Available on the CERN CDS information server CMS PAS BTV-15-001

CMS Physics Analysis Summary

Contact: cms-pog-conveners-btag@cern.ch 2016/05/13

# Identification of b quark jets at the CMS Experiment in the LHC Run 2

The CMS Collaboration

## Abstract

Many physics studies involving standard model processes as well as searches for physics beyond the standard model rely on the accurate identification of jets originating from bottom quarks. The b jet identification algorithms and their performance are presented using proton-proton collision data recorded by the CMS detector at a center of mass energy of $\sqrt{s} = 13$ TeV during the start of the LHC Run 2 in 2015. The efficiency to identify b quark jets and the probability to misidentify jets originating from non-b quark jets is measured as a function of the jet transverse momentum by selecting events with multiple jets, a Z boson or top quarks. Studies related to the b jet identification efficiency for wide jets with substructure are also presented, relevant for physics analyses targeting b quark jet identification in boosted topologies.

*This document has been revised with respect to the version dated March 23, 2016.*

# 1 Introduction

With the start of Run 2, the Large Hadron Collider (LHC) is providing proton-proton collisions at an unprecendented center-of-mass energy of $\sqrt{s} = 13\,\text{TeV}$. The new collision data recorded by the Compact Muon Solenoid (CMS) experiment reopen the opportunity to search for new physics phenomena at the TeV scale and to perform precision measurements of the standard model at a higher center-of-mass energy. Many searches for new physics as well as standard model measurements rely on the accurate identification of jets originating from b quarks. The CMS collaboration optimized the existing b jet identification techniques and measured their performance in both boosted and nonboosted event topologies.

The central feature of the CMS apparatus is a superconducting solenoid of 6 m internal diameter, providing a magnetic field of 3.8 T. Within the superconducting solenoid volume are a silicon pixel and strip tracker, a lead tungstate crystal electromagnetic calorimeter (ECAL), and a brass and scintillator hadron calorimeter (HCAL), each composed of a barrel and two endcap sections. Forward calorimeters extend the pseudorapidity coverage provided by the barrel and endcap detectors. Muons are measured in gas-ionization detectors embedded in the steel flux-return yoke outside the solenoid. A more detailed description of the CMS detector, together with a definition of the coordinate system used and the relevant kinematic variables, can be found in Ref. [1].

The trajectories of charged particles in the detector are reconstructed from the hits in the silicon tracking system using an iterative procedure referred to as Iterative Tracking [2]. The tracking efficiency is typically over 98% for tracks with a transverse momentum above 1 GeV and within the tracker acceptance [3]. The interaction vertex corresponding to the hard scattering is chosen as the one that maximizes the sum of the squared transverse momenta of the clustered physics objects associated to the vertex. The resolution of this primary vertex is around 20 $\mu$m in the transverse $(x, y)$ plane and around 30 $\mu$m along the beam axis.

Algorithms for b jet identification exploit the long life time of b hadrons present in jets originating from the hadronization of b quarks. This long life time results in a decay of the b hadron that is displaced with respect to the primary interaction vertex. This displacement of a few millimetres results in the presence of displaced tracks from which a secondary vertex may be reconstructed. In addition, b hadrons have a probability of around 20% to decay to a muon or electron. Hence, apart from the properties of the reconstructed secondary vertex or displaced tracks also the presence of these charged leptons can be exploited for b jet identification techniques and for measuring their performance with the collision data.

In Section 2 a brief description is given of the simulated proton-proton collisions as well as the recorded data analyzed for the performance measurement of the b jet identification algorithms. The algorithms are discussed in Section 3, together with a comparison of the observed collision data with the simulation for a number of variables relevant for b jet identification. Section 4 provides an overview of the performance measurements and their combination for standard jets. Finally, a summary for b jet identification efficiencies in topologies with boosted jets is given in Section 5.

# 2 Simulated and recorded proton-proton collisions at $\sqrt{s} = 13\,\text{TeV}$

We use proton-proton collisions at $\sqrt{s} = 13\,\text{TeV}$ delivered by the LHC with a bunch spacing of 25 ns during 2015 operations. The proton-proton collision data corresponds to an integrated luminosity of about 2.6 $\text{fb}^{-1}$, when considering only the collisions certified as good by the CMS detector subsystems relevant for objects reconstructed within the silicon tracker acceptance ($|\eta| < 2.4$).

The observed data are compared to dedicated simulated samples at the same center-of-mass energy. Proton-proton collisions producing top quarks are generated using POWHEG [4–6], both for top quark pair production [7] as well as electroweak production of a top quark in association with a W boson [8]. Events with two W bosons are also generated with POWHEG [9, 10], while the $Z + jets$ events are generated either with MADGRAPH 5 [11] or the aMC@NLO event generator [12]. The WZ diboson events are also generated with the aMC@NLO event generator. QCD multijet and ZZ events are generated with PYTHIA 8 [13], that is also used for the parton showering of all the simulated samples. In Run 1 we were comparing the collision data with PYTHIA 6 instead of PYTHIA 8. Studies comparing the two PYTHIA versions were performed between Run 1 and Run 2 and it was shown that both model the collision data equally well. The simulated samples are produced with the recently optimized CMS Underlying Event Tune for PYTHIA 8 (CUETP8M1) [14]. Additional simulated inelastic collisions generated with PYTHIA 8 are overlayed with the simulated events to model the contribution of these pileup events in the simulation. The CMS detector response is simulated using GEANT 4 [15].

## 3 Algorithms for b-quark jet identification

Stable particles are identified with the Particle Flow (PF) algorithm [16–18] that reconstructs each individual particle with an optimized combination of information from the various elements of the CMS detector. The energy of photons is directly obtained from the ECAL measurement, corrected for zero-suppression effects. The energy of electrons is determined from a combination of the electron momentum at the primary interaction vertex as determined by the tracker, the energy of the corresponding ECAL cluster, and the energy sum of all bremsstrahlung photons spatially compatible with originating from the electron track. The energy of muons is obtained from the curvature of the corresponding track. The energy of charged hadrons is determined from a combination of their momentum measured in the tracker and the matching ECAL and HCAL energy deposits, corrected for zero-suppression effects and for the response function of the calorimeters to hadronic showers. Finally, the energy of neutral hadrons is obtained from the corresponding corrected ECAL and HCAL energy. For the measurements presented in this document, jets are reconstructed by clustering PF particles using the anti-$k_T$ (AK) jet clustering algorithm [19], with a distance parameter $R = 0.4$ (AK4 jets), while during Run 1 a distance parameter of 0.5 was used to reconstruct the jets (AK5 jets). For the boosted topologies, jets are clustered with a larger opening angle corresponding to $R = 0.8$ (AK8 jets). When clustering the particles in jets, isolated electrons and muons as well as charged particles associated with other interaction vertices are removed. Jet momentum is determined as the vectorial sum of all particle momenta in the jet. Jet energies are calibrated to correct for the different detector response as a function of the transverse momentum and pseudorapidity of the jets. Furthermore, an offset correction is applied to jet energies to take into account the contribution from additional proton-proton interactions within the same bunch crossing [20]. For the studies presented here, jets should lie within the tracker acceptance, hence pseudorapidity $|\eta| < 2.4$, and have a transverse momentum exceeding 20 GeV. For some of the measurements and figures a tighter requirement on the jet transverse momentum is applied.

The flavour for jets in the simulated events is determined by re-clustering the jet contituents including also the generator-level hadrons and partons. The re-clustering is performed in such a way that the re-clustered jet four-momenta are identical to the original jets. The jet flavour is then determined based on the flavour of the clustered hadrons (or partons) inside a jet giving priority to the b flavour when at least one b hadron is present. In the absence of a clustered b hadron, priority is given to the c flavour in case a c hadron is found. If there are no b and c hadrons clustered in the jet, it is considered as light flavour unless a b (or c) quark is clustered in the jet in which case the jet will be considered as b (or c) jet.

For the figures presented in this and the following section, jets in three different event samples are studied. These samples are defined by selecting events in the following way:

- Inclusive multijet sample: events are selected if they pass an online selection requiring the presence of at least one jet with a transverse momentum exceeding a certain threshold. The figures shown in this section used the lowest online threshold of 40 GeV. A jet $p_{\mathrm{T}}$ range between 50 and 250 GeV is considered as an illustration. For the performance measurements events are selected with different online selection criteria with an increasing transverse momentum threshold. This topology is dominated by light-flavour jets and contains also contributions from jets from pileup collisions. When measuring the performance, jets from pileup collisions are considered as light-flavour jets.
- Muon enriched jets sample: events are considered when passing an online selection requiring at least two jets with transverse momenta exceeding a certain threshold (the lowest one being $p_{\mathrm{T}} > 20\,\mathrm{GeV}$) and among these jets at least one has to contain a muon with a transverse momentum above 5 GeV. The figures shown in this section use an online threshold of 40 GeV for the two leading jets. Only jets containing a muon and with a jet $p_{\mathrm{T}}$ range between 50 and 250 GeV are considered as an illustration. Due to the muon requirement, there is no visible contribution from jets from pileup collisions. For the performance measurements events are selected with different online requirements, at least one jet should contain a muon. This topology is enriched in b jets.
- Dilepton $\mathrm{t\bar{t}}$ sample: events are selected by requiring the presence of at least one isolated electron and at least one isolated muon in the online selection. Offline, the leading muon and electron should both have a transverse momentum exceeding 20 GeV and be identified as isolated leptons as is expected for leptonic W boson decays. Events are further considered if they contain at least two jets with a transverse momentum exceeding 20 GeV. In this event sample there is a contribution from jets from pileup collisions due to the higher acceptance for low-$p_{\mathrm{T}}$ jets.

## 3.1 Track selection and secondary vertex reconstruction

Tracks inside jets are used for b jet idenfication if they fulfill the following selection criteria. Only tracks with a transverse momentum of at least 1 GeV and a normalized $\chi^2$ of the trajectory fit below 5 are considered. The tracks should have at least 8 hits in the silicon tracker, of which at least 2 in the silicon pixel layers of the tracker detector. Additional requirements are imposed on the track impact parameter (IP), defined as the distance between the primary vertex and the track at their point of closest approach. The track impact parameter sign corresponds to the sign of the scalar product of the jet axis direction with the vector pointing from the primary vertex to the track at the point of closest approach. The transverse (longitudinal) impact parameter should be smaller than 0.2 (17) cm. These requirements ensure tracks are not too far away from the primary interaction vertex. The distance between the jet axis and the track at their point of closest approach should be smaller than 0.07 cm. This requirement reduces the contribution of tracks originating from additional pileup interaction vertices. The decay length is required to be less than 5 cm and is defined as the distance between the primary vertex and the point of closest approach between the jet axis and the track trajectory. These track selection requirements are used in all b jet identification observables and algorithms, except when vertices are reconstructed using the Inclusive Vertex Finder algorithm as described below. Figure 1 shows the impact parameter significance, defined as the impact parameter divided by its uncertainty, of the selected tracks for jets selected in three different topologies. A reasonable agreement between the data and simulated events is observed. For the simulation the distribution shows the jet-flavour composition. Small variations between the observed collision data and the simulation are related to differences in the alignment conditions in the simulation with respect to the observation. The underflow and overflow is included in the first and last bins, respectively, for all distributions. In addition, the total number of entries in the simulation is normalized to the observed number of entries in data.

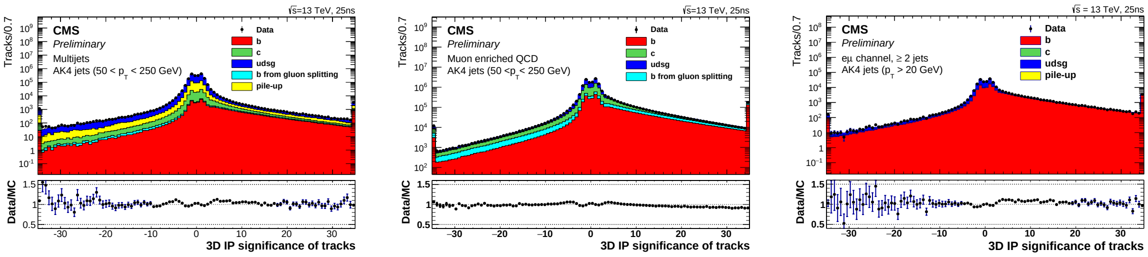

Figure 1: The impact parameter significance of the selected tracks for jets in inclusive multijet events (left), in muon enriched events (middle) and $t\bar{t}$ dilepton events (right). Underflow and overflow are added to the first and last bins, respectively. The total number of entries in the simulation is normalized to the observed number of entries in data.

Two algorithms for reconstructing secondary vertices are exploited. For the first algorithm, the tracks associated to jets and fulfulling the aforementioned selection requirements are used in the adaptive vertex reconstruction (AVR) algorithm [21] based on the adaptive vertex fitter [22]. This is the secondary vertex reconstruction algorithm used in the b jet identification algorithms of the CMS collaboration during LHC Run 1. A number of selection criteria are applied to remove vertices that are less likely to originate from a b hadron decay. Secondary vertices should have at least 2 tracks by construction and are rejected when sharing more than 65% of their tracks with the primary vertex or when the distance between the primary vertex to the secondary vertex in the transverse plane, the 2D flight distance, is more than 2.5 cm or less than 0.1 mm. The 2D flight distance divided by its uncertainty or so-called 2D flight distance significance should exceed 3. In addition, secondary vertices are only considered when they have a mass of less than 6.5 GeV and that is not compatible with the mass of the $K^0_S$ hadron in a window of 50 MeV. Additionally, the angular distance $\Delta R$ between the jet axis and the secondary vertex flight direction is required to be less than 0.4. When these requirements are fulfilled the jet contains a reconstructed AVR secondary vertex.

In contrast with the AVR algorithm, the Inclusive Vertex Finder (IVF), first introduced in Ref. [23], is not seeded from tracks associated to the reconstructed jets. The IVF algorithm uses as input the collection of reconstructed tracks in the event. In addition, looser selection requirements are applied by selecting tracks with at least 8 hits in the silicon tracker without making a requirement on the number of hits in the silicon pixel tracker. The transverse momentum of the tracks is required to exceed 0.8 GeV and the longitudinal impact parameter should be smaller than 0.3 cm. Among the selected tracks, displaced tracks are identified as seeds when having an impact parameter value of at least 50 $\mu$m and an impact parameter significance of at least 1.2. These seed tracks are then used to identify clusters of nearby tracks based on their minimum distance and the angles between them. The clusters are fitted with the adaptive vertex fitter and a cleaning procedure is applied removing vertices with a 2D (3D) flight distance significance below 2.5 (0.5). At this stage, tracks can appear in multiple vertices and therefore, one of the vertices is removed when it shares at least 70% of its tracks and the distance significance between the vertex and another one is less than 2. At this stage of the vertex reconstruction algorithm, a track can be assigned both to the primary and secondary vertex. To resolve this ambiguity, tracks are discarded from the secondary vertex when they have less than 1 hit in the pixel tracker or when they are more compatible with the primary vertex. This compatibility is decided by requirements on the (angular) distances between the secondary vertex and the track and between the primary vertex and secondary vertex. When there are at least 2 tracks associated to the secondary vertex after the track arbitration, the vertex is refitted. This time, a vertex is removed when it shares at least 20% of its tracks with another secondary vertex and the distance significance between the two vertices is less than 10. The selection criteria applied on the remaining IVF secondary vertices are mostly the same as in the case of the AVR vertices, with the exception that the fraction of shared tracks between the secondary vertex candidate and the primary vertex should not exceed 79%, the angular distance between the jet axis and the secondary flight direction $\Delta R$ is required to be smaller than 0.3 and the 2D flight distance significance should be at least 2. As an illustration, the secondary vertex flight distance significance and the secondary vertex mass are shown in Figure 2 for jets with an IVF secondary vertex (SV) and in the three different event topologies. The efficiency to reconstruct

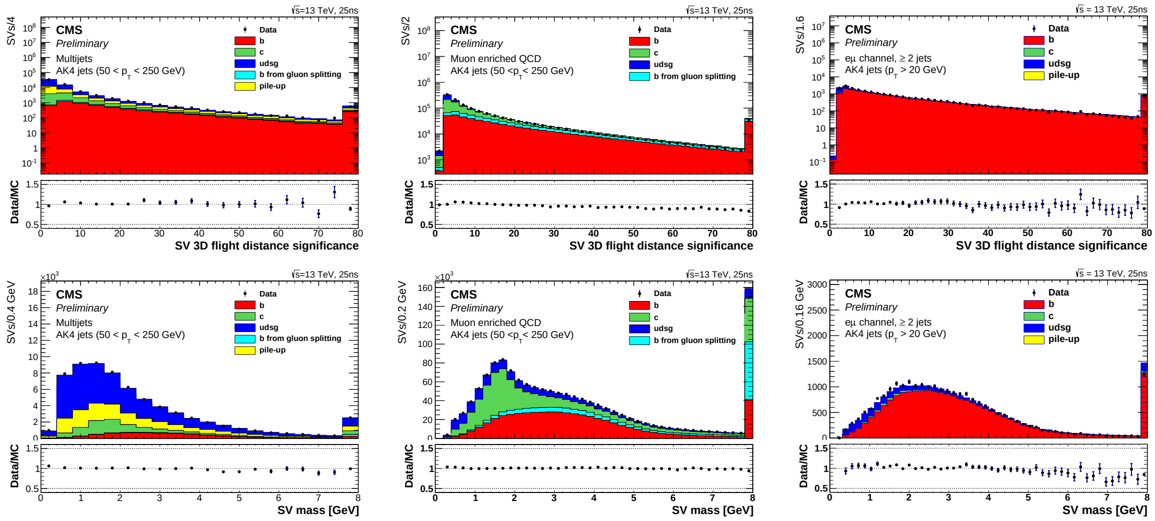

Figure 2: Significance of the secondary vertex flight distance (top) and mass of the secondary vertex (bottom). From left to right these distributions are shown for the inclusive multijet, muon enriched and dilepton $t\bar{t}$ topologies. Underflow and overflow are added to the first and last bins, respectively. The total number of entries in the simulation is normalized to the observed number of entries in data.

a secondary vertex for b (c) jets using the IVF algorithm is about 10% (15%) higher compared to the efficiency to reconstruct a secondary vertex with the AVR algorithm. However, for light-flavour jets the probability to find a secondary vertex also increases by about 8%. Since the vertex reconstruction efficiency increases for all jet flavours, the impact of using the IVF algorithm instead of AVR in a b jet identification algorithm is nontrivial. It should be noted that independently of the jet flavour, around 60% of the jets with an AVR vertex also have an IVF vertex.

## 3.2 Jet Probability and Combined Secondary Vertex algorithms

Among the different b jet identification algorithms defined by the CMS collaboration and used during Run 1, two are also used today during Run2. These are the Jet Probability (JP) and Combined Secondary Vertex (CSV) taggers [24]. The CSV algorithm was further optimized and the new version is referred to as CSVv2.

The JP algorithm computes the likelihood of the jet to originate from the primary vertex using the associated tracks. Tracks with a negative impact parameter are used to define resolution functions. The negative impact parameter is used since it characterizes the expectation for light flavour jets for which the signed impact parameter is symmetric around 0. For each track the probability to originate from the primary vertex is obtained by integrating the resolution function between the absolute track impact parameter value and infinity. If the track probability is smaller than 0.5%, it is set to a value of 0.5%. The track probabilities are then combined to obtain the jet probability. The resolution functions depend strongly on the track quality and are therefore calibrated both for the observed collision data and the simulated events. A variant of the JP algorithm also exists in which the four tracks with the highest impact parameter significance get a higher weight in the jet probability calculation. This algorithm is referred to as Jet B-Probability (JBP). For the JP algorithm, the angular distance between each track and the jet axis is restricted to $\Delta R = \sqrt{\Delta\phi^2 + \Delta\eta^2} < 0.3$, while the criterion is relaxed to $\Delta R < 0.4$ for the JBP algorithm.

The CSVv2 algorithm is based on the CSV algorithm described in Ref. [24] and combines the information of displaced tracks with the information of secondary vertices associated to the jet using a multivariate technique. The considered tracks are required to fulfill the selection requirements described in Section 3.1 and should have an angular distance with respect to the jet axis, $\Delta R$ smaller than 0.3. At least 2 such tracks per jet are requested. On top of these requirements, a number of additional criteria are applied, which have not changed with respect to the CSV algorithm used during Run 1. Any combination of two tracks compatible with the mass of the $\mathrm{K}^0_\mathrm{S}$ meson in a window of 30 MeV are rejected. At that stage, if there are no tracks associated to the jet, a negative value is assigned to the algorithm output to signify that there is no information for b jet identification. The training of the algorithm is performed in three independent vertex categories. The first vertex category contains jets with at least one associated reconstructed secondary vertex. When more than one reconstructed secondary vertex is associated to the jet, the vertices are sorted according to increasing uncertainty on the flight distance. Most of the discriminating variables relying on the presence of a secondary vertex (like the vertex mass and flight distance significance) are based on the first secondary vertex, hence the one with the smallest uncertainty on its flight distance. There are however also variables that combine the information from all available secondary vertices, for instance the number of secondary vertices present in the jet. The second vertex category contains jets with a so-called "pseudo-vertex". A jet in the pseudo-vertex category has at least two tracks not compatible with a $\mathrm{K}^0_\mathrm{S}$ meson in a window of 50 MeV and a signed impact parameter significance exceeding 2. No vertex fit is applied and hence there is no real vertex reconstruction. Therefore also the flight distance can not be calculated and the number of variables is reduced. The third vertex category is the complement of the two other cases; no reconstructed secondary vertex or pseudo-vertex is associated to the jet. In that case, only information related to the displaced tracks is exploited. A multilayer perceptron with one hidden layer is used to combine the discriminating variables in each vertex category. The information of the three vertex categories is combined with a likelihood ratio taking into account the fraction of jets of each flavour expected in $\mathrm{t\bar{t}}$ events. The main differences with the Run 1 version of the CSV algorithm are the different vertex reconstruction algorithm used, the number of input variables and the way those are combined. In the past the input variables were combined with a likelihood ratio instead of a multilayer perceptron. This limited the amount of input variables since correlation between those could not be taken into account properly. Examples of variables that were added are the number of secondary vertices, the angle between the secondary vertex and the jet axis, the ratio of the transverse momentum of the summed track four-momenta and the jet, the track decay length and the angle between the track and the jet. Two variants of the CSVv2 algorithm exist according to whether IVF or AVR vertices are used. For the CSVv2 algorithm IVF vertices are used, otherwise we refer to the algorithm as CSVv2 (AVR). Figure 3 shows the distributions of the discriminator values for the JP and CSVv2 algorithms. The simulation describes the observed data reasonably well, except at low and high discriminator values. These small discrepancies indicate the need for correction factors using the observed collision data as described in Section 4. The discontinuities in the distribution of the JP discriminator values are due to the minimum value of 0.5% for the individual track probabilities, while the small "bump" between 0.5 and 0.6 in the CSVv2 discriminator distribution for jets in the $t\bar{t}$ topology are due to tracks or jets from pileup collisions. This is not observed in the other topologies because of the higher threshold on the jet transverse momentum in that case.

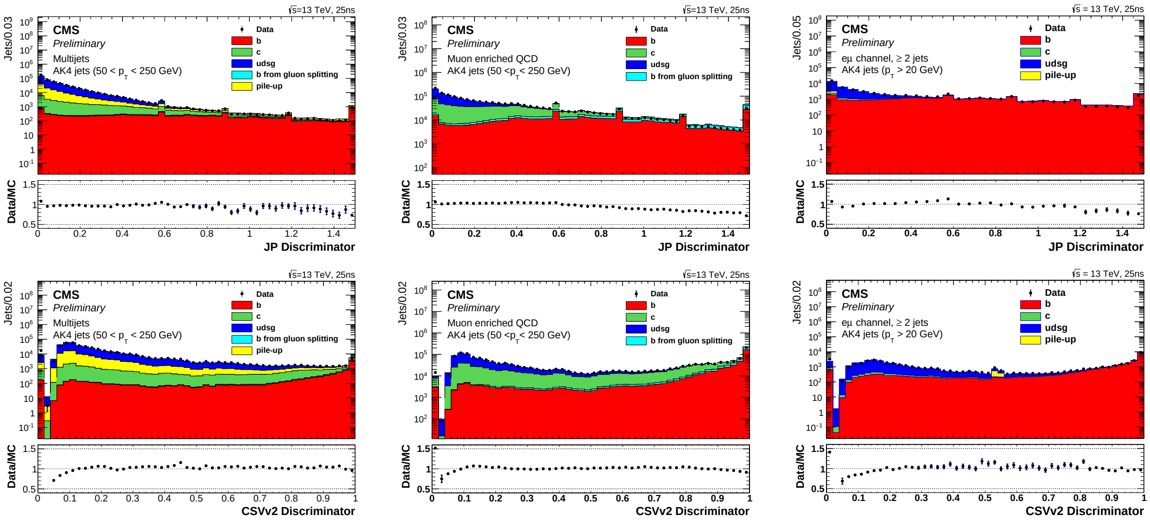

Figure 3: Discriminator values for the JP algorithm (top) and the CSVv2 algorithm (bottom). From left to right these distributions are shown for the inclusive multijet, muon enriched and dilepton $t\bar{t}$ topologies. Underflow and overflow are added to the first and last bins, respectively. The total number of entries in the simulation is normalized to the observed number of entries in data. The discontinuities in the distribution of the JP discriminator values are due to the minimum value of 0.5% for the individual track probabilities, while the small "bump" between discriminator values of 0.5 and 0.6 in the CSVv2 distribution for the jets in the $t\bar{t}$ topology are due to tracks or jets from pileup collisions. This is not observed in the other topologies because of the higher threshold on the jet transverse momentum in that case.

## 3.3 Combined MVA algorithm

A new b jet identification algorithm was developed combining the information from six different b jet identification discriminators with a Boosted Decision Tree (BDT) using the open-source scikit-learn package [25]. This combined multivariate algorithm (cMVAv2) for b jet identification is trained using the two variants of the JP algorithm as well as the two variants of the CSVv2 algorithm described in the previous section. Although the correlation between the two

CSVv2 discriminator values is very high, typically with values around 90%, a small improvement is expected for the cases in which the vertex finding algorithms reconstruct a different secondary vertex. In addition to the JP and CSVv2 algorithms also the Soft Electron (SE) and Soft Muon (SM) b jet identification discriminators are used as input variables for the cMVAv2 algorithm. The SE algorithm looks for a reconstructed electron [26] inside the jet cone ($\Delta R < 0.4$). Electrons corresponding to an associated track with too few associated hits or originating from conversions are rejected. The SM algorithm searches for a muon with a transverse momentum of at least 2 GeV among the jet constituents. For both algorithms the input variables are combined with a BDT. These discriminating variables are the 2D and 3D impact parameter significance of the lepton, the angular distance ($\Delta R$) between the jet axis and the lepton, the ratio of the transverse momentum of the lepton and that of the jet, and the transverse momentum of the lepton relative to the jet axis. In the case of the SE algorithm also an MVA-based electron identification variable is used as input. A comparison of the observed proton-proton collision data with the simulation is shown in Figures 4, 5 and 6 for the distributions of the four additional discriminators combined by the cMVAv2 algorithm, for the three jet selections respectively. The different range for the SE and SM algorithm values is related to different settings in the training when combining the input variables with a boosted decision tree.

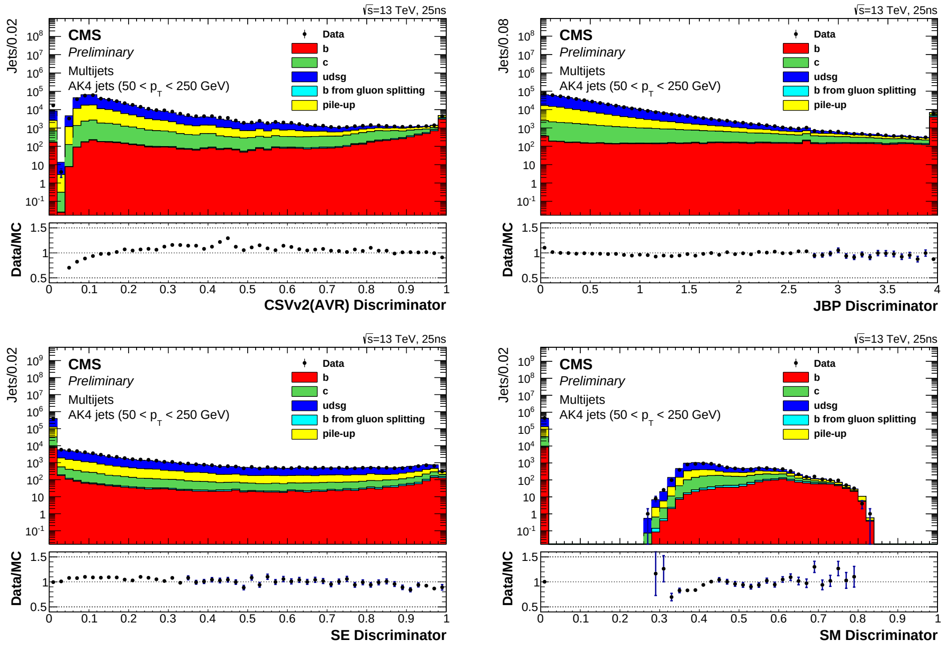

Figure 4: From top left to bottom right: distribution of the discriminator values of the CSVv2 (AVR), JBP, SE and SM taggers for jets selected in the inclusive topology. Underflow and overflow are added to the first and last bins, respectively. The total number of entries in the simulation is normalized to the observed number of entries in data. The different range for the SE and SM algorithm values is related to different settings in the training when combining the input variables with a boosted decision tree.

Figure 7 shows the cMVAv2 discriminator distribution and a reasonable agreement is observed between the observed data and the simulated events.

The performance of the JP, CSVv2 and cMVAv2 taggers is determined using simulated proton-proton collision events. The performance is presented in Figure 8 as the b jet identification efficiency versus the misidentification probability for jets in simulated $t\bar{t}$ events requiring a jet

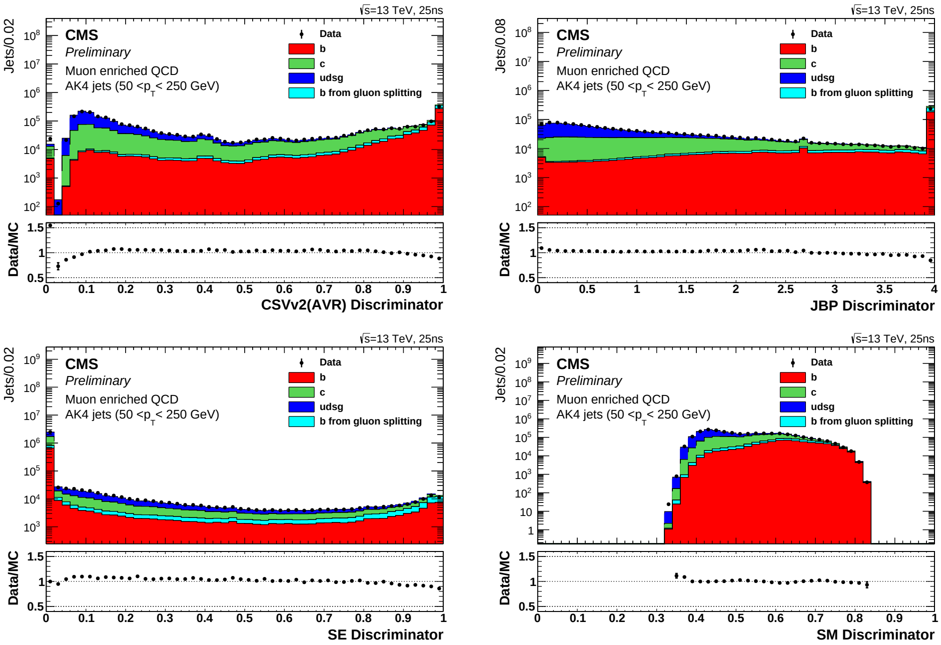

Figure 5: From top left to bottom right: distribution of the discriminator values of the CSVv2 (AVR), JBP, SE and SM taggers for jets selected in the muon enriched topology. Underflow and overflow are added to the first and last bins, respectively. The total number of entries in the simulation is normalized to the observed number of entries in data. The different range for the SE and SM algorithm values is related to different settings in the training when combining the input variables with a boosted decision tree.

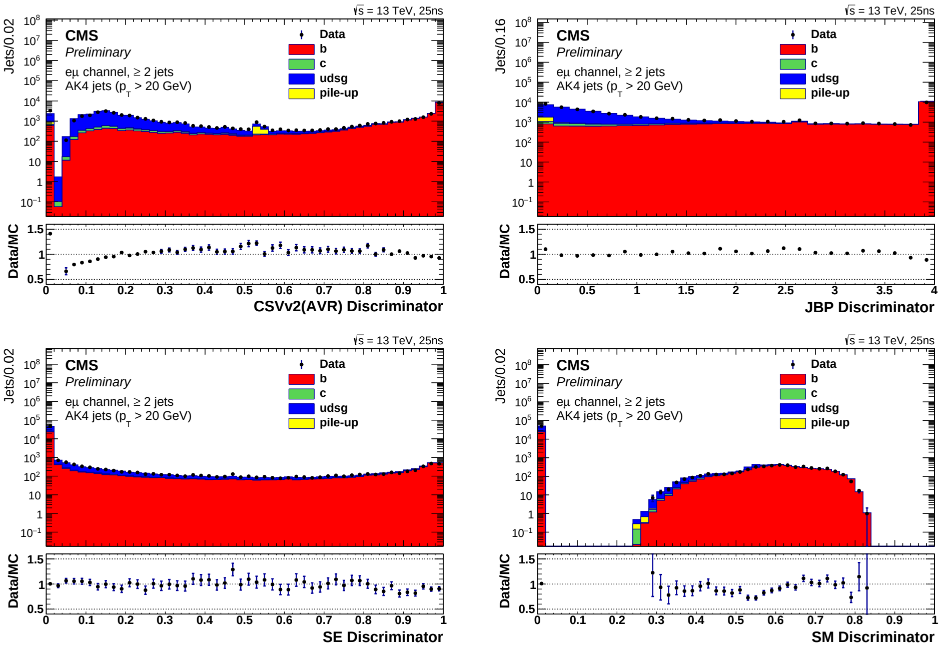

Figure 6: From top left to bottom right: distribution of the discriminator values of the CSVv2 (AVR), JBP, SE and SM taggers for jets selected in the dilepton $t\bar{t}$ topology. Underflow and overflow are added to the first and last bins, respectively. The total number of entries in the simulation is normalized to the observed number of entries in data. The different range for the SE and SM algorithm values is related to different settings in the training when combining the input variables with a boosted decision tree.

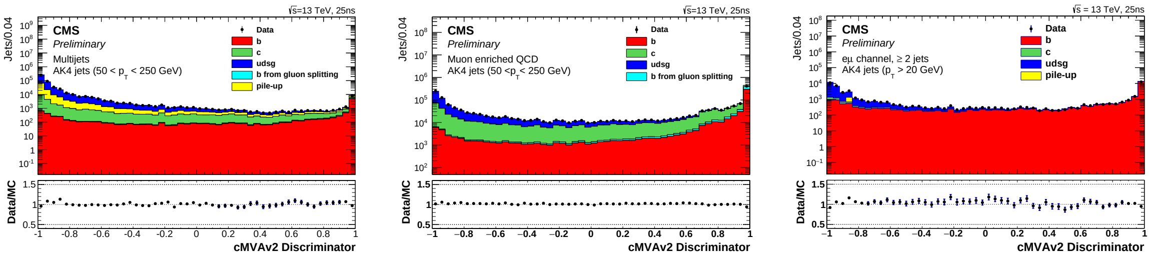

Figure 7: From left to right, the distribution of the discriminator value of the cMVAv2 algorithm for the inclusive multijet, the muon enriched and the dilepton $t\bar{t}$ topologies respectively. Underflow and overflow are added to the first and last bins, respectively. The total number of entries in the simulation is normalized to the observed number of entries in data.

transverse momentum exceeding 30 GeV. The performance in this figure serves as an illustration since the b jet identification efficiency depends on the $p_{\mathrm{T}}$ and $\eta$ distribution of the jets in the topology as well as the amount of b jets from gluon splitting in the sample. A comparison is shown with the Run 1 version of the CSV algorithm, without retraining this algorithm. The CSV algorithm was trained on 7 TeV proton-proton collisions using anti-$k_T$ jets clustered with a distance parameter $R = 0.5$. The absolute improvement of the CSVv2 (AVR) algorithm with respect to the CSV algorithm is of the order of 2 to 4% in b jet identification efficiency when comparing at the same misidentification probability for light-flavour jets. The improvement of using IVF vertices with respect to using AVR vertices in the CSVv2 algorithm is of the order of 1 to 2%. The cMVAv2 algorithm outperforms the other b jet identification algorithms for both c jets as well as light-parton and gluon jets. For light-flavour jets, the absolute efficiency improves by about 4% with respect to the CSVv2 algorithm. Three standard operating points are defined for each algorithm. These operating points, “loose” (L), “medium” (M) and “tight” (T), correspond to a threshold on the discriminator after which the misidentification probability is around 10%, 1% and 0.1% for light-flavour jets with a transverse momentum above 30 GeV. The value of the discriminator threshold for each algorithm and the corresponding efficiencies are summarized in Table 1. The numbers in Table 1 serve as an illustration since the b jet identification efficiency depends on the $p_{\mathrm{T}}$ and $\eta$ distribution of the jets as well as the amount of b

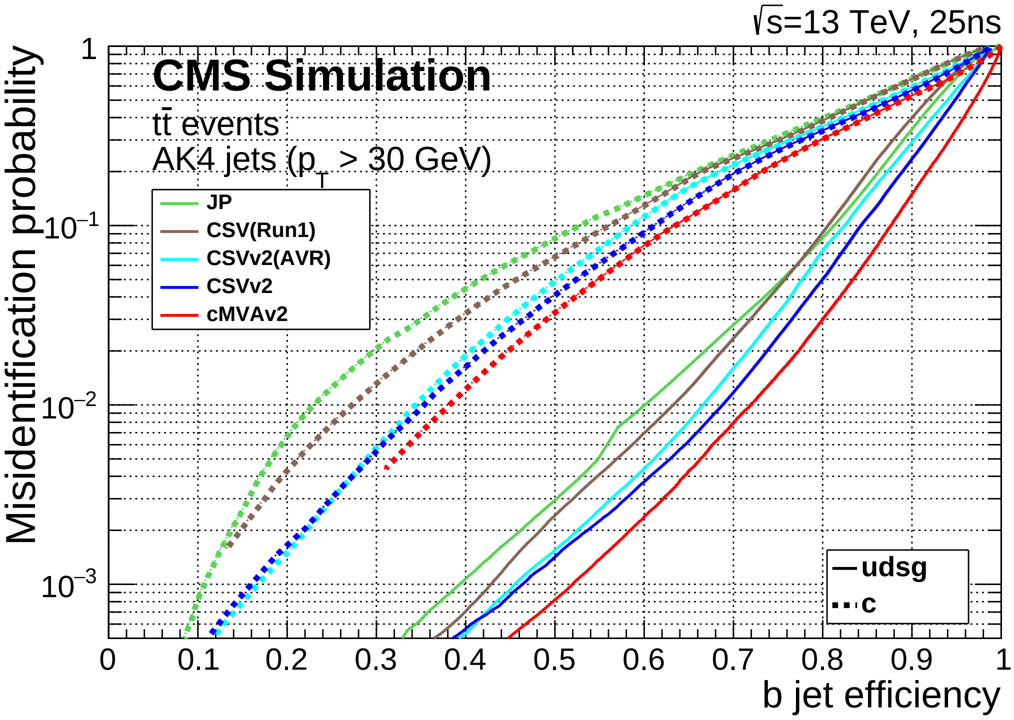

Figure 8: Performance of the b jet identification efficiency algorithms demonstrating the probability for non-b jets to be misidentified as b jet as a function of the efficiency to correctly identify b jets. The curves are obtained on simulated tt̄ events using jets with $p_T > 30$ GeV, b jets from gluon splitting to a pair of b quarks are considered as b jets. The cMVAv2 algorithm clearly outperforms the JP and CSVv2 algorithms for both c jets as well as light-parton and gluon jets. The improvement of the CSVv2 algorithm with respect to the Run 1 version of the algorithm is also shown. The performance in this figure serves as an illustration since the b jet identification efficiency depends on the $p_{\mathrm{T}}$ and $\eta$ distribution of the jets in the topology as well as the amount of b jets from gluon splitting in the sample.

jets from gluon splitting in the sample.

Table 1: Taggers, discriminator threshold and corresponding efficiency for b jets with transverse momentum above 30 GeV in simulated $\mathrm{t\bar{t}}$ events. b jets from gluon splitting to a pair of b quarks are considered as b jets. The numbers in this table serve as an indication since the b jet identification efficiency depends on the $p_{\mathrm{T}}$ and $\eta$ distribution of the jets in the topology as well as the amount of b jets from gluon splitting in the sample.

| Tagger | operating point | discriminator value | $\epsilon_b$ (%) |
|---|---|---|---|
| JetProbability (JP) | JPL | 0.245 | $\approx 82$ |
| | JPM | 0.515 | $\approx 62$ |
| | JPT | 0.760 | $\approx 42$ |
| Combined Secondary Vertex (CSVv2) | CSVv2L | 0.460 | $\approx 83$ |
| | CSVv2M | 0.800 | $\approx 69$ |
| | CSVv2T | 0.935 | $\approx 49$ |
| Combined MVA (cMVAv2) | cMVAv2L | -0.715 | $\approx 88$ |
| | cMVAv2M | 0.185 | $\approx 72$ |
| | cMVAv2T | 0.875 | $\approx 53$ |

# 4 Performance measurements for standard (AK4) jets

Several methods have been developed in the CMS collaboration to measure the b jet tagging efficiency, $\varepsilon_{\mathrm{b}}$, and the misidentification probability of light-parton jets [24, 27]. Here only the improvements or changes introduced to study the 13 TeV data are detailed, they concern mainly the dilepton-based methods.

## 4.1 b jet identification efficiency

Measurements of the b jet identification efficiency in data, and the associated *scale factors* $SF_b = {\varepsilon_{\mathrm{b}}}^{data}/{\varepsilon_{\mathrm{b}}}^{MC}$ to correct the efficiency in simulation, are performed in events enriched in b jets, either based on a muon enriched multijet selection or on the dilepton selection as defined in Section 3.

### 4.1.1 b jet identification efficiency from multijet events

During Run1, three different methods were used to measure the b jet identification efficiency, based on a sample of jets enriched in heavy flavour content. This enrichment is achieved by requiring a muon to be present within a cone $\Delta R < 0.4$ around the jet axis and is referred to as the *muon-jet sample*. The three methods using this muon-jet sample are the PtRel, Lifetime tagger (LT) and System8 methods [24, 27].

To estimate the b jet identification efficiency, the PtRel and LT methods rely on a variable for which the expected simulated distribution (*template*) is different for the various jet flavours. The fraction of b jets is then estimated by fitting the data distribution of that variable to the templates for the different jet flavours. The PtRel method uses the transverse momentum of the muon relative to the jet axis, $p_T^{rel}$, as the discriminating variable and the b tagging efficiency in data is measured as

$$\epsilon_b = \frac{N_b^{tagged}}{(N_b^{vetoed} + N_b^{tagged})}, \tag{1}$$

where $N_b^{tagged}$ and $N_b^{vetoed}$ are the estimated number of b jets from the fit in the subsample of muon-jets that pass or fail the tagging requirement, respectively. In order to reduce the light-flavour background, the presence of jet away from the first one ($\Delta R > 1.5$) (*away-jet*) and passing the threshold JBP > 2.06 is required in the event. Examples for the fitted $p_T^{rel}$ distributions for jets with $p_T$ between 70 and 100 GeV are shown in Figure 9 for muon-jets and for jets failing the CSVv2M requirement.

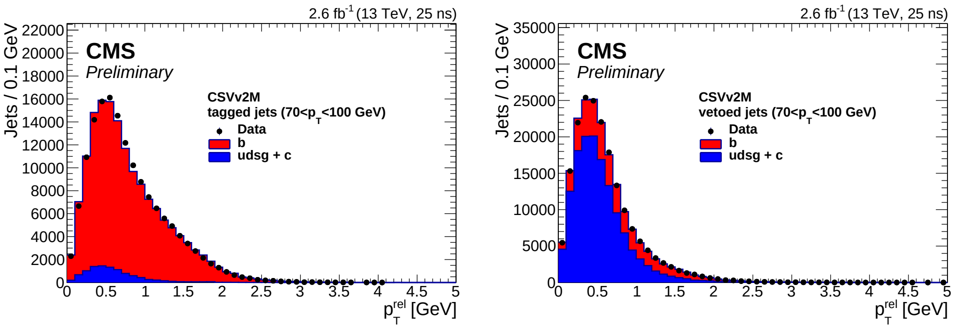

Figure 9: Comparison of the $p_T^{rel}$ distribution for the observed data with the sum of the fitted b and non-b templates for muon-jets passing (left) and jets failing the CSVv2M requirement (right). The distribution is shown for jets within a $p_T$ range of 70 to 100 GeV.

The LT method is similar to the PtRel method, but the fit is performed on the JP discriminator distribution. An away-jet is required but without applying a tagging requirement. The method is described in [27], but the treatment of the uncertainties has been improved. All shape systematic uncertainties are included in the fit as nuisance parameters and systematic correlations among b , c and light-flavour jets are taken into account. As an illustration, Figure 10 shows the fitted JP distributions for jets with a transverse momentum between 200 and 300 GeV before and after applying the CSVv2M requirement.

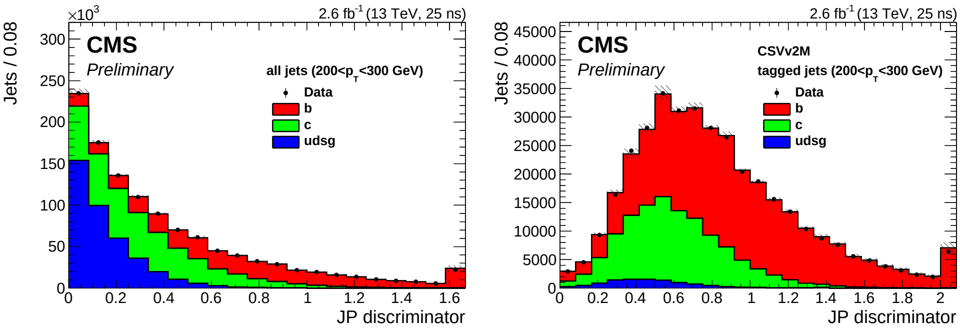

Figure 10: Comparison of the *JP* distribution for the observed data with the sum of the fitted b and non-b templates for muon-jets (left) and for the subsample of those jets passing the CSVv2M requirement (right). The distribution is shown for jets within a $p_T$ range of 200 to 300 GeV.

The System8 method uses the same away-jet tag requirement as the PtRel method and is based on the usage of two weakly correlated b taggers and two samples containing muons within jets. The first b tagger corresponds to the algorithm for which the efficiency has to be measured on the muon-jet; a second weakly correlated b tagger criterion is the request of $p_T^{rel} > 0.8$ GeV on the muon. The first sample consists of all those events with a muon-jet; the second sample is a subset where an away-jet is tagged. The correlations between the two b taggers and between the two samples are estimated using simulated events. The system of eight equations reflect the flavour contributions for each combination of sample and tagger taking into account the unknown tagging efficiencies and unknown number of b and non-b jets. The method has not changed with respect to the description in Ref. [27].

For the three methods, systematic uncertainties are taken into account. They correspond to effects related to gluon splitting into a b quark pair, the b quark fragmentation function, the choice of the b tagger operating point for the away jet. In case of the PtRel and System8 methods also the muon kinematics induces a systematic effect. For the LT method, systematic uncertainties also include the uncertainty on the fragmentation rate of a c quark to various D mesons, the branching ratios for c hadrons to muons, the $\mathrm{K_S^0}$ meson and $\Lambda$ baryon fraction and the jet energy scale. Also considered are sources of uncertainty related to the imperfect modeling of several variables that are strongly correlated to the JP algorithm's performance, such as the number of selected tracks in a jet, muon-in-jet $p_{\mathrm{T}}$ and the number of jets in the event. Sometimes it may happen that a jet has no associated tracks passing the selection requirement and if this is the case, the JP discriminator can not be calculated and no fit is performed for these jets. Therefore, an additional uncertainty to reflect the fraction of these jets in the selected sample is included for the measured scale factors. A last uncertainty is related to the away-tag requirement. This uncertainty is estimated after applying the CSVv2M requirement on the away-tag jet instead of no requirement at all. The PtRel and LT methods provide precise measurements on the full jet-$p_{\mathrm{T}}$ range from 30 to 670 GeV, while the sensitivity of the System8 method is limited to the lower part of the spectrum ($30 < p_{\mathrm{T}} < 140\,\mathrm{GeV}/c$).

### 4.1.2 b jet identification efficiency from dilepton $\mathrm{t\bar{t}}$ events

The b jet identification efficiency is measured using dilepton $\mathrm{t\bar{t}}$ events. The dilepton $\mathrm{t\bar{t}}$ events are selected by requiring the presence of exactly two isolated leptons $\ell$ (electron or muon) with opposite sign and with a dilepton invariant mass $M_{\ell\ell}$ larger than 12 GeV. Exactly two jets are required in the final state. The contribution from $Z + jets$ events is reduced by applying a veto around the $Z$ boson mass when the two leptons have the same flavour, i.e. ($|M_{\ell\ell} - M_Z| > 10\,\mathrm{GeV}$). In addition, for the events containing two same-flavour leptons, the missing transverse momentum in the event should be larger than 30 GeV. The missing transverse momentum is defined as the complement of the vectorial sum of the transverse momenta of all the reconstructed PF objects and assumes $p_{\mathrm{T}}$ balance in the transverse plane.

A first simple but robust method is a tag counting method ("TagCount"), which measures the b jet identification efficiency by counting the fraction of events with 2 b-tagged jets within a sample of events requiring exactly two selected jets with $p_T > 30\,\mathrm{GeV}$ and exactly one isolated electron and one isolated muon. The number of b-tagged jets is shown in Figure 11 after applying the CSVv2M criterion. The b tagging efficiency can be obtained from

$$\varepsilon_b = \sqrt{\frac{F_{2btag} - F_{2btag}^{non-bjet}}{f_{2b}}} \tag{2}$$

where $F_{2tag}$ is the fraction of events with two b-tagged jets, $F_{2btag}^{non-bjet}$ is the fraction of events with two b-tagged jets of which at least one is a non-b jet and $f_{2b}$ is the fraction of events with two true b jets. A closure test was performed to demonstrate that the method provides an unbiased estimate. In addition, since fractions of events are used, the approach does not depend on the $\mathrm{t\bar{t}}$ cross-section measurement and it is found to be less sensitive to experimental sources of systematics since most of these cancel. The method is however sensitive to the predicted fraction of events with non-b jets $F_{2btag}^{non-bjet}$. A conservative variation of 100% is used to estimate the systematic uncertainty on the fraction of non-b jets and represents the leading uncertainty on the final scale factor for the loose operating point of the b jet identification algorithms. The uncertainties on the background estimation for the data-driven $Z + jets$ background, as well as on the background estimated from simulation, are conservatively taken to be 50%. It constitutes the subleading source of uncertainty. Other sources of uncertainties are related to the $\mathrm{t\bar{t}}$ modelling uncertainties (factorisation and normalization scales), and to much lesser extent also the jet energy scale uncertainty and the electron and muon identification and isolation efficiencies. As a closure test Figure 11 also shows the number of b-tagged jets after applying the measured scale factors. The agreement clearly improves after applying the scale factors, thus demonstrating that the method behaves as expected.

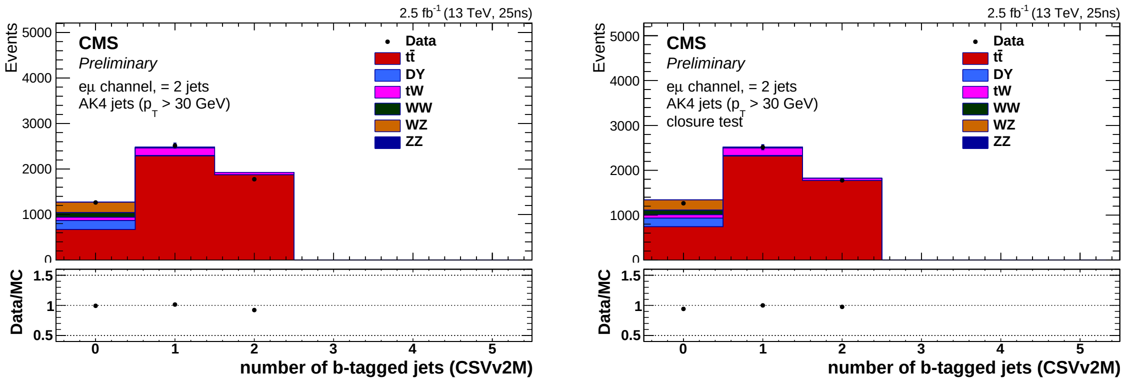

Figure 11: Number of b-tagged jets for the CSVv2M b tagging requirement, for events with exactly one electron, one muon and two jets. The distribution is shown before (left) and after (right) applying the measured scale factors. Clearly the agreement is much better after applying the measurement scale factors, demonstrating the method works well.

A second method based on the dilepton event selection is a “reweighting” technique aiming to calibrate the full b tagging discriminator shape and first described in Ref. [28]. This calibration technique (Reweight method) is designed to meet the needs of analyses in which the distribution of the b tagging discriminator values is used instead of a requirement on one (or more) of the b tagger operating points. Therefore, the distribution using simulated events has to be corrected to match the one observed in data. A scale factor for both b and light-flavour jets is derived as a function of the discriminator value and the jet kinematics, $p_\mathrm{T}$ and $\eta$. An iterative procedure is used based on a tag and probe method to measure the scale factor for both b and light-flavour jets simultaneously. The scale factor for b jets is derived from events passing the dilepton selection described earlier with exactly two jets and requiring a *tag* jet to pass the medium operating point of the tagger for which the scale factor is being measured. The scale factor for light-flavour jets is derived from a $Z + jets$ enriched selection obtained by requiring two leptons with the same flavour and an invariant mass close to that of the $Z$ boson ($|M_{\ell\ell} - M_Z| < 10\,\mathrm{GeV}$) and inverting the requirement on the missing transverse momentum. Also in this case exactly two jets are required and a b veto is applied on the tag jet using the loose operating point of the tagger for which the scale factor is being measured.

To measure the scale factor for b jets the contribution from non-b jets is subtracted using the simulated events. Similary, when measuring the scale factor for light-flavour jets, the expected contribution from b and c jets is subtracted. The scale factors are then extracted by comparing the b tagging discriminator distribution observed in data to that obtained from the simulated events. The scale factor is determined separately for exclusive bins of b tagging discriminator distribution, $p_{\mathrm{T}}$, and (in the case of the light-flavour jet scale factor) $\eta$ of the jets. Since the scale factors for light-flavour jets have an impact on the measured scale factors for the b jets, an iterative procedure is applied. In the first iteration no scale factor is applied, while for the next iteration the background is subtracted while using the scale factors obtained in the previous iteration. The iterative procedure stops once the scale factors obtained in the current iteration are stable with respect to the scale factors obtained in the previous iteration. This convergence is achieved after three iterations. When estimating the scale factor for b jets and light-flavour jets, the scale factor for c jets is set to unity with an uncertainty that is twice the one of the b scale factor. Systematic uncertainties include the jet energy scale uncertainty, the sample purity, i.e. the contamination of jets with a flavour complementary to the one for which the scale factor is measured, for which a variation of 20% is assumed, the uncertainty on the scale factor for c jets and the uncertainty due to the finite number of events in the simulation resulting in statistical fluctuations. Figures 12 and 13 show an example of the distribution for the CSVv2 tagger and the derived scale factors for jets with $p_{\mathrm{T}}$ between 30 and 40 GeV in a b jet enriched and in a non-b enriched topology, respectively.

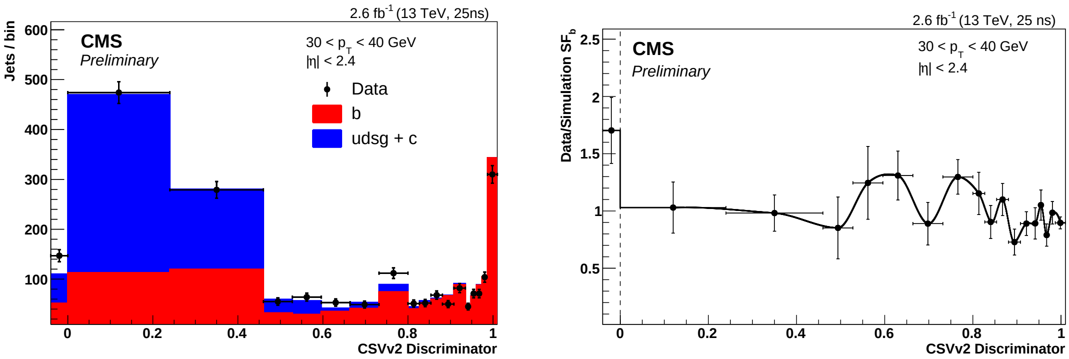

Figure 12: Distribution of the CSVv2 discriminator values for jets with $p_{\mathrm{T}}$ between 30 and 40 GeV before the scale factors are applied in the $\mathrm{t\bar{t}}$ dilepton enriched sample (left). The simulation is normalized to the number of entries for data. Measured scale factor for b jets as a function of the CSVv2 discriminator value along with the interpolation (right).

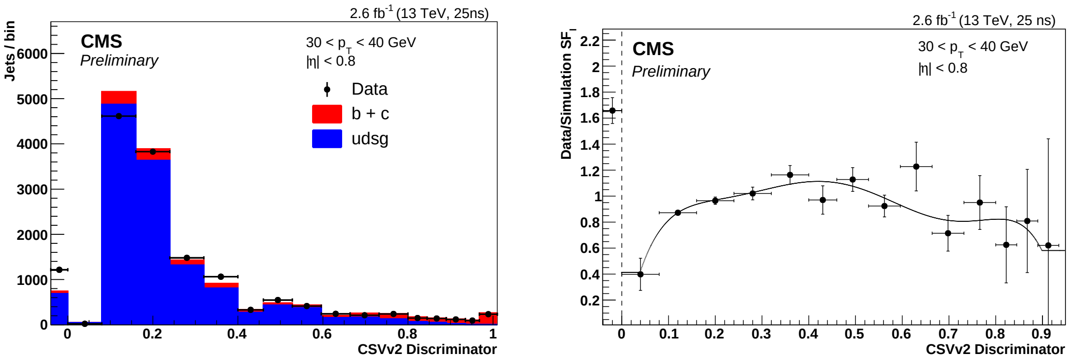

Figure 13: Distribution of the CSVv2 discriminator values for jets with $p_{\mathrm{T}}$ between 30 and 40 GeV and $|\eta| < 0.8$ before the scale factors are applied in the $Z + jets$ enriched sample (left). The simulation is normalized to the number of entries for data. Measured scale factor for light-flavour jets as a function of the CSVv2 discriminator value along with the interpolation (right).

The scale factors obtained with the Reweight method have been validated in various control regions. One example is the validation in a control region for semileptonic $\mathrm{t\bar{t}}$ events. The flavour composition in this control region is very different from both the dilepton $\mathrm{t\bar{t}}$ and $Z +$ *jets* topologies used to derive the scale factors, therefore providing a powerful cross-check. Events are selected requiring an isolated electron or muon with $p_{\mathrm{T}}$ above 30 GeV and $|\eta| < 2.1$, exactly four jets with $p_{\mathrm{T}}$ above 30 GeV, of which exactly two b-tagged according to the CSVv2M criterion. The distribution of the CSVv2 discriminator values is shown in Figure 14 for all the jets in the control region. The agreement between the observed data and simulation improves significantly after applying the measured scale factors using the Reweight method and the remaining fluctuations are covered by the systematic uncertainties.

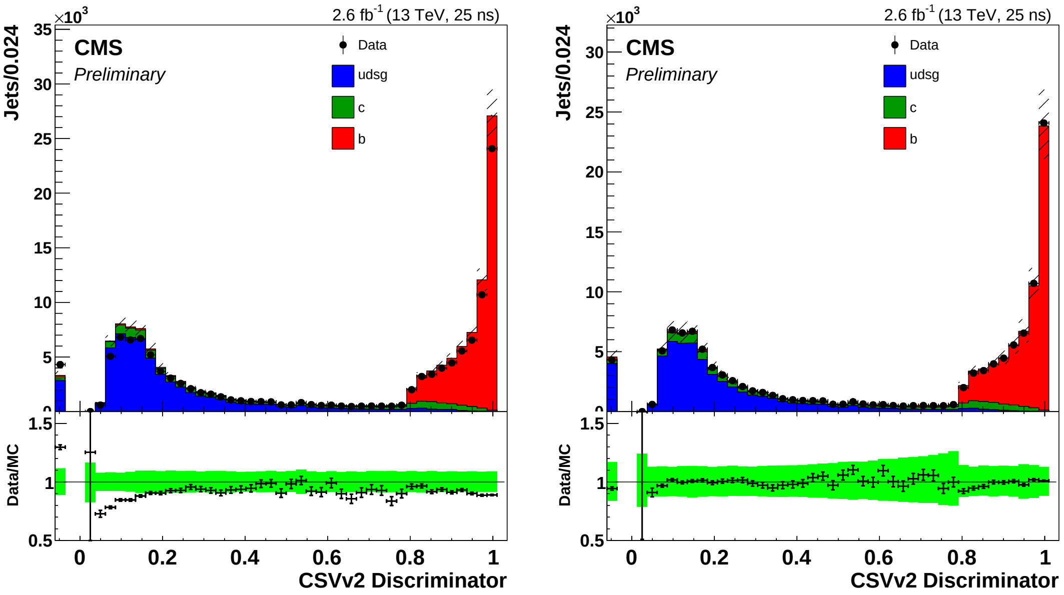

Figure 14: Distribution of the CSVv2 discriminator values for events with exactly four selected jets in semileptonic $\mathrm{t\bar{t}}$ events. Exactly two jets are required to pass the CSVv2M criterion. The values of the discriminator are shown before (left) and after (right) applying the scale factors derived with the Reweight method. Clearly the simulation describes the observed data better after applying the scale factors. The green band around the data-to-simulation ratio includes both statistic and systematic uncertainties with the latter including the uncertainty on the measured scale factors in case of the plot on the right.

### 4.1.3 Combination of the measurements for the b jet identification efficiency

Several methods have been described in the previous sections to measure the data-to-simulation scale factors of the b tagging efficiency. In this section, a combination of the measurements performed on the multijet events is presented for the b jet scale factors as a function of the jet transverse momentum ranging from 30 GeV up to 670 GeV. Jets with a higher transverse momentum are included in the last bin.

The combination is based on a weighted average, taking into account the full covariance matrix for the uncertainties, using the so-called BLUE method [29]. This technique was also used in Run 1 [24, 27], but was extended to fit all the jet $p_{\mathrm{T}}$ bins simultaneously [30], treating the bin-to-bin correlations for the systematic uncertainties more correctly.

The PtRel, System8 and LT methods are applied on the same events. However, the requirement on the away jet is different for each method. The fraction of events with b quarks in common between each pair of methods can be obtained from the simulation, and is used to estimate the statistical correlations in the combination of results. Common systematic uncertainties are treated as correlated between the three methods. Some of the systematic effects that induce a large uncertainty are however uncorrelated.

Results of $SF_\mathrm{b}$ measurements from the PtRel, System8 and LT methods for CSVv2M are compared in the upper panel of Figure 15 as a function of the b jet $p_T$. For each measured $SF_\mathrm{b}$, the thick error bar corresponds to the statistical error and the narrow one to the overall statistical plus systematic uncertainties. The combined value is displayed as a hatched area in both panels with its overall uncertainty. In the lower panel the result of a fit function is superimposed. The function used in the fit is $SF_\mathrm{b}(p_\mathrm{T}) = \alpha\frac{1+\beta p_\mathrm{T}}{1+\gamma p_\mathrm{T}}$, where $\alpha$, $\beta$, $\gamma$ are free parameters. The combined statistical plus systematic uncertainties are centered around the fit result.

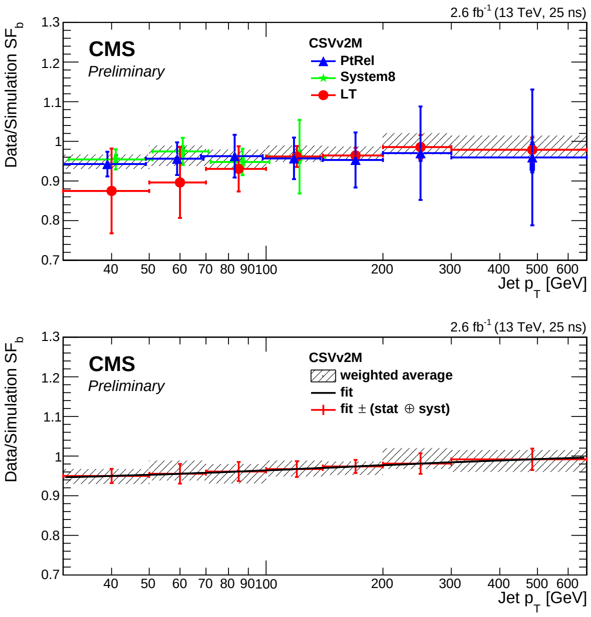

Figure 15: (upper panel) Data-to-simulation scale factor of the b tagging efficiency for CSVv2M as measured with the three methods in muon-jet events, with (thick error bar) statistical error and (narrow error bar) combined statistical and systematic uncertainties. The combined $SF_\mathrm{b}$ value with its overall uncertainty is displayed as a hatched area. (lower panel) Same combined $SF_\mathrm{b}$ value with the result of a fit function superimposed (solid curve). The combined statistical and systematic uncertainty is centered around the fit result (points with error bars). The last bin includes the overflow.

The scale factors obtained by the combination are compared to the ones measured in $\mathrm{t}\bar{\mathrm{t}}$ events. For this comparison, the combined scale factors measured in multijet events are averaged over the observed $p_\mathrm{T}$ spectrum of the b jets from $\mathrm{t}\bar{\mathrm{t}}$ decays. Results are shown in Figure 16 for the CSVv2 and cMVAv2 b jet identification algorithms. For the Reweight method, a cumulative scale factor for jets with $p_\mathrm{T}$ above 30 GeV is extracted to allow a comparison.

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

The negative tag method [27] is applied to derive scale factors for light-flavour jets. The negative tag method is based on the definition of positive and negative taggers, which are identical

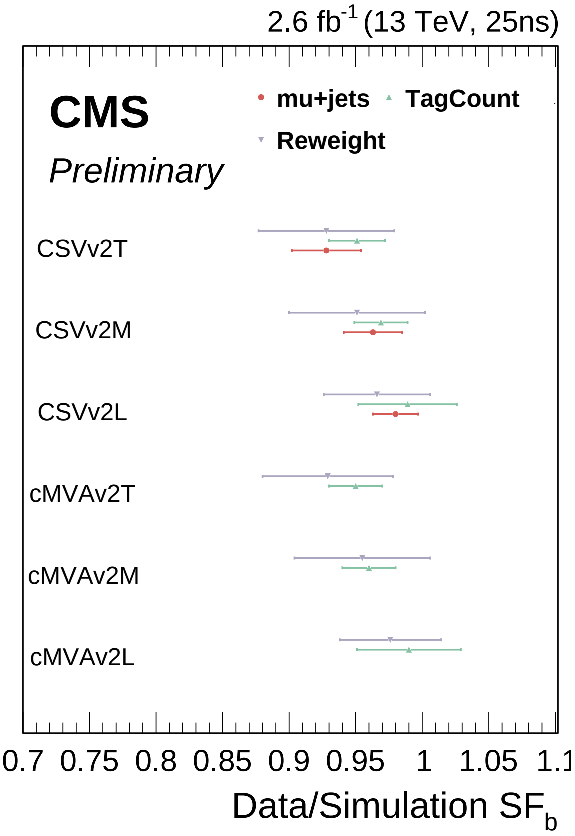

Figure 16: Comparison between the combined scale factors obtained from the muon enriched sample and the scale factors obtained in $t\bar{t}$ events. The scale factors measured in the muon enriched sample are averaged over the observed $p_T$ spectrum of the b jets from $t\bar{t}$ decays. For the Reweight method a cumulative scale factor for jets with $p_T$ above 30 GeV is extracted to allow a comparison.

to the default algorithms, except that only tracks with positive/negative impact parameter values or secondary vertices with positive/negative decay lengths are used. 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 in track reconstruction. We can therefore derive the misidentification probability from the rate, $\varepsilon^{-}$, of negative-tagged jets in an inclusive multijet sample. A correction factor, $R_{\mathrm{light}}$, is evaluated from the simulation in order to correct for second-order asymmetries in the negative and positive tag rates of light-flavour jets, and for the heavy-flavour contribution to the negative tags: $\varepsilon^{\mathrm{misid}}_{\mathrm{data}} = \varepsilon^{-}_{\mathrm{data}} \cdot R_{\mathrm{light}}$ where $R_{\mathrm{light}} = \varepsilon^{\mathrm{misid}}_{\mathrm{MC}} / \varepsilon^{-}_{\mathrm{MC}}$. The positive and negative b discriminator distributions are presented in Figure 17 for jets with $p_T > 40\,\mathrm{GeV}$. For convenience, the discriminator values of the negative taggers are shown with a negative sign. Note that since cMVAv2 discriminator values range between $-1$ and 1, a shift was introduced such that the positive cMVAv2 discriminator is defined between 0 and 2, while the negative discriminator is shown with a negative sign and gets values between $-2$ to 0. The distribution obtained from simulation is normalised to the total number of jets in data. The data agree with the simulation within 20%. 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 18 shows a summary of the measured misidentification probability and scale factor for the CSVv2M and cMVAv2M tagging requirements.

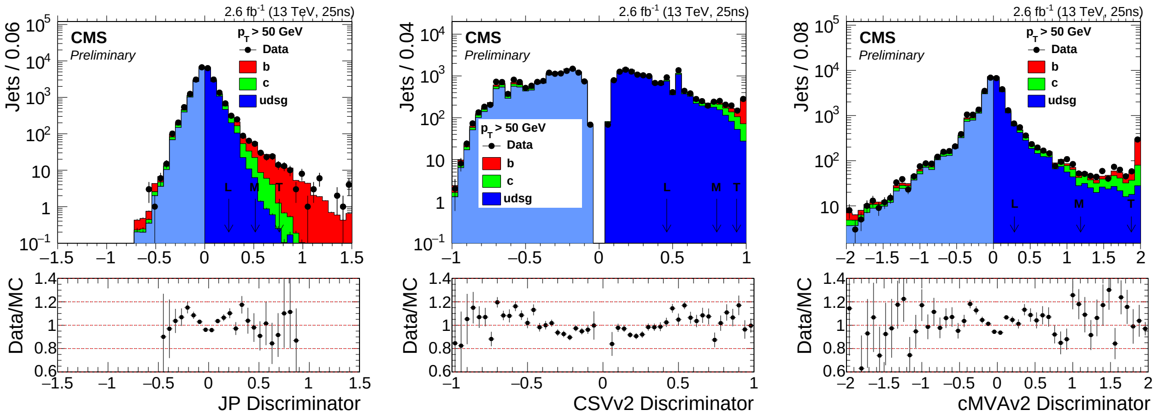

Figure 17: Comparison of the discriminator distributions for negative (negative side of the distribution) and positive taggers (positive side) in data (dots) and simulation for light-flavour jets (blue histogram, with a lighter colour for the discriminator values from negative taggers), c jets (green histogram), and b jets (red histogram) for the (left) JP, (middle) CSVv2, (right) cMVAv2 algorithms. A jet-trigger $p_{\mathrm{T}}$ threshold of 40 GeV is required for both data and simulation. The simulation is normalised to the number of entries in the data. Underflow and overflow entries are added to the first and last bins, respectively.

## 5 Identification of b jets in boosted topologies

"Boosted b tagging" aims to identify b quarks arising from boosted particles, such as the decay of highly Lorentz-boosted top quarks via t→Wb, with the W boson decaying hadronically to $q\bar{q}$, or for instance boosted Higgs or Z bosons to a b quark pair. As a result of the boost of the parent particle the decay products are collimated and subsequently merged into a single jet after hadronization. These jets are typically reconstructed with a wider distance parameter $R = 0.8$ than "standard" jets. Jet substructure techniques can be used to resolve the substructure or subjets corresponding to the hadronized partons insides these jets [27, 31–33]. When the decay

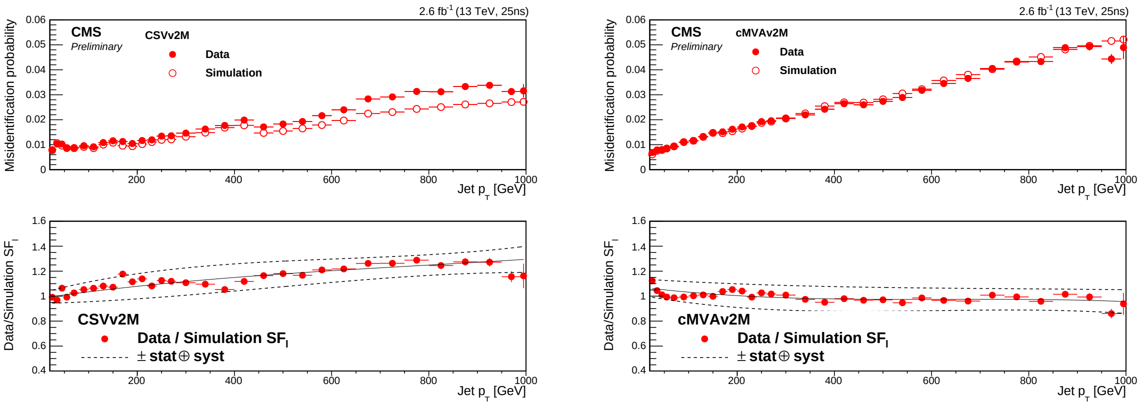

Figure 18: For the CSVv2M (on the left) and cMVAv2M (on the right) tagging requirement as a function of the jet $p_T$: (top) misidentification probability in data and simulation; (bottom) data-to-simulation scale factor of the misidenfication probability. The solid curve is the result of a fit to the observed data, the dashed curves represent the combined statistical and systematic uncertainties on the measurements.

of boosted particles contain a b quark, b tagging the subjets provides an additional handle in the boosted top quark or Higgs/$Z$ boson identification, in order to distinguish them from other jets from light-flavour partons. The concept of boosted b tagging was demonstrated in Run 1 CMS analyses [27].

## 5.1 Validation of b jet identification observables in boosted topologies

The natural choice of using events with Higgs or $Z$ boson decays into b quark pairs cannot be exploited due to limited statistics and the difficulty to select a pure sample. Therefore, QCD multijet events containing gluon splitting to b$\overline{\text{b}}$ (GSP) are used as a substitute. A muon enriched multijet sample is used to compare b tagging related observables in muon-tagged AK8 jets and their subjets. Collision events are selected using single jet triggers with $p_T$ threshold of 400 GeV. Only jets within the tracker acceptance $|\eta| < 2.4$ are selected. The soft drop algorithm [34] was used to resolve the AK8 jets into substructures, henceforth called soft drop subjets. The AK8 jets are required to have the nsubjettiness parameters [35] $\tau_1$ and $\tau_2$ satisfy $\tau_2/\tau_1 < 0.5$ to be consistent with an AK8 jet having two subjets. Offline a $p_T$ threshold of $> 425$ GeV was used for the AK8 jets. Tracks associated to AK8 jets were selected as described in Section 3. A track is associated to a subjet if the charged PF particle associated to the track is clustered in the subjet. A comparison between data and simulation for a number of b tagging observables for the AK8 jets and their subjets are shown in Figure 19 and 20. Underflow and overflow are added to the first and last histogram bins, respectively. For the figures in this section, the total number of entries in the simulation is normalized to the observed number of entries in data. The simulation describes the observed data reasonably well, with variations up to 20%.

The identification of boosted hadronically decaying top quarks is performed using t tagging algorithms [36]. Due to the large Lorentz boost, the decay products of boosted top quarks are clustered inside one single large jet, called fat-jet. The t tagging algorithms rely on the mass of the jet and on the clustering history of the jet constituents to decide whether a fat jet is likely to come from a top quark decay. The performance of t tagging algorithms can be significantly improved exploiting subjet b tagging to identify the b quark from the top quark decay, as discussed in [27]. When applying b tagging to subjets in boosted top quark events, it is necessary to confirm that the properties of the tracks and of the secondary vertices associated

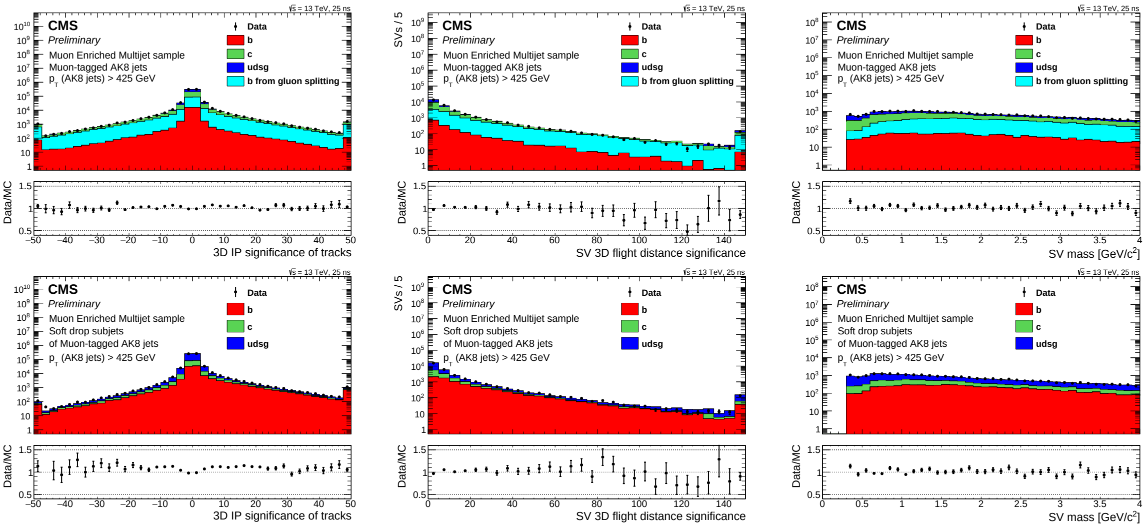

Figure 19: The impact parameter significance of the selected tracks (left), significance of the secondary vertex flight distance (middle) and mass of the secondary vertex (right). The top (bottom) panels show the distributions for (the soft drop subjets of) AK8 jets in the muon enriched sample. Underflow and overflow are added to the first and last bins, respectively. The total number of entries in the simulation is normalized to the observed number of entries in data.

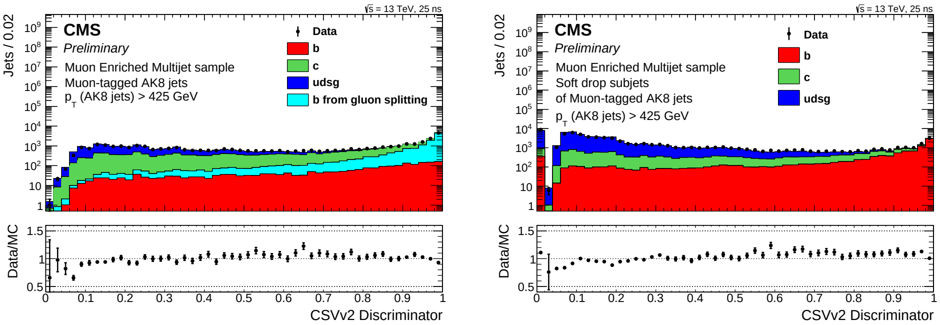

Figure 20: Discriminator values for the CSVv2 algorithm for AK8 jets (left) and soft drop subjets thereof (right). Underflow and overflow are added to the first and last bins, respectively. The total number of entries in the simulation is normalized to the observed number of entries in data.

to the subjets are described properly by the simulated events. A comparison between data and simulation for a number of b tagging observables for subjets is shown in Figure 21. Events are selected corresponding to the semileptonic $\mathrm{t\bar{t}}$ decay. Exactly one isolated muon with $p_T >$ 50 GeV and $|\eta| < 2.1$ and at least one AK4 jet are required in the same hemisphere of the event, using $|\phi_{jet} - \phi_\mu| < \frac{2}{3}\pi$. The hadronic decay of the top quark is selected requiring a t-tagged AK8 jet in the other hemisphere. This AK8 jet is required to have $p_T >$ 400 GeV, $|\eta| < 2.4$ and fulfilling the t tagging criteria using the soft drop algorithm. The nsubjettiness parameters $\tau_2$ and $\tau_3$ should satisfy $\tau_3/\tau_2 < 0.86$ and the soft drop mass $m_{SD}$ should be consistent with the top quark mass, with $110 < m_{SD} < 210$ GeV.

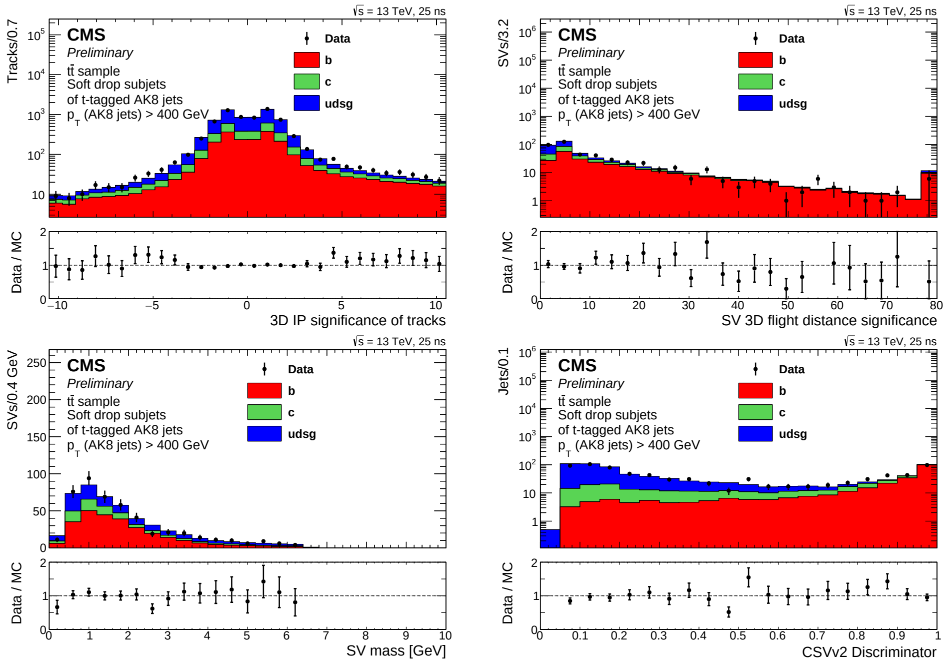

Figure 21: The impact parameter significance of the selected tracks (upper left), the significance of the secondary vertex flight distance (upper right), the mass of the secondary vertex (lower left) and the discriminator values for the CSVv2 algorithm (lower right) for soft drop subjets corresponding to the hadronic decay of the top quark selected using a t-tagged AK8 jet. Underflow and overflow are added to the first and last bins, respectively. The total number of entries in the simulation is normalized to the observed number of entries in data.

## 5.2 b jet identification measurements for subjets

For b jets in boosted topologies, the b tagging efficiency scale factor for the CSVv2 algorithm is measured for softdrop subjets of AK8 jets using the same selection requirements as in the previous section. To increase the number of selected jets, AK8 jets are required with $p_\mathrm{T}$ exceeding 200 GeV. Only subjets containing a muon are selected and the LT method presented in Section 4.1 is applied to estimate the scale factors. An example for the resulting fitted JP distribution is presented in Figure 22 for softdrop subjets with a transverse momentum between 180 and 240 GeV before and after applying the CSVv2M requirement.

The LT method is performed in bins of softdrop subjet transverse momentum for the CSVv2L and CSVv2M requirements. The obtained scale factors for b subjets for the CSVv2L and CSVv2M requirement are shown in Figure 23. The total uncertainty on the measured scale factors for softdrop subjets includes the systematic uncertainties also considered for AK4 jets. An addi-

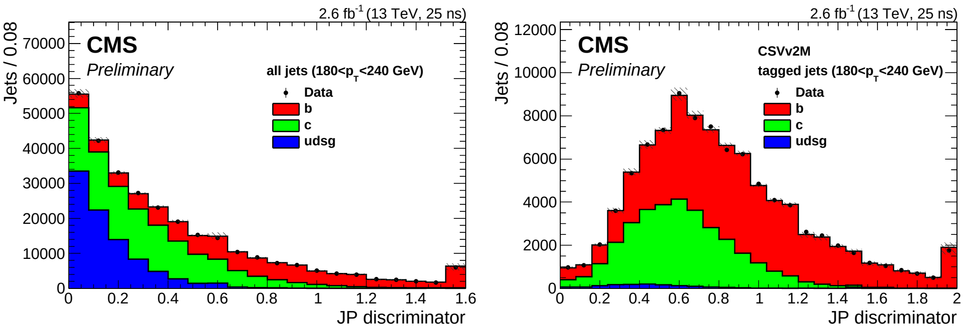

Figure 22: Comparison of the $JP$ distribution for the observed data with the sum of the fitted b and non-b templates for softdrop subjets containing a muon (left) and for the subsample of those jets passing the CSVv2M requirement (right). The distribution is shown for subjets within a $p_{\mathrm{T}}$ range of 180 to 240 GeV.

tional uncertainty is added to cover the variation in the scale factor central value when selecting subjets without requiring the presence of a muon. On the same figure also the scale factors for AK4 jets obtained with the LT method are shown for comparison. The scale factors are compatible in both cases, providing confidence in the measurement.

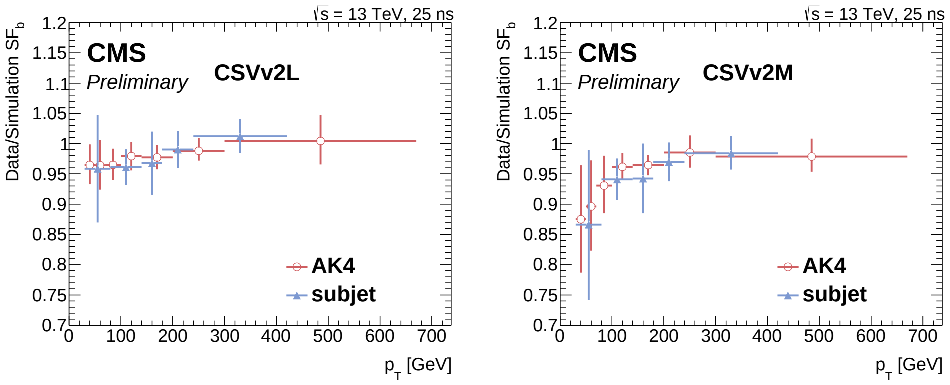

Figure 23: Comparison for the CSVv2L (left) and CSVv2M (right) scale factors obtained with the LT method for AK4 b jets and softdrop subjets of AK8 jets as a function of the (sub)jet transverse momentum. The uncertainty shows the combined statistical and systematic uncertainty. The scale factors derived for the two different jet types agree well.

### 5.3 Measurement of the misidenfication probability for subjets

For the scale factor measurement of the misidentification probability, the inclusive multijet sample is used with a selection requirement on the fat-jet pruned mass of $50 < m_{pruned} < 200\,\mathrm{GeV}$. The same negative tag method is applied as for the AK4 jets. The measurement is performed in bins of pseudorapidity for both the CSVv2L and CSVv2M requirement. As an example Figure 24 shows the measured scale factor for the CSVv2M requirement as a function of the subjet transverse momentum in an inclusive bin of the pseudorapidity. As a comparison the equivalent AK4 jet scale factor is shown and the measurements agree well within their uncertainty.

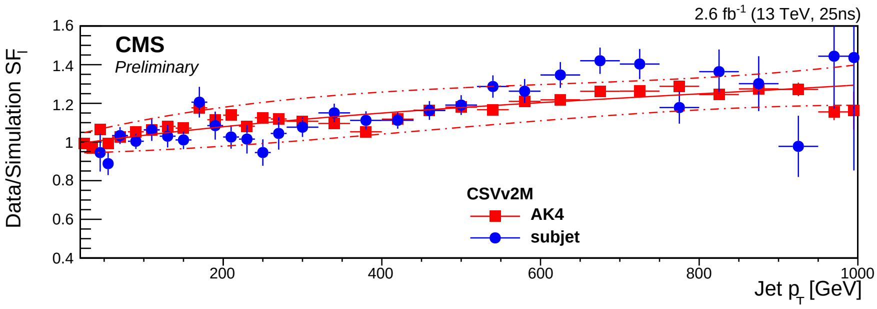

Figure 24: Comparison for the CSVv2M scale factor of the misidentification probability obtained with the negative tag method for AK4 jets and softdrop subjets of AK8 jets as a function of the (sub)jet tranvserse momentum. Within the uncertainty, the scale factor measurements are compatible.

## 6 Conclusions

Results for performance of b tagging algorithms with 2.6 fb$^{-1}$ of proton-proton collision data collected in 2015 at 13 TeV with 25 ns bunch spacing are presented. Data-to-simulation scale factors for the b tagging efficiency for the CSVv2, cMVAv2 and JP b tagging algorithms are measured in both muon enriched multijet events as well as in $t\bar{t}$ events. The scale factors for the misidentification probability of light-flavour jets as b jets are performed using multijet and $Z + jets$ events. The b jet identification algorithms and variables are also studied in boosted multijet and top quark pair events for AK8 jets and soft drop subjets. Scale factors are also derived for softdrop subjets of AK8 jets for the CSVv2 algorithm.

## References

[1] CMS Collaboration, “The CMS experiment at the CERN LHC”, *JINST* **3** (2008) S08004, doi:10.1088/1748-0221/3/08/S08004.

[2] CMS Collaboration, “Description and performance of track and primary-vertex reconstruction with the CMS tracker”, *Journal of Instrumentation* **9** (2014), no. 10, P10009.

[3] CMS Collaboration, “Tracking and Vertexing Results from RunII First Collisions”, CMS Physics Analysis Summary CMS-PAS-TRK-15-001, CERN, 2016.

[4] P. Nason, “A New method for combining NLO QCD with shower Monte Carlo algorithms”, *JHEP* **11** (2004) 040, doi:10.1088/1126-6708/2004/11/040, arXiv:hep-ph/0409146.

[5] S. Frixione, P. Nason, and C. Oleari, “Matching NLO QCD computations with Parton Shower simulations: the POWHEG method”, *JHEP* **11** (2007) 070, doi:10.1088/1126-6708/2007/11/070, arXiv:0709.2092.

[6] S. Alioli, P. Nason, C. Oleari, and E. Re, “A general framework for implementing NLO calculations in shower Monte Carlo programs: the POWHEG BOX”, *JHEP* **06** (2010) 043, doi:10.1007/JHEP06(2010)043, arXiv:1002.2581.

[7] J. M. Campbell, R. K. Ellis, P. Nason, and E. Re, “Top-pair production and decay at NLO matched with parton showers”, *JHEP* **04** (2015) 114, doi:10.1007/JHEP04(2015)114, arXiv:1412.1828.

[8] E. Re, “Single-top Wt-channel production matched with parton showers using the POWHEG method”, *Eur. Phys. J.* **C71** (2011) 1547, doi:10.1140/epjc/s10052-011-1547-z, arXiv:1009.2450.

[9] T. Melia, P. Nason, R. Rontsch, and G. Zanderighi, “W+W-, WZ and ZZ production in the POWHEG BOX”, *JHEP* **11** (2011) 078, doi:10.1007/JHEP11(2011)078, arXiv:1107.5051.

[10] P. Nason and G. Zanderighi, “$W^+W^-$ , $WZ$ and $ZZ$ production in the POWHEG-BOX-V2”, *Eur. Phys. J.* **C74** (2014), no. 1, 2702, doi:10.1140/epjc/s10052-013-2702-5, arXiv:1311.1365.

[11] J. Alwall et al., “MadGraph 5 : Going Beyond”, *JHEP* **06** (2011) 128, doi:10.1007/JHEP06(2011)128, arXiv:1106.0522.

[12] J. Alwall et al., “The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations”, *JHEP* **07** (2014) 079, doi:10.1007/JHEP07(2014)079, arXiv:1405.0301.

[13] T. Sjostrand, S. Mrenna, and P. Z. Skands, “A Brief Introduction to PYTHIA 8.1”, *Comput. Phys. Commun.* **178** (2008) 852–867, doi:10.1016/j.cpc.2008.01.036, arXiv:0710.3820.

[14] CMS Collaboration, “Event generator tunes obtained from underlying event and multiparton scattering measurements”, arXiv:1512.00815.

[15] GEANT4 Collaboration, “GEANT4: A Simulation toolkit”, *Nucl. Instrum. Meth.* **A506** (2003) 250–303, doi:10.1016/S0168-9002(03)01368-8.

[16] CMS Collaboration, “Particle-Flow Event Reconstruction in CMS and Performance for Jets, Taus, and MET”, Technical Report CMS-PAS-PFT-09-001, CERN, 2009. Geneva, Apr, 2009.

[17] CMS Collaboration, “Commissioning of the Particle-Flow reconstruction in Minimum-Bias and Jet Events from pp Collisions at 7 TeV”, Technical Report CMS-PAS-PFT-10-002, CERN, Geneva, 2010.

[18] F. Beaudette, “The CMS Particle Flow Algorithm”, (2013). arXiv:1401.8155.

[19] M. Cacciari, G. P. Salam, and G. Soyez, “The anti-$k_t$ jet clustering algorithm”, *JHEP* **04** (2008) 063, doi:10.1088/1126-6708/2008/04/063, arXiv:0802.1189.

[20] CMS Collaboration, “Determination of jet energy calibration and transverse momentum resolution in CMS”, *Journal of Instrumentation* **6** (2011), no. 11, P11002.

[21] W. Waltenberger, “Adaptive Vertex Reconstruction”, Technical Report CMS-NOTE-2008-033, CERN, Geneva, Jul, 2008.

[22] R. Frühwirth, W. Waltenberger, and P. Vanlaer, “Adaptive Vertex Fitting”, Technical Report CMS-NOTE-2007-008, CERN, Geneva, Mar, 2007.

[23] CMS Collaboration, “Measurement of $B\bar{B}$ Angular Correlations based on Secondary Vertex Reconstruction at $\sqrt{s} = 7$ TeV”, *JHEP* **03** (2011) 136, doi:10.1007/JHEP03(2011)136, arXiv:1102.3194.

[24] CMS Collaboration, "Identification of b-quark jets with the CMS experiment", *JINST* **8** (2013) P04013, doi:10.1088/1748-0221/8/04/P04013.

[25] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python", *Journal of Machine Learning Research* **12** (2011) 2825–2830.

[26] CMS Collaboration, "Performance of electron reconstruction and selection with the CMS detector in proton-proton collisions at $\sqrt{s} = 8$ TeV", *J. Instrum.* **10** (Feb, 2015) P06005. 63 p. Comments: Replaced with published version. Added journal reference and DOI.

[27] CMS Collaboration, "Performance of b tagging at sqrt(s)=8 TeV in multijet, ttbar and boosted topology events", CMS Physics Analysis Summary CMS-PAS-BTV-13-001, CERN, 2013.

[28] CMS Collaboration, "Search for Higgs Boson Production in Association with a Top-Quark Pair and Decaying to Bottom Quarks or Tau Leptons", Technical Report CMS-PAS-HIG-13-019, CERN, Geneva, 2013.

[29] L. Lyons, D. Gibaut, and P. Clifford, "How to combine correlated estimates of a single physical quantity", *Nucl. Instrum. Meth. A* **270** (1988) 110, doi:10.1016/0168-9002(88)90018-6.

[30] A. Valassi, "Combining correlated measurements of several different physical quantities", *Nucl. Instrum. Meth. A* **500** (2003) 391–405, doi:10.1016/S0168-9002(03)00329-2.

[31] CMS Collaboration, "Identifying Hadronically Decaying W Bosons Merged into a Single Jet", CMS Physics Analysis Summary CMS-PAS-JME-13-006, CERN, 2013.

[32] CMS Collaboration, "Boosted Top Jet Tagging at CMS", CMS Physics Analysis Summary CMS-PAS-JME-13-007, CERN, 2013.

[33] CMS Collaboration, "Top Tagging with New Approaches", CMS Physics Analysis Summary CMS-PAS-JME-15-002, CERN, 2015.

[34] A. J. Larkoski, S. Marzani, G. Soyez, and J. Thaler, "Soft Drop", *JHEP* **05** (2014) 146, doi:10.1007/JHEP05(2014)146, arXiv:1402.2657.

[35] J. Thaler and K. Van Tilburg, "Identifying Boosted Objects with N-subjettiness", *JHEP* **1103** (2011) 015, doi:10.1007/JHEP03(2011)015, arXiv:1011.2268.

[36] CMS Collaboration, "Top Tagging with New Approaches", CMS Physics Analysis Summary CMS-PAS-JME-15-002, CERN, 2016.

# A Efficiency of b jet identification as a function of the jet transverse momentum

To facilitate phenomenological studies relying on b jet identification, we provide b jet identification efficiencies as a function of the jet transverse momentum for the three operating points of the cMVAv2 algorithm. These efficiencies are shown in Figure 25 and obtained from a simulated $t\bar{t}$ sample. The dependency of the efficiency as a function of the transverse momentum is fitted by the polynomial functions given in Table 2.

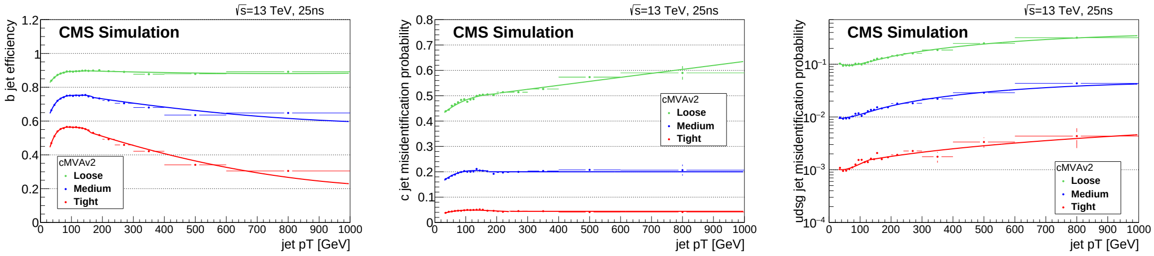

Figure 25: Efficiency to identify jets as originating from b quarks for the three different operating points (loose, medium and tight) of the cMVAv2 algorithm. The efficiencies are shown as a function of the jet transverse momentum for jets with $p_T > 30\,\mathrm{GeV}$ for b jets (left), c jets (middle) and light jets (right). The last bin includes the overflow. The dependency of the efficiency on the jet transverse momentum is fitted and the fitted function is shown. The contributions of b and c jets from gluon splitting are not considered in these figures. The efficiencies are obtained from simulated $t\bar{t}$ events.

Table 2: Fitted functions for the efficiency of the cMVAv2 algorithm at its three different operating points as a function of the jet $p_{\mathrm{T}}$.

| Jet flavour | operating point | jet $p_{\mathrm{T}}$ range | function |
|---|---|---|---|
| b | Loose | $30 \leq p_T < 150\,\mathrm{GeV}$ | $0.707 + 5.6 \cdot 10^{-3} \cdot p_T - 6.27 \cdot 10^{-5} \cdot p_T^2 + 3.10 \cdot 10^{-7} \cdot p_T^3 - 5.63 \cdot 10^{-10} \cdot p_T^4$ |
| | | $150 \leq p_T$ | $0.906 - 6.39 \cdot 10^{-5} \cdot p_T + 4.11 \cdot 10^{-8} \cdot p_T^2$ |
| | Medium | $30 \leq p_T < 175\,\mathrm{GeV}$ | $0.421 + 0.0107 \cdot p_T - 1.314 \cdot 10^{-4} \cdot p_T^2 + 7.268 \cdot 10^{-7} \cdot p_T^3 - 1.523 \cdot 10^{-9} \cdot p_T^4$ |
| | | $175 \leq p_T$ | $0.79 - 3.17 \cdot 10^{-4} \cdot p_T + 1.24 \cdot 10^{-7} \cdot p_T^2$ |
| | Tight | $30 \leq p_T < 160\,\mathrm{GeV}$ | $0.127 + 0.01578 \cdot p_T - 2.126 \cdot 10^{-4} \cdot p_T^2 + 1.273 \cdot 10^{-6} \cdot p_T^3 - 2.88 \cdot 10^{-9} \cdot p_T^4$ |
| | | $160 \leq p_T$ | $0.634 - 6.74 \cdot 10^{-4} \cdot p_T + 2.69 \cdot 10^{-7} \cdot p_T^2$ |
| c | Loose | $30 \leq p_T < 205\,\mathrm{GeV}$ | $0.40 + 1.23 \cdot 10^{-3} \cdot p_T - 4.60 \cdot 10^{-6} \cdot p_T^2 + 5.71 \cdot 10^{-9} \cdot p_T^3$ |
| | | $205 \leq p_T$ | $0.478 + 1.573 \cdot 10^{-4} \cdot p_T$ |
| | Medium | $30 \leq p_T < 170\,\mathrm{GeV}$ | $0.13 + 1.48 \cdot 10^{-3} \cdot p_T - 1.00 \cdot 10^{-5} \cdot p_T^2 + 2.65 \cdot 10^{-8} \cdot p_T^3 - 2.36 \cdot 10^{-11} \cdot p_T^4$ |
| | | $170 \leq p_T$ | $0.20$ |
| | Tight | $30 \leq p_T < 240\,\mathrm{GeV}$ | $0.024 + 5.27 \cdot 10^{-4} \cdot p_T - 3.72 \cdot 10^{-6} \cdot p_T^2 + 9.87 \cdot 10^{-9} \cdot p_T^3 - 8.83 \cdot 10^{-12} \cdot p_T^4$ |
| | | $240 \leq p_T$ | $0.044$ |
| light | Loose | $30 < p_T < 130\,\mathrm{GeV}$ | $0.124 - 1.0 \cdot 10^{-3} \cdot p_T + 1.06 \cdot 10^{-5} \cdot p_T^2 - 3.18 \cdot 10^{-8} \cdot p_T^3 + 3.13 \cdot 10^{-11} \cdot p_T^4$ |
| | | $130 \leq p_T$ | $0.055 + 4.53 \cdot 10^{-4} \cdot p_T - 1.6 \cdot 10^{-7} \cdot p_T^2$ |
| | Medium | $30 \leq p_T < 170\,\mathrm{GeV}$ | $9.59 \cdot 10^{-3} - 1.96 \cdot 10^{-5} \cdot p_T + 4.53 \cdot 10^{-7} \cdot p_T^2 - 1.08 \cdot 10^{-9} \cdot p_T^3 + 7.62 \cdot 10^{-13} \cdot p_T^4$ |
| | | $170 \leq p_T$ | $5.07 \cdot 10^{-3} + 6.02 \cdot 10^{-5} \cdot p_T - 2.3 \cdot 10^{-8} \cdot p_T^2$ |
| | Tight | $30 \leq p_T < 130\,\mathrm{GeV}$ | $1.24 \cdot 10^{-3} - 1.27 \cdot 10^{-5} \cdot p_T + 1.98 \cdot 10^{-7} \cdot p_T^2 - 7.46 \cdot 10^{-10} \cdot p_T^3 + 8.35 \cdot 10^{-13} \cdot p_T^4$ |
| | | $130 \leq p_T$ | $1.08 \cdot 10^{-3} + 3.54 \cdot 10^{-6} \cdot p_T$ |