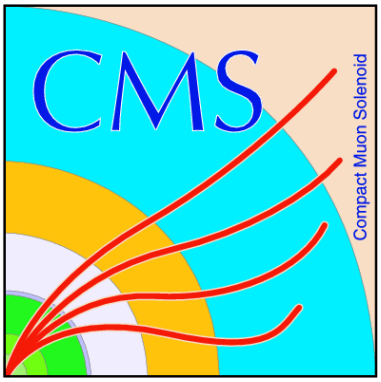
Development of an online high-level b-tag trigger data quality monitor

DESY Summer Student Programme, 2012

D. Oliinychenko

MIPT, Russia

Supervisor
W. Lohmann, I.Marfin



31th of August 2012

Abstract

A data quality monitor for the high level B-trigger is explained. Motivations for its development are given. An online data quality monitor for HLT is developed, described and results of its application are demonstrated.

Contents

1 Introduction

1.1 Large Hadron Collider

The Large Hadron Collider (LHC) is a circular particle collider hosted at the European Organization for Nuclear Research (CERN), Geneva (Switzerland). Bunches of protons circulate in the opposite directions and collide at four points. There CMS, ATLAS, LHCb and ALICE detectors are placed. Currently (July, 2012) the beam energy is 7 TeV. Bunches collide each 50 ns. Luminosity is $\mathcal{L} = 1.0 \times 10^{34} cm^{-2} s^{-1}$ [1].

1.2 CMS

CMS (Compact Muon Solenoid) [2] is a general purpose detector located at one of the LHC collision points. A sketch is shown on Fig. 1.2. The experiment comprises a 13m long, 6m diameter, 4T superconducting solenoid providing large bending power (12 T-m) for track measurements. The return field is large enough to saturate the 1.5 m iron plates in the return yoke, used for muon track reconstruction. The gaps between the plates provide slots for the four muon tracking stations, each of which consists of several layers of aluminum drift tubes (DT) in the barrel region and cathode strip chambers

(CSCs) in the endcap region.The system is complemented by resistive plate chambers (RPCs) used for a trigger. The bore of the magnet is large enough to accommodate the inner tracker and the calorimetry systems. The tracker is contained in a cylinder of 5.8m length and 2.6 m diameter. Ten layers of silicon microstrip detectors, which provide the required granularity and precision to reconstruct efficiently high multiplicity events and provides excellent momentum resolution. In addition three layers of silicon pixel detectors in the barrel region, complemented by two forward disks at each end, seed the tracks for reconstruction and improve the impact parameter resolution allowing to reconstruct secondary vertices.

The electromagnetic calorimeter (ECAL) provides coverage up to $|\eta| = 3$ and uses lead tungstate ($PbWO_4$) crystals whose scintillation light is detected by silicon avalanche photodiodes (APDs) in the barrel and vacuum phototriodes (VPTs) in the endcaps. A preshower system is installed in front of the endcap ECAL for π_0 rejection. The ECAL is surrounded by a brass/scintillator sampling hadron calorimeter (HCAL) with coverage up to $|\eta| = 3$. The light is converted by wavelength shifting (WLS) fibres embedded in the scintillator tiles and channeled via clear fibres to hybrid photodiodes (HPDs) which can operate in high axial magnetic fields. The central calorimetry is complemented by a "tail-catcher" (HO) in the barrel region insuring that hadronic showers are sampled over nearly eleven interaction lengths. Coverage from $\eta = 3$ to $\eta = 5$ is provided by an iron/quartz-fibre calorimeter (HF). The Cherenkov light emitted in the quartz fibres is detected by photomultipliers. The HF ensures large geometric coverage for measurement of the transverse energy in the event. Two additional calorimeters, called CASTOR and the Zero Degree Calorimeter (ZDC), not shown in Fig. 1.2, provide coverage at even larger rapidities than the HF.

The CMS uses an axial coordinate system with axes coinciding with the beam line, zero at collision point, azimuthal angle ϕ and polar angle θ. For convenience instead of θ pseudorapidity $\eta = -ln\left(tan\frac{\theta}{2}\right)$ is used.

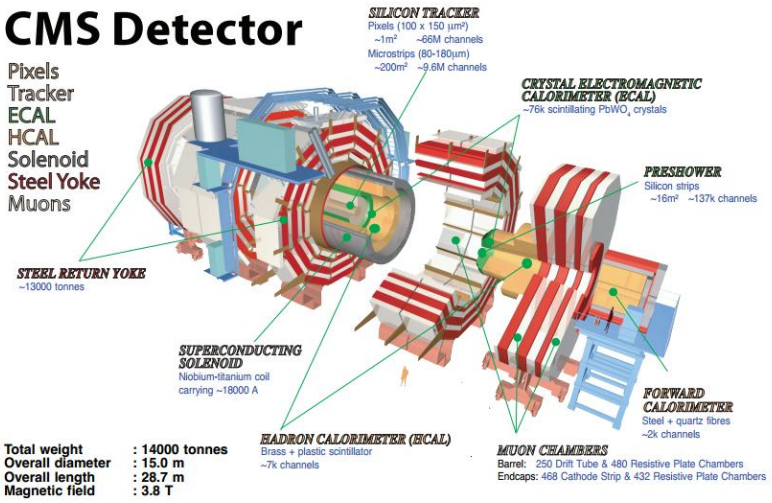


Figure 1: CMS detector

1.3 Jets

A large number of particles flying into a small solid angle is called a jet. Jets are identified in experiments as large number of hits per unit area in the tracking system and considerable energy deposit in small area of the calorimeter. Qualitatively jet formation can be explained as follows:

- Each beam particle a certain partonic substructure.
- Hard scattering in the collision point, called primary vertex, involves quarks and gluons. It is described by QCD Feynmann diagrams.
- Hard parton produces gluons and quark - anti-quark pairs born from vacuum. Gauge bosons can also be produced. Finally all created constituents hadronize and form a jet.
- Unstable hadrons in a jet decay and produce more particles and displaced vertices.

In the data analysis a jet is an associated object with the following characteristics:

- Direction: ϕ and η
- Energy E, momentum $\overrightarrow{p}$
- Transverse momentum $P_t = \sqrt{P_x^2 + P_y^2}$
- Initiating parton flavour

None of these characteristics is rigorously defined, because there is no unique recipe to decide which particle belongs to a jet and which does not. But there are several algorithmic definitions that give similar values of jet characteristics. They are divided into two groups:

- **Cone algorithms** Take advantage of the fact that a jet is a large number of particles in a small solid angle. All particles flying within the "cone" $\sqrt{\Delta\phi^2 + \Delta\theta^2} < \Delta R$ are assigned to the jet. In practice ΔR is usually taken 0.5, however the only demand on ΔR is that different jet definitions should give similar jet parameters.
- **Clustering algorithms** Backtrace successive decays that formed jet. Well-known algorithms are the k_T algorithm, the Cambridge/Aachen and the Anti-k_T algorithm. Each of this algorithms combines the four-vector of two particles, i and j, pair-wise according to the distance of the two particles d_{ij} and their individual distance to the beam d_{iB}. For all particles the smallest distance among d_{ij} is determined. If $min_j(d_{ij}) < d_{iB}$ then i and j are combined, adding their four-momenta, otherwise particle i is rejected as candidate for this jet and is considered for another jet. The procedure is repeated until all particles are clustered in jets.

1.4 B-tagging

Jet tagging is assigning flavour to a jet. B-tagging is assigning the flavour "b" to the jet, implying that the initiating parton was a b-quark. A jet can be tagged as 'b-jet', 'c-jet' or 'other jet'. There is no 't-jet', because t-quarks quickly decay into b-quarks without forming hadrons. Other jets are not tagged by flavour, e.g. as 's-jet' or 'u-jet'.

The reason is that if jet starts from, e.g. $u\bar{d}$, one can not say if it is a 'u-jet' or 'd-jet'. However, if jet starts from meson containing a c- or b-quark, one can tag this jet as 'c' or 'b' respectively, because probability to bare $c\bar{c}$ or $b\bar{b}$ pair from vacuum is negligible. $u\bar{u} : d\bar{d} : s\bar{s} : c\bar{c} = 1 : 1 : 0.3 : 10^{-11}$ [3]. Thus, if a jet originates from a b-hadron it is very likely originating from b-quark.

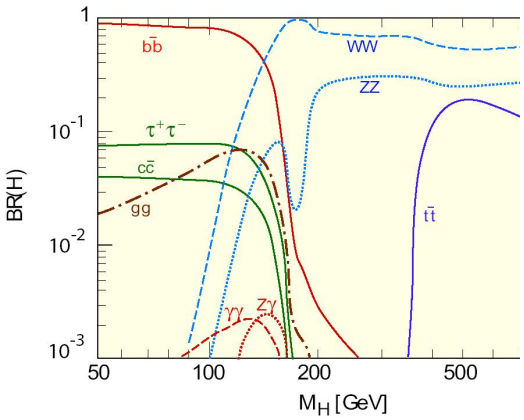


Figure 2: The branching ratios of a Higgs boson as a function of its mass, m_H

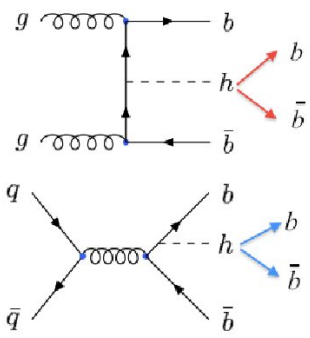


Figure 3: Feynmann diagrams including a Higgs boson and 4 b-jets in the final state

B-jets are of particular interest for top-quark physics, Higgs boson investigation and testing physics beyond the Standard Model. In particular, $H \to b\bar{b}$ is the dominant decay for a light (<130 GeV) Higgs boson, as can be seen in Fig. 2. In the MSSM final states with 4 b-jets, depicted in Fig. 3, are even enhanced compared to SM. That gives us the chance to discover supersymmetry or put limits on it.

The idea of detecting b-jet is the following. B-mesons have a comparatively large lifetime (1.5 ps) because they can decay only weakly. Therefore B-mesons travel several millimeters before the decay and the secondary vertex (SV) in the jet is displaced from the pp interaction point, see Fig. 4. It is possible to detect displaced vertices using the high-precision vertex detector like the CMS silicon pixel tracker, which reconstructs tracks and their impact parameters with very high precision. These tracks together with information from calorimeter are used to reconstruct primary vertex and secondary vertices.

In Fig. 4 one can see that if SV is not displaced then distance from tracks to the primary vertex is more likely to be small. On the contrary, if the SV is displaced, then track extensions tend to lie far from PV (primary vertex). It is convenient to introduce a significance $= d_0/\sigma$, where d_0 is the impact parameter - the minimal distance of the track to PV and σ is uncertainty of the measurement of d_0. Large significance of jet constituents means that the jet does not start in PV and it is likely a b-jet. The significance of the jet constituents together with additional parameters such as presence of lepton in a jet are used to build b-tag discriminator. If discriminator exceeds some predefined threshold then jet is b-tagged. Here is list of algorithms used to obtain the discriminator [7]:

- **Track counting** The signed impact parameter significance of all good tracks is calculated. Tracks are ordered by decreasing significance. The b tag discriminator is defined as the significance of the N'th track. Two choices are used: $N = 2$ (high efficiency) or $N = 3$ (high purity).

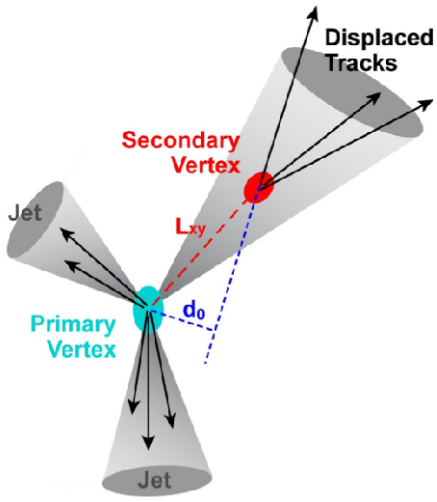


Figure 4: To the idea of B-tagging

- **Jet Probability** This is a more sophisticated algorithm. Its b tag discriminator is equal to the negative logarithm of the confidence level that all the tracks in the jet are consistent with originating from the primary vertex. This confidence level is calculated from the signed impact parameter significances of all good tracks.

- **Simple secondary vertex** These class of algorithms reconstructs the B decay vertex using an adaptive vertex finder, and then uses variables related to it, such as decay length significance to calculate its b tag discriminator. It has been found to be more robust to Tracker misalignment than the other lifetime-based tags.

- **Soft muon and soft electron** These two algorithms tag b jets by searching for the lepton from a semi-leptonic B decay, which typically has a large P_t with respect to the jet axis.

- **Combined secondary vertex** This sophisticated and complex tag exploits all known variables, which can distinguish b from non-b jets. Its goal is to provide optimal b tag performance, by combining information about impact parameter significance, the secondary vertex and jet kinematics. Currently lepton information is not included. The variables are combined using a likelihood ratio technique to compute the b tag discriminator.

In Fig. 5 distributions of the discriminator obtained from simple track counting algorithm are shown for MC $t\bar{t}$-events both for b-jets and udsg-jets. One can see that jets with high discriminator are mostly b-jets. There is no perfect discriminator. Any known discriminator tags non b-jets as b-jets and vice versa with some probability. To characterize the tagger performance tagger efficiency is introduced. The tagger efficiency is a ratio of flavour-tagged jets to all jets of this flavour, $\epsilon_{flav} = \frac{tagged_{flav}}{all_{flav}}$. Various algorithms and calibrations are used for the b-tag trigger, their efficiencies, discriminator thresholds and discriminators are different. This makes trigger performance monitoring necessary.

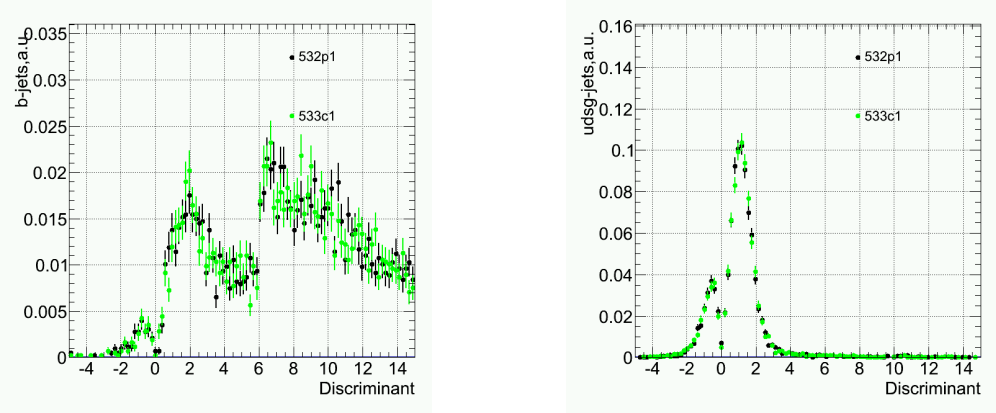


Figure 5: Discriminator distribution. Left: b-jets, right: dusg-jets. Green and black points stand for different versions of CMS software

1.5 HLT DQM (High Level Trigger Data Quality Monitor)

LHC is currently producing one collision per 50 ns, i.e. with the rate of 20 MHz. The size of one event is 1 MB. Data of 20 TB/s is too large to store. Only small part of events is considered to be "interesting" and saved for further analysis. Rejection of non-interesting events is done by triggers in two steps: low-level hardware trigger does a preselection of events and decreases the rate by O(1000) and a high-level software-implemented trigger decreases the rate by O(1000) [4]. But the selection algorithms for high level trigger are more complicated and time-consuming.

Data Quality Monitor(DQM) is a system of HLT monitoring. DQM produces and stores plots characterizing the performance of HLT. DQM can work in online or offline regime. In offline regime it stores data on local host and in online regime it publishes data each predefined time interval.

2 My task: B-Tagging HLT online DQM on data

2.1 Motivation

All the software used by CMS collaboration is maintained together, the code is open-source and can be seen at [5]. CMSSW (CMS software) is being constantly changed and developed, as a result new versions appear. Not only new software is produced and bugs are corrected, but new algorithms are implemented or some parameters are changed. To guaranty the reliable work, the software is constantly checked. Procedure of comparing performance of different versions is called "validation". New versions of CMSSW are mandatory to be validated.

Up to the current moment (31.08.2012) the validation of b-tag HLT was performed offline on MC samples. However, it is important to validate software on data. There is also a need in the online control of the b-tag HLT. The primary goal of my summerstudent work was to develop software for validation on data and for online b-tag HLT DQM.

2.2 Offline validation

My first task was to perform the validation of CMSSW5.3.2c1 vs. CMSSW5.3.3p1 offline as an exercise. To validate a new version(offline) the following has to be done [6]:

1. Generate a sample of MC events
2. Simulate the detector response for both versions
3. Reconstruct events for both versions
4. Compare results of reconstruction for both versions

Plots used for the offline comparison are kinematic distributions, flavour efficiency vs. discriminator cut, non b-jet vs. b-jet efficiency and discriminator distribution for all jets, b-jets, c-jets and usdg-jets.

Validation was done in the following steps: $t\bar{t}$-events were generated using Pythia 6 [3], HLT was simulated via CMSSW, jets were reconstructed, matched to partons and validation plots were done as shown in Fig. 6, 7. From the plots we conclude that versions are consistent with each other.

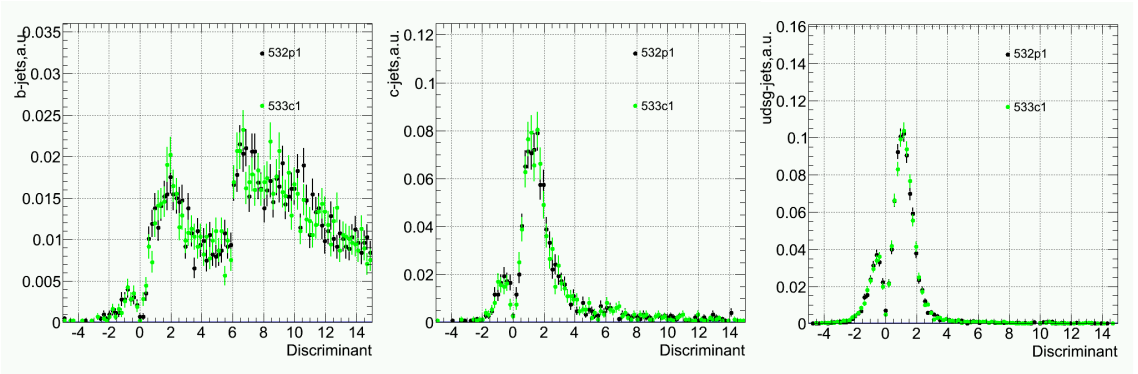


Figure 6: Discriminator distribution for different jet flavours: left - b-jets, middle - c-jets, right - usdg-jets. Green and black points correspond to different versions of CMSSW.

2.3 Validation on data and online DQM

The main difference when using data is that for MC-generated events one knows the jet flavour while for data one can only assign probabilities. HLT reconstructs jets and b-tags them if a discriminator is higher than a given value. Thus, any plot depending on flavour is biased by bTag HLT performance. Most of plots used for offline version comparison are flavour-dependent and can not be used for HLT comparison on data.

Consequently, for HLT comparison on data flavour-blind plots will be used: kinematic and discriminator distribution for all jets, discriminator distribution for jets with the largest transverse momentum in event, for the jet with second transverse momentum in event, etc.

For this purpose the analyzer software was rewritten to build only flavour-blind plots. The configuration file was changed. Parton reconstruction, jet-parton and jet-flavour matching were removed. The examples of flavour-blind plots for online HLT DQM are shown in Fig. 8.

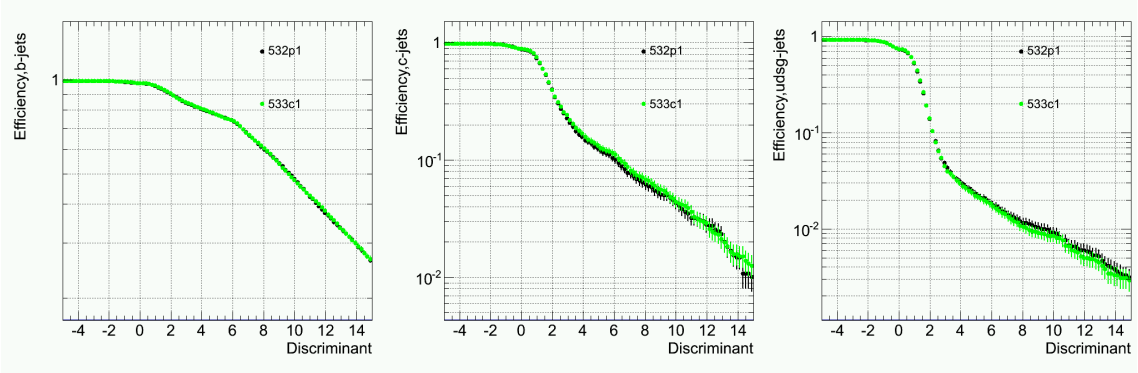


Figure 7: Efficiency as a function of the discriminator cut for different jet flavours: left - b-jets, middle - c-jets, right - usdg-jets. Green and black points correspond to different versions of CMSSW.

3 Acknowledgments

I would like to thank Igor Marfin for his explanations and patience while working with me. I also appreciate help of Roberval Walsh. And I would like to thank DESY for hospitality.

References

[1] J. Beringer et al. (Particle Data Group), Phys. Rev. D86, 010001 (2012), http://pdg.lbl.gov/2012/reviews/rpp2012-rev-hep-collider-params.pdf

[2] CMS Collaboration, CMS TriDAS project: Technucal Design Report; 1, The Trigger Systems. Technical Design Report CMS, CERN, 2000

[3] T. Sjostrand, S.Mrenna, P.Skands Pythia 6.4 Physics and Manual, 2006

[4] The CMS Trigger and DATA Acquisition Group, EPJ, Nov. 2005 The CMS High Level Trigger

[5] cmslxr.fnal.gov/lxr

[6] twiki.cern.ch/twiki/bin/view/CMSOublic/SWGiudeBtagValidation

[7] twiki.cern.ch/twiki/bin/view/CMSPublic/SWGuideBTagging

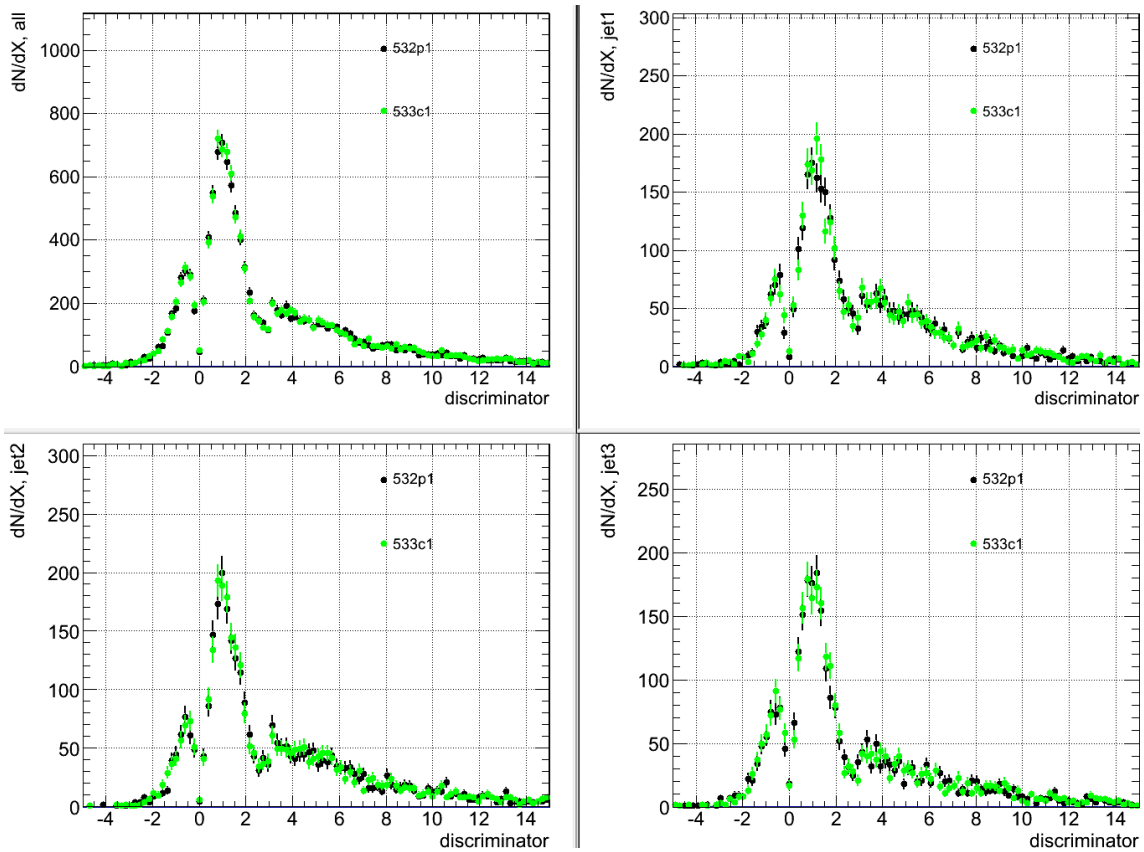


Figure 8: Discriminator distribution. Left,top: all jets. Right,top: jets with largest transverse momentum in event, "first jets". Left, bottom: jets with second large transverse momentum in the event. Right, bottom: jets with third large transverse momentum in the event. Green and black points correspond to different versions of CMSSW.