CMS Physics Analysis Summary

Contact: cms-pog-conveners-btag@cern.ch 2009/07/06

Algorithms for b Jet Identification in CMS

The CMS Collaboration

Abstract

This note describes the algorithms CMS has developed to tag b-jets. The algorithms rely on the long lifetime of the b quark and the consequent occurence of displaced secondary vertices, tracks with substantial transverse impact parameter and leptons with high transverse momentum with respect to the jet axis. The note focuses on the algorithmic part of the taggers, and presents results for the case of a perfectly aligned tracker.

1 Introduction

The identification of b-jets is crucial to characterize a variety of Standard Model (SM) and discovery channels like the measurement of bottom or top pair production, the search for Higgs bosons, and other New Physics scenarios. The hard fragmentation, long lifetimes and high masses of B hadrons, and the relatively high fraction of semileptonic decays distinguish these jets from those originating from gluons, light quarks and - to a lesser extent - from c quarks. Due to its precise inner tracking system and its lepton identification capabilities the CMS experiment is well positioned to exploit these features, and many algorithms have been developed. They range from comparatively simple and robust approaches based on the presence of leptons to complex multi-variate techniques extracting lifetime and kinematic information from displaced vertices.

The goal of this note is to provide an overview of the different algorithms and their performance on simulated events, using the most recent CMS software versions.

2 Samples and Software

The analyses described in this note are based on CMS software and Monte Carlo samples current as of 2008. The results are obtained on fully simulated samples: most of the plots use PYTHIA[1] QCD events, generated with a $\hat{p}_{\mathrm{T}} > 80\,\mathrm{GeV/c}$ for a total of 3 million events[1]; the plots exploring the high jet p_{T} part of the spectrum use also the samples with $\hat{p}_{\mathrm{T}} > 470$ GeV/c, $\hat{p}_{\mathrm{T}} > 1400$ GeV/c, $\hat{p}_{\mathrm{T}} > 3000$ GeV/c, for a total of another 3 million events. The events do not include any pile-up simulation, and use standard settings for b-fragmentation and B hadron decays. Also, no misalignment is applied since the scope of the present note is to study the intrinsic separation power. For the same reason, jet to parton association is done by matching to the heaviest parton in the jet. In this way, jets initiated by a parton other than a b but containing $g \to b\bar{b}$ are flagged as b-jets.

3 Inputs to b-tagging

The algorithms described in this note allow one to tag the flavour of a b-jet, using the specific decay characteristics of B hadrons.

Jets are reconstructed with calorimetric only information, by clustering energy deposits in the electromagnetic (ECAL) and hadronic (HCAL) calorimeters. Several algorithms are present in literature to choose which deposits to cluster, and to extract information such as jet energy and jet direction from the used deposits; the ones available in CMS are documented in [2]: they are the "Iterative Cone" (IC) algorithm, the k_T (KT) algorithm and the "Seedless Infrared Safe Cone" (SC) algorithms; in addition to choosing an algorithm, one must specify a configurable variable which represents the cone size around the jet direction used to collect the deposits. The standard b-tagging setup uses the Iterative Cone with $\Delta R = 0.5$, named from this point IC05, and is the default for this note unless explicitly stated otherwise. Jet energy corrections are needed since the reconstructed energies are biased and not centered around the true jet energies; those corrections range from 50% for jets with low energy, to a few percent. A hard cut on jet $p_{\mathrm{T}} > 20\,\mathrm{GeV/c}$ is applied as a jet preselection, to reduce sensitivity to calorimeter noise.

Tracks are the most powerful ingredient to b-tagging. Since the tracking in the vicinity of the

[1] $\hat{p}_{\mathrm{T}}$ is the transverse momentum in the rest frame of the hard interaction

interaction vertex contains most of the discriminating power, b-tagging studies rely heavily on the presence of hits in the pixel system, which allow a precise extrapolation close to the primary vertex. Tracks are reconstructed using a standard Kalman Filter based method [3]; to minimize fake and badly reconstructed tracks, basic track quality requirements are imposed by the b-tagging code:

- total number of silicon (pixel + strip) hits ≥ 8
- number of pixel hits ≥ 2
- transverse impact parameter $d_{xy} < 0.2\,\text{cm}$
- longitudinal impact parameter $d_z < 17\,\text{cm}$
- transverse momentum $> 1.0\,\text{GeV/c}$
- $\chi^2/ndof$ of the track fit < 5.0
- distance ($\Delta R \equiv \sqrt{\Delta\eta^2 + \Delta\phi^2}$) to the jet axis < 0.5

The primary vertex (see [4]) is reconstructed from all available tracks in the event satisfying the above requirements. The Adaptive Vertex Fitter [5] algorithm is used to perform the vertex fit, using a sample of tracks from which those with $d_{xy}/\sigma_{d_{xy}} > 5$ have been removed. Among the primary vertices found in this way, the one with the highest $\sum_{Tracks} p_{\mathrm{T}}^2$ is selected.

Reconstructed muons are also used to select b-jets. These muons are seeded from the CMS muon chambers, and are then linked to tracker tracks to form so called global muons[6], which allows for a very low fake rate.

4 B-tagging Observables

The observables exploited to discriminate between b- and c- and light jets, where light jets are those originating from u, d, or s quarks, or gluons, are discussed in the following.

The most powerful single-track observable is the impact parameter (IP) - the distance between the track and the vertex at the point of closest approach. The geometrical interpretation of the IP for a single track is depicted in Figure 1. For B hadrons with finite lifetime, the IP is Lorentz invariant and the typical scale is set by $c\tau \sim 480\mu$m. The IP can be calculated either in the transverse plane or in 3D. In CMS, the good z resolution provided by the pixel system allows 3D reconstruction to be used. Given that the uncertainty can be of the same order of magnitude as the IP, a better observable for b-tagging is the impact parameter significance defined as IP/σ_{IP} .

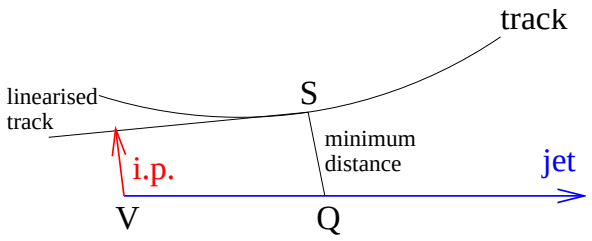

Figure 1: Schematic representation of the impact parameter of a track with respect to the vertex.

The IP is "life time signed". The IP sign is obtained from the sign of the scalar product of the IP segment with the jet direction. A "sign flip" can happen due to differences between the reconstructed jet axis and the true B hadron flight direction. For decays without a sizeable lifetime, the IP is expected to be symmetric with respect to zero; for B hadrons decaying weakly,

it is mostly positive.

The IP significance is already an estimator able to disentangle tracks from b- or lighter jets. One can also use it to define a track by track probability (P_{tr}), by extracting the probability density function for tracks not coming from b-jets. For this purpose we use tracks with $IP < 0$, which as described before, are symmetric in the IP shape around zero.

For a set of tracks, one can directly search for a secondary vertex from the B hadron decay. The same tool used for the primary vertex finding is used on all the tracks associated to the jet. Vertices with at least 65% of tracks shared with the primary vertex are removed from the list.

5 B-tagging Algorithms

All the b-tagging algorithms discussed are part of the b-tagging framework in CMS, which demands that the unique output of any algorithm be a "discriminator", defined as a single number which the user can cut on to select different regions in the efficiency versus purity phase space. The discriminator can be a simple physical quantity like the IP significance for some taggers, or a complex variable like the output of likelihood ratio or neural network.

The simplest way of producing a discriminator based on track impact parameters is an extension of the so-called track counting algorithm. The track counting approach identifies a jet as a b-jet if there are at least N tracks each with a significance of the impact parameter exceeding S. This algorithm has two major parameters (N and S). The way of producing a continuous discriminator for this algorithm is to fix the value of N, and consider as discriminating variable the impact parameter significance of the Nth track (ordered in decreasing significance). If one is interested in a high efficiency for b-jets, the second track can be used; for higher purity selections the third track is a better choice. The discriminators obtained in this way are plotted for QCD events in Figure 2, and are simply the IP significance shapes for the chosen track.

The jet probability algorithms are a natural extension of the track counting algorithms. The idea is to combine the information coming from all selected tracks. In order to do so the track probability prevously defined is used. Two discriminators are provided; the first labelled "jet probability" is strictly related to the combined probability that all the the tracks in the jet come from the primary vertex[2], defined as $P_{jet} = \Pi \cdot \sum_{j=0}^{N-1} \frac{(-\ln \Pi)^j}{j!}$ where $\Pi = \prod_{i=1}^{N} P_{tr}(i)$. The second, labelled "jet B probability" estimates how likely it is that the four most displaced tracks are compatible with the primary vertex; the selection comes from the fact that the average charged track multiplicity in weak b hadron decay is ~ 5, and from the average track reconstruction efficiency, around 80% for tracks in jets. The shapes of the discriminant variable are presented in Figure 3.

The presence of a muon close to the jet is already a hint of a weak decay of a B hadron. This can be complemented with some additional quantity, in order to build a discriminator. In the "soft muon by $p_{\mathrm{T}_{rel}}$" algorithm the p_{T} of the muon with respect to the jet axis is used; harder cuts yield higher purities. In the "soft muon by IP significance" the IP significance of the muon is used instead, but only when found to be positive. In all the cases, when more than one muon is reconstructed, the one with the highest discriminator value is used. Figure 4 shows normalized discriminator shapes for those taggers.

Secondary vertices can be used to select jets from B hadrons with high purity. A simple version,

[2] A minimum probability of 0.5% is forced for highly displaced tracks, to avoid badly reconstructed tracks to drive the global probability too low.

called "simple secondary vertex"[7] tagging algorithm is based upon the reconstruction of at least one secondary vertex. If no such vertex is found, the algorithm returns no discriminator, limiting its maximum b-jet efficiency to the probability of finding a vertex in the presence of weak B hadron decay (around 60-70%). The significance of the $3D$ flight distance is used as a discriminating variable for this tagger. The distribution of this discriminator is shown on the left side of Figure 5.

A more complex approach involves the use of secondary vertices, together with other lifetime information, like the IP significance or decay lengths. By using these additional variables, the "combined secondary vertex" algorithm provides discrimination even when no secondary vertices are found, so the maximum possible b-tagging efficiency is not limited by the secondary vertex reconstruction efficiency. In many cases, tracks with an IP significance > 2 can be combined in a so-called "pseudo vertex", allowing for the computation of a subset of secondary vertex based quantities even without an actual vertex fit. When even this is not possible, a "no vertex" category reverts simply to track based variables similarly to the "jet probability" algorithm.

The list of variables used is:

- the vertex category (real, "pseudo," or "no vertex");
- 2D flight distance significance;
- vertex mass;
- number of tracks at the vertex;
- ratio of the energy carried by tracks at the vertex with respect to all tracks in the jet;
- the pseudo-rapidity of the tracks at the vertex with respect to the jet axis;
- 2D IP significance of the first track that raises the invariant mass above the charm threshold of 1.5 GeV when subsequently summing up tracks ordered by decreasing IP significance.
- number of tracks in the jet;
- 3D signed IP significances for all tracks in the jet.

These variables are used as input to a Likelihood Ratio, used twice to discriminate between b- and c-jets and between b- and light jets, and then combined additively with a factor of 0.75 and 0.25 respectively. The discriminator shapes for the "combined secondary vertex" taggers are presented in Figure 5 (right).

6 Performance

We show the performance for b-tagging algorithms in the form of b-jet versus c- or light jet efficiency, additionally dividing light jets into uds- and gluon jets. Results are presented here for a QCD sample, with the only requirements of $\hat{p}_{\mathrm{T}} > 80\,\mathrm{GeV/c}$ and a jet reconstructed with corrected $p_{\mathrm{T}} > 20\,\mathrm{GeV/c}$; the jets are also required to be within the tracker acceptance, with $|\eta| < 2.4$.

Figures 6 and 7 show the performance for track based algorithms. Please note that in the "track counting high purity" tagger the b-tag efficiency does not reach 100%, due to the requirement of at least three good tracks in the jet.

Figures 8 and 9 show the performance for secondary vertex based and muon based algorithms, respectively.

An overview plot with all the algorithms is presented in Figure 10.

In some analyses it is important to use taggers which have the same performance, defined either as the efficiency for b-jets or the rejection for non-b jets, in the whole range of η or p_T. The taggers' performance vary considerably with those two variables due to the CMS tracker design and the collimation of the jets. For the purpose of showing the behavior when looking at different detector regions or p_T spectra, Figures 11 and 12 show how the mistag rate varies as a function of p_T and η, for the "track counting high purity" algorithm working at 50% b-jet efficiency. Plots with the other taggers and different settings for the performance do not differ sensibly and are not shown here.

Figures 13 and 14 show the b-jet efficiency as a function of p_T and η, for the "track counting high purity" algorithms working at a given uds-jet efficiency.

7 Additional Studies

Many variations on the input quantities have been explored to optimize the b-tagging discrimination power.

Tracking quality cuts have been varied in order to

- lower the p_T cut to 500 MeV/c; or to
- use the "high purity" selection defined by the tracking group [3] instead of ad-hoc b-tagging cuts.

As shown in Figure 15, no significant difference is found.

Another tunable parameter is the choice of the algorithm used to reconstruct the jet direction. The default setting is to use the direction from the calorimetric deposit to the primary vertex, but one can also

- use the direction from the jet and the associated tracks (0.5 times the vectorial energy of the calorimetric jet plus the vectorial momentum of all tracks inside the jet);
- use the direction secondary vertex $-$ primary vertex (when a secondary jet is not found, fall-back to the default);

The resulting performance for the "track counting" algorithms, which have proven to be most sensitive to the jet axis definition, can be seen in Figure 16.

While currently b-tagging algorithms use by default IC05 jet reconstruction, they can also be tested with the other available algorithms in CMS, SC05 and KT04. Figure 17 shows the efficiency versus the mistag rate of the "track counting high purity" b-tagging algorithm for different choices of calorimeter-based clustering algorithms. Very small differences can be seen, but not enough to suggest that one clustering algorithm is superior to the others.

8 Conclusions

We have described the set of standard b-tagging techniques used in CMS and evaluated their optimal performance in a full detector simulation. The algorithms range from very simple, Impact Parameter based approaches to multi-dimensional likelihoods combining all available information from tracks and vertices. The efficiencies have been studied for different tracking and jet reconstruction parameters. The results are Monte Carlo based and do not cover the problem of the evaluation of b-tagging performance from data, which instead can be found in

[8] and [9]. The present results are obtained with a perfect detector; performance in presence of misalignment has already been studied in [7].

References

[1] T. Sjostrand, S. Mrenna, and P. Skands, "PYTHIA 6.4 physics and manual," *JHEP 05* **026** (2006) arXiv:hep–ph/0603175.

[2] **CMS** Collaboration, "Performance of Jet Algorithms in CMS," *CMS PAS JME_07_003* (2007).

[3] C. Collaboration, "CMS Physics TDR: Volume I, Detector Performace and Software," *CMS PTDR* **1** (2008).

[4] P. Vanlaer, L. Barbone, N. D. Filippis, T. Speer, O. Buchmuller, and F.-P. Schilling, "Impact of CMS Silicon Tracker Misalignment on Track and Vertex Reconstruction," *CMS NOTE 2006_029* (2006).

[5] W. Waltenberger, "Adaptive Vertex Reconstruction," *CMS NOTE 2008_033* (2008).

[6] **CMS** Collaboration, S. Chatrchyan et al., "The CMS experiment at the CERN LHC," *JINST* **3** (2008) S08004.

[7] **CMS** Collaboration, "Performance of b-Tagging Algorithms with Realistic Detector Scenarios at CMS," *CMS PAS BTV_07_003* (2008).

[8] **CMS** Collaboration, "b tag efficiency from System 8 and Ptrel Method," *CMS PAS BTV_07_001* (2007).

[9] **CMS** Collaboration, "Measuring uds mistag rate of b tag using negative tags," *CMS PAS BTV_07_002* (2008).

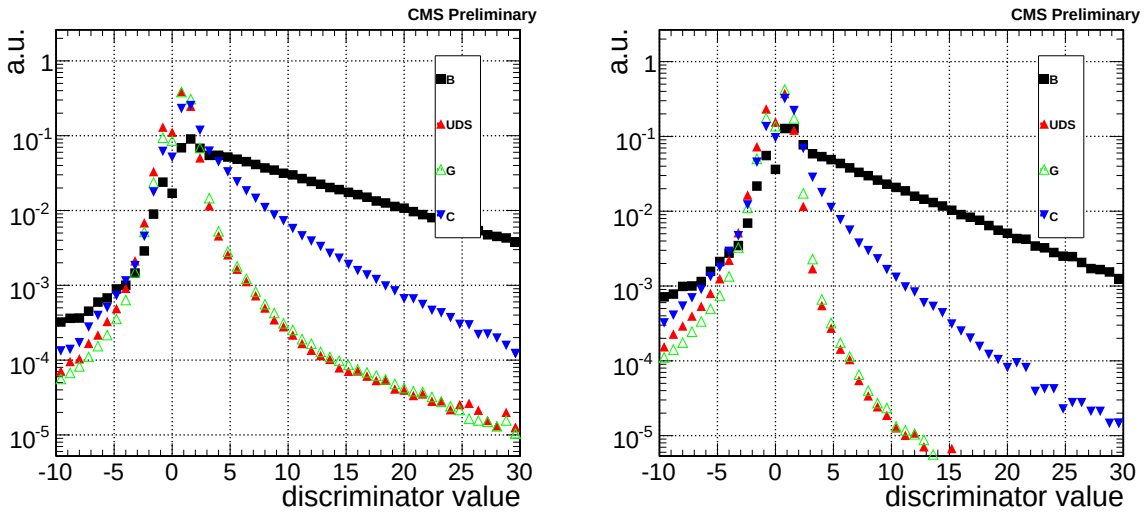

Figure 2: The normalized distribution of the discriminator for the "track counting high efficiency " algorithm on the left and for the "track counting high purity" algorithm on the right, for different jet flavours.

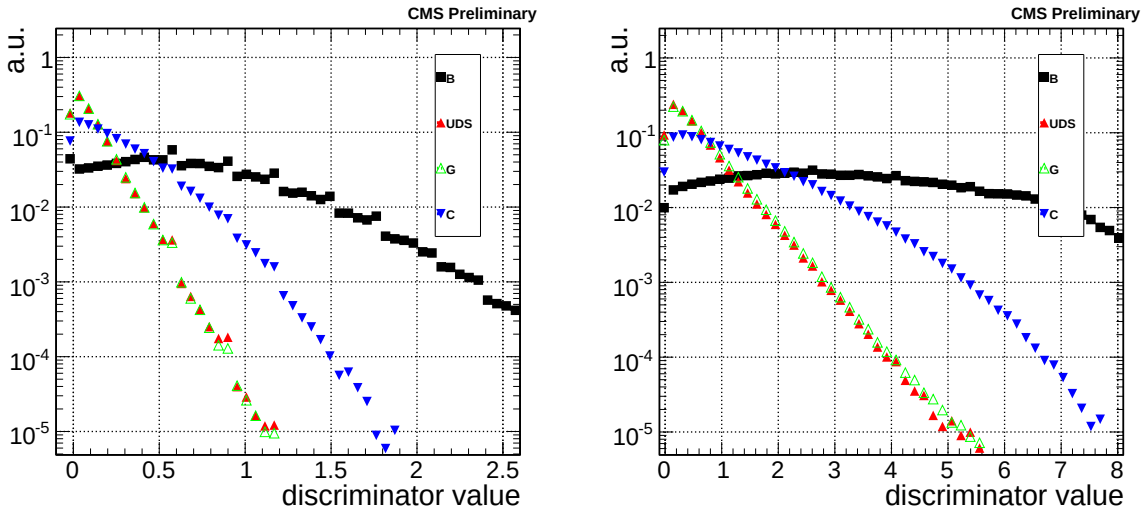

Figure 3: The normalized distribution of the discriminator for the "jet probability" algorithm on the left and for the "jet B probability" algorithm on the right, for different jet flavours.

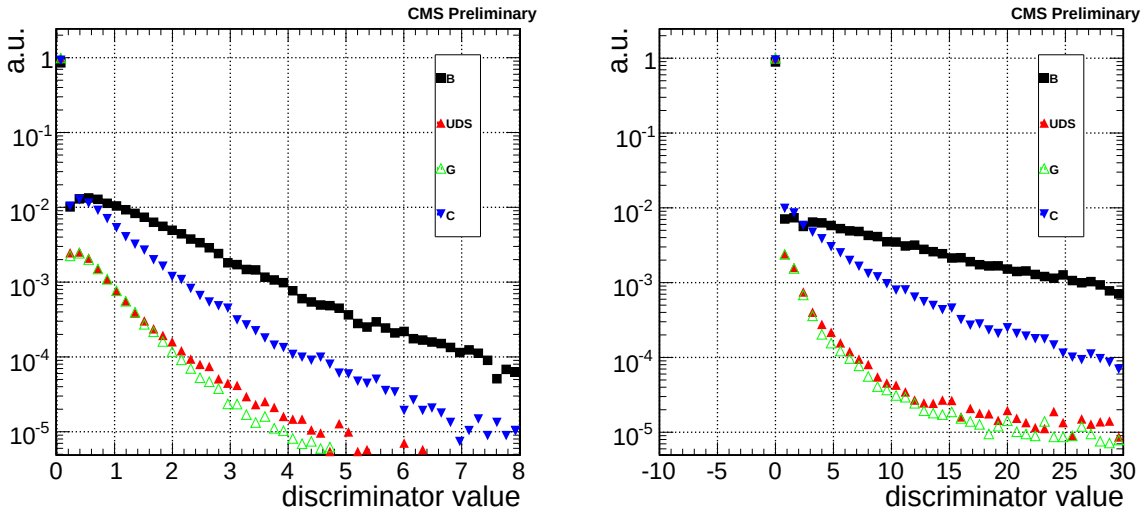

Figure 4: The normalized distribution of the discriminator for the "soft muon by $p_{\mathrm{T}_{rel}}$" algorithm on the left and for the "soft muon by IP significance" algorithm on the right, for different jet flavours. In both cases the spike at zero accounts for jets where no close muon was reconstructed.

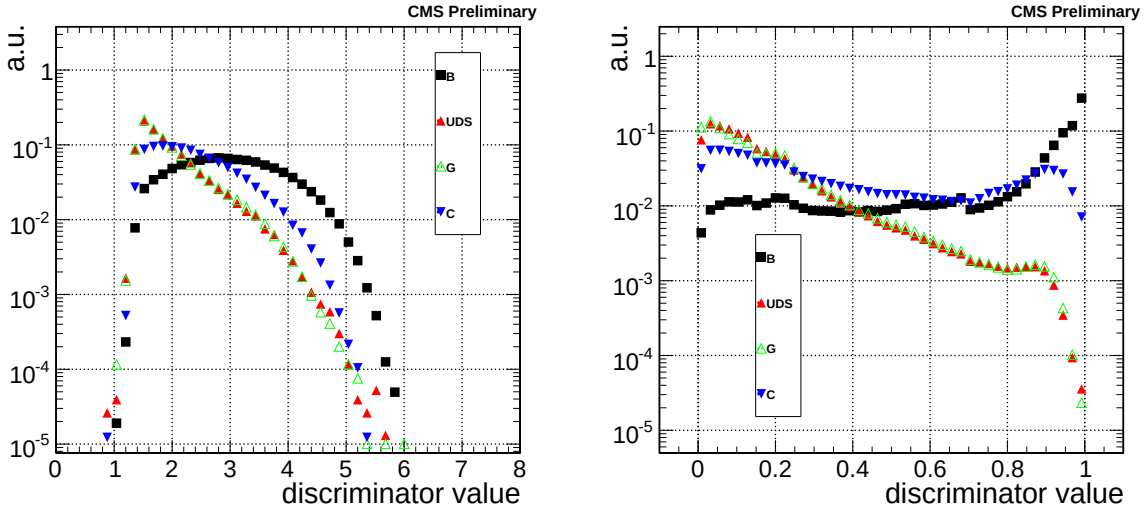

Figure 5: The normalized distribution of the discriminator for the "simple secondary vertex" algorithm on the left and for the "combined secondary vertex" algorithm on the right, for different jet flavours. For the "simple secondary vertex", the $\approx 30\%$ of jets without a reconstructed secondary vertex are assigned negative discriminator values and are not shown.

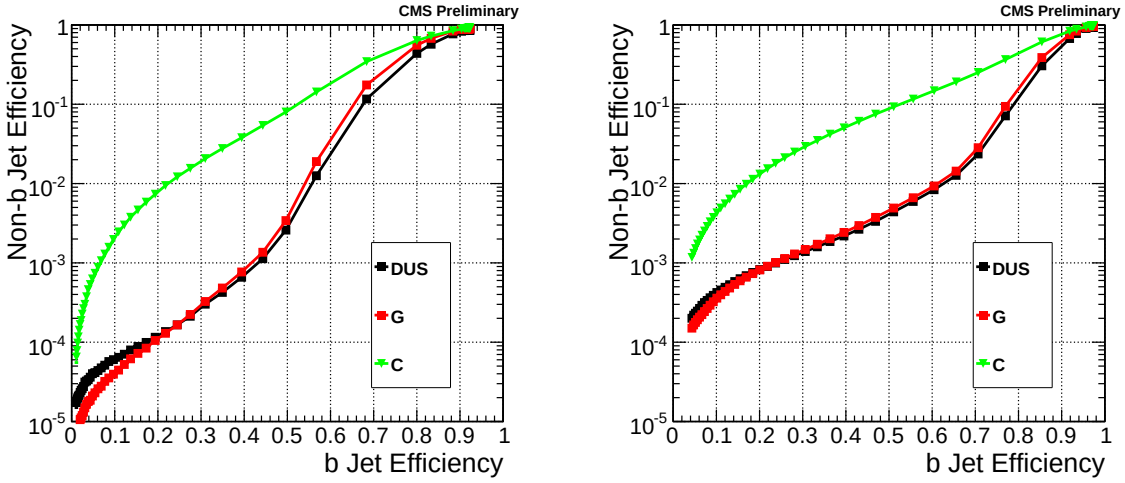

Figure 6: Mistag rate versus efficiency for the "track counting high purity" (left) and "track counting high efficiency" (right) taggers.

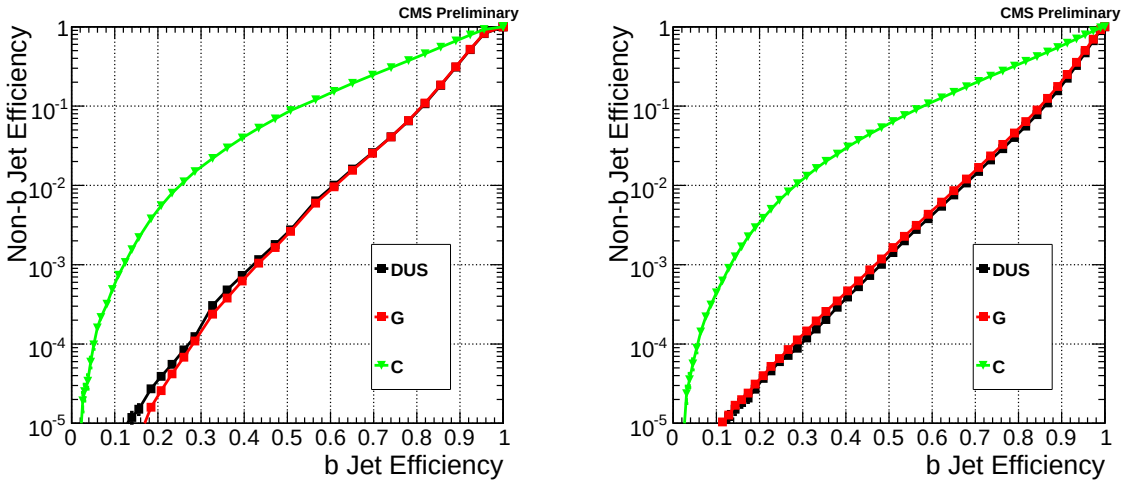

Figure 7: Mistag rate versus efficiency for the "jet probability" (left) and "jet B probability" (right) taggers.

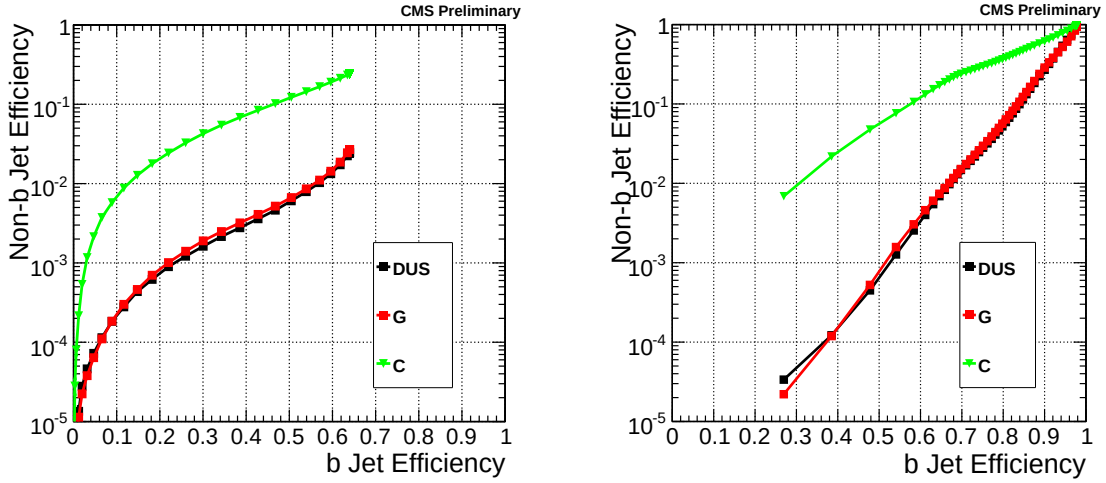

Figure 8: Mistag rate versus efficiency for the "simple secondary vertex" (left) and "combined secondary vertex" (right) taggers.

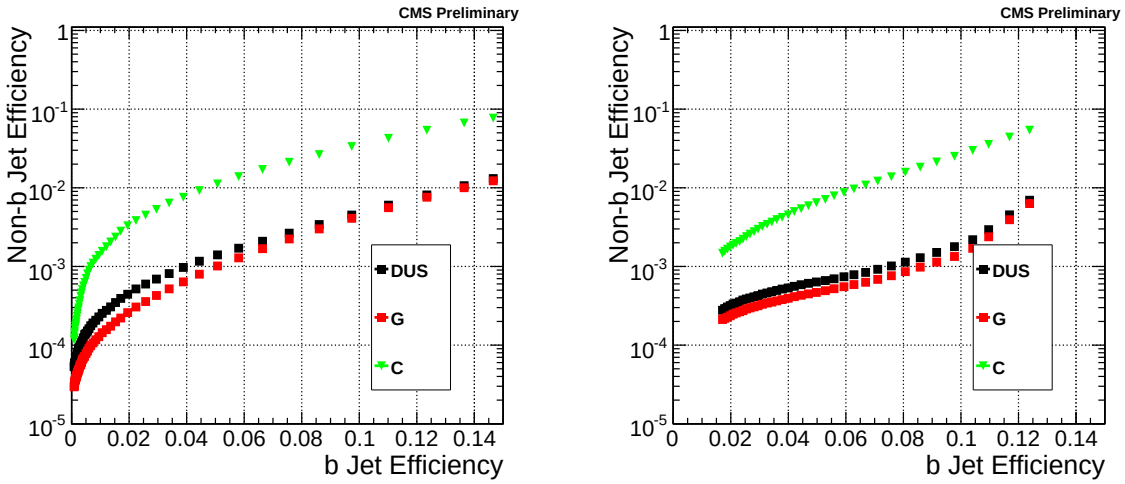

Figure 9: Mistag rate versus efficiency for the "soft muon by $p_{T_{rel}}$" (left) and "soft muon by IP" (right) taggers.

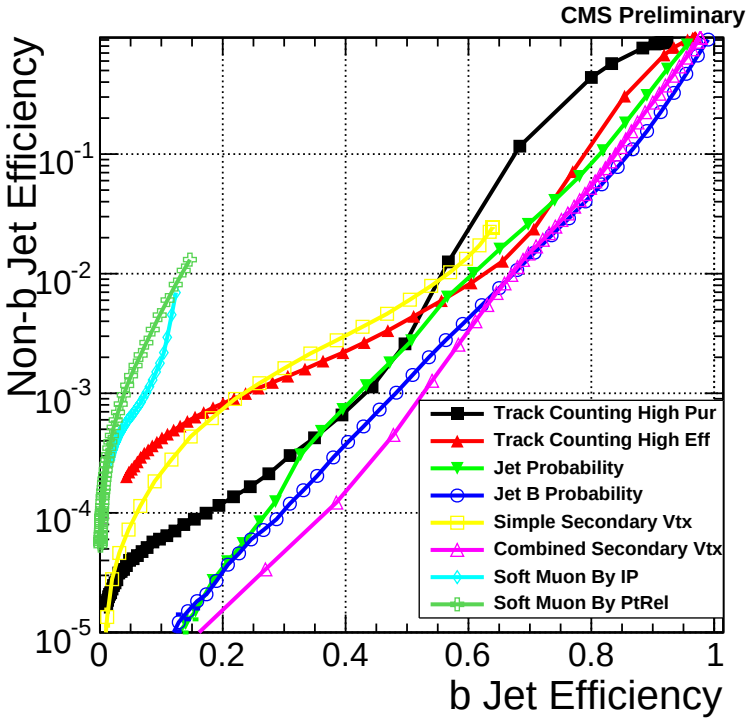

Figure 10: Comparison of the uds mistag rates versus b-jet efficiencies for all the taggers.

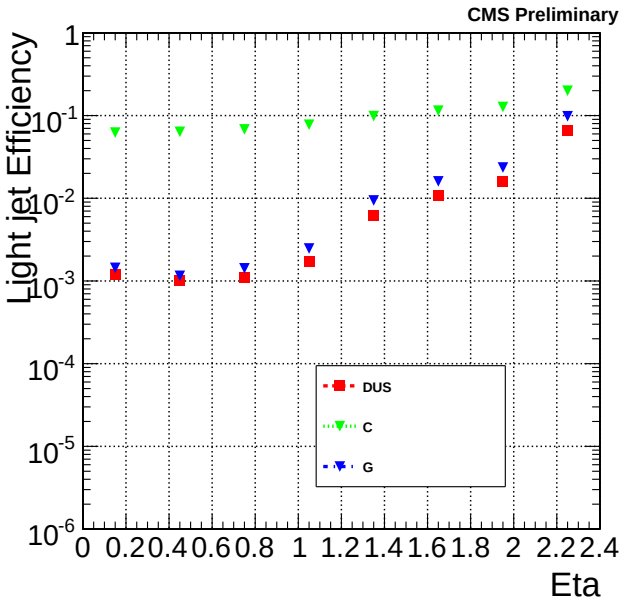

Figure 11: Charm and uds-jet efficiencies as a function of jet η for the "track counting high purity" tagger at 50% b-jet efficiency.

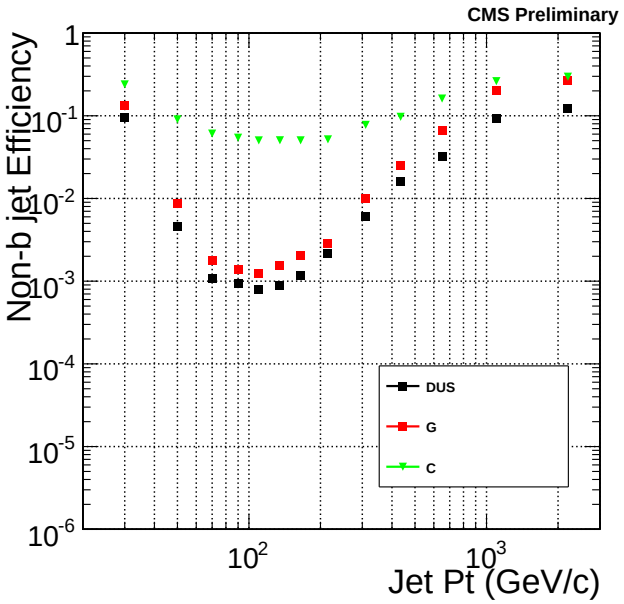

Figure 12: Charm and uds jet efficiencies as a function of jet p_{T} for the "track counting high purity" tagger at 50% b-jet efficiency.

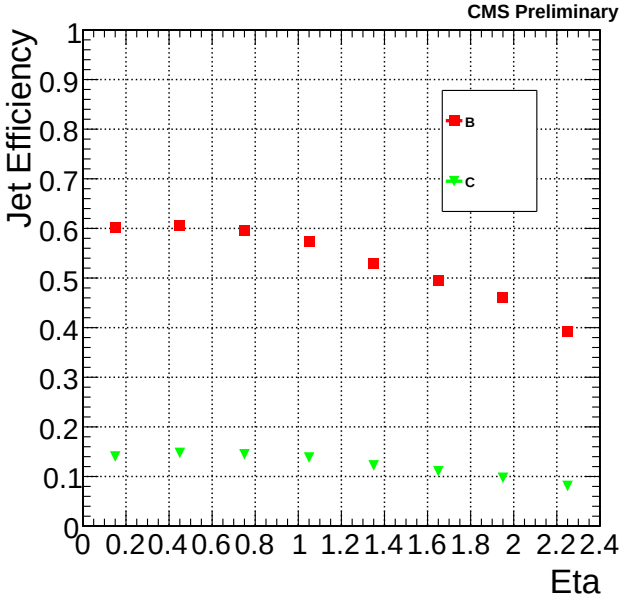

Figure 13: Efficiencies for b and c as a function of jet η for the "track counting high purity" tagger at 1% uds-jet efficiency.

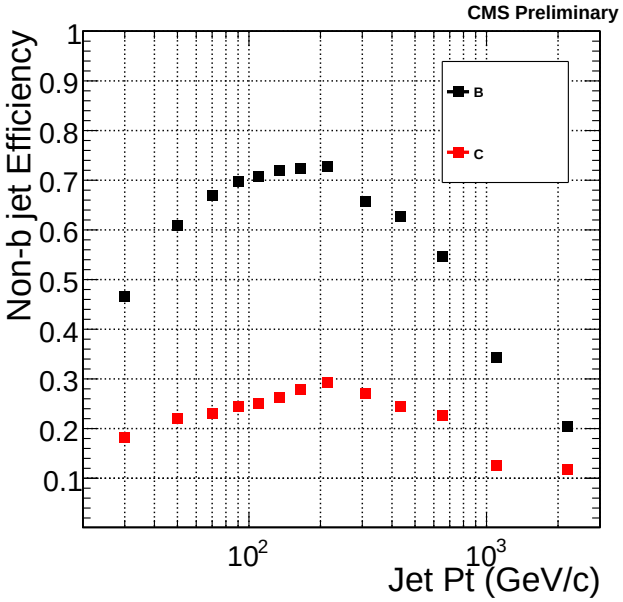

Figure 14: Efficiencies for b and c as a function of jet p_{T} for the "track counting high purity" tagger at 5% uds-Jet efficiency. Please note that a 1% uds efficiency was not used since in the last p_{T} bin the efficiency is never so low.

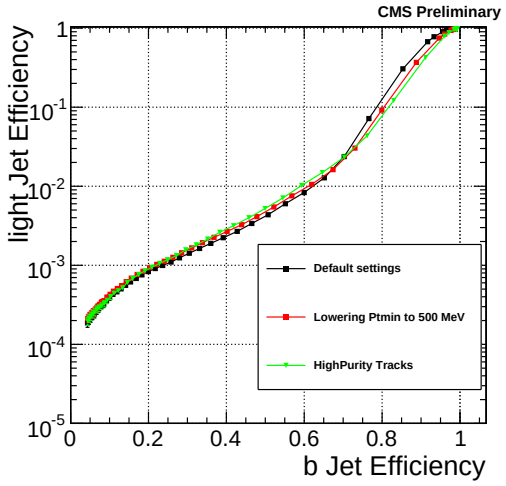

Figure 15: Light jet mistag rate versus efficiency for the "track counting high efficiency" algorithm, using different track quality selection criteria.

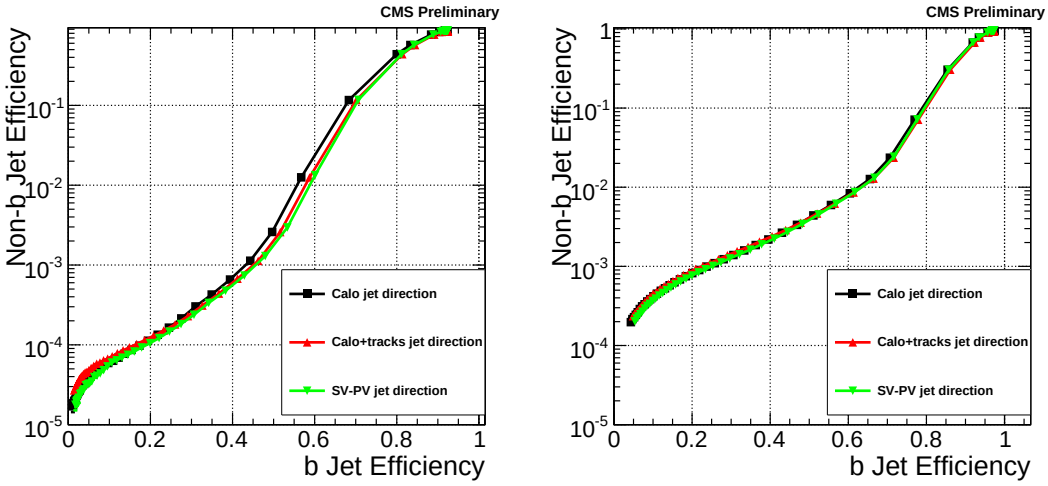

Figure 16: uds mistag rate versus efficiency for the "track counting" algorithms, with different options for the choice of the jet axis used for the track quality selection and the computation of the IP sign.

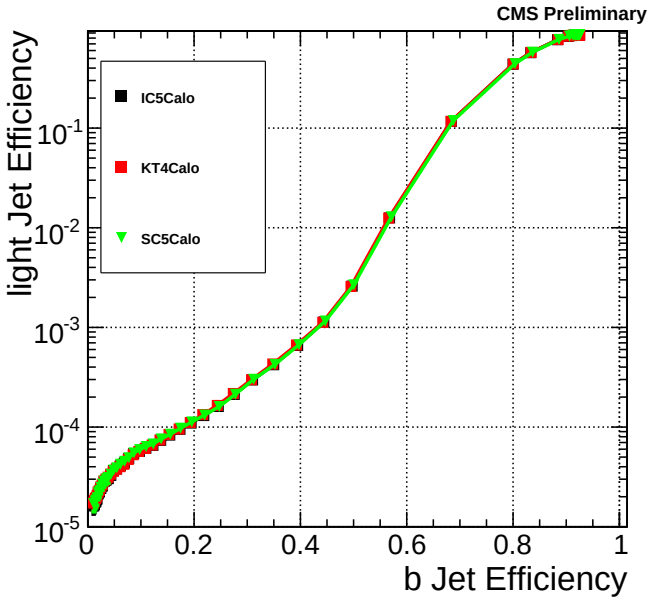

Figure 17: uds mistag rate versus efficiency for the "track counting high purity" algorithm, with different jet clustering algorithms.