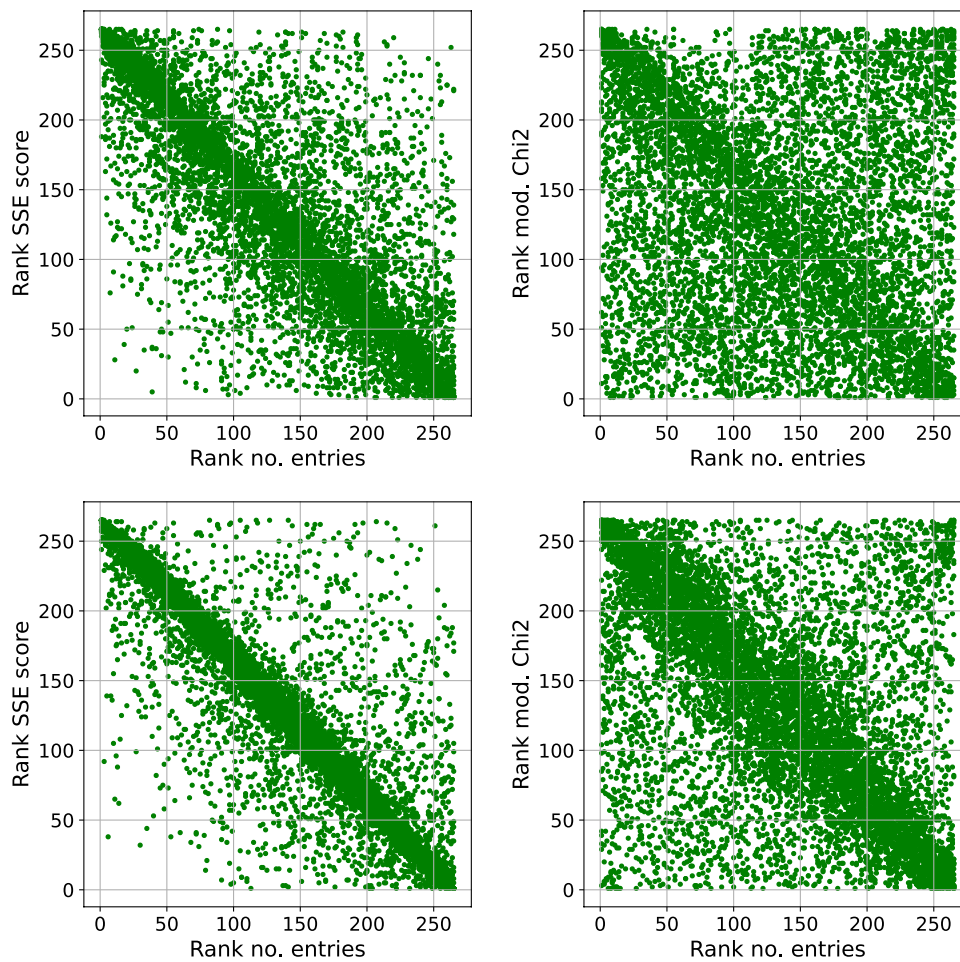
### Assessment results

The beta-binomial and PCA algorithms both show strong discrimination between good and bad runs. When the mean number of histogram flags is low ( $< 3$ ) for good runs, the HF ROC plots show 3 – 4 times more flags in bad runs. The RF ROC plots show similar behavior for the fraction of good and bad runs with  $N \geq 1, 3, \text{ and } 5$  histogram flags: with less than 12% of good runs flagged, 35 – 50% of bad runs exceed the threshold. It is worth noting that we do not expect AutoDQM (or any anomaly detection approach) to identify 100% of bad runs in this data set. In many cases, the relevant issue had no impact on the L1T, or was simply not visible in the online DQM histograms. We also do not expect to achieve a 0% flagging rate for good runs. In fact, many good runs have true anomalies which *should* be flagged. Nevertheless, AutoDQM applied to L1T monitoring alone was able to detect half of the serious issues affecting CMS data quality in 2022, with less than 12% of good runs flagged as anomalous.

The beta-binomial  $\chi^2$  and  $Z'_{max}$  tests perform significantly better when using a larger number of reference runs for the comparison (Fig. 5). This is expected, as reconstructed particle occupancy distributions in the detector are sensitive to the number of simultaneous collisions (“pileup”), so runs taken with different amounts of pileup yield DQM histograms with different shapes. When using multiple reference runs, at least one of them is likely to have similar pileup conditions to the data run being tested. The pileup dependence is taken into account naturally in the PCA algorithms (Fig. 6), which are trained on a large number of runs spanning the full range of pileup conditions.



**Fig. 10** 2D correlation of PCA derived SSE (left),  $\chi^2$  (middle), and  $\chi^2'$  (right) score values to the number of input data histogram entries for 1D (top) and 2D (bottom) LIT DQM histograms from good runs collected by CMS in 2022. The  $\chi^2'$  test has the smallest correlation to the number of entries, and the range of  $\chi^2'$  values is orders of magnitude smaller than the SSE and  $\chi^2$  ranges

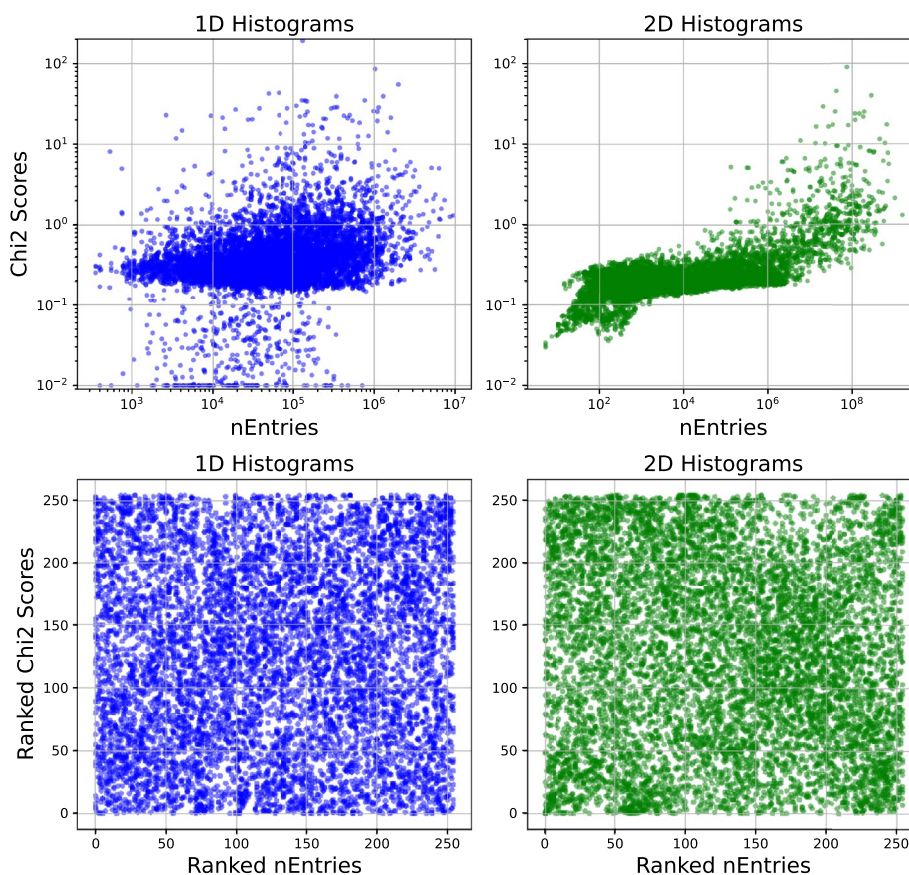


**Fig. 11** 2D correlation of PCA derived SSE (left) and  $\chi^2$ ' (right) score ranks to the rank of the number of input data histogram entries for 1D (top) and 2D (bottom) L1T DQM plots from 265 good runs collected by CMS in 2022. The rank is computed by comparing each histogram to histograms of the same type from other good runs, with 62 histogram types in total. The  $\chi^2$ ' test has much lower correlation to the number of entries than SSE, and high-rank scores appear for histograms with both high- and low-rank number of entries

While there is some variation among the algorithms, none appears to be decisively superior to the others. The best performance is achieved by applying all three quality tests simultaneously (Fig. 7). In this case, the HF ROC plot shows 4 – 6 times more flags in bad runs than in good runs, and over 55% of bad runs have at least 3 flags for a threshold where only 13% of good runs have 3 flags. For the combined tests, the number of distinct *flags* is counted, i.e. if the same histogram is flagged by 2 tests, that counts as 2 anomaly flags.

#### Application to Muon Detector Monitoring

While the L1T online DQM histograms provide a good test case for global AutoDQM performance with inputs from multiple subdetectors, the AutoDQM tool has also been applied successfully in muon detector monitoring. Of the 43 bad runs used in the studies above, only 1 was flagged as bad by PPD due to issues in the muon detectors.

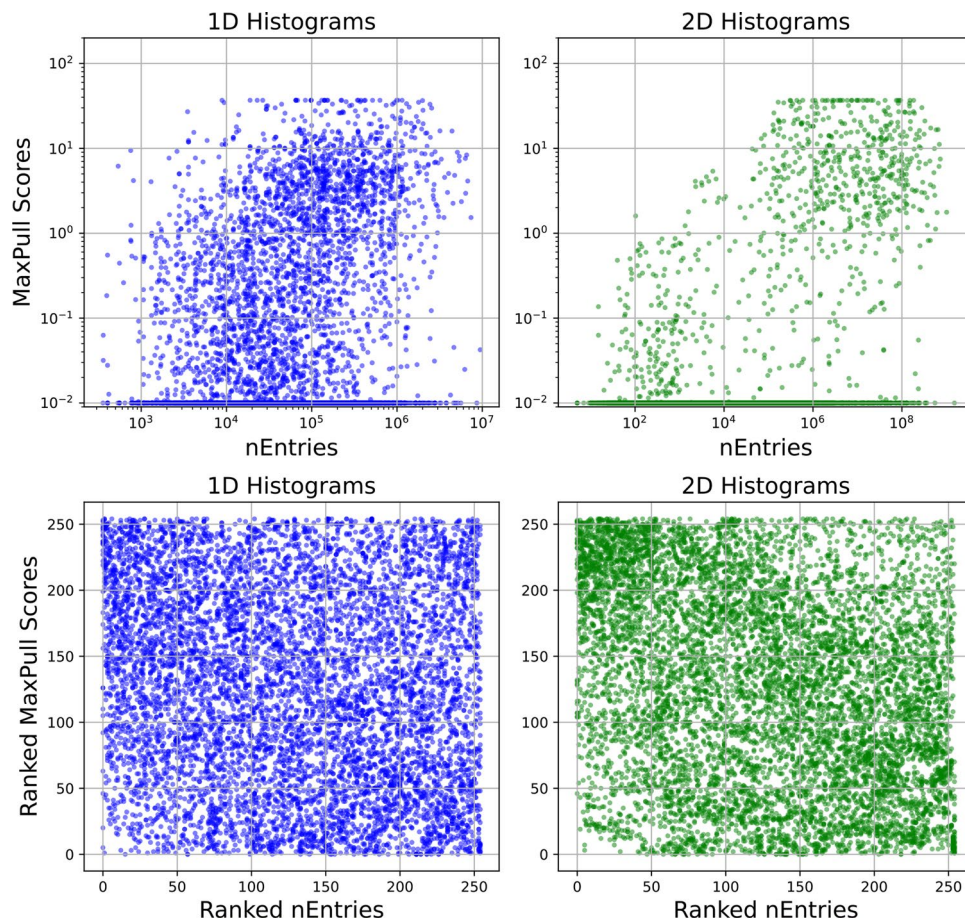


**Fig. 12** 2D correlation of beta-bimomial  $\chi^2$  statistical test score values (top) and score ranks (bottom) to the number of input data histogram entries for 1D (left) and 2D (right) L1T DQM histograms from 265 good runs collected by CMS in 2022. The rank is computed by comparing each histogram to histograms of the same type from other good runs, with 62 histogram types in total. No strong correlation between rank and number of entries is observed

Nevertheless, it is important to closely monitor and identify any significant changes in the muon detector performance, e.g. chambers which occasionally malfunction, requiring expert intervention. The CSCs in the CMS endcaps contain a total of 540 chambers, of which a handful may be disabled at any given time. Very rarely, a dozen or more chambers temporarily malfunction simultaneously. In this case, the AutoDQM webpage flags numerous CSC DQM histograms as anomalous (Fig. 8). Furthermore, the individual AutoDQM plots clearly show the geometrical regions with new deficits of muon tracks in blue (Fig. 9). This allows CSC experts to quickly assess the scope and identify the source of new detector issues, enabling prompt intervention.

### Summary and Outlook

Data quality monitoring (DQM) presents an immense challenge to particle physics experiments, which will only grow as the data collected increases in volume and complexity. The AutoDQM system for generalized, automated DQM dramatically augments the ability of physicists to quickly identify and localize anomalous behavior in the CMS detector. Using a set of monitoring histograms from the CMS Level-1 Trigger system



**Fig. 13** 2D correlation of beta-bimomial maximum single-bin pull ( $Z'_{max}$ ) statistical test score values (top) and score ranks (bottom) to the number of input data histogram entries for 1D (left) and 2D (right) L1T DQM histograms from 265 good runs collected by CMS in 2022. The rank is computed by comparing each histogram to histograms of the same type from other good runs, with 62 histogram types in total. No strong correlation between rank and number of entries is observed

covering the entire CMS proton-proton collision data set from 2022, AutoDQM's combined statistical and machine learning tests successfully identified over 50% of all “bad” data with significant detector malfunction, while flagging less than 15% of “good” data as anomalous. AutoDQM also demonstrates its effectiveness in visually highlighting changes in CMS muon detector performance. Application to additional CMS subdetector systems will allow for more rapid, accurate identification of important issues affecting collision data in the future.

### Appendix A : Anomaly Score Dependence on Histogram Occupancy

In order effectively monitor data quality, anomaly tests should be unbiased with respect to the number of entries in the histogram, which correlates directly to the length of the data-taking run and the collision rate. As shown in Fig. 10 (left), the SSE values (Eq. 2) comparing the normalized input histograms and the PCA transformed output histograms in good runs show a strong linear anti-correlation with the number of entries

in the original data histograms. This means that histograms with lower statistics (from shorter runs) will have a much higher probability to be flagged as anomalous. Comparing the unnormalized input data histogram to a high-statistics scaling of the PCA transformed output histogram using a beta-binomial  $\chi^2$  test (Eqs. 1 and 3), the trend is opposite: only high-statistics histograms get very high  $\chi^2$  scores (Fig. 10 middle). After testing numerous approaches, the modified  $\chi^{2'}$  value (Eq. 4), which is scaled by the number of entries to the  $-1/3$  power, shows the least correlation with the number of entries (Fig. 10 right). This is verified in Fig. 11, where the score rank within a given histogram class (similar to the threshold logic in Sec. 3.2) is plotted against the rank of the number of entries. Almost all of the high-rank SSE scores cluster at low-rank number of entries (left), while high-rank  $\chi^{2'}$  scores are found across the spectrum of ranked number of entries (right). The 1D and 2D histograms plot different quantities related to different types of particles, and the 2D histograms generally have a larger number of bins, so we do not expect their score distributions to be the same.

Figures 12 and 13 show that the beta-binomial  $\chi^2$  and maximum single-bin pull ( $Z'_{max}$  from Sec. 2.1) statistical tests (Eq. 1), performed with 8 reference runs, are not correlated to the number of entries in the data histogram being tested.

## Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1007/s41781-025-00147-2>.

Supplementary file 1.

## Acknowledgements

We congratulate our colleagues in the CERN accelerator departments for the excellent performance of the LHC and thank the technical and administrative staffs at CERN and at other CMS institutes for their contributions to the success of the CMS effort. In addition, we gratefully acknowledge the computing centers and personnel of the Worldwide LHC Computing Grid and other centers for delivering so effectively the computing infrastructure essential to our analyses. Finally, we acknowledge the enduring support for the construction and operation of the LHC, the CMS detector, and the supporting computing infrastructure provided by the following funding agencies: SC (Armenia), BMBWF and FWF (Austria); FNRS and FWO (Belgium); CNPq, CAPES, FAPERJ, FAPERGS, and FAPESP (Brazil); MES and BNSF (Bulgaria); CERN; CAS, MoST, and NSFC (China); Minciencias (Colombia); MSES and CSF (Croatia); RIF (Cyprus); SENESCYT (Ecuador); ERC PRG, RVTT3 and MoER TK202 (Estonia); Academy of Finland, MEC, and HIP (Finland); CEA and CNRS/IN2P3 (France); SRNSF (Georgia); BMBF, DFG, and HGF (Germany); GSRI (Greece); NKFIH (Hungary); DAE and DST (India); IPM (Iran); SFI (Ireland); INFN (Italy); MSIP and NRF (Republic of Korea); MES (Latvia); LMTLT (Lithuania); MOE and UM (Malaysia); BUAP, CINVESTAV, CONACYT, LNS, SEP, and UASLP-FAI (Mexico); MOS (Montenegro); MBIE (New Zealand); PAEC (Pakistan); MES and NSC (Poland); FCT (Portugal); MESTD (Serbia); MCIN/AEI and PCTI (Spain); MOSTR (Sri Lanka); Swiss Funding Agencies (Switzerland); MST (Taipei); MHESI and NSTDA (Thailand); TUBITAK and TENMAK (Turkey); NASU (Ukraine); STFC (United Kingdom); DOE and NSF (USA).

## Author Contributions

This project is a collaborative project, supported by many members of the CMS Collaboration over the years. This effort was coordinated by Andrew Brinkerhoff, Indara Suarez, Chad Freer, and Samuel May. Andrew Brinkerhoff, Chosila Sutantawibul, Indara Suarez, Robert White, Caio Daumann, Jonathan Guiang, Chad Freer, Samuel May, and Bennett Marsh are the main developers of the autoDQM software. Andrew Brinkerhoff, Chosila Sutantawibul, Indara Suarez, Robert White, and Caio Daumann compiled the manuscript. Darin Acosta, Alex Aubuchon, Emanuela Barberis, Aaron Bundock, Claudio Campagnari, Evan Collins, Preston Epps, Johannes Erdmann, Henning Flaecher, Junshen Huang, Vivan Nguyen, Ryan Nie, Sudarshan Paramesvaran, John Rotter, Kaitlin Salyer, Siddhesh Sawant, Tanvi Sheokand, and Darien Wood have completed important studies and helped in the development of the autoDQM software. The CMS Muon Detector Collaboration supported the work by providing experts in each detector subsystem and generous resources for the development. The names of all members of the CMS Muon Detector Collaboration are included in the supplementary material. All authors reviewed the manuscript.

## Data Availability

No datasets were generated or analysed during the current study.

## Declarations

### Competing interests

The authors declare no Conflict of interest.

Received: 11 April 2025 Accepted: 30 September 2025

Published online: 09 February 2026

## References

1. Collaboration CMS (2008) The CMS experiment at the CERN LHC. JINST 3:08004. <https://doi.org/10.1088/1748-0221/3/08/S08004>
2. LHC Study Group (1995) The Large Hadron Collider: Conceptual design. LHC Yellow Book <https://doi.org/10.17181/CERN-AC-95-05-LH>
3. Collaboration CMS (2012) Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC. Phys Lett B 716:30. <https://doi.org/10.1016/j.physletb.2012.08.021>. arXiv:1207.7235
4. Aad G et al (2012) Observation of a new particle in the search for the standard model Higgs boson with the ATLAS detector at the LHC. Phys Lett B 716:1. <https://doi.org/10.1016/j.physletb.2012.08.020>. arXiv:1207.7214
5. Dainese A, et al (eds.) (2019) Report on the Physics at the HL-LHC, and Perspectives for the HE-LHC. CERN Yellow Reports: Monographs, 7/2019. CERN, Geneva, Switzerland. <https://doi.org/10.23731/CYRM-2019-007>
6. Sirunyan AM et al (2017) Particle-flow reconstruction and global event description with the CMS detector. JINST 12:10003. <https://doi.org/10.1088/1748-0221/12/10/P10003>. arXiv:1706.04965 [physics.ins-det]
7. Chatrchyan S et al (2014) Description and performance of track and primary-vertex reconstruction with the CMS tracker. JINST 9:10009. <https://doi.org/10.1088/1748-0221/9/10/P10009>. arXiv:1405.6569 [physics.ins-det]
8. Sirunyan AM et al (2021) Electron and photon reconstruction and identification with the CMS experiment at the CERN LHC. JINST 16(05):05014. <https://doi.org/10.1088/1748-0221/16/05/P05014>. arXiv:2012.06888 [hep-ex]
9. Sirunyan AM et al (2018) Performance of reconstruction and identification of  $\tau$  leptons decaying to hadrons and  $\nu_\tau$  in pp collisions at  $\sqrt{s} = 13$  TeV. JINST 13(10):10005. <https://doi.org/10.1088/1748-0221/13/10/P10005>. arXiv:1809.02816 [hep-ex]
10. Khachatryan V et al (2017) Jet energy scale and resolution in the CMS experiment in pp collisions at 8 TeV. JINST 12:02014. <https://doi.org/10.1088/1748-0221/12/02/P02014>. arXiv:1607.03663 [hep-ex]
11. Sirunyan AM et al (2018) Performance of the CMS muon detector and muon reconstruction with proton-proton collisions at  $\sqrt{s} = 13$  TeV. JINST 13:06015. <https://doi.org/10.1088/1748-0221/13/06/P06015>. arXiv:1804.04528 [physics.ins-det]
12. Sirunyan AM et al (2020) Performance of the CMS Level-1 Trigger in proton-proton collisions at  $\sqrt{s} = 13$  TeV. JINST 15:10017. <https://doi.org/10.1088/1748-0221/15/10/P10017>. arXiv:2006.10165 [hep-ex]
13. Collaboration CMS (2019) Performance of the CMS High Level Trigger during LHC Run-2. JINST 19:11021. <https://doi.org/10.1088/1748-0221/19/11/P11021>. arXiv:2410.17038
14. Azzolini V et al (2019) The Data Quality Monitoring software for the CMS experiment at the LHC: past, present and future. EPJ Web Conf 214:02003. <https://doi.org/10.1051/epjconf/201921402003>
15. Borisyak M, Ratnikov F, Derkach D, Ustyuzhanin A (2017) Towards automation of data quality system for CERN CMS experiment. J. Phys. Conf. Ser. **898**(9) <https://doi.org/10.1088/1742-6596/898/9/092041> arXiv:1709.08607
16. Pol AA, Cerminara G, Germain C, Pierini M, Seth A (2018) Detector monitoring with artificial neural networks at the CMS experiment at the CERN Large Hadron Collider. Comput. Softw. Big Sci. <https://doi.org/10.48550/arXiv.1808.00911> arXiv:1808.00911
17. CMS ECAL Collaboration (2024) Autoencoder-based Anomaly Detection System for Online Data Quality Monitoring of the CMS Electromagnetic Calorimeter. Comput. Softw. Big Sci. **8**(11) <https://doi.org/10.1007/s41781-024-00118-zarXiv:2309.10157>
18. Asres M (2023) the CMS HCAL Collaboration, et al Spatio-temporal anomaly detection with graph networks for data quality monitoring of the hadron calorimeter. Sensors 23(24):9679. <https://doi.org/10.3390/s23249679>. arXiv:2311.04190
19. SciPy-Contributors: SciPy beta-binomial implementation, accessed 2025-01-23. Accessed: 2025-01-23. <https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.betabinom.html>
20. Pearson K (1901) On lines and planes of closest fit to systems of points in space. Lond Edinb Dublin Philos Mag J Sci 2(11):559–572. <https://doi.org/10.1080/14786440109462720>
21. Pedregosa F et al (2011) Scikit-learn: Machine learning in Python. J Mach Learn Res 12:2825–2830 arXiv:1201.0490
22. Glorot X, Bordes A, Bengio Y (2011) Deep sparse rectifier neural networks. In: Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, 15, 315–323. PMLR, ??? <https://proceedings.mlr.press/v15/glorot11a.html>

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.