Identification of c-quark jets at the CMS experiment

The CMS Collaboration

Abstract

An accurate identification of jets originating from b quarks is of primary importance in many measurements and searches at the LHC. The development of a charm tagger, identifying jets initiated by charm jets, would be of similar importance. In this note a technique where an MVA–based discriminator is used in order to select charm jets is presented, together with its expected performance on simulations, and its calibration on W+c and top quark pairs. The datasets used for the calibration of the algorithm are from proton-proton collisions at 13 TeV, recorded by the CMS experiment during the first year of the LHC Run II.
1 Introduction

Many interesting physics processes include charm quarks in their final states, like supersymmetric models in which the lightest scalar squark decays into charm quarks.

So far, the possibility to observe such processes in CMS has been limited by the absence of a dedicated tool. A charm tagger has been developed starting from the pre-existing b tagging [1] algorithms, which provided a solid base upon which build a dedicated MVA classifier, implemented as a set of two boosted decision trees (BDTs) [2]. Like b jets, charm jets contain long-lived hadrons with a relatively large branching fraction into final states including leptons, even though such features are less marking than in the b jets case. It is therefore possible to use the same features used in the b tagging, such as track displacement, secondary vertex mass, flight distance and significance, to achieve discrimination against both light and b jets, as typically the distributions of such features for charm jets lie in between the ones of b and light jets.

The most important detector in CMS for charm tagging is the silicon tracker, made of several layers of silicon pixels and strips, immersed in a 3.8 T magnetic field provided by a superconducting solenoid of 6 m internal diameter. A lead tungstate crystal electromagnetic calorimeter (ECAL), and a brass and scintillator hadron calorimeter (HCAL) are housed within the magnet as well. Forward calorimeters provide additional coverage beyond the pseudorapidity reach of the calorimeters previously mentioned. Muons are measured in gas–ionisation detectors embeded in the return yoke of the magnet. A more detailed description of the CMS detector, including the definition of the coordinate system used, can be found in Ref. [3].

In Section 2 we briefly describe the simulated proton-proton collisions and the recorded data used for the training of the tool and its performance measurement. A detailed description of the algorithm and its training procedure is given in Section 3, together with expected performance computed on Monte Carlo simulations and comparison of relevant input variables of the algorithm between simulations and recorded data. Section 4 reviews the methods used to measure efficiencies and mistagging rates on real data and their combination.

2 Simulated and recorded events

The datasets used in this analysis, corresponding to 2.6 fb⁻¹ of proton–proton collisions at $\sqrt{s} = 13$ TeV, were recorded by the CMS detector during 2015. The collisions were delivered by the LHC with a bunch spacing of 25 ns.

Monte Carlo simulations (MC) of specific processes have been used in order to compare the observed spectra of physical observables and as input dataset for the supervised training of the BDTs. POWHEG (v2) [4–6] has been used to simulate top quark pair production [7] and single top associated production (tW) [8]. AMC@NLO [9] has been used to simulate vector boson production in association with jets, and t-channel single top. QCD multi–jet production has been simulated with PYTHIA [10]. All samples use PYTHIA version 8 to simulate parton showering; the tool is tuned with the recently optimised CMS Underlying Event Tune (CUETP8M1) [11]. Additional pileup proton-proton collisions are injected in the simulated events to model this effect.
3 Algorithm

A dedicated algorithm has been developed in order to separate c jets from jets initiated by other quark flavours. For convenience, we divide the background jets into light jets, initiated by u, d, s quarks and gluons, and b jets, initiated by bottom quarks. 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, and soft leptons to produce a discriminating output variable referred to as discriminator. Those classification algorithms are trained on labelled input data.

This Section describes in more detail the reconstruction of the jets and secondary vertices, the software, and the classification algorithm used.

3.1 Jet reconstruction and secondary vertex finding

The particle flow (PF) [12–14] algorithm identifies all stable particles inside the CMS detector by combining in an optimal way the various sub-detector elements. Photon energies are obtained from deposits in the ECAL, corrected for zero-suppression effects. Electron energy is determined by combining the measurements from its track in the tracker system and its energy deposit in the ECAL as well as by summing up all the energy of bremsstrahlung photons that are spatially compatible with the electron. Muon energy is determined from the track curvature measured both in the central tracker and the muon chambers. Charged hadrons energies are measured combining the momentum measurements of their tracks inside the tracker system and the deposits in the ECAL and HCAL, again correcting for zero-suppression effects and for the response function of the calorimeters to hadronic showers. Finally, the neutral hadron energies correspond to the remaining corrected ECAL and HCAL energies.

The primary interaction vertex (PV), corresponding to the hard scattering, is selected by looking for the combination of clustered physics objects which maximize the sum of the squared transverse momenta associated to the vertex.

Particle flow jets (PFJets) are built by clustering the PF objects using an anti-$k_T$ jet clustering algorithm [15] with a distance parameter of $R = 0.4$ (AK4 jets). The jet energy is corrected for energy deposits from pile-up charged particles and for detector-response effects [16]. The jets considered for the training and performance validation of the algorithm have the minimal requirements: $p_T > 15$ GeV and $|\eta| < 2.4$. However, when explicitly mentioned, tighter requirements are applied.

Jet flavours are determined in simulated events by adding to the list of generated stable particles the heavy hadrons produced in the event, scaled down to a negligible momentum. By re-clustering the jets with this newly added particles it is possible to assign a jet flavour based on these probe hadrons content. Jets containing b hadrons are defined as b jets; the ones containing charmed hadrons, but no b hadrons, are defined c jets; the remaining jets are considered light jets. Given the hadrons have their momentum scaled down, the total jet momentum is not affected by this process.

The tracks used to compute the input features of the tagging algorithms are required to pass basic trajectory fit quality and displacement requirements. The former ensures optimal momentum and spatial resolution while the latter mitigates the contribution of pile-up tracks, while still retaining flavour-discriminating capabilities. The displacement requirements enforce the longitudinal (transverse) impact parameter of the track with respect to the primary vertex to be smaller than 17 (0.2) cm, the distance between the track and the jet axis at their point of closest approach not to exceed 0.07 cm, and the distance between said point of closest approach and the primary vertex to be smaller than 5 cm.
Low-energetic, non-isolated electrons and muons in jets, together referred to as soft leptons (SL), are also used to compute input features for the taggers. Soft muons are first searched for amongst the PF jet constituents, and if none are found the particle is matched to a collection of reconstructed muons. Soft electrons are directly matched to a collection of reconstructed electrons on which extra track quality requirements are imposed. The information provided by soft leptons can be very relevant in the distinction between jet flavours due to the high semi-leptonic branching fraction of heavy hadrons.

Like b tagging algorithms, the c tagging exploits the lifetime of heavy hadrons, which is on average long enough to allow them to travel a few millimetres before decaying. The trajectories reconstructed from the decay products of such heavy hadrons will be displaced with respect to the PV and can be clustered around a secondary vertex (SV). The Inclusive Secondary Vertex Finder (IVF) [17] is used to find such secondary vertices. The IVF is independent of the jet clustering itself and only uses the reconstructed tracks in the event, after applying basic quality and $p_T$ requirements on them. An additional upper threshold of 0.3 cm on the track longitudinal impact parameter is applied to mitigate the effect of pile–up, while still retaining the ability to detect heavy hadrons decays. The IVF procedure starts by selecting seed-tracks based on the large impact parameter significance and then proceeds to cluster additional tracks into the same SV. The compatibility between the tracks and the seed is evaluated based on their spatial distance, angle, and separation relative to the track impact parameter with respect to the PV. The clustered tracks are then fitted with the outlier-resistant Adaptive Vertex Fitting algorithm (AVF) [18], forming a collection of secondary vertices which is then filtered requiring a minimum two– and three–dimensional flight distance significance with respect to the PV.

Given the shorter life–time of charm hadrons, which is roughly half of their bottom counterparts, the values for these thresholds have been halved to 0.25 (3D) and 1.25 (2D) with respect to the requirements used in the b tagging procedure [19]. Duplicate secondary vertices are then removed based on the fraction of shared tracks and distance significance between them. The tracks associated to the remaining secondary vertices are then arbitraged, removing those more compatible with the PV than the SV, and adding additional tracks compatible with the SV which were initially discarded due to track selection criteria. After this procedure, the SV is re–fitted. Finally, the duplicate removal is repeated with looser compatibility requirements and the passing vertices are required to have a flight distance significance larger than 1.5 and share no more than 79% of the tracks with the PV.

A SV is associated to a jet when the angular distance $\Delta R = \sqrt{\Delta\phi^2 + \Delta\eta^2}$ between the jet axis and the flight direction of the SV is smaller than 0.3.

Based on the presence or absence of a SV as reconstructed by the IVF algorithm inside a jet, three secondary vertex categories are defined:

- **RecoVertex**, in which the jet contains one or more secondary vertices as reconstructed by the IVF algorithm;
- **PseudoVertex**, in which the IVF algorithm did not reconstruct any secondary vertices, but a set of at least two tracks with a 2D impact parameter significance above 2 (highly displaced) and a combined invariant mass that lies at least 0.05 GeV away from the $K^0_S$ mass were found;
- **NoVertex**, containing any jet not assigned to one of the previous two categories.

The presence or absence of a soft lepton, as discussed in the previous paragraph, leads to the definition of three soft lepton categories, independent of the SV ones:

- **NoSoftLepton**, including jets without soft leptons found inside the jet;
• **SoftMuon**, where a soft muon was found inside the jet;
• **SoftElectron**, where no soft muon, but a soft electron was found inside the jet.

A total of 27 different combinations are possible by selecting three flavours, three SV categories and three SL categories. Even though the 9 different SV-SL categories are combined in the training of the multivariate classification algorithm, this distinction is useful to test the effects of the separate categories and to keep track of which jet observables are defined, which differs for each one of the categories. This will further be discussed in the following section.

### 3.2 Algorithm workflow and sample preparation

In order to limit the dependency on the training sample, it was decided to use different samples for training the algorithm and for validating its performance, thus mitigating the effect of overfitting. The training of the c tagger was done using simulated QCD multi-jet samples from proton-proton collisions at $\sqrt{s} = 13$ TeV: the events in these samples host a large variety of jets in terms of $p_T$ and flavour. It is therefore a good and generic training sample. QCD events are however not very representative of the majority of present-day analyses and therefore the validation is done using a simulated $t\bar{t}$ sample at the same center-of-mass energy.

From these training and validation samples, a list of jet observables are extracted. The c tagging algorithm combines information from displaced tracks, secondary vertices and soft leptons inside the jets. Some observables are only defined when a SV was reconstructed or when a SL was found inside the jet. Whenever an observable is not available, a default value is assigned to it. The values of the first three tracks (if available) and the first three SV (if available), for which the ordering is based on decreasing 2d impact parameter significance of the tracks and increasing error on the 3d flight distance of the SV, are used in the classification algorithm.

The agreement between data and simulated samples has been checked for all the input variables in events with exactly one isolated electron or muon ($p_{lep}^T > 30$ GeV and $|\eta_{lep}| < 2.1$) and exactly four jets ($p_{jet}^T > 25$ GeV and $|\eta_{jet}| < 2.4$), as shown in Fig. 1 for a sample of input features.

The training samples are pre-processed to better fit the data model and to apply two additional weights to each jet entry used for training. These weights take care of flattening the jet $p_T$ and $\eta$ distributions in the whole training sample, to avoid introducing any unwanted dependence in the tagger, and to simultaneously skew the relative contribution of the different SV categories in the QCD sample to fit the observed ones in the $t\bar{t}$ sample, as a large difference may affect the final performance of the tagger.

### 3.3 Training and Validation of the charm-tagging algorithm

A c tagging algorithm is built to most optimally identify jets from charm quarks, while rejecting either b or light jets. Typically, the observed distributions for c jet observables show an intermediate behaviour between the b and light jets ones. A BDT was chosen as a classification algorithm. Two separate classifiers have been trained in order to achieve classification into three jet flavours: one for discriminating c jets from light jets (CvsL) and one for discriminating c jets from b jets (CvsB). A two-dimensional cut on the discriminators is applied in order to identify jets from charm quarks.

The training of the classifiers has been performed with TMVA [20], a software package for multivariate analysis (MVA) classification and regression techniques inside the ROOT [21] data analysis framework. TMVA provides a variety of classification techniques to be applied within
Figure 1: Agreement between data and simulated samples for some of the input variables: 3D flight distance significance of the secondary vertex (top left), the number of secondary vertices associated to the jet (top right), the secondary vertex mass (bottom left) and a variable called “massVertexEnergyFraction” which is defined as the product of the vertex mass and the fraction of the vertex energy with respect to the jet energy (bottom right).
ROOT and can be used as a standalone application.

3.3.1 Optimization studies on CvsL discrimination

Optimisation studies were only performed on the CvsL discriminator, which is the main focus of this study, as CvsB discrimination can be already achieved by using a b tagging algorithm. The CvsL discrimination performance has been optimised by adjusting both the IVF and the BDT training settings.

The optimisation of the IVF reconstruction was performed by varying its internal parameters. The shorter lifetime of charm hadrons required a loosening of the flight distance requirements of the SV to allow for smaller separation between the PV and the SV. Extended optimisation tests performed on other IVF configuration parameters yielded no significant improvement in the performance of the tagger and were therefore left out from the final algorithm.

The optimisation of the BDT settings has shown more significant improvements for the CvsL training. All of the parameters probed in the optimisation process were varied over a wide range of values, to ensure the optimal setting was contained within the scanned range. The settings were probed separately, leaving the others to default values. Table 1 shows the final outcome of the optimisation process under the CvsL column. Since the CvsB discrimination was not optimised, the set of default values can be read from the same table under the CvsB column. Some of the optimized values did not change the performance visibly when being varied, but they were chosen to reduce the computation time without a loss in performance.

Table 1: Explanation of the BDT options that are explicitly set during the c tagger training and their used values. To ease the reproducibility, the setting names used by TMVA are listed under the BDT Option column.

<table>
<thead>
<tr>
<th>BDT Option</th>
<th>Description</th>
<th>CvsL</th>
<th>CvsB</th>
</tr>
</thead>
<tbody>
<tr>
<td>NTrees</td>
<td>The number of trees used in the boosting algorithm to build up the forest of decision trees.</td>
<td>2000</td>
<td>1000</td>
</tr>
<tr>
<td>nCuts</td>
<td>The number of points in the input variable range to find the most optimal cut in the splitting of a node.</td>
<td>50</td>
<td>80</td>
</tr>
<tr>
<td>MinNodeSize</td>
<td>The minimum fraction of jets (with respect to the full sample set) required in each node. Once a node contains less than this fraction the node splitting stops and it becomes a final leaf.</td>
<td>5%</td>
<td>1.5%</td>
</tr>
<tr>
<td>BoostType</td>
<td>The type of boosting used for the trees in the forest.</td>
<td>Grad</td>
<td>Grad</td>
</tr>
<tr>
<td>Shrinkage</td>
<td>Learning rate for the gradient boosting (Grad) algorithm.</td>
<td>0.5</td>
<td>0.1</td>
</tr>
<tr>
<td>UseBaggedGrad</td>
<td>Use bagging within the Gradient boosting algorithm. Each tree in the forest will use only a subsample of all the jets.</td>
<td>True</td>
<td>True</td>
</tr>
<tr>
<td>GradBaggingFraction</td>
<td>The (stochastically chosen) fraction of events used in each tree in the forest when using bagging.</td>
<td>0.3</td>
<td>0.5</td>
</tr>
<tr>
<td>MaxDepth</td>
<td>The maximum depth of each tree in the forest. This can be seen as the maximal amount of subsequent node splittings before constructing a final leaf of the decision tree.</td>
<td>8</td>
<td>2</td>
</tr>
</tbody>
</table>

The individual variations yield only very small changes in performance. Only the MaxDepth
showed a visible improvement. Figure 2 compares the performance of the CvsL discrimination before and after the optimisation process. These performance curves (or ROC curves) display the relation between the c jet selection efficiency, on the x-axis, and the light jet mis-tag efficiency, on the y-axis, as a function of the threshold imposed to the classifier output. The closer the curve is to the right lower corner, the better the performance. The combined effect of optimising the BDT parameters shows a clear improvement in performance with respect to the default BDT settings over the entire range of charm efficiencies.

Figure 2: Performance of the CvsL tagger compared between the default BDT options (in red) and the optimized BDT options (in green). The dashed grey line (the diagonal in this plane) represents a tagger with a random choice.

### 3.3.2 Performance measurements on simulated samples

With the BDT setting from Tab. 1, the final CvsL and CvsB trainings have been performed and validated. Figure 3 shows the output discriminator distributions as well as the corresponding ROC performances compared with the currently available cMVAv2 and CSVv2 b tagger algorithms [1]. From this comparison it is possible to see that the CvsL tagger has better discrimination power between light and charm jets than currently available tools, while the CvsB tagger is performing in general worse than the already existing b tagging algorithms in discriminating between charm and b jets. This lack of performance can be attributed to the non-optimized tuning of the tagger and, possibly, the different tuning of the IVF, which is more focused on charm–light separation than charm–b.

The discriminator shapes also show spikes in their distributions for both trainings. A detailed investigation showed that these spikes arise only in the NoVertex categories, suggesting that they originate from default values for many input features in that category, and are caused mainly by jets in which no track passes the selection criteria discussed in Section 3.1. It is...
expected that the CvsB training is less capable of handling these jets because its individual decision trees are at most two layers deep\(^1\), which might explain the multiple peaks compared to the single peak in the CvsL training, which has a maximum depth of 8. The agreement between simulations and data for the discriminator distributions of the CvsL and CvsB trainings can be found in Fig. 4.

\[\text{Observed}\  c \ | \ b \ | \ u\bar{d}g \ | \ pile-up\]

\[\text{CMS Preliminary}\ \text{tt}\ (13\ TeV)\]

CvsL Discriminator

\[\text{Observed/MC}\]

\[\text{J}\ e/0.05\]

\[\text{CvsL Discriminator}\]

Figure 3: Left: Overlay of the BDT discriminator output for the different flavours for the CvsL (top) and for the CvsB (bottom) discriminators (normalized for each flavour). Right: ROC curves showing the final performance of the CvsL (blue full line and axis) and CvsB (red full line and axis) trainings, validated on the tt sample. For comparison the performance of the CSVv2 and cMVAv2 [1] algorithms (used for b tagging) is also shown by the dotted lines.

Figure 4: Agreement between data and simulated samples for the CvsL and CvsB discriminator distributions.

In order to build a working charm tagger, the outputs of the two classifiers have to be combined in a two dimensional plane and, correspondingly, two-dimensional cuts need to be applied to evaluate the performance. Figure 5 shows the distribution of jets of different flavours in the plane formed by the two discriminators. The BDT classifiers used output a value close to 1 for signal–like jets and -1 for background–like ones, therefore c jets will be located towards

\[\text{1If the individual training trees are only two layers deep they might fail to detect how many default values occur always together and create multiple peaks instead of one peak every time these defaults occur.}\]
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. In order to isolate c jets from the background,
a rectangular cut is placed to isolate the upper right corner of this phase space.

The performance curve of the classifier consist in a three dimensional surface. It is easier,
though, to visualise the performance by drawing constant-charm-efficiency contour lines of
such surface in the plane representing the light and b jets mistag efficiencies, as shown on Fig. 5
for jets with $p_T > 20$ GeV and $|\eta| < 2.4$. For a constant predefined charm efficiency, there is
freedom to tune the light and b jets efficiencies. Three working points (WP) have been defined
to evaluate the performance of the tool on real data: the loose, medium and tight WP. The WP
threshold definitions and global efficiencies are summarised in Tab. 2. The loose WP is special-
ized in rejecting b jets, whereas the tight WP is specialized in rejecting light jets. The medium
WP rejects both b jets and light jets.

![Figure 5: 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: relation between bottom and light mistag efficiency for different values of a constant charm efficiency.](image)

Table 2: Definitions of the three working points with the corresponding cuts on the discrimi-
nator values and the global efficiencies, obtained from simulated $t\bar{t}$ samples, for each flavour.

<table>
<thead>
<tr>
<th>WP</th>
<th>$\epsilon_c$</th>
<th>$\epsilon_b$</th>
<th>$\epsilon_{\text{light}}$</th>
<th>CvsL</th>
<th>CvsB</th>
</tr>
</thead>
<tbody>
<tr>
<td>$c$-tagger $L$</td>
<td>0.9</td>
<td>0.45</td>
<td>0.99</td>
<td>$&gt; -0.67$</td>
<td>$&gt; -0.23$</td>
</tr>
<tr>
<td>$c$-tagger $M$</td>
<td>0.39</td>
<td>0.26</td>
<td>0.19</td>
<td>$&gt; 0.05$</td>
<td>$&gt; -0.16$</td>
</tr>
<tr>
<td>$c$-tagger $T$</td>
<td>0.2</td>
<td>0.24</td>
<td>0.02</td>
<td>$&gt; 0.45$</td>
<td>$&gt; -0.35$</td>
</tr>
</tbody>
</table>

4 Calibration on data

In order to validate the charm tagging as a tool for physics, its performance needs to be val-
idated on data. The tagging performance achieved in MC simulations has been compared to
data, computing scale factor corrections as a function of the working point, the jet flavour, the
jet transverse momentum and, in some cases, the jet pseudo-rapidity. In this paper we present
one method to measure the light-to-charm mistag rate based on negative tags, and two methods to measure the charm-tagging efficiency scale factors. While the first method has been already introduced for measuring the similar quantity in jet b tagging, the last two measurements are new. The first uses a very pure sample of charm jets derived from a W+c selection, while the second is based on semi-leptonic top quark pair events.

4.1 Measurement of the misidentification probability for light-quark jets from multijet events

The negative tag method [19] is applied to derive scale factors for light-flavour jets, based on the definition of positive and negative taggers, which are identical to the default algorithms, except that only tracks with positive/negative impact parameter values or secondary vertices with positive/negative decay lengths are used; soft leptons are treated as normal tracks. The discriminator values for negative and positive taggers are expected to be symmetric for light-flavour jets, since negative or positive impact parameter values or decay lengths for light-flavour jets are mostly due to resolution effects and mis-reconstructed tracks. Remaining asymmetries in the distributions are due to long lived strange hadrons, which are anyhow a minority of the sample. We can therefore derive the misidentification probability from the rate, \( \varepsilon^{-} \), of negative-tagged jets in an inclusive multijet sample. A correction factor, \( R_{\text{light}} \), is evaluated from the simulation in order to correct for asymmetry effects in the negative and positive tag rates of light-flavour jets, and for the heavy-flavour contribution to the negative tags:

\[
\varepsilon_{\text{misid data}} = \varepsilon_{\text{data}} \cdot R_{\text{light}},
\]

where \( R_{\text{light}} \) is extracted from simulation. The positive and negative b discriminator distributions are shown in Fig. 6 for jets with \( p_T > 40 \text{ GeV} \). For convenience, the discriminator values of the negative taggers are shown with a negative sign. Given that both CvsL and CvsB discriminator outputs are defined in the [-1, 1] range, a shift of +1 and -1 was introduced to positive and negative tags, respectively, to avoid overlap in the distributions. The distribution obtained from simulation is normalised to the total number of jets in data; deviations are related to the modeling of the heavy-flavour content by the generators and to nonidentical detector conditions in simulation and observed data. Figure 7 shows a summary of the measured misidentification probability and scale factor for the charm-tagging working points.

4.2 Measurement of charm jets identification scale factor with W+c

We evaluate the performance of the charm tagging tool on data, using a sample enriched in c jets using events with a W boson produced in association with a c quark.

4.2.1 W plus c jet events selection

The analysis uses a sample consisting of charm jets produced in association with a W-boson. The W + c events are selected according to the criteria used in [22].

The production of a W boson in association with a c quark proceeds at LO via the processes \( sg \rightarrow W^{-} + c \) and \( sg \rightarrow W^{+} + \bar{c} \). A key property of the \( qg \rightarrow W + c \) reaction is the presence of a charm quark and a W boson with opposite-sign (OS) charges. Background processes deliver evenly OS and same-sign (SS) events, whereas \( qg \rightarrow W + c \) is always OS, as shown in Fig. 8, and a very pure sample of c jets can thus be obtained by OS-SS subtraction [22].

The leptonic decay of a W boson into a muon or an electron and a neutrino is characterized by the presence of a high-transverse-momentum, isolated lepton. The neutrino escapes detection causing an imbalance in the transverse energy of the event; therefore, candidates for \( W \rightarrow \mu \nu \) and \( W \rightarrow e \nu \) events are selected requesting a high \( p_T \), well-identified, and isolated electron or
muon. The lepton offline $p_T$ threshold is set to 25 GeV due to the thresholds set for the online selection of the single lepton triggers. The reconstructed transverse mass between the isolated lepton and the missing transverse energy $E_{T miss}$, $M_T = \sqrt{2 p_T^\ell E_{T miss} \times (1 - \cos(\phi^\ell - \phi^{E_{T miss}}))}$, is required to be larger than 55 GeV to ensure the presence of a $W$ boson in the event.

Charm jets are selected starting from a collection of jets with $p_{jet}^T > 25$ GeV and $|\eta_{jet}| < 2.4$ which are at least $\Delta R (jet, lepton) > 0.5$ apart from the lepton from the $W$ decay. The jets are then required to contain a well-identified, non-isolated muon among the jet constituents. The muon candidate inside the jet is required to have $p_{\mu}^T < 25$ and $|\eta_{\mu}| < 2.4$.

The charge of the muon determines the charge of the charm quark. OS (SS) events are defined as events for which the muon has opposite (same) charge as the charge of the lepton from the $W$ decay.

The expected signal purities, computed from MC simulations, are $\sim 70\%$ for $W^+ \rightarrow \mu^+\nu$ events and $\sim 85\%$ for $W^+ \rightarrow e^+\nu$ events. The major sources of background are Drell Yan and $t\bar{t}$ events for $W^+ \rightarrow \mu^+\nu$ and $t\bar{t}$ events for $W^+ \rightarrow e^+\nu$.

Figure 9 shows the $c$ tagger discriminator distributions after $W$+charm selection.

### 4.2.2 $c$-tagging efficiency

Efficiencies for the $c$ tagging algorithm are obtained independently for the observed data and for $W$+jets simulated events as the ratio between the number of $W + c$ events tagged by the $c$
systematic uncertainties on the measurements. The solid curve in the bottom plots represents a fit to the observed data, while the dashed curves show the combined statistical and systematic uncertainties on the measurements.

Figure 1: Misidentification probability in data and simulation (top) breakdown of systematic uncertainties contributions (middle) data-to-simulation scale factor of the light mis-identification probability (bottom) for the loose (left), medium (center) and tight (right) working points. The solid curve in the bottom plots represents a fit to the observed data, while the dashed curves show the combined statistical and systematic uncertainties on the measurements.
4.2 Measurement of charm jets identification scale factor with W+c

The c jet tagging efficiency in simulation is computed as the fraction of W+charm events that are tagged. In this case both numerator and denominator are evaluated directly from generator truth information.

Figure 8: Left and middle: leading order production of W+c signal with opposite sign charges (OS). Right: production of W+charm final state through gluon splitting process. In gluon splitting there is an additional charm quark with the same sign as the W boson (SS).

Figure 9: Distribution of the c tagger discriminators (CvsB and CvsL) after applying W+charm selection and OS-SS subtraction. The plots corresponds to the sum of both electron and muon decays of the W.
Data-to-simulation scale factors, $SF_c$, are then computed as the ratio between the $c$ jet tagging efficiencies in data and simulation:

$$SF_c = \frac{\epsilon_c^{\text{data}}}{\epsilon_c^{\text{MC}}}.$$ 

4.2.2.1 Bias study: As the $c$ tagger uses soft lepton information as an input, the jet selection, which requires a muon in the jet, may potentially introduce a bias in the $SF_c$ measurement. This bias is estimated by repeating the measurement using a modified version of the tagger which treats leptons as normal tracks and assigns to the SL input features their default values. The difference between the values measured with the modified tagger and the default one is taken as systematic uncertainty.

4.2.2.2 Systematic uncertainties: While the analysis is dominated by statistical uncertainties, several sources of systematic uncertainties have been considered:

- **Pile-up:** The nominal 2015 pile-up distribution, assuming a $p$–$p$ inelastic cross section of 69 mb, has been used to reweigh simulated events, as a function of the number of interactions recorded at generator level. Possible systematic effects due to the wrong pile-up profile choice have been evaluated by shifting the $p$–$p$ inelastic cross section by $\pm 5\%$, resulting in a variation of the measured $SF_c$ of $\sim 1\%$.

- **Charged tracks in a muon-jet:** Uncertainties in the modelling of the number of charged tracks in a jet have been considered by reweighing such distribution in simulated events in order to match the observed one. The difference observed in $SF_c$ is 1%.

- **Background subtraction:** The ratio $f_{\text{bkg}}^{\text{sim}}$ has been varied by $\pm 50\%$. The effect of this variation is $< 1\%$ on the $SF_c$ measurement.

- **Jet Energy scale:** Modifying the jet $p_T$ according to the jet energy scale uncertainties results in a $SF_c$ variation of $< 1\%$.

- **Branching ratios of $D \to X\mu$ hadrons and fragmentation $c \to D$:** In order to account for discrepancies in the branching ratios and fragmentation between PYTHIA and the PDG, the simulation has been reweighed to match the PDG by modifying $BR(D^+ \to \mu X)$ by $-4\%$, $BR(D^0 \to \mu X)$ by $+13\%$, $BR(D_s \to \mu X)$ by $+13\%$, and by modifying the fragmentation by $+23\%$ if $c \to D^+$, $-9\%$ if $c \to D^0$, and $+42\%$ if $c \to D_s$. All these changes have been performed simultaneously. The effect of this variation on the $SF_c$ measurement is $< 1\%$.

- **Muon and electron efficiency:** The systematic uncertainty associated with the lepton reconstruction, identification, and isolation has been estimated by omitting the data-to-MC corrections connected to such objects, resulting in a variation of $< 1\%$.

- **SL vs NOSL:** As explained in the previous paragraph, the analysis has been repeated with a $c$ tagger constructed without taking into account any soft lepton information input. The effect of this variation is $< 2\%$.

4.2.3 Results:

The $c$ jet tagging efficiency $\epsilon_c$ (in data and in simulated events) and the scale factor $SF_c$ are computed both inclusively and as a function of jet $p_T$. Figure 10 shows the efficiencies and scale factors obtained as a function of the $p_T^{\text{jet}}$. The scaling factors are summarized in Tab. 3.

The efficiency decreases (increases) slightly for the Loose (Medium and Tight) working point. No strong dependence on $p_T^{\text{jet}}$ of the $SF_c$ is observed.
The event selection closely follows, with some small but relevant discrepancies, the event selection used to measure the differential $\bar{t}t$ cross section in semi-leptonic decays [23]. Events are selected by requiring only one well-identified and isolated lepton and at least four jets. Fol-
Figure 11: Data and MC distributions of the invariant mass of the hadronically-decaying W (left), top quark (middle), and of combined discriminant $\lambda_M$ for every permutation in preselected events (i.e. those passing trigger and objects selection). The lower panels show the distribution of the MC over data ratio. These distributions were obtained processing $2.3 \text{fb}^{-1}$ of 2015 data.

Following this selection, all the possible four-jets permutations are computed and each of them, together with the lepton, represents a possible top quark pair candidate. In each permutation, jets are assigned specific roles linked to the final state such as “leptonic top b jet”, “hadronic top b jet” and two “W-jets”. The best possible candidate in the event is selected by choosing the lowest mass discriminant $\lambda_M$. The mass discriminant combines the invariant mass of the two “W-jets” and the invariant mass of the jets belonging to the hadronic top into a single value which has the same properties of the negative logarithm of a likelihood ratio. The distributions of $\lambda_M$ and of the two input variables are shown in Fig. 11. Finally, the two jets assigned as “leptonic top b jet” and “hadronic top b jet” by the best permutation are required to be b tagged [1]. Events in which the best permutation does not comply to this requirement are discarded. The b tagging requirement is applied as the last step in the selection not to impose any bias on the probe jets.

4.3.2 Efficiency measurement

After the final selection, the distribution of $\lambda_M$ still retains discriminating power to separate the different components observed in the data, as clearly visible in Fig. 12. For the purpose of this study, the top quark pair simulated sample can be divided into three sub-samples:

- $t\bar{t}$, right $W_{h}$: in which the permutation jets assigned as “W–jets” are properly matched to the decay partons of the hadronically-decaying W.
- $t\bar{t}$, wrong $W_{h}$: in which one or both of the permutation jets assigned as “W–jets” are not matched to the decay partons of the hadronically-decaying W.
- Other $t\bar{t}$ decay: in which the generated decay of the top quark pair was not semi-leptonic.

The selected events are additionally divided into four categories according to which of the “W–jets” pass the charm–tagging working point under study:

- notag: in which neither of the probed jets pass the c-tagging working point under study
- leadtag: in which the most energetic jet passes the c-tagging working point, but the
Figure 12: Data and MC distributions of $\lambda_M$ after full selection, including the selection of the best permutation and the b tagging identification of two of the jet of the permutation. The different simulated processes contributing are shown with different colours. The major contributions are semi-leptonic top quark pair decays with the hadronic W properly matched to the generator particles (violet), wrongly matched semi-leptonic top quark decays (red), and non semi-leptonic 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 multi-jet production (“QCD”, in blue).
A simultaneous maximum likelihood fit is performed in order to disentangle the contributions in large biases.

This constraint gives rise to ambiguities in the likelihood that cannot be resolved by the fit, resulting in a loss of precision and stability of the result. Toy studies have in fact shown that the lack of such a constraint would require changing Eq. 1 to consider only leading and subleading jets within a certain range, leading to the introduction of $n^2$ categories for $n$ jet $p_T$ bins considered, which is beyond the statistical power of the sample at disposal.

The light-to-charm mistagging scale factor measured in Section 4.1 is applied in this measurement in order to replace the free-floating light efficiency in Equation 1 with the constrained measurement provided by the negative tags method. Given the strong correlation between the light and charm scale factors in this model, this solution is extremely beneficial for the final precision and stability of the result. Toy studies have in fact shown that the lack of such a constraint gives rise to ambiguities in the likelihood that cannot be resolved by the fit, resulting in large biases.

A simultaneous maximum likelihood fit is performed in order to disentangle the contributions of the correctly–matched $\bar{t}t$ permutations from the other background components as well as to infer the value of $S_{F_c}$. The fit is performed on the binned $\lambda_M$ distributions using signal and background templates derived from MC simulations.

The uncertainties on jet energy scale and resolution, pile-up reweighting, $b$ tagging data-to-Monte Carlo corrections, light-to-charm mistag-rate, and cross-sections of the physical processes involved have been considered as systematic uncertainties and mapped to nuisance parameters which are profiled during the fit. The effect of a deviation from Monte Carlo expectations for the tagging efficiency is modelled on the other samples entering the fit with a simpler linear model and is assigned as an additional systematic.
Table 4: Measured scale factors for the efficiency of tagging charm jets in semi-leptonic $t\bar{t}$ events. The quoted uncertainty includes both statistical and systematic uncertainty. The last column shows the statistical-only uncertainty.

<table>
<thead>
<tr>
<th>working point</th>
<th>charm SF</th>
<th>stat only unc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>c-tagger L</td>
<td>1.008$^{+0.041}_{-0.042}$</td>
<td>+0.014 −0.013</td>
</tr>
<tr>
<td>c-tagger M</td>
<td>0.899$^{+0.072}_{-0.072}$</td>
<td>+0.029 −0.029</td>
</tr>
<tr>
<td>c-tagger T</td>
<td>0.923$^{+0.063}_{-0.061}$</td>
<td>+0.034 −0.034</td>
</tr>
</tbody>
</table>

The scale factor values inferred from the fit are summarised in Tab. 4. The uncertainty resulting from the fit is a combination of statistical and systematic uncertainties.

4.4 Tagging efficiency scale factor combination

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 [24]. Systematic uncertainties shared by the two measurements have been properly correlated in the process. The averaging is done using the finer jet $p_T$ binning of the W+c topology, while considering the measurement in the $t\bar{t}$ region as a convolution of different measurements. The relative population, and hence contribution to the observed value, of the different $p_T$ bins has been estimated from simulation and taken into account when combining the results.

The combined results are shown in Fig. 13 and summarized in Tab. 5 for the three working points.

<table>
<thead>
<tr>
<th>WP</th>
<th>$p_T$ (jet) [GeV]</th>
<th>$SF_c$</th>
</tr>
</thead>
<tbody>
<tr>
<td>c-tagger L</td>
<td>$25 &lt; p_T &lt; 35$</td>
<td>0.994 ± 0.058</td>
</tr>
<tr>
<td></td>
<td>$35 &lt; p_T &lt; 50$</td>
<td>0.999 ± 0.063</td>
</tr>
<tr>
<td></td>
<td>$50 &lt; p_T &lt; 70$</td>
<td>1.005 ± 0.086</td>
</tr>
<tr>
<td></td>
<td>$70 &lt; p_T &lt; 200$</td>
<td>1.03 ± 0.11</td>
</tr>
<tr>
<td>c-tagger M</td>
<td>$25 &lt; p_T &lt; 35$</td>
<td>0.948 ± 0.071</td>
</tr>
<tr>
<td></td>
<td>$35 &lt; p_T &lt; 50$</td>
<td>0.945 ± 0.075</td>
</tr>
<tr>
<td></td>
<td>$50 &lt; p_T &lt; 70$</td>
<td>0.941 ± 0.098</td>
</tr>
<tr>
<td></td>
<td>$70 &lt; p_T &lt; 200$</td>
<td>0.92 ± 0.14</td>
</tr>
<tr>
<td>c-tagger T</td>
<td>$25 &lt; p_T &lt; 35$</td>
<td>0.955 ± 0.082</td>
</tr>
<tr>
<td></td>
<td>$35 &lt; p_T &lt; 50$</td>
<td>0.953 ± 0.086</td>
</tr>
<tr>
<td></td>
<td>$50 &lt; p_T &lt; 70$</td>
<td>0.95 ± 0.11</td>
</tr>
<tr>
<td></td>
<td>$70 &lt; p_T &lt; 200$</td>
<td>0.94 ± 0.16</td>
</tr>
</tbody>
</table>

5 Conclusions

A new tool for identifying charm jets is presented. The tagger has been developed using two boosted decision trees, trained and tested on simulated input datasets. The light-to-charm mistagging rate has been measured in multi-jet events. The performance of the tagger has
Figure 13: (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.
been measured on W+c and semi-leptonic top quark pairs events, deriving scale factors for the charm identification. The precision achieved with the first method ranges between 6% and 21%, while the second provides a more precise evaluation of the scale factor, in the order of 5%, but without jet $p_T$ dependence. The W+c jet population is strongly skewed towards low-$p_T$ jets. The datasets for the efficiency calibrations have been extracted from proton–proton collisions collected by the CMS experiment in 2015.

References


