PROCEEDINGS OF SCIENCE

Charm jet identification at the CMS experiment

Seth Moortgat*
Vrije Universiteit Brussel
E-mail: semoortg@vub.ac.be

on behalf of the CMS Collboration

Identification of jets originating from c quarks is becoming more and more important for a wide variety of standard model physics as well as searches for physics beyond the standard model. Recently, the CMS Collaboration developed a charm jet identification algorithm, trained to discriminate against b jets and light-quark or gluon jets. In this report, this newly developed algorithm will be described and its performance on the proton-proton collision data recorded by the CMS detector at a center-of-mass energy of $\sqrt{s}$ = 13 TeV during 2015 is presented.

VIII International Workshop On Charm Physics
5-9 September, 2016
Bologna, Italy

*Speaker.

PoS(CHARM2016)060

1. Introduction

At the Large Hadron Collider protons collide at a center-of-mass energy of up to 13 TeV. These collisions, and especially the particles that emerge from them are recorded by the Compact Muon Solenoid (CMS) detector [1]. If quarks[1] are formed in these collisions, they will hadronize and undergo a fragmentation process, leading to showers of particles in the detector which are called jets. The identification of jets from bottom quarks [2] has proven to be very useful in many analyses carried out by the CMS Collaboration. Since many interesting physics processes include charm quarks in their final states, also an algorithm to identify jets from charm quarks (a charm tagger) can be beneficial for such analyses. Interesting examples include supersymmetric models in which scalar squarks decay into charm quarks [3, 4], models with charged Higgs bosons that decay into charm quarks [5] or searches for standard model and beyond the standard model flavour-changing neutral current processes [6, 7, 8, 9]. Such an algorithm to identify jets from charm quarks (c jets) and distinguish them from jets originated by bottom quarks (b jets) or by up, down, strange quarks or gluons (collectively referred to as light jets) has been developed by the CMS Collaboration [10].

The CMS detector is made up from several subdetectors. The most important one for charm tagging is the silicon tracker, made of several layers of silicon pixels and strips and used for the identification of charged particle tracks. A 3.8 T magnetic field provided by a superconducting solenoid of 6 m internal diameter causes the direction of flight of the charged particles to bend, which allows for a very precise reconstruction of their momentum. A lead tungstate crystal electromagnetic calorimeter (ECAL), and a brass and scintillator hadron calorimeter (HCAL) are housed within the magnet as well and are of vital importance for the jet reconstruction. Outside of the magnet, muon chambers measure muon tracks using gas–ionization detectors. The above mentioned subdetectors are all embedded in the cylindrical barrel part of the detector. Extensions of these subdetectors can be found in the endcaps of the detector to provide additional coverage beyond the pseudorapidity reach of the barrel. A more detailed description of the CMS detector, including the definition of the coordinate system used, can be found in [1].

In Section 2 the algorithm for charm jet identification and its performance on simulations are discussed and in Section 3 the methods used for calibrating the algorithm to proton-proton collision data recorded by CMS are reviewed.

2. Algorithm for c jet identification

A dedicated algorithm [10] has been developed in order to separate c jets from jets initiated by other parton flavours. The tagging of c jets is achieved using a set of multivariate classification algorithms that combine input observables from the jets related to displaced tracks, secondary vertices (SV) and soft leptons to produce a discriminating output variable referred to as discriminator. Jets from charm quarks are on average heavier and their tracks are more displaced with respect to the primary interaction vertex compared to light jets, but they are lighter and their tracks are less displaced compared to b jets. This is why the distinction in the background flavours is needed when using binary classification algorithms. SV are reconstructed using the Inclusive Vertex Finding [11] algorithm. In order to allow for more reconstructed SV in c jets, the reconstruction criteria

[1] For top quarks, the story is different as they will decay almost immediately in a bottom quark and a W boson.

are less stringent compared to those used in the SV reconstruction for the existing bottom tagging techniques.

The identification of c jets is achieved using machine learning algorithms, namely boosted decision trees (BDT). One such BDT is used for discrimination between charm and light jets and another one is used for discriminating charm jets from bottom jets. Such machine learning algorithms are trained on simulated jets which contain information on the originating partons. The charm tagger is trained on simulated QCD multijet events. After the training of the algorithm one can test its prediction power by validating the performance of the algorithm on an independent sample, for which a sample of simulated top pair events is used. The training is performed using the TMVA [12] software package.

Two BDTs are thus trained, with so far a focus put on optimizing the discrimination between c jets and light jets (CvsL), since the discrimination between c jets and b jets (CvsB) is in some sense already possible with the existing b jet identification algorithms. Therefore the performance of the CvsB discrimination is not yet optimized, but it will be in the next version of the charm tagger.

In order to test the performance of the c tagging algorithm, the outputs of the two BDTs have to be combined in a two–dimensional plane and, correspondingly, two–dimensional selections need to be made to identify most optimally the c jets while the background flavours are rejected. Figure 1 on the left shows the distribution of jets of different flavours in the plane formed by the two discriminators. The BDT classifiers output a value close to 1 for signal–like jets and -1 for background–like ones, therefore c jets will be located towards the upper right corner of this plot whereas b jets and light jets are located more towards the bottom right and the top left corners, respectively. A rectangular selection toward the upper right corner of this phase space is made in order to isolate c jets from the background. The corresponding performance curves are presented by drawing constant charm efficiency (ε^c) contour lines in the plane representing the light and b jet mistag efficiencies (ε^{light} and ε^b respectively), as shown in Figure 1 on the right for jets with transverse momentum of $p_T > 20$ GeV and a pseudorapidity range of $|\eta| < 2.4$. This two–dimensional structure introduces the freedom to tune the light and b jet mistag efficiencies for a given constant predefined charm efficiency.

The calibration of the algorithm on data can however not be done for every possible selection in the two–dimensional phase space of discriminators. Therefore, three working points (WP) have been defined on which the calibration is performed: the loose (L), medium (M) and tight (T) WP. The WP threshold definitions and the corresponding global efficiencies are summarised in Table 1. The loose WP has been chosen such that it is rejecting rather well the b jets (but with a very high mistag rate for light jets), whereas the tight WP is specialized in rejecting light jets (but with a high mistag rate for b jets). The medium WP rejects both b jets and light jets.

WP	ε^c	ε^b	ε^{light}	**CvsL**	**CvsB**
c-tagger L	0.9	0.45	0.99	$>$ -0.67	$>$-0.23
c-tagger M	0.39	0.26	0.19	$>$ 0.05	$>$ -0.16
c-tagger T	0.2	0.24	0.02	$>$ 0.45	$>$ -0.35

Table 1: Definitions of the three working points with the corresponding selections on the discriminator values and the global efficiencies, obtained from simulated top pair samples, for each flavour.

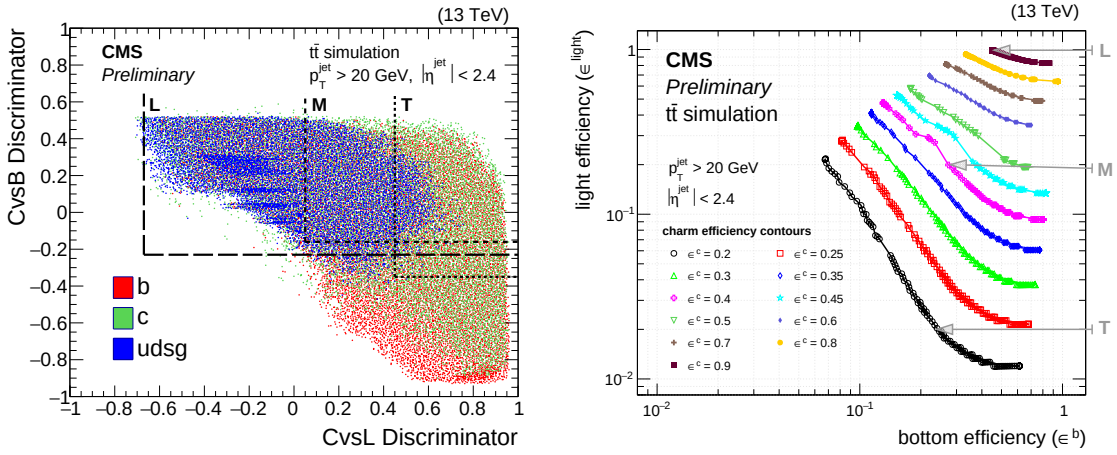

Figure 1: Left: two–dimensional scatter overlay of the BDT discriminators for b (red), c (green), and light jets (blue). The CvsL discriminator is shown on the x-axis and the CvsB discriminator is shown on the y-axis. Right: Distribution of the bottom versus the light mistag efficiency for different values of a constant charm efficiency.

3. Calibration of the algorithm on data

To correct for possible discrepancies between simulations and data, scale factors (SF) are measured for proton–proton collisions collected by the CMS detector in 2015 for the three WPs from Table 1. Such SF are defined as the ratio of the measured selection efficiency in data for a certain jet flavour f ($\varepsilon_f(DATA)$) to the selection efficiency of that jet flavour found in simulation ($\varepsilon_f(SIM)$), as defined in Equation (3.1).

$$SF_f = \frac{\varepsilon_f(DATA)}{\varepsilon_f(SIM)} \qquad f \in \{\text{c,b,light}\} \tag{3.1}$$

For the current algorithm and at the time of writing, only light- and c jet SF are measured. For light jets the negative tag method [2], which is also used to measure the mistag rates for the b tagging algorithms, is used. For c jets, two new methods for measuring the scale factors have been developed, one using W+charm events and another one using semileptonic top pair events. These two methods are discussed in the following Sections.

3.1 Measurement of the charm jet identification scale factors using W+charm events

The production of a W boson in association with a c quark proceeds at leading order via the processes shown in Figure 2 in the two left diagrams. A key property of the this process is the presence of a charm quark and a W boson with opposite–sign (OS) electric charges, whereas background processes are expected to deliver evenly OS and same–sign (SS) events, of which an example is shown in Figure 2 on the right for W+c$\bar{\text{c}}$(b$\bar{\text{b}}$), where the pair of heavy flavour quarks in produced through gluon splitting. Exploiting this difference a very pure sample of c jets can thus be obtained by the OS-SS subtraction method [13].

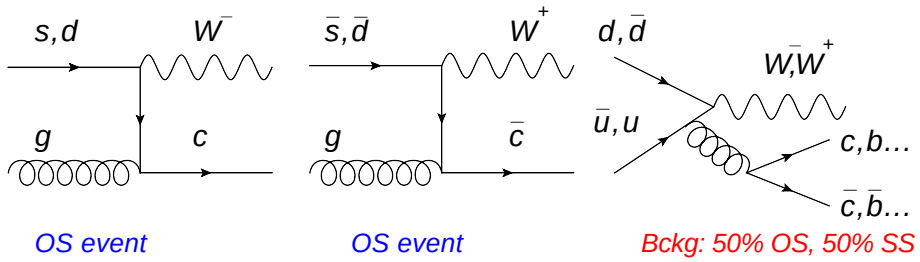

Figure 2: Left and middle: leading order production of W+charm signal with opposite–sign charges (OS). Right: production of W+charm final state through gluon splitting. In the gluon splitting process, two c quarks are produced, one with an opposite–sign (OS) electric charge with respect to the W boson, and one with the same–sign (SS) electric charge, leading to an equal probability to select an OS pair (W,c) or a SS pair.

The leptonic decay of a W boson into a muon or an electron is characterized by the presence of a high transverse momentum, isolated lepton. The charge of this isolated electron or muon identifies the electric charge of the corresponding leptonically decaying W boson. The charge of the charm quark is deduced by requiring a well-identified, non-isolated muon among the jet constituents and identifying the charge of that muon as the charge of the charm quark. Events with OS (SS) electric charge are defined as events for which the non-isolated muon has opposite (same) charge as the charge of the isolated electron or muon from the W decay. Dedicated studies have shown that there is no significant bias (within uncertainties) in the SF measurement due to the requirement of having a muon inside the jet. The SFs for the c–tagging algorithm are determined using selected jets with a non–isolated muon, as a function of the jet transverse momentum using Equation (3.1) for the three WPs from Table 1. The results are shown in red in the top panels of Figure 4.

3.2 Measurement of the charm jet identification scale factors using semileptonic top pair events

Semileptonic top quark pair decays contain a great amount of c jets due to the hadronic decays of the W boson (roughly 25% of the jets). An event selection following closely the one from Reference [14] results in events with an isolated lepton and four jets. The jets are associated to the final state particles, namely to the b jet from the leptonic top quark decay, the b jet from the hadronic top quark decay and the two jets from the hadronic W decay. From all possible permutations, the best one is selected by choosing the smallest value of a mass discriminant λ_M that combines the invariant mass of the two jets from the hadronic W decay and the invariant mass of the jets belonging to the hadronically decaying top quark into a single value which has the same properties of the negative logarithm of a likelihood ratio. The distribution of λ_M after the full event selection is shown in Figure 3 and shows a clear distinction between correct permutations and wrong ones or background events.

A simultaneous maximum likelihood fit on the binned λ_M distributions (using signal and background templates derived from simulations) is performed in order to disentangle the contributions of the correctly–matched top pair jet permutations from the other background components as well

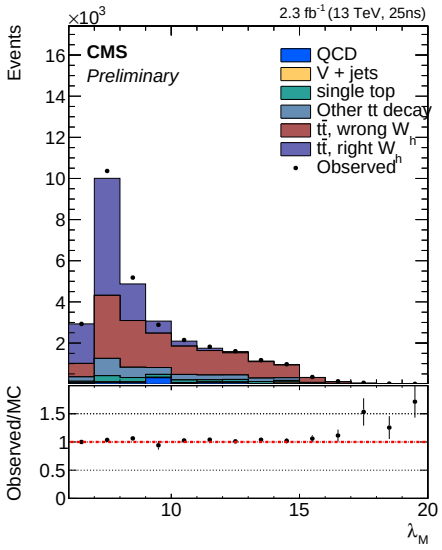

Figure 3: Distributions of λ_M (on data and simulations) after the full selection. The different simulated processes contributing are shown with different colours. The major contributions are semileptonic top quark pair decays with the hadronic W properly matched to the generator particles (violet), wrongly matched semileptonic top quark decays (red), and non semileptonic top quark decays (azure). Minor contributions are also present due to single top events (green), vector-boson plus jets (labeled as "V+jets", in yellow), and multijet production ("QCD", in blue).

as to infer the value of SF_c, which is considered as a free parameter in the fit function. The results obtained from the negative tag method for the light jet SFs are also used in this fit. A detailed description of this fitting procedure, including a treatment of the systematics can be found in Reference [10]. The resulting SFs (not binned in transverse momentum of the jet due to statistical limitations) can be seen in blue in the top panels of Figure 4 for the three WPs.

3.3 Scale factor combinations

The two SF_c measurements described in the previous Sections can be combined via a weighted average, taking into account the full covariance matrix of the uncertainties, using the so–called BLUE method [15]. The results from the combined measurement are shown in the bottom panels of Figure 4 for the three WPs. For all the three WPs the two methods agree well within uncertainties and the combination is mostly consistent with a SF of 1 within uncertainties of 5 to 15 percent depending on the range of transverse momentum that is considered.

4. Conclusion and prospects

A new tool for identifying jets originating from charm quarks has been developed for the first time in the CMS Collaboration. The algorithm uses two boosted decision trees, trained and tested on simulated QCD multijet and top pair events respectively. The mistag rate for light jets has been measured in multijet events. Two new methods have been developed to measure the charm

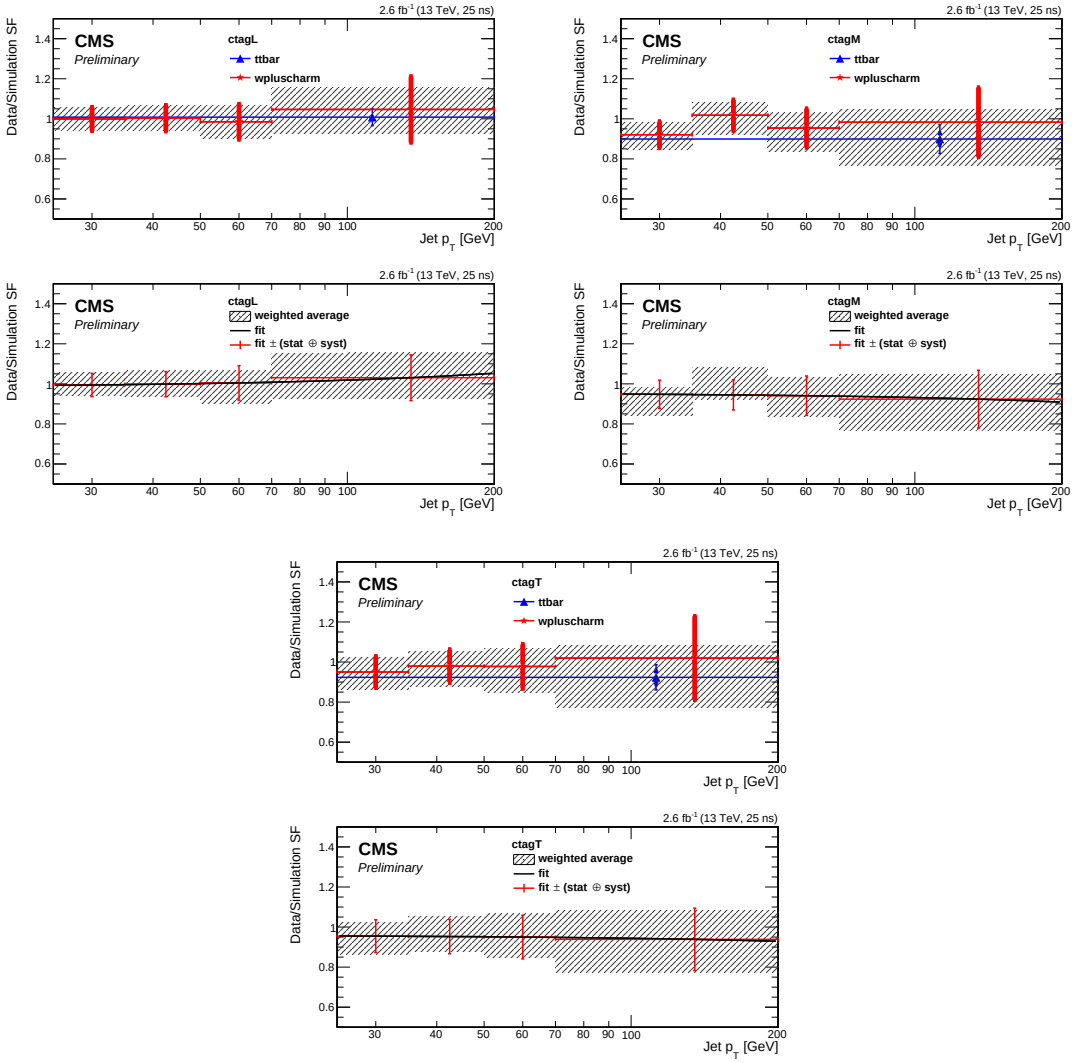

Figure 4: (upper panels) Data-to-simulation scale factor of the charm tagging efficiency for the c-tagging WP (loose on the top left, medium on the top right, tight on the bottom) as measured with the two methods, with (thick error bar) statistical error and (narrow error bar) combined statistical and systematic uncertainties. The combined SF value with its overall uncertainty is displayed as a hatched area. (lower panels) Same combined SF value with the result of a linear fit function superimposed (solid curve). The combined statistical and systematic uncertainty is centred around the fit result (points with error bars). The last bin includes the overflow.

tagging efficiency on the proton–proton collision data collected by CMS in 2015. By comparing the efficiencies obtained from data to those obtained in simulations for three predefined working points, a set of dedicated data to simulation corrections was derived. For the data collected in 2016, a more optimized version of the algorithm has already been developed and scale factors will also be measured for this newer version.

References

[1] CMS Collaboration. The CMS experiment at the CERN LHC. *JINST*, 3:S08004, 2008.

[2] CMS Collaboration. Identification of b quark jets at the CMS Experiment in the LHC Run 2. Technical Report CMS-PAS-BTV-15-001, CERN, Geneva, 2016.

[3] CMS Collaboration. Searches for third-generation squark production in fully hadronic final states in proton-proton collisions at $\sqrt{s} = 8$ TeV. *JHEP*, 06:116, 2015.

[4] ATLAS Collaboration. Search for Scalar Charm Quark Pair Production in pp Collisions at $\sqrt{s} = 8$ TeV with the ATLAS Detector. *Phys. Rev. Lett.*, 114(16):161801, 2015.

[5] CMS Collaboration. Search for a light charged Higgs boson decaying to $c\bar{s}$ in pp collisions at $\sqrt{s} = 8$ TeV. *JHEP*, 12:178, 2015.

[6] CMS Collaboration. Search for associated production of a Z boson with a single top quark and for tZ flavour-changing interactions in pp collisions at sqrt(s) = 8 TeV. Technical Report CMS-PAS-TOP-12-039, 2016.

[7] CMS Collaboration. Search for anomalous single top quark production inassociation with a photon. Technical Report CMS-PAS-TOP-14-003, 2014.

[8] CMS Collaboration. Search for anomalous Wtb couplings and top FCNC in t-channel single-top-quark events. Technical Report CMS-PAS-TOP-14-007, CERN, Geneva, 2014.

[9] CMS Collaboration. Combined multilepton and diphoton limit on t to cH. Technical Report CMS-PAS-HIG-13-034, CERN, Geneva, 2014.

[10] CMS Collaboration. Identification of c-quark jets at the CMS experiment. Technical Report CMS-PAS-BTV-16-001, CERN, Geneva, 2016.

[11] CMS Collaboration. Measurement of $B\bar{B}$ Angular Correlations based on Secondary Vertex Reconstruction at $\sqrt{s} = 7$ TeV. *JHEP*, 03:136, 2011.

[12] Andreas Hoecker, Peter Speckmayer, Joerg Stelzer, Jan Therhaag, Eckhard von Toerne, and Helge Voss. TMVA: Toolkit for Multivariate Data Analysis. *PoS*, ACAT:040, 2007.

[13] CMS Collaboration. Measurement of associated W + charm production in pp collisions at $\sqrt{s}$ = 7 TeV. *JHEP*, 02:013, 2014.

[14] CMS Collaboration. Measurement of the inclusive and differential tt production cross sections in lepton + jets final states at 13 TeV. Technical Report CMS-PAS-TOP-16-008, CERN, Geneva, 2016.

[15] L. Lyons, D. Gibaut, and P. Clifford. How to combine correlated estimates of a single physical quantity. *Nucl. Instrum. Meth. A*, 270:110, 1988.