Machine Learning Tools for the CMS Tracker Data Quality Monitoring and Certification

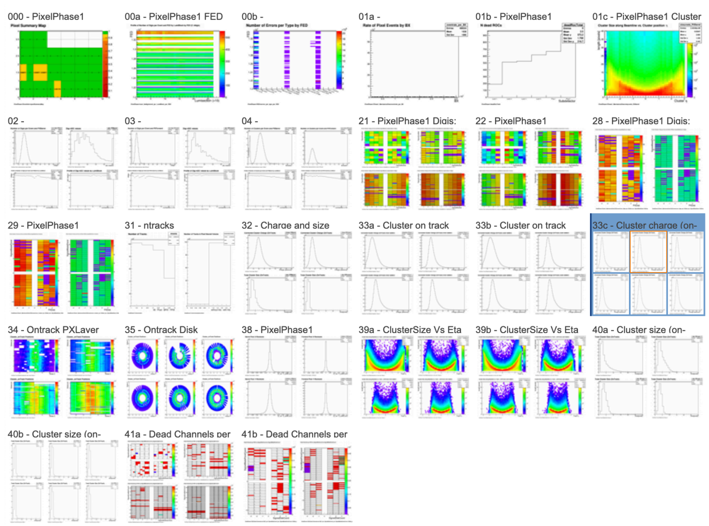
Gabriele Benelli (Brown University) for the CMS Collaboration

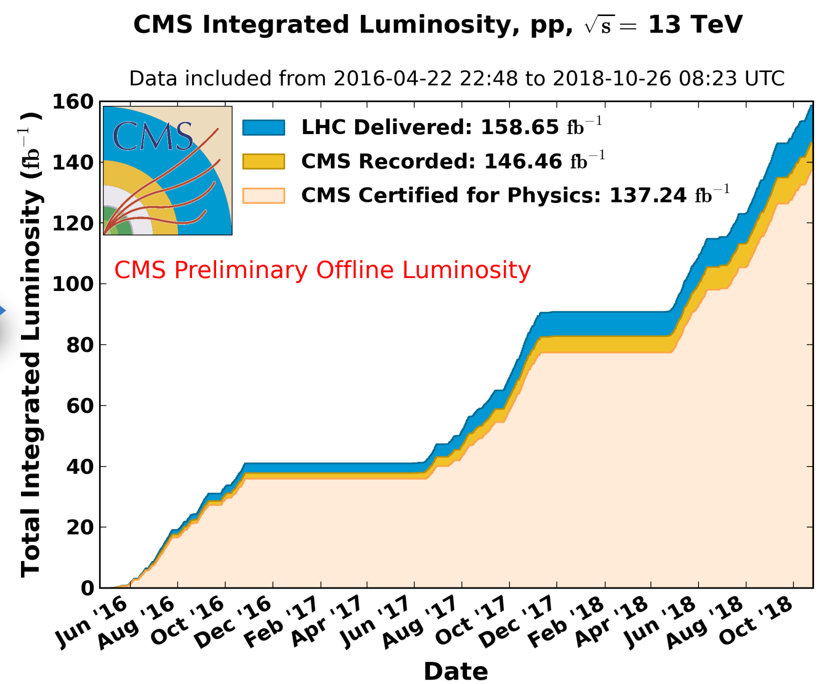
Outline

- A brief introduction to CMS Data Quality Monitoring (DQM) and Data Certification (DC)
  - Online and Offline DQM
  - Challenges
- Machine Learning Tools for DQM/DC
  - Taking on the challenges with anomaly detection
  - Dedicated dataset
  - Data exploration and integration
  - ML studies
- Prospects and Summary

CMS Data Quality Monitoring and Certification

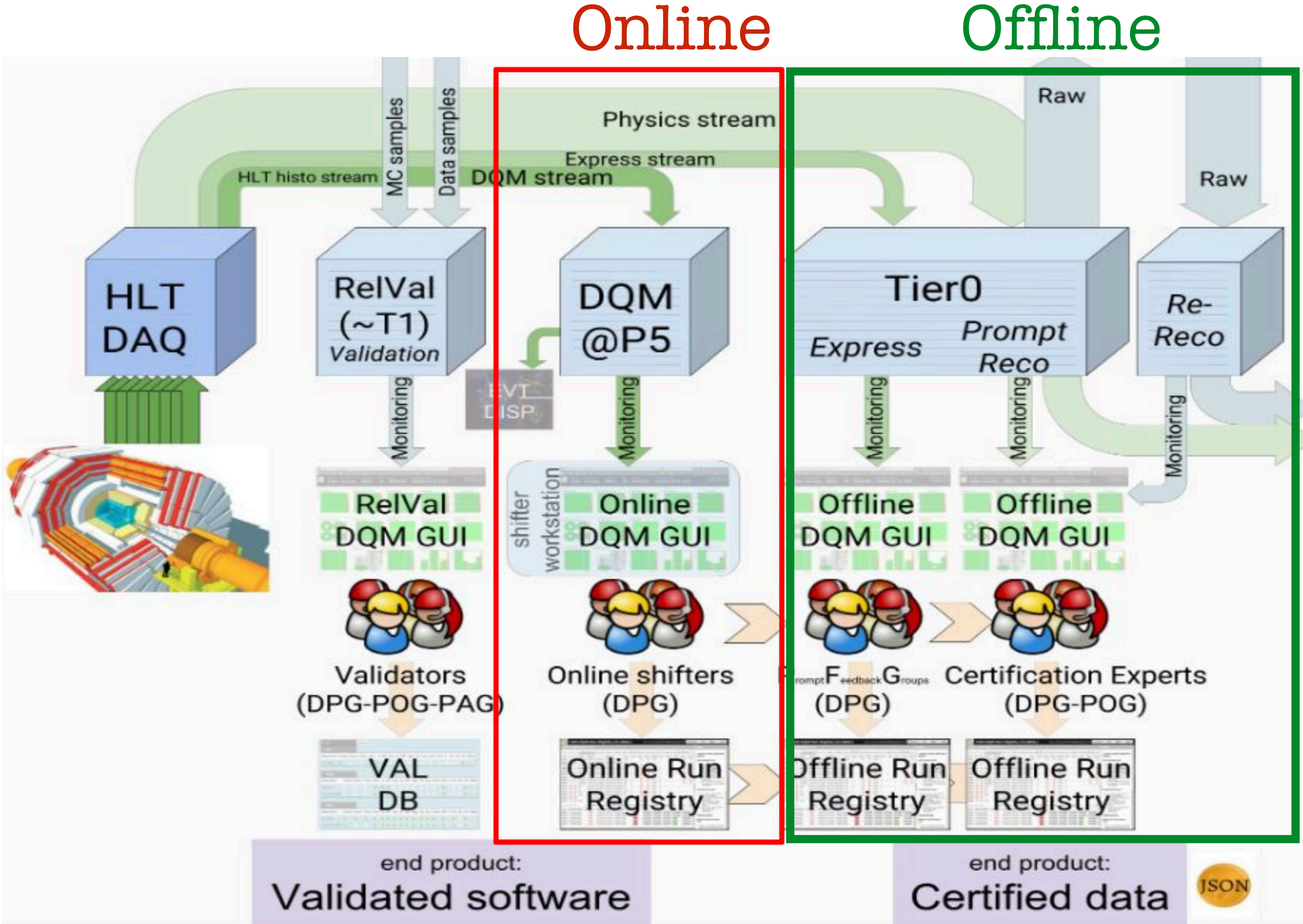
- Ultimate goal: highest data quality for CMS publications
- Shifters, on-call personnel and experts scrutinize hundreds of histograms for each of the CMS subsystems to provide data quality (sub)flag(s)





- Typical run lasts several hours and made of several hundreds of lumi-sections (LS), the atomic unit of data-taking, lasting about 23 seconds
- For offline data certification, DQM histograms harvested with 'run' granularity
- Beam and detector data sources complement the information from DQM histograms

CMS DQM System



- Two offline data reconstruction tiers:
  - Express (available within a few hours) used for calibrations
  - PromptReco (available within 48hrs) used for physics analysis

CMS DQM and DC processes

- Online DQM: monitor detector performance live to ensure efficient data collection
  - Reduced set of histograms, central shifters, real-time reaction time
- **Offline DQM**: data certification to ensure good quality for physics analysis
  - Full set of histograms
  - Each subsystem handles certification for their flag
  - Additional sources of data (beam and detector status, calibrations)

- Challenges:
  - Large number of people involved
  - Large number of histograms and other data to inspect and correlate
  - Human error
  - **Time granularity**
  - Changing beam and detector(s) conditions
  - Anomalies may be unexpected

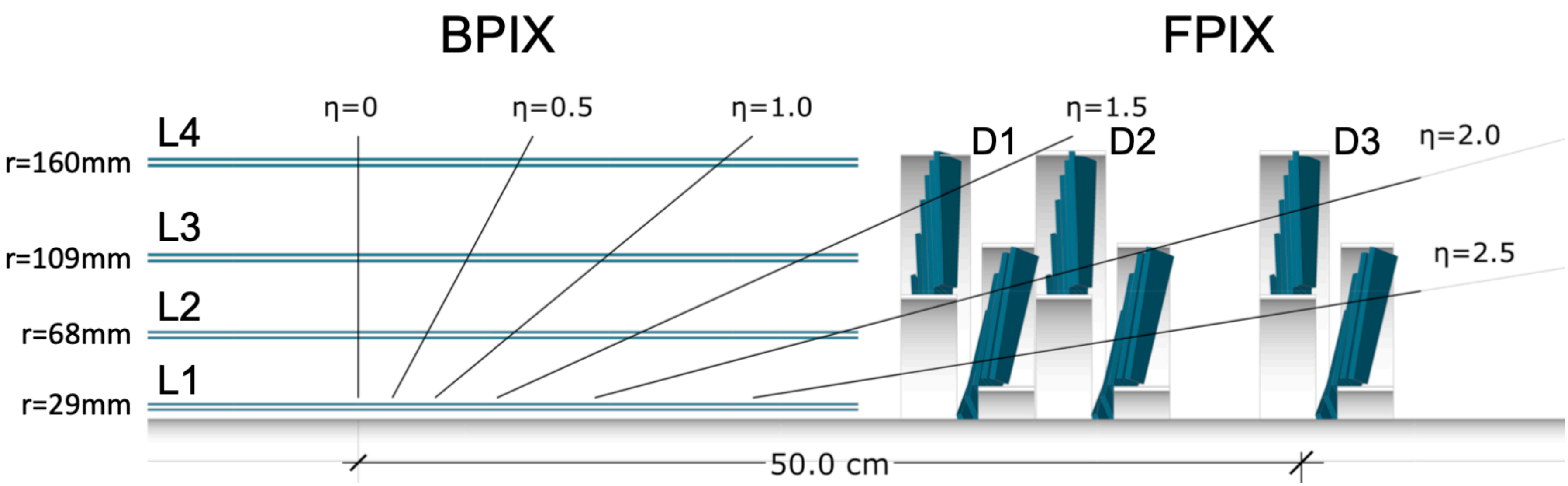
Machine Learning for DQM/DC

- Explore the use of ML for anomaly detection, automating part of the process
  - **Tools to facilitate certification** tasks for shifters and experts
  - **Automate** the evaluation of DQM histograms
    - Alarms/flags
    - Provide outputs robust against changing conditions and low statistics
    - Enable scaling to larger number of histograms
    - Standardize certification easing human interface issues
  - **Leverage other data sources integration and anomaly detection techniques**
  - Provide algorithmic **insight** to experts

- Basic anomaly detection idea:
  - "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

Machine Learning for DQM/DC

- **Offline dataset**:
  - Run 2 data (specifically **2017** and 2018, after the Pixel Phase-1 upgrade installation)
  - PromptReco (ReReco) datasets for **Pixel**, Strip, Tracking, Muon and JetMET certification
  - Subset of histograms (~100)
  - Highest time granularity DQM harvesting: per lumi-section (LS), ~23 seconds
- Several studies were conducted on the tracker dataset:
  - Data exploration and integration at the LS level
  - Principal Component Analysis (PCA), t-distributed Stochastic Neighbor Embedding (t-SNE), Convolutional Neural Networks (CNNs), **1D AutoEncoders, 1D Non-negative Matrix Factorization (NMF), 2D Residual Network (ResNet) AutoEncoders**
- The following featured studies are all based on **pixel (Phase-I)** [1] tracker inputs

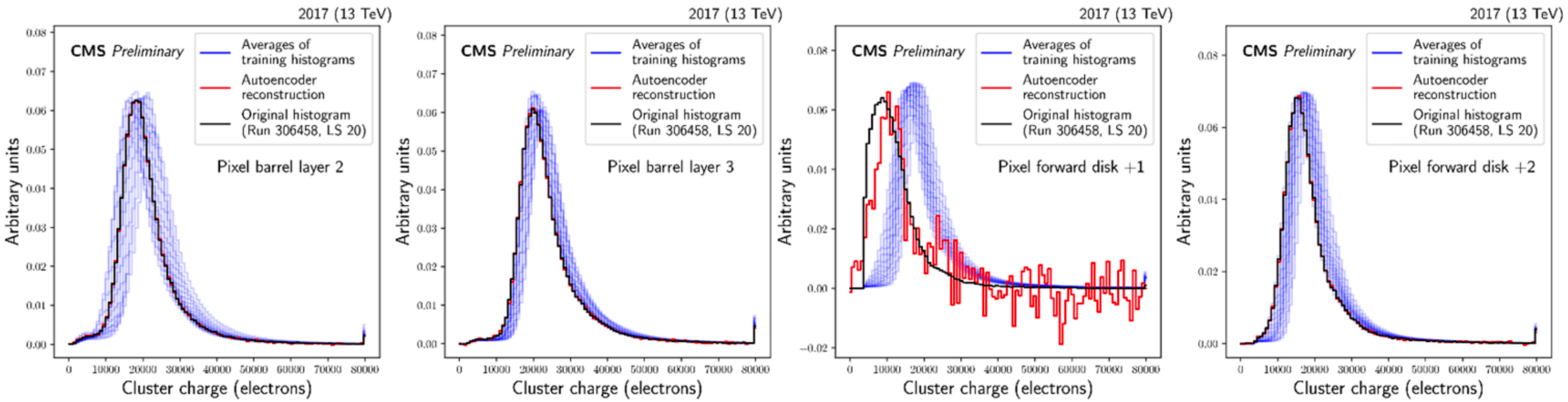


Data Exploration and Integration

- **Inputs**:
  - CSV files with **DQM histograms** bin contents, run number/LS number
  - Data integration of **additional sources**
    - Luminosity, pileup, trigger rate, beam conditions
    - Certification information
      - Official certification flags (**AND** of all subsystems/histos, perLS)
      - Tracker certification information (reference runs, problem classification)
- Common tools to streamline access to data, facilitate development and data exploration with Jupyter notebooks (SWAN+EOS at CERN)
- Caveats about training datasets (statistics, homogeneity)
  - Lack of **histogram-level** labelled data:
    - Preference for **semi-supervised/unsupervised approach**
  - Resampling tool developed to address scarcity of data flagged "bad"
    - Use certification information to extract "bad" data
    - Key for both unsupervised and supervised approaches
- Standardize:
  - Training and testing of algorithms
  - Handling of inputs (data cleaning)
  - Anomaly definition and reporting
  - MSE of the given input histogram with a reference run used as ML performance benchmark

Pixel 1D Histograms AutoEncoder

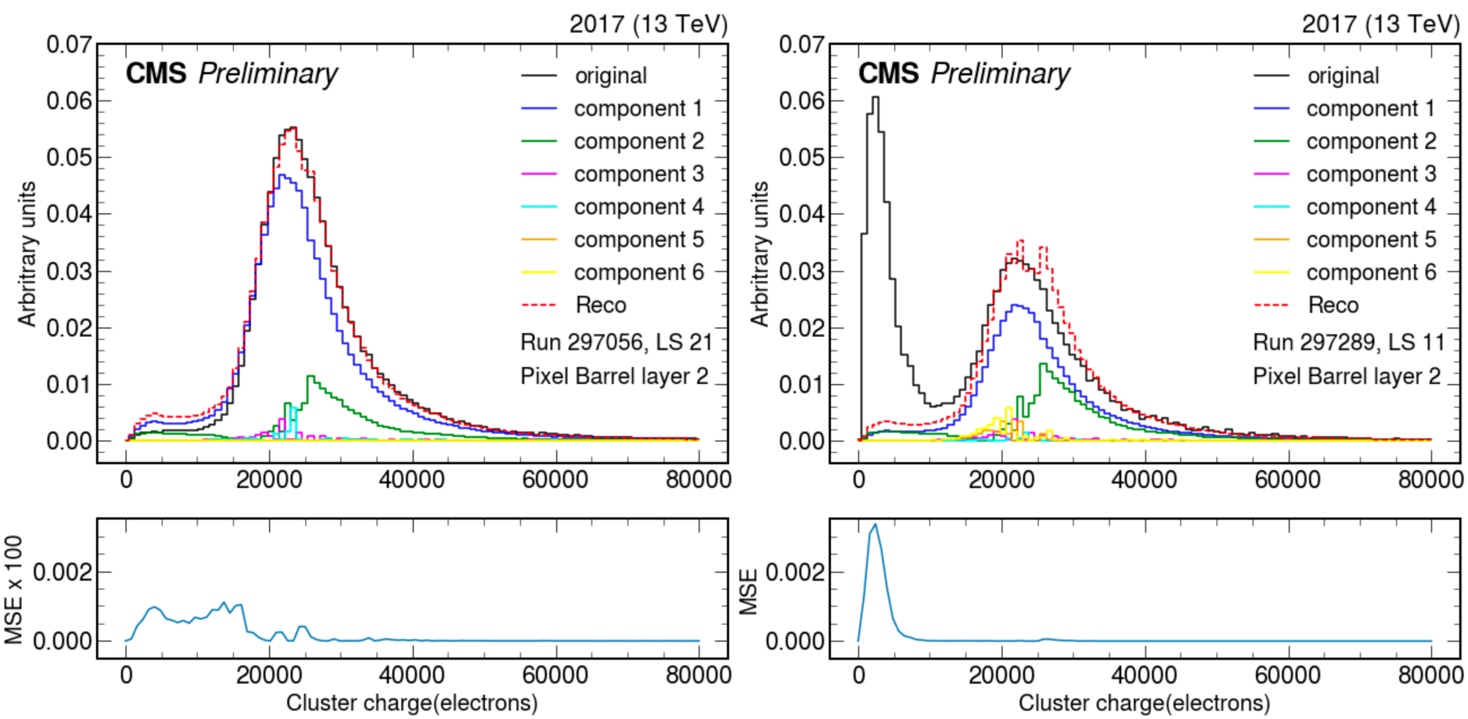
- Input histograms: **cluster charge in barrel and forward pixel layers** [2]
- AutoEncoders trained separately on each histogram type:
  - One hidden layer with half as many nodes as input bins ($100 \rightarrow 50$)
  - Training on all lumi-sections from 2017, with filters (HV on, minimal statistics)



- Changing detector/beam conditions produce a **spectrum of shape variations**
- Anomalous lumi-sections **input** histograms get badly **reconstructed** resulting in large MSE
- Anomalies may affect individual histograms, in this case an on-going HV scan for pixel forward disk +1, and at different lumi-sections during a given run
- Multiple histograms MSEs may be combined to flag anomalies

Pixel 1D Histograms NMF

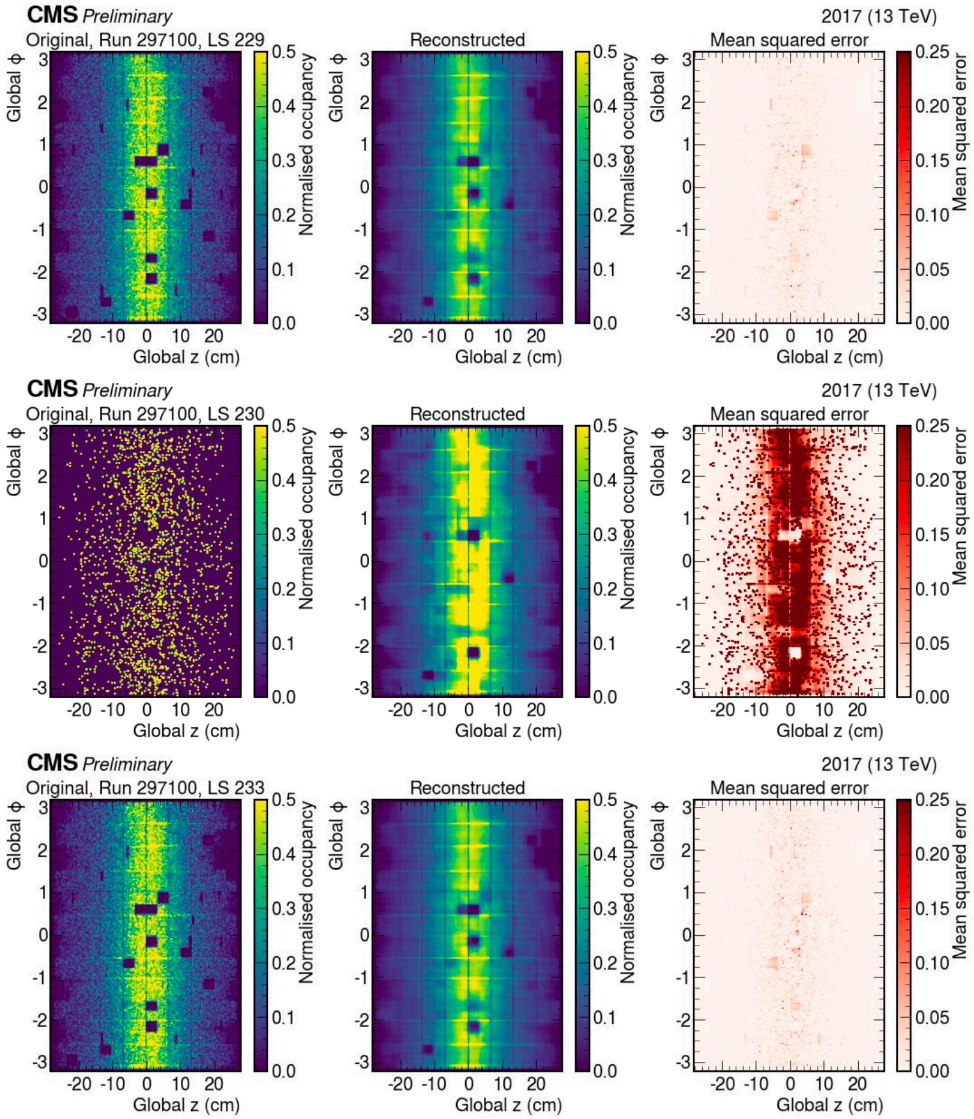
- Input histogram: **cluster charge in barrel pixel layer 2** [2]
- NMF [4] algorithm trained using all LSs from reference run used for human certification:
  - 6-component model
  - Two individual LSs are reconstructed as a linear combination of the 6 components



- Anomalous lumi-sections **input histograms** cannot be **reconstructed** satisfactorily resulting in large MSE (note the factor 100 scale difference in the lower MSE panels)
- The anomalous LS in the right plot is due to a pixel timing scan affecting barrel layer 2
- NMF component contribution provides extra insight that could be used by experts and in further classification of anomalies

Pixel 2D Histograms ResNet AutoEncoder

- Input histogram: **2D pixel occupancy in barrel pixel layer 1** [2]
- ResNet [5] used as AutoEncoder, trained on 2017 data from same data-taking era with statistics preselection:



- 200x140 input, 4 encoding blocks with increasing number of filters, hidden layer with 1000 nodes, 4 decoding blocks with decreasing number of filters, 200x140 output
- Pixel barrel layer 1 occupancy for 3 different lumi-sections (one per row) in the same run
- First column: input histograms
- Second column: ResNet reconstructed histograms
- Third column: MSE between input and reconstructed (darker red for larger MSE)
- ResNet not able to reconstruct anomalous LS input histogram, resulting in large MSE
- Anomaly due to transient lower trigger rate

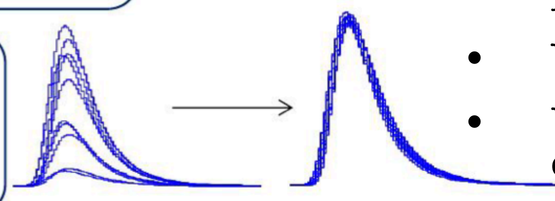
ML for DC method illustration

[3]

- Training, Good and Anomalous test sets all obtained from the 2017 perLS dataset
- No reliable labels available per histogram perLS
- Very few bad LSs (used with resampling)
- Used MSE of individual LSs to filter outliers for Training and Good test sets

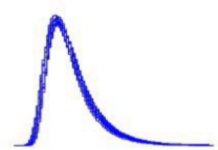
Inputs
- Set of histograms per luminosity section.
- Cluster charge distributions split by region: BPIX L2, L3, L4, FPIX- D1, D2, D3, FPIX+ D1, D2, D3.

Preprocessing
- Normalization.
- Filter on detector status.
- Filter on sufficient number of entries.

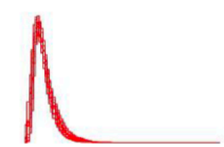


Training set
- Large body of unlabeled data.

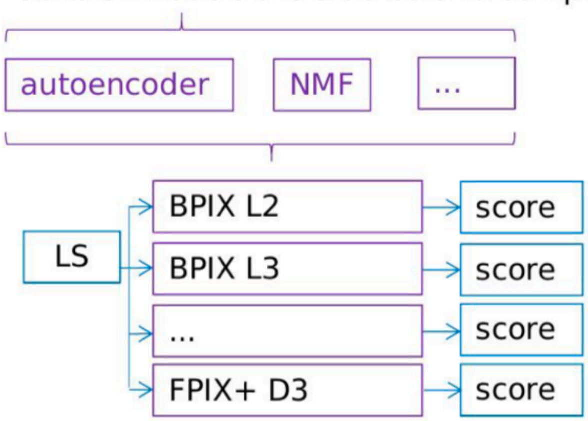
Good test set
- Additional filters.



Anomalous test set
- Predefined bad runs.
- Resampled to increase statistics.

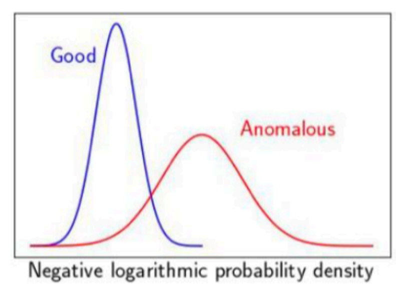


Single histogram models
- Unsupervised anomaly detection methods.
- Several models are studied and compared.

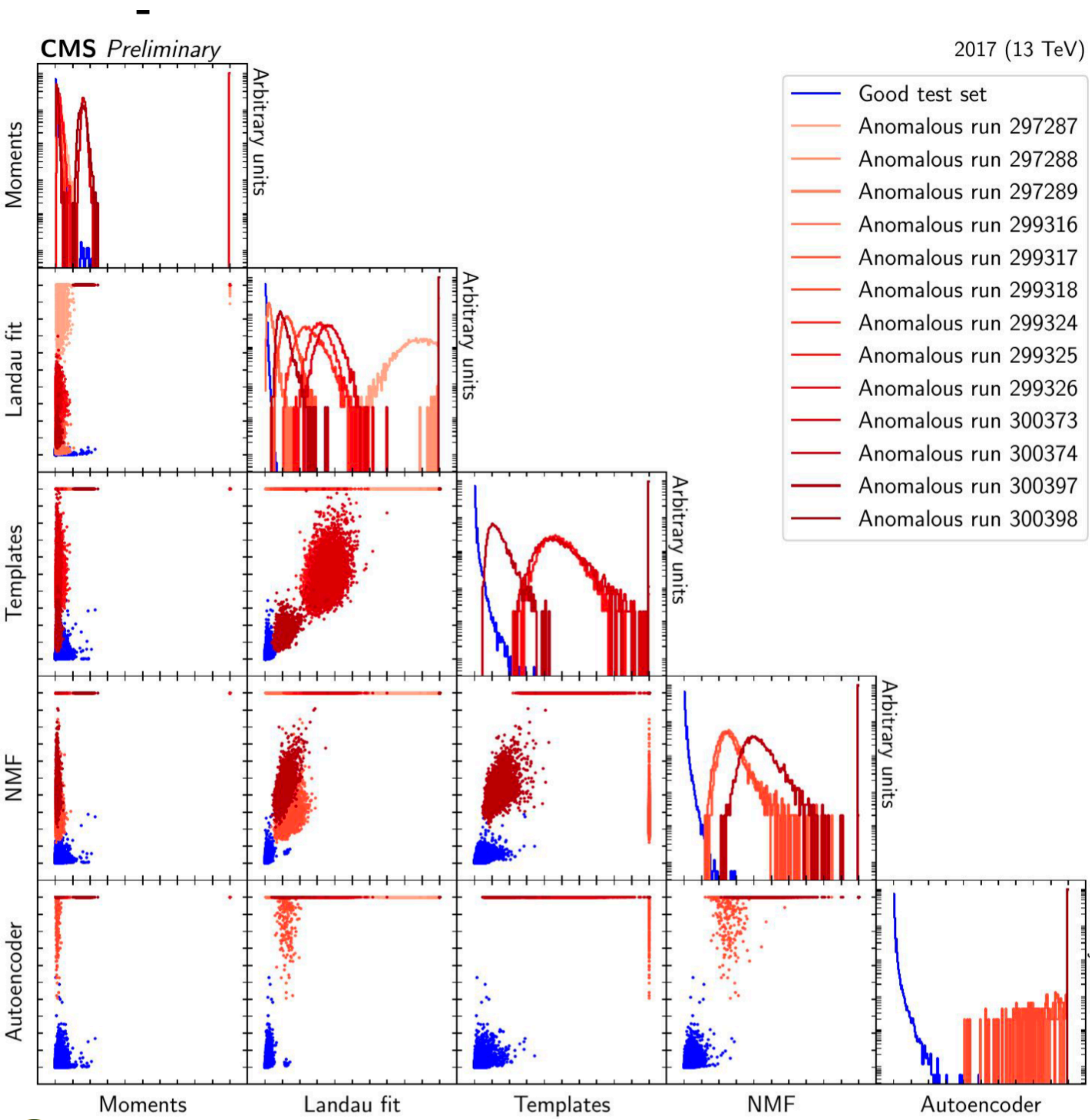


Combination
- Fit multidimensional probability density to scores for training set.
- Low probability density indicates anomaly

Testing and results



ML/statistical methods comparison

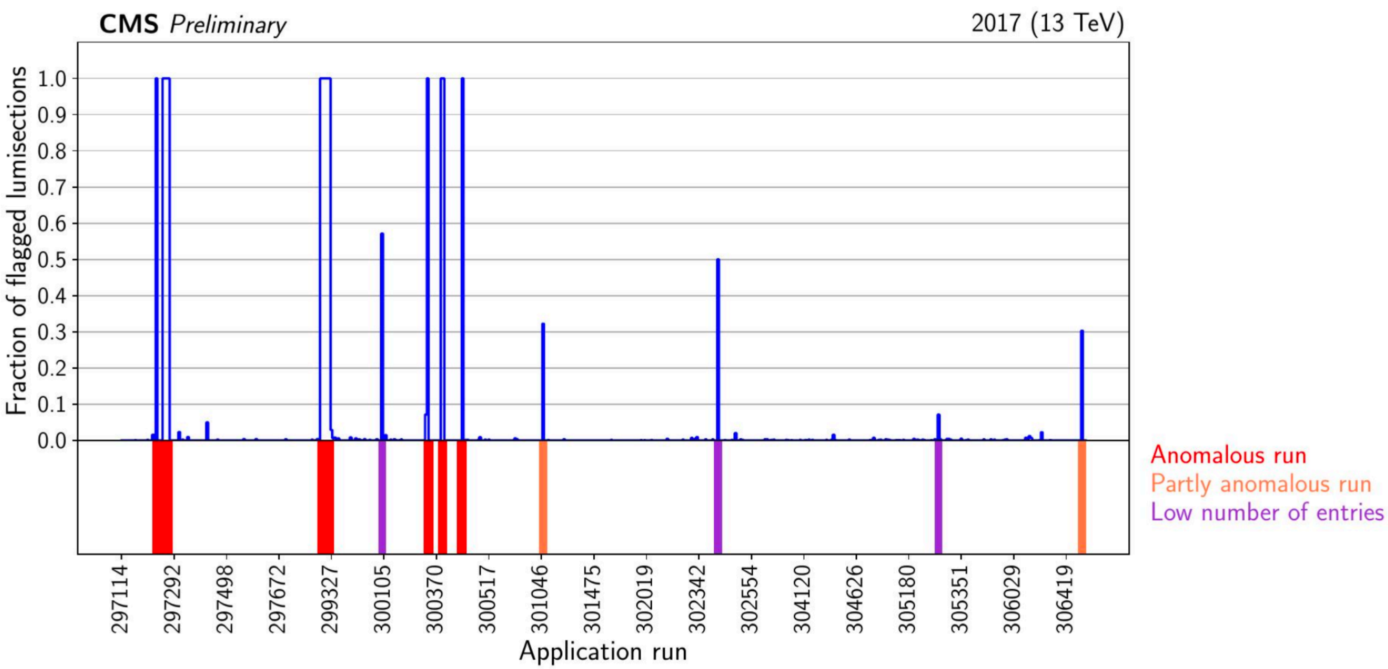


- Nine cluster charge (BPIX L2, L3, L4, FPIX+D1,D2,D3, FPIX-D1,D2,D3) perLS 1-D histograms input [3]
- Moments: 1^{st} and 2^{nd} moments w.r.t. average and standard deviation of training dataset
- Landau fit: MSE between input and best fit
- Templates: minimum MSE between input and a set of reference histograms
- NMF: MSE between input and NMF reconstruction
- AutoEncoder: MSE between input and AutoEncoder reconstruction

- Diagonal plots: global score for **good test set** and **anomalous runs** for all models
- Off-diagonal plots: correlations between models (one LS per dot, scores normalized between 0-1)
- Sensitivity vs. fake rate can be tuned and optimized

Operational performance example

- A variation to the ML for DC method illustrated above (same input histograms), reflecting the situation during regular data-taking [3]:
  - Model is updated with dedicated training on LSes immediately preceding the run whose LSes are under consideration



- Fraction of flagged LSs is low for good runs
- **Anomalous runs** typically are runs with HV bias or timing scans (most of the time affecting globally the whole, at times only **partially**)
- **Low luminosity or low trigger rate runs** typically results in flagged LSs

- First runs in a fill not included in the study above,
- Studies conducted on exploiting reference runs (based on instantaneous luminosity, beams and detector conditions) to get homogeneous training datasets

Prospects and Summary

- Very promising studies for ML tools to assist data quality monitoring and certification shifters and experts
- ML may help address some of the challenges that make DQM and DC such a labour intensive process
- Several ongoing efforts:
  - Streamline access to various sources of data
  - Data exploration and data cleaning
  - Facilitate ML training and testing
  - Streamline metrics for comparisons and optimization
  - Extend studies to other inputs
  - Extend from semi-supervised to fully supervised approaches (from anomaly detection to classification)
  - Integrate some of these tools as aids to the certification team and later as aids to online data quality monitoring
- Prospects to deploy some of these tools together with DQM perLS harvesting in the ongoing Run3 to improve DC and DQM

References

[1] CMS Collaboration, "**The CMS Phase-1 Pixel Detector Upgrade**", CERN-CMS-NOTE-2020-005, https://cds.cern.ch/record/2745805

[2] CMS Collaboration, "**Tracker DQM Machine Learning studies for data certification**", CERN-CMS-DP-2021-034, https://cds.cern.ch/record/2799472

[3] CMS Collaboration, "**Prospects for computer-assisted data quality monitoring at the CMS pixel detector**", CERN-CMS-DP-2022-013, http://cds.cern.ch/record/2812026

[4] Lee, D.D. and Seung, H.S., "**Algorithms for Non-negative Matrix Factorization**", **Adv. Neural Inf. Process. Syst.13556-562 (2000)**

[5] K. He, X. Zhang, S. Ren, and J. Sun, "**Deep Residual Learning for Image Recognition**", https://arxiv.org/abs/1512.03385

Back-up

CMS Run Registry

Run Registry | ONLINE | OFFLINE | ML | JSON PORTAL | LOG — Gabriele Benelli | Log out

WORKSPACES: GLOBAL, BTAG, CASTOR, CMS, CSC, CTPPS, DC, DT, ECAL, EGAMMA, GEM, HCAL, HLT, JETMET, L1T, LUMI, MUON, RPC, TAU, TRACKER

Offline / Datasets / Global /

Show Cycles

Feedback is welcome! (JIRA) | Configuration

Datasets waiting to appear in DQM GUI (1096):

Export to CSV | Get API Call | Open in HDQM | Open in GUI

AND | +Rule | +Group

Run Number	Dataset Name	Class	Manage / LS	GUI Link	OMS	global State	LS Duration	btag	castor	cms	csc	ctpps	lowlumi	dt	ecal	es	egamma	hcal	hlt	jetmet	l1t	l1tmu	l1tcalo	lumi	muon	rpc	tau	track	pixel	strip	gem
367132	/PromptReco/C...	Collisions23	Manage / LS	Not in GUI yet	OMS	waiting dqm gui / move	4	NOTSET	EXCLUDED	GOOD	GOOD	GOOD	NO VALUE F	GOOD	GOOD	GOOD	NOTSET	GOOD	GOOD	NOTSET	GOOD	NOTSET	NOTSET	GOOD	NOTSET	GOOD	NOTSET	NOTSET	GOOD	GOOD	GOOD
367131	/PromptReco/C...	Collisions23	Manage / LS	Not in GUI yet	OMS	waiting dqm gui / move	152	NOTSET	EXCLUDED	B G	GOOD	GOOD	NO VALUE F	GOOD	GOOD	GOOD	NOTSET	GOOD	GOOD	NOTSET	GOOD	NOTSET	NOTSET	GOOD	NOTSET	GOOD	NOTSET	NOTSET	GOOD	GOOD	GOOD
367130	/PromptReco/C...	Collisions23	Manage / LS	Not in GUI yet	OMS	waiting dqm gui / move	172	NOTSET	EXCLUDED	GOOD	GOOD	GOOD	NO VALUE F	GOOD	GOOD	GOOD	NOTSET	GOOD	GOOD	NOTSET	GOOD	NOTSET	NOTSET	GOOD	NOTSET	GOOD	NOTSET	NOTSET	GOOD	GOOD	GOOD
367129	/PromptReco/C...	Collisions23	Manage / LS	Not in GUI yet	OMS	waiting dqm gui / move	10	NOTSET	EXCLUDED	GOOD	GOOD	GOOD	NO VALUE F	GOOD	GOOD	GOOD	NOTSET	GOOD	GOOD	NOTSET	GOOD	NOTSET	NOTSET	GOOD	NOTSET	GOOD	NOTSET	NOTSET	GOOD	GOOD	GOOD
367128	/PromptReco/C...	Collisions23	Manage / LS	Not in GUI yet	OMS	waiting dqm gui / move	20	NOTSET	EXCLUDED	B G	GOOD	GOOD	NO VALUE F	GOOD	GOOD	GOOD	NOTSET	GOOD	GOOD	NOTSET	GOOD	NOTSET	NOTSET	GOOD	NOTSET	GOOD	NOTSET	NOTSET	GOOD	GOOD	GOOD

Previous | Page 1 of 220 | 5 rows | Next

Editable datasets (31347):

Export to CSV | Get API Call | Open in HDQM | Open in GUI

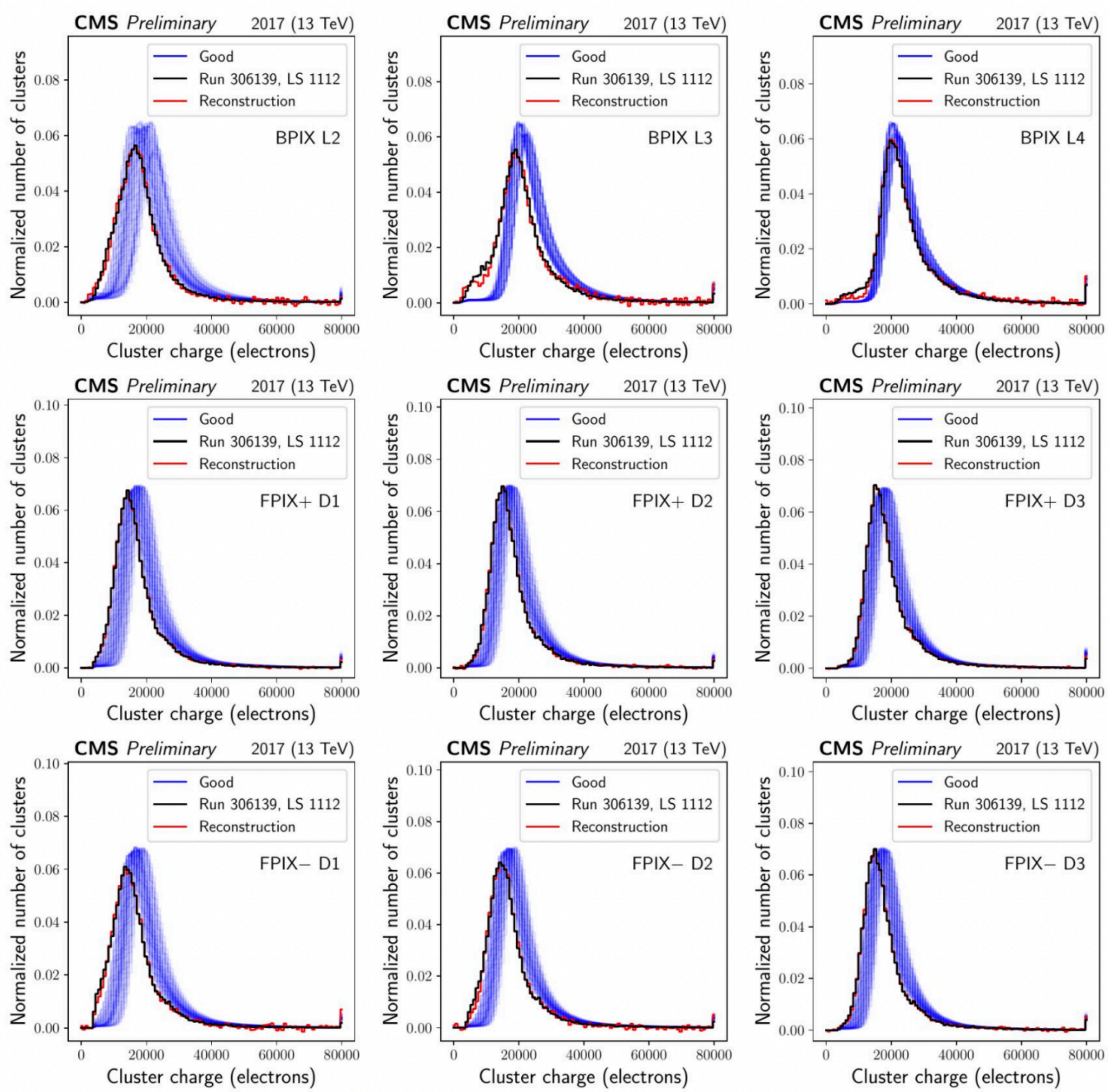
AND | +Rule | +Group

Run Number	Dataset Name	Class	Manage / LS	GUI Link	OMS	global State	LS Duration	btag	castor	cms	csc	ctpps	lowlumi	dt	ecal	es	egamma	hcal	hlt	jetmet	l1t	l1tmu	l1tcalo	lumi	muon	rpc	tau	track	pixel	strip	gem
366956	/PromptReco/C...	Commissioning2	Manage / LS	GUI Link	OMS	OPEN / move	13	NOTSET	EXCLUDED	BAD	GOOD	STANDBY	NO VALUE F	GOOD	EXCLUDED	EXCLUDED	NOTSET	GOOD	GOOD	NOTSET	GOOD	NOTSET	NOTSET	GOOD	NOTSET	GOOD	NOTSET	NOTSET	STANDBY	STANDBY	GOOD
366953	/PromptReco/C...	Commissioning2	Manage / LS	GUI Link	OMS	OPEN / move	37	NOTSET	EXCLUDED	BAD	GOOD	STANDBY	NO VALUE F	GOOD	EXCLUDED	EXCLUDED	NOTSET	GOOD	GOOD	NOTSET	GOOD	NOTSET	NOTSET	GOOD	NOTSET	GOOD	NOTSET	NOTSET	STANDBY	STANDBY	GOOD
366936	/PromptReco/C...	Cosmics23	Manage / LS	GUI Link	OMS	OPEN / move	40	NOTSET	EXCLUDED	BAD	EXCLUDED	STANDBY	NO VALUE F	GOOD	GOOD	GOOD	NOTSET	GOOD	GOOD	NOTSET	GOOD	NOTSET	NOTSET	GOOD	NOTSET	EXCLUDED	NOTSET	NOTSET	GOOD	GOOD	EXCLUDED
366933	/PromptReco/C...	Collisions23	Manage / LS	GUI Link	OMS	OPEN / move	1325	NOTSET	EXCLUDED	B G	GOOD	G S	NO VALUE F	G S	GOOD	GOOD	NOTSET	GOOD	GOOD	NOTSET	GOOD	NOTSET	NOTSET	GOOD	NOTSET	G S	NOTSET	GOOD	G S	G S	GOOD
366932	/PromptReco/C...	Commissioning2	Manage / LS	GUI Link	OMS	OPEN / move	47	NOTSET	EXCLUDED	BAD	STANDBY	STANDBY	NO VALUE F	STANDBY	GOOD	GOOD	NOTSET	GOOD	GOOD	NOTSET	GOOD	NOTSET	NOTSET	GOOD	NOTSET	STANDBY	NOTSET	NOTSET	STANDBY	STANDBY	STANDBY
366931	/PromptReco/C...	Commissioning2	Manage / LS	GUI Link	OMS	OPEN / move	181	NOTSET	EXCLUDED	BAD	STANDBY	STANDBY	NO VALUE F	STANDBY	GOOD	GOOD	NOTSET	GOOD	GOOD	NOTSET	GOOD	NOTSET	NOTSET	GOOD	NOTSET	STANDBY	NOTSET	NOTSET	STANDBY	STANDBY	STANDBY
366928	/PromptReco/C...	Commissioning2	Manage / LS	GUI Link	OMS	OPEN / move	21	NOTSET	EXCLUDED	BAD	G S	STANDBY	NO VALUE F	STANDBY	GOOD	GOOD	NOTSET	GOOD	GOOD	NOTSET	GOOD	NOTSET	NOTSET	GOOD	NOTSET	STANDBY	NOTSET	NOTSET	STANDBY	STANDBY	STANDBY
366927	/PromptReco/C...	Cosmics23	Manage / LS	GUI Link	OMS	OPEN / move	35	NOTSET	EXCLUDED	BAD	GOOD	STANDBY	NO VALUE F	G S	GOOD	GOOD	NOTSET	GOOD	GOOD	NOTSET	GOOD	NOTSET	NOTSET	GOOD	NOTSET	G S	NOTSET	NOTSET	G S	G S	G S
366926	/PromptReco/C...	Cosmics23	Manage / LS	GUI Link	OMS	OPEN / move	27	NOTSET	EXCLUDED	BAD	GOOD	STANDBY	NO VALUE F	GOOD	GOOD	GOOD	NOTSET	GOOD	GOOD	NOTSET	GOOD	NOTSET	NOTSET	GOOD	NOTSET	GOOD	NOTSET	NOTSET	GOOD	GOOD	GOOD
366925	/PromptReco/C...	Cosmics23	Manage / LS	GUI Link	OMS	OPEN / move	14	NOTSET	EXCLUDED	BAD	GOOD	STANDBY	NO VALUE F	GOOD	GOOD	GOOD	NOTSET	GOOD	GOOD	NOTSET	GOOD	NOTSET	NOTSET	GOOD	NOTSET	GOOD	NOTSET	GOOD	GOOD	GOOD	GOOD
366923	/PromptReco/C...	Cosmics23	Manage / LS	GUI Link	OMS	OPEN / move	18	NOTSET	EXCLUDED	BAD	GOOD	STANDBY	NO VALUE F	GOOD	GOOD	GOOD	NOTSET	GOOD	GOOD	NOTSET	GOOD	NOTSET	NOTSET	GOOD	NOTSET	GOOD	NOTSET	BAD	G S	G S	GOOD
366922	/PromptReco/C...	Commissioning2	Manage / LS	GUI Link	OMS	OPEN / move	315	NOTSET	EXCLUDED	BAD	GOOD	STANDBY	NO VALUE F	GOOD	GOOD	GOOD	NOTSET	GOOD	GOOD	NOTSET	GOOD	NOTSET	NOTSET	GOOD	NOTSET	GOOD	NOTSET	NOTSET	STANDBY	STANDBY	GOOD
366921	/PromptReco/C...	Commissioning2	Manage / LS	GUI Link	OMS	OPEN / move	122	NOTSET	EXCLUDED	BAD	GOOD	STANDBY	NO VALUE F	GOOD	GOOD	GOOD	NOTSET	GOOD	GOOD	NOTSET	GOOD	NOTSET	NOTSET	GOOD	NOTSET	GOOD	NOTSET	NOTSET	STANDBY	STANDBY	GOOD
366920	/PromptReco/C...	Commissioning2	Manage / LS	GUI Link	OMS	OPEN / move	97	NOTSET	EXCLUDED	BAD	G S	STANDBY	NO VALUE F	G S	GOOD	GOOD	NOTSET	GOOD	GOOD	NOTSET	GOOD	NOTSET	NOTSET	GOOD	NOTSET	G S	NOTSET	NOTSET	STANDBY	STANDBY	G S
366919	/PromptReco/C...	Commissioning2	Manage / LS	GUI Link	OMS	OPEN / move	266	NOTSET	EXCLUDED	BAD	GOOD	STANDBY	NO VALUE F	GOOD	GOOD	GOOD	NOTSET	GOOD	GOOD	NOTSET	GOOD	NOTSET	NOTSET	GOOD	NOTSET	STANDBY	NOTSET	NOTSET	STANDBY	STANDBY	STANDBY
366911	/PromptReco/C...	Commissioning2	Manage / LS	GUI Link	OMS	OPEN / move	113	NOTSET	EXCLUDED	BAD	STANDBY	STANDBY	NO VALUE F	STANDBY	GOOD	GOOD	NOTSET	GOOD	GOOD	NOTSET	GOOD	NOTSET	NOTSET	GOOD	NOTSET	STANDBY	NOTSET	NOTSET	EXCLUDED	STANDBY	STANDBY
366910	/PromptReco/C...	Commissioning2	Manage / LS	GUI Link	OMS	OPEN / move	426	NOTSET	EXCLUDED	BAD	STANDBY	STANDBY	NO VALUE F	STANDBY	GOOD	GOOD	NOTSET	GOOD	GOOD	NOTSET	GOOD	NOTSET	NOTSET	GOOD	NOTSET	STANDBY	NOTSET	NOTSET	EXCLUDED	STANDBY	STANDBY
366895	/PromptReco/C...	Collisions23	Manage / LS	GUI Link	OMS	OPEN / move	1515	NOTSET	EXCLUDED	B G	GOOD	G S	NO VALUE F	G S	GOOD	GOOD	NOTSET	GOOD	GOOD	NOTSET	GOOD	NOTSET	NOTSET	GOOD	NOTSET	G S	NOTSET	G S	G S	G S	G S
366893	/Express/Collisi...	Collisions23	Manage / LS	GUI Link	OMS	SIGNOFF / move	26	NOTSET	EXCLUDED	GOOD	GOOD	GOOD	NO VALUE F	GOOD	GOOD	GOOD	NOTSET	GOOD	GOOD	NOTSET	GOOD	NOTSET	NOTSET	GOOD	NOTSET	GOOD	NOTSET	GOOD	GOOD	GOOD	GOOD
366893	/PromptReco/C...	Collisions23	Manage / LS	GUI Link	OMS	SIGNOFF / move	26	NOTSET	EXCLUDED	GOOD	GOOD	GOOD	NO VALUE F	GOOD	GOOD	GOOD	NOTSET	GOOD	GOOD	NOTSET	GOOD	NOTSET	NOTSET	GOOD	NOTSET	GOOD	NOTSET	GOOD	GOOD	GOOD	GOOD

Previous | Page 3 of 1568 | 20 rows | Next

Show workspace state columns

Pixel 1D histograms AutoEncoder



- Input 1-D histograms for ML/ Statistical methods comparison
- Example anomalous LS plots due to beam dump