PROCEEDINGS OF SCIENCE

Heavy-flavour jet identification at the CMS experiment for Run 2

Mauro Verzetti, on behalf of the CMS Collaboration*
University of Rochester
E-mail: mauro.verzetti@cern.ch

Identification of jets originating from b or c quarks is important for a wide variety of standard model physics (e.g. final states involving top quarks and Higgs bosons) as well as searches for physics beyond the standard model. Due to the increased center-of-mass energy during the LHC Run 2, final states with boosted b quark jets became increasingly important.

The CMS Collaboration improved on the existing b jet identification algorithms, and developed the first charm jet tagger. To address boosted topologies a boosted double-b tagger has been developed as well. A review of these new developments is presented, together with their performance on proton-proton collision data recorded by the CMS detector at a centre-of-mass energy of $\sqrt{s} = 13$ TeV during 2015 and 2016.

38th International Conference on High Energy Physics
3-10 August 2016
Chicago, USA

*Speaker.

PoS(ICHEP2016)288

1. Heavy-flavour jet identification in CMS

The Compact Muon Solenoid (CMS) detector [1] is one of the two multi-purpose experimental apparatus that gathers data from the LHC accelerator, where proton beams collide at the center-of-mass energy of $\sqrt{s} = 13$ TeV. Quarks and gluons emitted in the hard scattering process hadronize, producing showers of particles that are reconstructed in the detector. These particles are subsequently clustered by dedicated algorithms to form jets which aim to model the original shower. Showers initiated by a heavy flavour quark contain a heavy hadron, which can be used to tag the jet through the presence of highly displaced tracks and secondary vertices, given by the long lifetime of the hadron, or the presence of leptons in the jet, given by the high branching fraction of such hadrons to final states involving leptons. In the boosted regime decay products from heavy particles can be highly Lorentz-boosted leading to final states with overlapping and merged jets (known as fat jets)

Heavy flavour tagging in CMS starts from clustered jets, where the tracks are selected in order to remove contributions from neighbouring collisions within the same bunch crossing (pileup) and misreconstructed tracks. From the set of selected tracks, it is possible produce a track-based tagger exploiting quantities linked to the displacement of the tracks from the primary interaction vertex. One example of such tagger is the jet probability (JP) [2]. A more effective approach, however, is to cluster the selected tracks into secondary vertices (SV), displaced from the primary one. This clustering is performed in CMS using two algorithms: the adaptive vertex reconstruction (AVR) [3], operating on all the tracks associated to the jet, and the inclusive vertex fitter (IVF) [4], which is applied to all the tracks in the event, matching the SV to the jet in a second moment. The IVF has been chosen as the default SV clustering algorithm for Run 2 analyses. Finally, it is possible to tag heavy-flavour jets by the presence of a low p_T lepton within the jet constituents (soft lepton taggers).

These three tagging methods exploit different characteristics of heavy-flavour jets and can therefore be combined to obtain a better-performing classifier. Three notable examples in CMS are the combined secondary vertex version 2 (CSVv2) [5], the combined MVA (cMVAv2) [5] and the charm tagger [6]. The CSVv2 algorithm combines the tracks and the SV's information in the jet by means of a multi-layer perceptron. The cMVAv2 algorithm takes as input features the outcome of the CSVv2 tagging both with AVR and IVF vertices, the output value of the two available JP taggers and from the soft lepton taggers. These features are combined in a boosted decision tree (BDT) to obtain maximal separation. The performances of the taggers discussed so far can be found in Fig. 1.

2. Charm tagging in CMS

The approach followed by the charm tagger is slightly different from the one used by the cMVAv2. In the case of the charm tagger all the discriminating features coming from the different sources (tracks, secondary vertices and soft leptons) are directly combined into a BDT in order to minimize the information lost. The features used for the discrimination show a behaviour for c jets that is in between the distributions for light and b jets. Given this peculiarity and the implementation used for the BDT classifier, which does not contemplate more than two classes, it has been decided

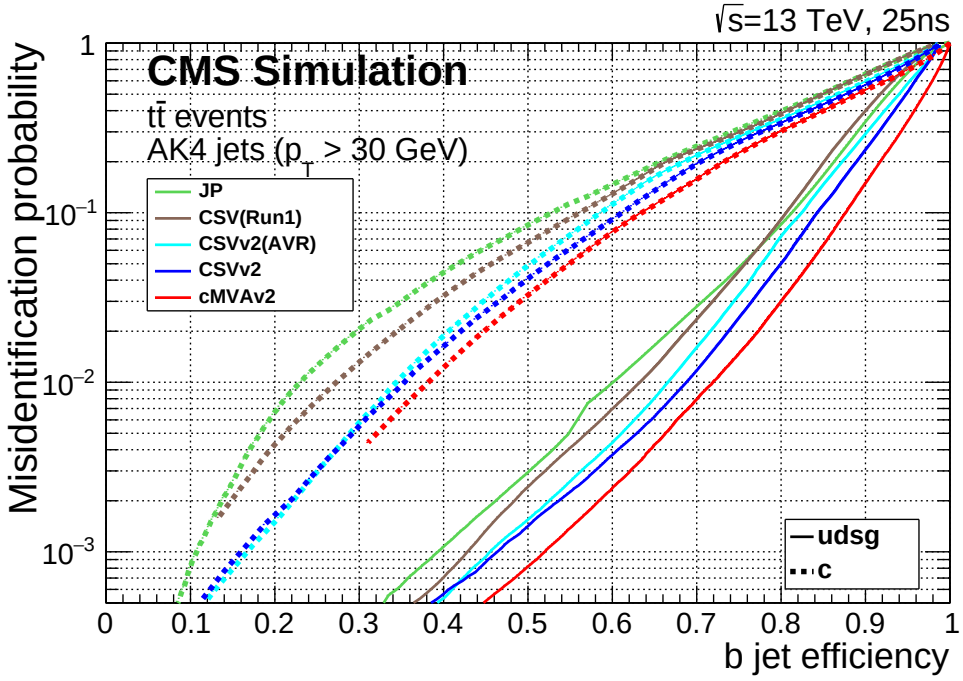

Figure 1: Performance of the different b tagging algorithms with respect to charm-initiated jets (dashed) and other jets (solid). The CSV lines refer to the first implementation of the CSV tagger used during Run 1.

to train two separate, but strongly correlated, classifiers: one to separate c from light jets (CvsL), and one to separate c from b jets (CvsB). For this tagger, a greater care has been devoted to tuning the light jet rejection as opposed to the b jet one.

In order to achieve simultaneous discrimination from b and light jets, three working point are by cutting simultaneously on the two discriminators. The working points have been chosen to focus on b jet rejection (loose, L), light jet rejection (tight, T), or a compromise between the two (medium, M). The distribution of the jets in the 2D plane of the discriminators and its 1D projections are shown in Fig. 2.

Being there two background classes, the performance of the tagger can be visualized with a receiver operating characteristic (ROC) plane. To ease the visualization of the performance, such plane has been plotted in Fig. 3 (left) in the form of constant-efficiency (ε_c) contour lines in the (b jet efficiency ε_b, light jet efficiency ε_l) plane. In the same figure on the right a comparison with the current b taggers is presented.

3. Identification of b jets in boosted environments

CMS has a wide physics programme covering boosted topologies, and specific flavour tagging algorithms have been developed for this purpose [5]. The standard CSVv2 discriminator can be applied to larger cone jets (a.k.a. fatjets) to tag the presence of B hadrons in the jet. Another option is to apply the tagger to subjets, obtained with any substructure algorithms, and b tag the subjets. This method is applied when tagging boosted top quark decays. Finally, a dedicated training has been performed for fat jets originating from a boosted resonance decaying into a $b\bar{b}$ pair [7]. The

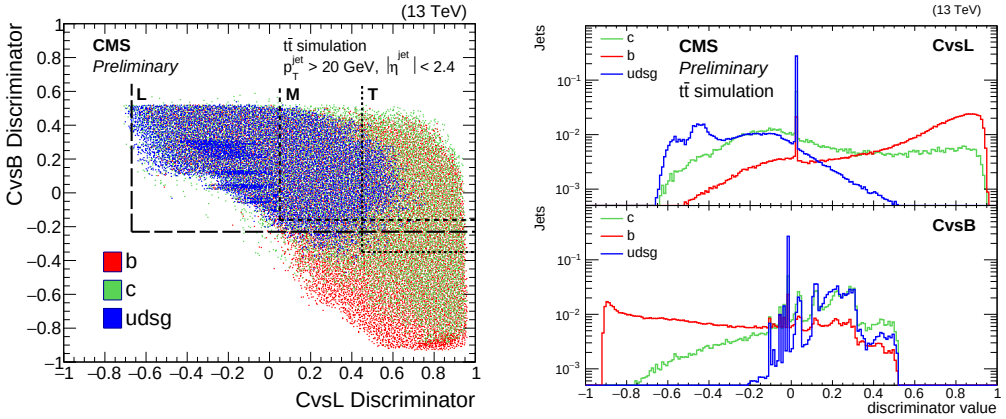

Figure 2: Left: jet distribution in the 2D plane of the discriminators, divided by jet flavour. Right: CvsL (top) and CvsB (bottom) outputs for jets of different flavours

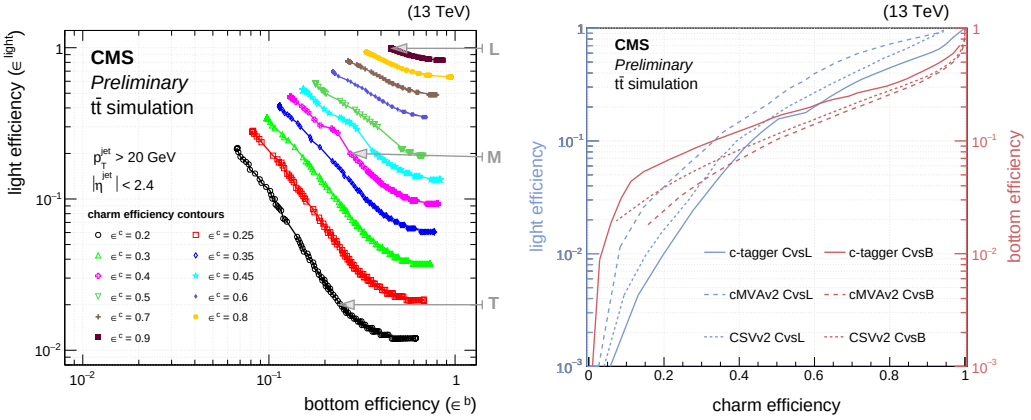

Figure 3: Left: ROC plane for the charm tagger. The plane is represented as constant ε_c contours in the $\varepsilon_b \times \varepsilon_l$ plane. Right: ε_l (blue lines and axis) and ε_b (red lines and axis) as function of the charm efficiency for the dedicated charm tagger training (solid line), CSVv2 (short dashes) and cMVAv2 (long dashes)

need for such training originated from the fact the none of the previous two methods is superior to the other in all the resonance p_T range. The better performance of the tagger is shown in Fig. 4.

4. Performance measurement on data

The performance of the taggers observed in data is used to correct the one from simulated Monte Carlo events by computing scale factors: $SF_f = \frac{\varepsilon_{DATA}}{\varepsilon_{MC}}$. Scale factors are computed as a function of the jet flavour and jet p_T.

The efficiency on b jets is computed both in a multi-jet sample and in $t\bar{t}$ events as described in Refs. [2] and [5]. The results from the different methods are combined with the BLUE method [8] and shown in Fig. 5 (left) for the first 7.7 fb^{-1} of data collected in 2016.

The mistagging efficiency on light jets is computed on multi-jet QCD events with the "negative tags" method described in Refs. [2] and [5].

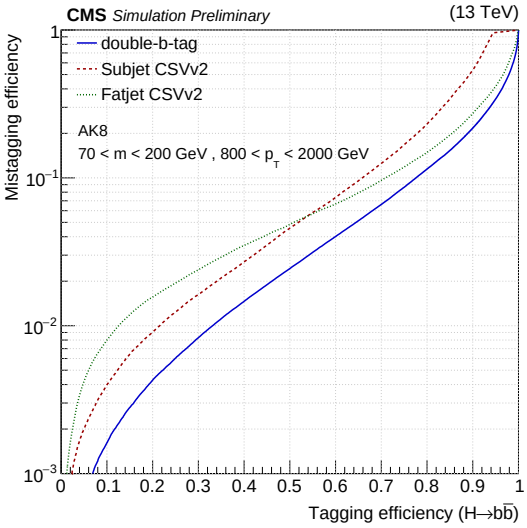

Figure 4: ROC curves for fatjet b tagging (green dotted line), subjet b tagging (red dashed line) and double-b tagging (blue solid line) for a resonance of mass between 70 and 200 GeV and p_T between 800 GeV and 2 TeV decaying into a $b\bar{b}$ pair. The mistagging is computed on QCD multi-jet fatjets.

The measurement of the efficiency of the newly developed charm tagger required the development of new techniques to enrich the data with charm jets. Two methods were devised to carry out such measurement [6].

The first focuses on a $W + c$ sample. The c jet is identified by the presence of a muon with opposite charge to the lepton from the W boson decay among the jet constituents, as at tree-level the signal is produced only with opposing charges. Events with same sign leptons are used for background subtraction, producing a very clean sample of charm jets on which it is possible to measure the tagger efficiency directly.

The second method relies on semileptonic $t\bar{t}$ events, in which $\sim 25\%$ of the partons from the W decay are charm quarks. The peculiar kinematic of the double weak decay of the top quark makes the up-type W products on average more energetic than their down-type counterparts. In each event a single, fully-reconstructed, $t\bar{t}$ candidate is selected and a likelihood discriminant is built to separate the signal form the backgrounds. The events are categorized according to which of the jets associated to the W decay is tagged (none, leading, subleading, both). The different contribution of the charm quarks to the leading and subleading jets is leveraged by a simultaneous fit of the likelihood discriminant in these four categories to infer the value for the tagging efficiency scale factor. The results from these two methods are combined with the same method used for the b jet efficiency and are shown in Fig. 5 (right) for the 2.6 fb^{-1} of data collected in 2015.

5. Conclusion

The physics programme that the CMS Collaboration is carrying out on a wide range of subjects is supported by a strong commitment in the development and maintaining of reliable and performant identification tools. Among those, heavy-flavour taggers play a very important role. In

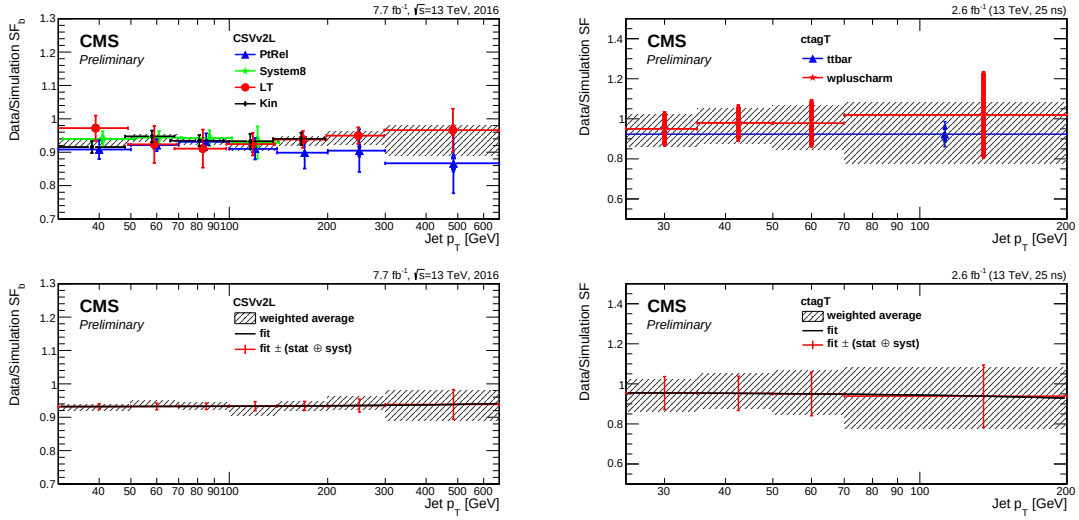

Figure 5: Combination (hatched area) of the different scale factors measurements as function of jet p_T for the CSVv2 loose working point (left) and charm tagging tight working point (right). The bottom pad shows the combination result with a fitting function superimposed.

the beginning of Run 2, the collaboration extended the breadth of the taggers, adding a dedicated double-b tagger for boosted topologies, a new, more performing, b tagger and a charm tagger. The performance of these algorithms has been measured on multiple final states to ensure the best possible precision for the final physics usage. The introduction of the charm tagger required the implementation of dedicated efficiency measurement methods, one of which has been developed for the first time.

References

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

[2] S. Chatrchyan *et al.* [CMS Collaboration], "Identification of b-quark jets with the CMS experiment," JINST **8** (2013) P04013 doi:10.1088/1748-0221/8/04/P04013 [arXiv:1211.4462 [hep-ex]].

[3] W. Waltenberger, "Adaptive vertex reconstruction," CERN-CMS-NOTE-2008-033.

[4] V. Khachatryan *et al.* [CMS Collaboration], "Measurement of $B\bar{B}$ Angular Correlations based on Secondary Vertex Reconstruction at $\sqrt{s} = 7$ TeV," JHEP **1103** (2011) 136 doi:10.1007/JHEP03(2011)136 [arXiv:1102.3194 [hep-ex]].

[5] CMS Collaboration, "Identification of b quark jets at the CMS Experiment in the LHC Run 2," CMS-PAS-BTV-15-001.

[6] CMS Collaboration, "Identification of c-quark jets at the CMS experiment," CMS-PAS-BTV-16-001.

[7] CMS Collaboration, "Identification of double-b quark jets in boosted event topologies," CMS-PAS-BTV-15-002.

[8] 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