# Beauty Production and Identification at CMS

Alexander Schmidt[a]
On behalf of the CMS Collaboration

[a]Universität Zürich, Physik-Institut,
Winterthurerstr. 190, CH-8057 Zürich, Switzerland

At the LHC, b-hadrons will be produced in a very high quantity at a yet unreached center-of-mass energy, enabling precision measurements to improve our understanding of the physics of b-quarks. The capability to measure the inclusive b-hadron production cross section is discussed on the basis of realistic detector simulations. Effects which are limiting the accessible range of differential cross sections are outlined.

The identification of b-jets is a crucial tool for a large number of topics, like top quark physics, and potential discoveries of the Higgs Boson and Supersymmetry. The applied methods of identifying b-jets are addressed and a discussion of the expected performance in terms of efficiencies and misidentification probabilities under realistic conditions is given. Approaches to measure these efficiencies with real data are also presented.

## 1. Introduction

B-hadrons are involved in a huge variety of physics studies at the LHC. This is not only true for the b-physics focused LHCb experiment but also for the general purpose experiments CMS and ATLAS. The precise knowledge of the b production cross-section is of crucial importance for an understanding of backgrounds in many searches for physics beyond the Standard Model (BSM). This measurement also provides a good test of the theoretical QCD predictions that have recently been subject to many discussions. The b-quark production cross-section was first measured in 1988 by the UA1 experiment at CERN [1]. Later measurements at the Tevatron [2–6] showed significant deviations from the theoretical predictions, even when running at UA1 energies [7]. Recent calculations [8] with improved fragmentation schemes lead to a better agreement so that the differences are not alarming anymore. The kinematical range accessible by experiments has been quite limited so far and it will be significantly extended by the LHC experiments. Studies for CMS, based on Monte Carlo simulations, suggest that the differential $b\bar{b}$ production cross-section can be measured up to $p_t < 1.5$ TeV [9]. This will be discussed in more detail in Section 3. An important part of this analysis is the application of methods to identify hadronic jets stemming from b-quarks.

Processes with b-quarks in the final state need to be identified efficiently and inclusively for a wide range of other studies, for example in top quark and Higgs physics. For instance, the Standard Model Higgs Boson decays preferably into b-quarks with a branching ratio larger than 50% for low masses ($m_H < 130$ GeV). Recent studies [10] showed that the discovery of the $H \to b\bar{b}$ decay is extremely difficult due to the limited ability to suppress backgrounds from mis-identified jets. Further topics in which b-tagging of hadronic jets is a crucial ingredient are top physics and BSM searches. Technical details for these inclusive b-tagging methods are given in Section 2.

Information about the CMS detector is available in [11].

## 2. B-Tagging in CMS

The primary goal of "b-tagging" is to identify the presence of b-quarks as efficiently as possible while keeping a reasonable purity. Two main properties of b-hadron decays are exploited for this purpose: the lifetime of about $\tau = 1.6$ ps with $c\tau = 490$ $\mu$m which leads to a measurable

separation between the primary event vertex and the b decay vertex. Secondly, the semileptonic decay modes with a branching ratio of about 10% per lepton flavour provide a clean signature that can be used for b-identification.

Production of quarks, either in the hard interaction or as decay products, leads to a a bundle of particles in the final state (leptons and hadrons) which are emitted in about the same direction. This jet of particles can be detected by the tracker and calorimeters. The direction defined by the jet is subsequently used to associate tracks to the jet. The association is done simply by using a cone around the jet direction, defined in pseudorapidity $\eta$ and azimuthal angle $\phi$. Tracks are then extrapolated back to the interaction point where the calculation of the impact parameter [12] is performed. The impact parameter (IP) is defined as the distance between the primary vertex and the point of closest approach of the track to this vertex. As shown in Section 2.1, the IP already provides a powerful discrimination between b-jets and non-b-jets. Since the measurement of the IP can be distorted by various effects like multiple scattering and wrong hit-to-track associations, it is important to take the measurement's error into account. Therefore, the variable of choice is the IP significance, defined as the ratio $value/error$. The distribution of the transverse impact parameter significance is shown in Figures 1 and 2 for light flavour jets and b-jets, respectively, in various detector misalignment scenarios which are discussed in Section 2.1. In these figures, the second track, ordered by the IP significance itself is shown. The first track has a higher probability to suffer from mismeasurements which causes the purity to decrease. It is also possible to use the third track which results in an even better purity at the cost of a lower efficiency.

The most simple algorithms to distinguish b-jets from other jets use this IP significance distribution as discriminator. More advanced algorithms make use of secondary vertex information [13]. Secondary vertices can be reconstructed in b-jets using the associated tracks, passing well defined quality criteria. A vertex fit [14] is performed and the point of the b-hadron decay is reconstructed with very high precision. The dis-

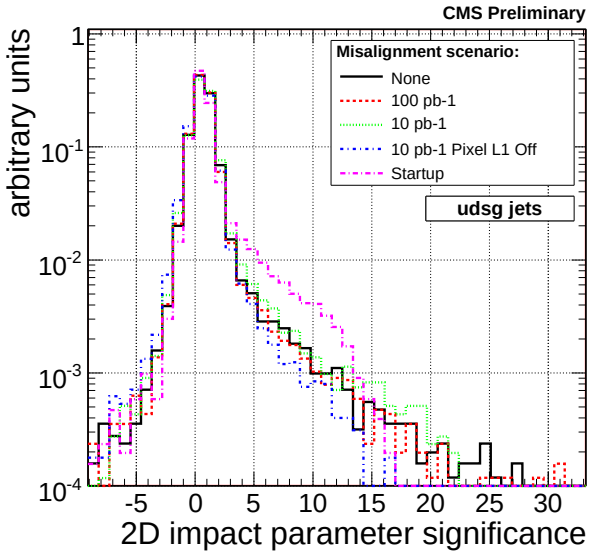


Figure 1. Distribution of the transverse impact parameter significance of the second track (ordered by I.P. significance) for various misalignment scenarios and jet flavours. Only light flavour jets are shown here. The term "light flavour jets" corresponds to jets originating from u, d and s quarks as well as from gluons. Inclusive $t\bar{t}$ events have been used.

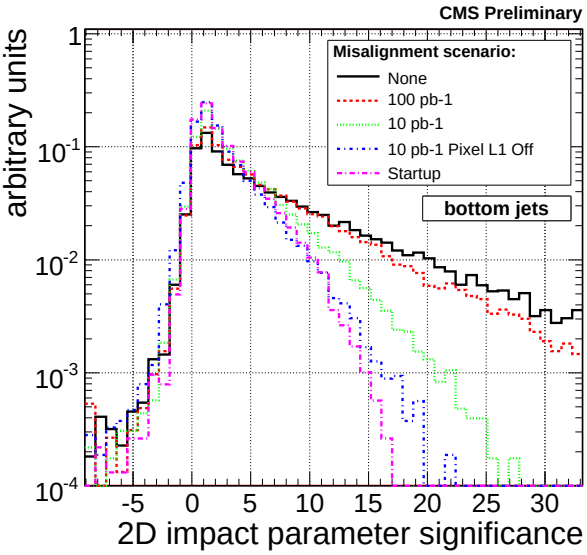


Figure 2. As Figure 1 but for b-jets instead of light flavour jets.

tance between the primary event vertex and the secondary vertex serves as a powerful and robust discriminator for the identification of b-jets [15].

### 2.1. Detector Alignment and Performance

The calibration and alignment of the detector are of major importance during the startup phase of the experiment. Since physicists want to do their analyses as early as possible, it is important to understand the impact of a misaligned detector on physics results. For this purpose, the following set of scenarios has been defined and the effect on b-tagging has been studied [15]:

- startup: only information from survey measurements, the Laser Alignment System and cosmic muon tracks can be used to perform a detector alignment.
- 10 pb$^{-1}$: the tracker can be aligned by using cosmic muon data and a sample of collision tracks, mainly being isolated hadrons in minimum bias events and muons from the decays of low mass resonances like J/$\psi$ and Upsilon. For the pixel detector, it is assumed that the alignment of its larger structures can be improved by a factor of 5, but that there is no improvement of the module-level alignment. For the strip tracker, it is assumed that the sub-detector positions can be aligned with an accuracy of 100 $\mu$m.
- 100 pb$^{-1}$: high $p_T$ muons from Z and W boson decays are available in significant quantity. The misalignment of the pixel tracker is expected to be $\mathcal{O}(20\ \mu\text{m})$, and that of the strip tracker $\mathcal{O}(30 \ldots 50\ \mu\text{m})$.

The tracker can be considered to be aligned with an integrated luminosity of 1000 pb$^{-1}$ which can probably be achieved within 1 year of detector operation.

The effects on the IP significance are shown in Figures 1 and 2. The distribution for b-jets becomes narrower due to the increased error of the measurement. The distribution for light flavour jets is not as much affected but it becomes broader in the startup scenario. This is due to the increased rate of fake tracks or tracks with wrongly associated hits in the first detector layers, which cause pathologically high IP values. The distributions for the different flavours become more similar with increasing misalignment and therefore the separation power is reduced. Obviously, this is a rather large effect and it can be expected that purely impact parameter based tagging algorithms are not optimal for early data taking. The performance in terms of misidentification rates depending on b-tagging efficiency is shown for the "track counting" algorithm [12,15] in Figure 3. This algorithm simply uses the transverse IP significance of the second track as a discriminator and a continuous cut on this variable is applied.

As mentioned, single tracks and the associated impact parameters can be quite sensitive to mismeasurements of any kind and it is therefore worthwhile to investigate more robust taggers. Soft lepton based taggers proved to be very resistant against detector effects [15], but they are limited by the small fraction of semileptonic b decays (about 10% per lepton flavour). Promising alternatives are secondary vertex based taggers.

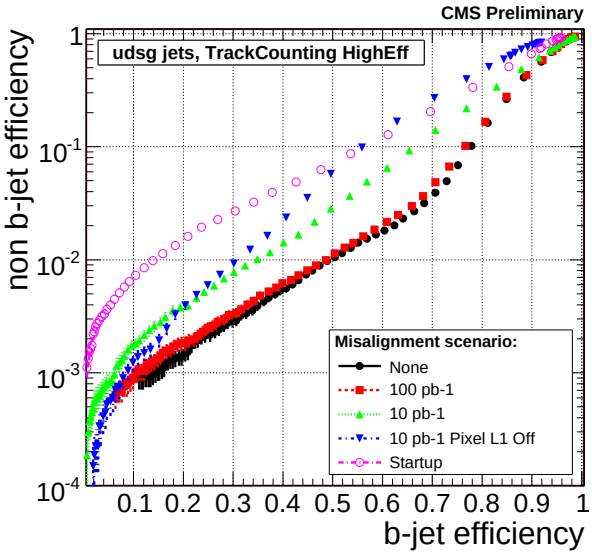


Figure 3. b-jet efficiency versus non b-jet efficiency for the various misalignment scenarios in case of the TrackCounting (high efficiency) algorithm, for light flavour jets. Inclusive $t\bar{t}$ events have been used.

Studies showed that the observables associated to a reconstructed secondary vertex are indeed more robust [15], since several tracks are combined and outlying tracks are included with smaller weights. For instance, the so-called "simple secondary vertex" tagger, which only uses the flight distance significance of the secondary vertex as discriminator, turned out to be a promising candidate. After optimizations, this algorithm reaches a b-jet identification efficiency of 65% at a light flavour misidentification rate of 2% in an aligned detector. Its robustness is visible in Figure 4 where the relative performance decrease is shown. It is defined as the ratio of misidentification efficiencies in a misaligned detector with respect to an aligned detector ($\epsilon^{mistag}_{misaligned}/\epsilon^{mistag}_{aligned}$) for various b-tagging algorithms. In the 10 pb$^{-1}$ scenario, the simple secondary vertex algorithm has a performance decrease of a factor of less than two, while the most powerful algorithm (in an ideal detector) loses more than a factor of 6 in a misaligned detector.

## 3. Measurement of the $b\bar{b}$ Cross-Section

The evaluation of the ability to measure the $b\bar{b}$ cross-section [9] is based upon a data sample produced with the PYTHIA [16] event generator using QCD jets, followed by a complete Monte Carlo simulation of the CMS detector. Conditions corresponding to the low-luminosity LHC run with $\mathcal{L} = 2 \cdot 10^{33}\mathrm{cm}^{-2}\mathrm{s}^{-1}$ have been applied. A total accumulated integrated luminosity of 10 fb$^{-1}$ is assumed.

Events are required to pass the Level-1 trigger selection for single muons with $p_t > 14$ GeV/$c$. At the High-Level trigger a muon+b-jets cross channel trigger is used with jet $E_t > 50$ GeV resulting in a trigger rate of 6.1 Hz. The offline selection requires a b-tagged jet to be present in the event using the most performant "combined" b-tagging algorithm [17] mentioned above. The average b-tagging efficiency in this sample is 65% in the barrel region and 10% less in the endcap region. The efficiency degrades to less than 50% for higher jet energies ($E_t > 500$ GeV) due to a worse track momentum resolution, increased track multiplicity from fragmentation and more

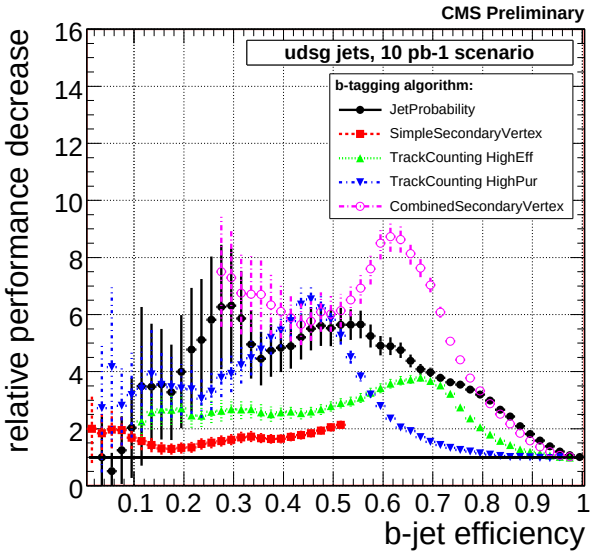


Figure 4. Relative performance decrease ($\epsilon^{mistag}_{misaligned}/\epsilon^{mistag}_{aligned}$) in light flavour mistagging rate for several tagging algorithms compared to a perfectly aligned tracker. The plot shows the 10 pb$^{-1}$ scenario . "TrackCounting HighEff" ("HighPur") refers to the case in which the second (third) track's IP significance is used as discriminator. The "combined secondary vertex" algorithm is the most powerful of all algorithms in an ideal detector. It combines all available information from track and vertex observables (Section 2.1) by applying a likelihood method.

difficult pattern recognition in dense jets.

For the measurement of the cross-section, four basic ingredients are necessary: the number of events passing the selection, the integrated luminosity, the signal selection efficiency and the event sample purity (signal fraction). The selected events and the luminosity are directly accessible by measurement. The selection efficiency and purity need to rely on simulation to a certain extent. The absolute predictions given by simulation suffer from large uncertainties, therefore methods relying as little as possible on simulation are used for this purpose. In the analysis presented here, the signal fraction is determined by using the signal and background shapes of the relative transverse momentum of muons with respect to the b-jet. A fit of the muon $p_t$ spectrum by the expected shapes for contributions from b, c and light quark events is shown in Figure 5. The normalisations of the three contributions, determined in the fit, represent the corresponding flavour fractions in the sample. The true flavour fractions are well reproduced within statistical errors. Further methods to measure efficiencies and the flavour content in event samples are discussed in Section 4.

The contribution of background from $\mathrm{t\bar{t}}$ events has been estimated to contribute less than 1% to the selected events. It becomes more pronounced in the high momentum region ($p_t > 500$ GeV/$c$) of the b-spectrum with a contribution of 2.4%.

Several sources of systematic uncertainties have been taken into account. The largest one is due to a 3% error on the jet energy scale, which leads to a 12% uncertainty on the cross-section measurement at $E_t > 50$ GeV. For very high energies, this error drops to 4%. The most important remaining uncertainties are:

- event selection procedure and Monte Carlo modeling of the detector, like lepton identification and angular and energy resolutions. This results in a 6% uncertainty.
- b-tagging uncertainties, 5%
- luminosity uncertainty, 5%
- fragmentation modeling, 9%

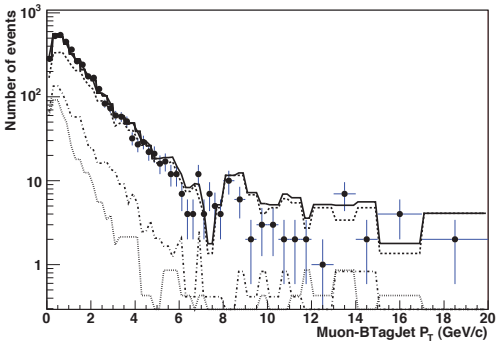


Figure 5. Fit of the muon $p_t$ spectrum with respect to the closest b-tagged jet. The contributions of tagged muons from b-events (dashed curve), c-events (dot-dashed curve) and light quark events (dotted curve) as defined by the fit are shown. The solid curve is the sum of the three contributions. [9]

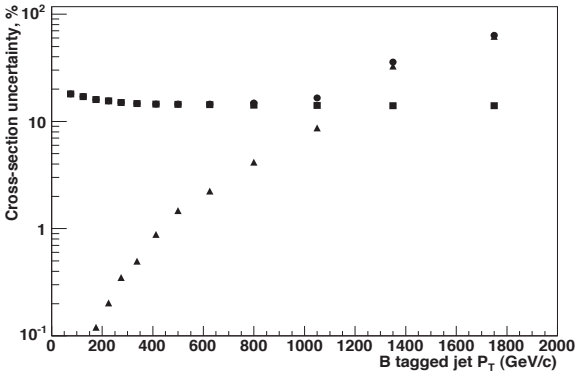


Figure 6. The statistical uncertainty for the cross section measurement (triangles), systematic (squares) uncertainty and total (dots) uncertainty as function of the b-tagged jet transverse momentum for 10 fb$^{-1}$. [9]

The total systematic and statistical errors are shown in Figure 6. It is visible that the analysis is limited by the systematic error of around 10% in the low $p_t$ region, while the statistical error becomes dominant in the high $p_t$ range. The analysis becomes statistically limited at 1.5 TeV for an assumed integrated luminosity of 10 fb$^{-1}$.

## 4. Performance Measurement with Data

Absolute predictions for tagging efficiencies given by the simulation are usually not very reliable. Methods to measure the event content and therefore the tagging efficiencies from data have been developed, for example the template fits to the relative muon $p_t$ as discussed above. Another example is the "System8" method developed by the D0 collaboration, which is also being applied in CMS [18]. This method exploits three uncorrelated identification criteria for b-jets:

- the working point of the lifetime tagging algorithm under study
- the cut on the relative transverse momentum of the muon with respect to the b-jet
- the presence of a second b-jet due to the preferred mode of b-quark pair production in QCD events

The various criteria are applied individually or all together leading to a system of eight linear equations, with eight unknowns which are the tagging efficiencies. The method assumes that the efficiency for tagging a jet with both the lifetime tag and the relative muon $p_t$ cut can be calculated as the product of the individual efficiencies. The only input from simulation are the correlation factors between the lifetime tag and the muon requirement as well as the ratio of the lifetime tagging efficiencies for b and c+light jets, respectively, corresponding to the two different data samples used to solve the system of equations. More details can be found in [18]. The result in terms of b-jet tagging efficiency determined with this method is shown in Figure 7. It is visible that the values determined with this method agree well with the true values within the

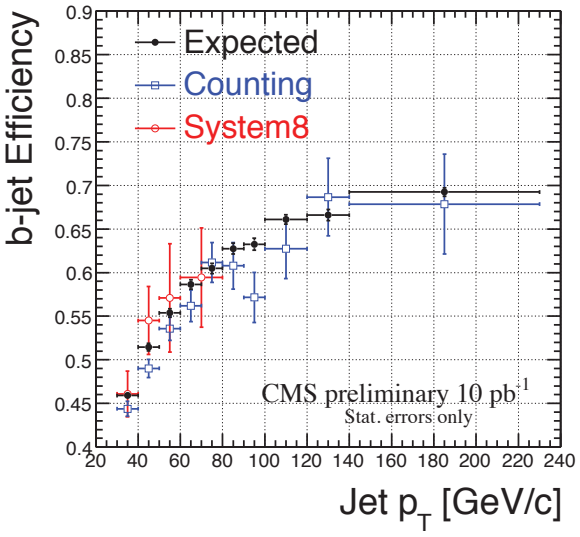


Figure 7. b-tagging efficiency as a function of jet $p_t$ as measured with the Counting and System8 methods. More details on the Counting method can be found in [18]. Results are shown with statistical errors for a corresponding integrated luminosity of about 10 pb$^{-1}$.

indicated errors.

Another important method to measure the misidentification rates is the use of the negative tail of the impact parameter distribution. The impact parameter is labelled positive (negative) if the track originates upstream (downstream) with respect to the primary vertex [19]. In the ideal case, the IP distribution would be symmetric around zero for light flavour jets and only positive for b-jets, since tracks from b-jets are really displaced in the direction of the jet, while tracks from light flavour jets originate at the primary vertex. High IP values of tracks originating at the primary vertex are mostly caused by instrumental mismeasurements or multiple scattering and therefore have an equal opportunity to cause positive or negative signs. In reality however, the distribution for b-jets does also have negative values due to mismeasurements. In addition the distribution for light flavours is not perfectly symmetric because it is affected by displaced processes such as long lived particles ($K^0_S, \Lambda$), photon conversions and other effects. The mistag efficiency can then be evaluated using the negative tag rate from multijet processes and a correction factor which needs to be obtained from simulation. Studies [19] show that this method can be applied with a relative precision of 7% for a light flavour mistag rate of 1% in case of an integrated luminosity of 100 pb$^{-1}$.

Another possibility is the use of the large number of top quark events expected at the LHC. It is possible to extract highly enriched b-jet samples from these events based on kinematical constraints exploiting the clear top quark signatures [20]. The expected precision of this method for 1 fb$^{-1}$ (10 fb$^{-1}$) is a relative accuracy of 6% (4%) on the b-jet identification efficiency in the barrel region and 10% (5%) in the forward region.

## 5. Conclusions

A large number of physics studies like Higgs searches and searches for physics beyond the Standard Model (BSM) depend on efficient and clean identification of b-quark jets, mainly to distinguish these processes from backgrounds. However, not all backgrounds are reducible, so the

precise knowledge of the $b\bar{b}$ production cross-section is critical for these searches. CMS will be able to measure the differential $b\bar{b}$ production cross-section up to 1.5 TeV. Technical details about these methods to identify b-jets have been presented and their behavior under realistic conditions has been studied. It is necessary to determine the performance of these b-identification algorithms in terms of efficiency and purity using data only, relying as little as possible on the Monte Carlo simulation. The CMS collaboration managed to develop and evaluate all necessary tools to achieve these tasks and is ready for first LHC collisions.

## REFERENCES


1. UA1 Collaboration, "Measurement of the bottom quark production cross section in proton-antiproton collisions at $\sqrt{s}$ = 0.63 TeV", Phys. Lett B **213**, 1988, 405-412.
2. CDF Collaboration, F. Abe et al., "Measurement of the $B^+$ total cross section and $B^+$ differential cross section $d\sigma/dp_T$ in $p\bar{p}$ collisions at $\sqrt{s}$ = 1.8 TeV", Phys. Rev. D **65** (2002) 052005.
3. CDF Collaboration, F. Abe et al., "Measurement of the B Meson Differential Cross Section $d\sigma/dp_T$ in $p\bar{p}$ Collisions at $\sqrt{s}$ = 1.8 TeV", Phys. Rev. Lett. **75** (1995).
4. D0 Collaboration, S. Abachi et al., "Inclusive $\mu$ and $b$-quark production cross-sections in $p\bar{p}$ collisions at $\sqrt{s} = 1.8$ TeV", Phys. Rev. Lett. **74** (1995) 3548.
5. D0 Collaboration, B. Abbott et al., "The $b\bar{b}$ production cross section and angular correlations in $p\bar{p}$ collisions at $\sqrt{s} = 1.8$ TeV", Phys. Lett. B **487** (2000) 264.
6. D0 Collaboration, B. Abbott et al., "Cross section for $b$-Jet production in $\bar{p}p$ collisions at $\sqrt{s}$ = 1.8 TeV", Phys. Rev. Lett. 85 (2000) 5068.
7. CDF Collaboration, F. Abe et al., "Measurement of the ratio of $b$ quark production cross sections in $\bar{p}p$ collisions at $\sqrt{s} = 630$ GeV and $\sqrt{s}$ = 1800 GeV", Phys. Rev. D 66 (2002) 032002.
8. M. Cacciari and P. Nason, "Is There a Significant Excess in Bottom Hadroproduction at the Tevatron?", Phys. Rev. Lett. **89** (Aug, 2002) 122003.
9. V. P. Andreev, D. B. Cline, S. Otwinowski, "Measurement of open beauty production at LHC with CMS", CMS Note **2006/120** (2006).
10. D. Benedetti, S. Cucciarelli, C. Hill, J. Incandela, S. A. Koay, C. Riccardi, A. Santocchia, A. Schmidt, P. Torre, and C. Weiser, "Observability of Higgs produced with top quarks and decaying to bottom quarks", J. Phys. G: Nucl. Part. Phys. **34** (2007), no. 5, N221-N250.
11. The CMS Collaboration, "The CMS experiment at the CERN LHC", J. Inst. **3** (2008) S08004.
12. A. Rizzi, F. Palla, and G. Segneri, "Track impact parameter based b-tagging with CMS", CMS Note **2006/019** (2006).
13. T. Müller, et. al., "Inclusive Secondary Vertex Reconstruction in Jets", CMS Note **2006/027** (2006).
14. T. Speer, K. Prokofiev, R. Fruehwirth, W. Waltenberger, P. Vanlaer, "Vertex Fitting in the CMS Tracker", CMS Note **2006/032** (2006).
15. CMS Collaboration, "Impact of Tracker Misalignment on the CMS b-Tagging Performance", CMS PAS **BTV 07 003** (2008).
16. T. Sjöstrand et al., "High-energy-physics event generation with PYTHIA 6.1", Comput. Phys. Commun. **135** (2001) 238.
17. Christian Weiser, "A Combined Secondary Vertex Based B-Tagging Algorithm in CMS", CMS Note **2006/014** (2006).
18. CMS Collaboration, "Performance Measurement of b-tagging Algorithms Using Data containing Muons within Jets", CMS PAS **BTV 07 001** (2008).
19. CMS Collaboration, "Evaluation of udsg Mistags for b-tagging using Negative Tags", CMS PAS **BTV 07 002** (2008).
20. Steven Lowette, Jorgen D'Hondt, Jan Heyninck, Pascal Vanlaer, "Offline Calibration of b-Jet Identification Efficiencies", CMS Note **2006/013** (2006).