PROCEEDINGS OF SCIENCE

Heavy Flavour Jet Identification with the CMS Experiment in Run 2

Andrzej Novak* **for the CMS Collaboration**
RWTH Aachen
E-mail: andrzej.novak@cern.ch

Identification of bottom and charm quarks is crucial for most physics analyses at the CMS Experiment. Advancement and proliferation of deep learning techniques as well as hardware developments have facilitated their use in high energy physics and CMS is successfully employing them to classify jets originating from bottom quarks with unprecedented performance. Furthermore, the improvements have been sufficient to begin meaningfully identifying charm jets as well.

European Physical Society Conference on High Energy Physics - EPS-HEP2019 -
10-17 July, 2019
Ghent, Belgium

*Speaker.

PoS(EPS-HEP2019)146

1. Introduction - Jets at CMS

Identification of bottom and charm quarks at CMS [1] is important for most physics analyses, either directly or as means of background rejection for analyses with leptonic final states. Quarks themselves cannot exist freely. However, through the process of hadronization, they form more stable particles with integer electric charge.

Bottom and charm quarks produce B and D hadrons respectively, which have a measurable lifetime (flight distance) and mass. These hadrons undergo further fragmentation, creating particle showers, which are reconstructed as jets within the CMS detector.

The jet reconstruction at CMS uses the anti-kT algorithm [2] with cone size (ΔR) 0.4 for regular jets (AK4) and 0.8 for so-called "fat jets" (AK8). The decay of a hadron forms a secondary vertex (SV) and a set of tracks, displaced from the primary collision point (PV). The displacement is characterized by the impact parameter (IP). Soft leptons can also be produced in the particle shower. This provides three handles on flavour identification: secondary vertex information, displaced track information, and soft lepton information.

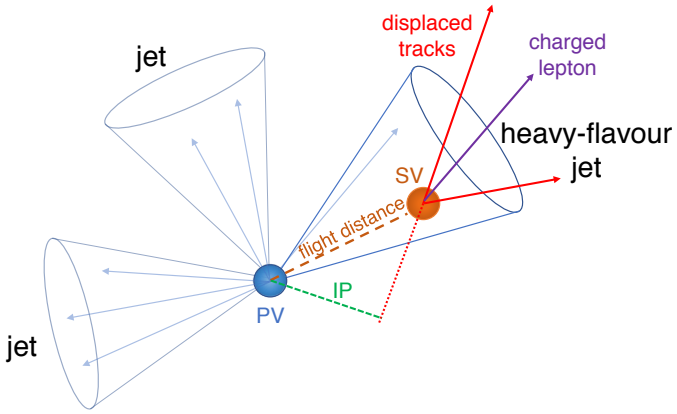

Figure 1: Illustration of decay with a secondary vertex with a reconstructed jet, showing various features of interest [3].

2. Neural Networks and Deep Learning

Neural networks (NN) were used at CMS for b-tagging since the beginning of Run 2 (13TeV).

1. *CSVv2* (Combined-SV-version2) is a shallow NN with SV and track inputs, hence combined. Successfully deployed in many CMS 2016/2017 analyses.

2. *cMVAv2* is a boosted decision tree (BDT) combining other classifiers based on soft leptons and hadron lifetime with *CSVv2*.

3. *DeepCSV* is a multi-classifier dense NN of 5 fully connected layers of 100 nodes each, trained with the same inputs as *CSVv2*. It provides 3 probability scores as an output for bottom, charm and light flavour.

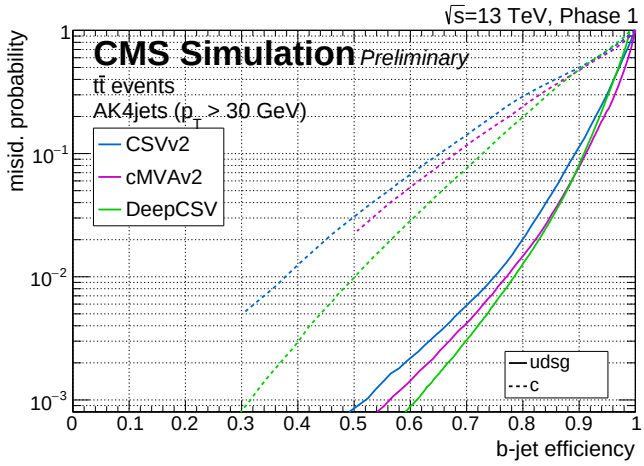

Figure 2: Performance of b-tagging algorithms with the phase-1 CMS detector [4].

The *DeepCSV* classifier played an instrumental role in the observation of Higgs boson decay to bottom quarks by the CMS experiment in 2018 [5] and it remains the standard in use by many analyses.

3. Convolutional and Recurrent Layers

Dramatic improvement was achieved by incorporating deep learning advances from other fields like image recognition (convolutional) and natural language processing (recurrent). The *DeepJet* NN [6] uses 1D Convolutional and LSTM layers to build abstract features from low-level information for three input collections:

1. Secondary vertices
2. Charged particles (tracks)
3. Neutral particles

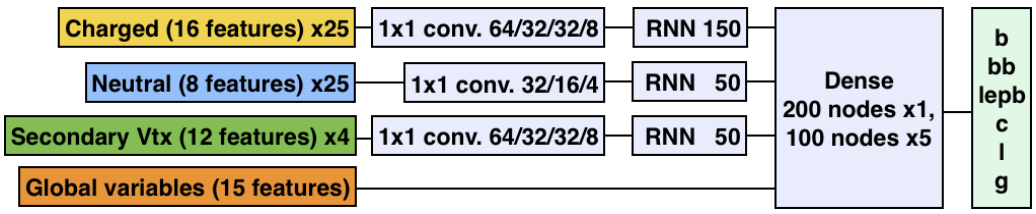

Figure 3: *DeepJet* network architecture [6].

The outputs are combined with global jet features in several dense layers to give a final probabilities for each jet flavor. Scale Factors (SF), adjusting for MC mismodeling are derived with two independent methods in different topologies - di-lepton top quark pairs (Kin) and multi-jet (LTSV) - for three working points [3]. These are labeled loose, medium, and tight, corresponding to misidentification probabilities of 10%, 1% and 0.1% respectively.

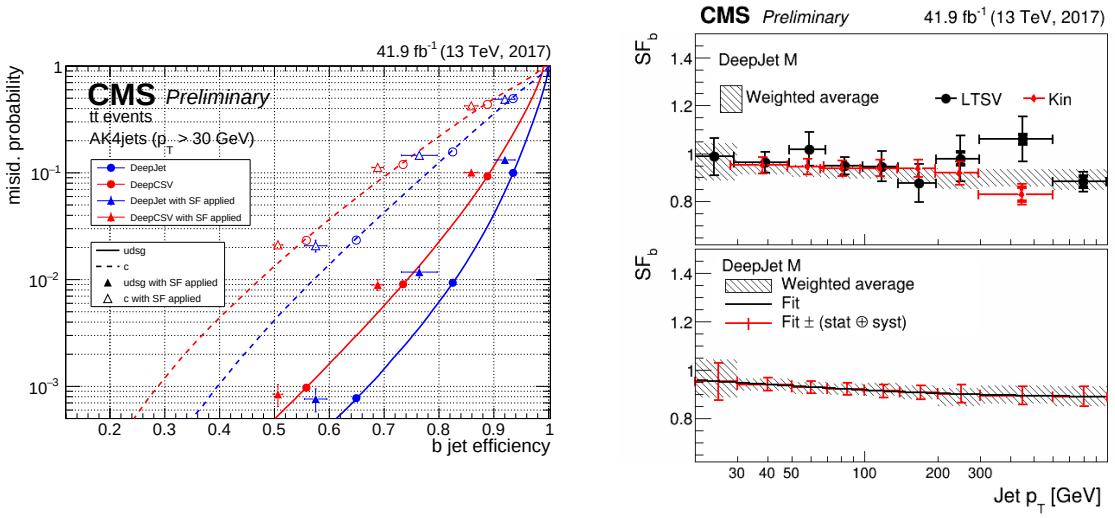

Figure 4: Performance of the *DeepJet* classifier with the phase-1 CMS detector [6] (Left) ROC curves are overlaid with WPs, as well as the efficiencies of these WPs, after being adjusted by measured SFs. (Right) Scale Factors measured for medium WP as a function of jet momentum.

4. Charm Identification

Identification of charm jets is inherently difficult due to their "intermediate" nature compared to bottom and light jets as can be seen for example from the 2D $\frac{IP}{\sigma}$ distribution in Fig. 4 (right) [3], which is one of the input features for both *DeepCSV* and *DeepJet*. Nevertheless, it is possible to define new discriminators from the multi-classifier scores for identifying charm vs. light (CvL) and charm vs. bottom (CvB) flavour jets.

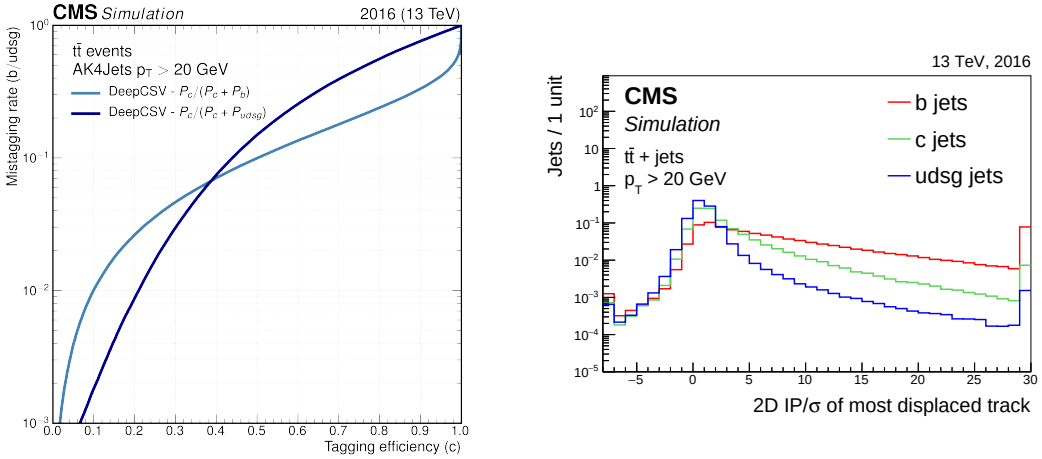

Figure 5: (Left) Performance of the *DeepCSV* classifier distinguishing charm jets from bottom and light jets. (Right) 2D $\frac{IP}{\sigma}$ of the most displaced track for different jet flavours.

The performance of *DeepCSV* proves itself to be sufficient to drive a competitive improvement in exclusion limit of the Higgs to charm branching ratio in the vector boson associated production mode down to 37xSM expectation [7].

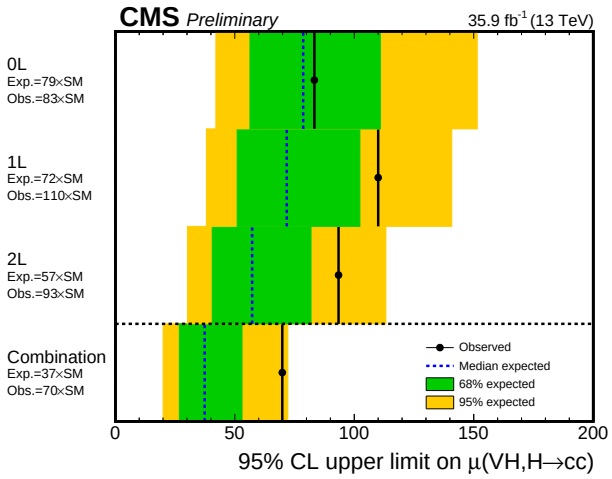

Figure 6: 95% and 68% confidence level upper limits on μ for the $VH(H \to cc)$ process in the first CMS search for Higgs to charm decay [7].

5. Boosted Topologies

Identification of quarks produced at high transverse momenta is difficult as it results in collimated or overlapping decay products. For that reason, dedicated classifiers are used, taking into account the specific topology. The *double-b* is a BDT trained to identify boosted resonance decays into pairs of bottom quarks, using tracking and vertexing variables reconstructed along two subjettiness axes [3]. The *DeepDoubleX* (BvL, CvL, CvB) [8] is a set of three dedicated trainings for distinguishing pairs of bottom, charm, and light quarks from each other. It improves upon the *double-b* approach by combining the same jet level inputs with low-level information using a similar NN structure as *DeepJet* for a major performance gain.

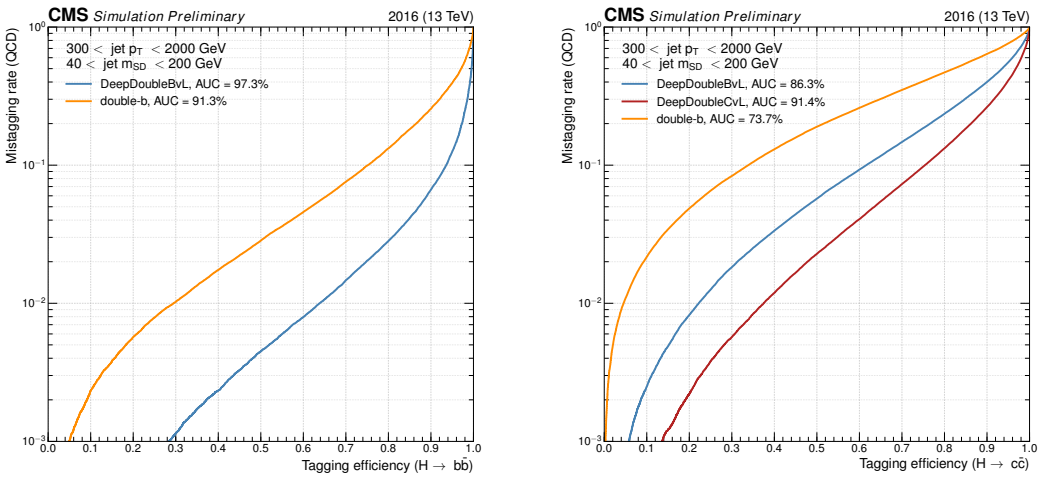

Figure 7: (Left) Performance of the *double-b* and *DeepDoubleX* classifier distinguishing bottom from light jets. (Right) Performance of the dedicated *DeepDoubleX* classifier distinguishing charm from light jets. Performance of *double-b* and *DeepDoubleX* (BvL configuration) for the same is shown as a reference.

References

[1] CMS collaboration, *The CMS Experiment at the CERN LHC*, *JINST* **3** (2008) S08004.

[2] M. Cacciari, G. P. Salam and G. Soyez, *The anti-ktjet clustering algorithm*, *Journal of High Energy Physics* **2008** (2008) 063.

[3] CMS collaboration, *Identification of heavy-flavour jets with the CMS detector in pp collisions at 13 TeV*, *Journal of Instrumentation* **13** (2018) P05011.

[4] CMS collaboration, *CMS Phase 1 heavy flavour identification performance and developments*, CMS-DP-2017-013, May, 2017.

[5] CMS collaboration, *Observation of higgs boson decay to bottom quarks*, *Phys. Rev. Lett.* **121** (2018) 121801.

[6] CMS collaboration, *Performance of the DeepJet b tagging algorithm using 41.9/fb of data from proton-proton collisions at 13TeV with Phase 1 CMS detector*, CMS-DP-2018-058, Nov, 2018.

[7] CMS collaboration, *Search for the standard model Higgs boson decaying to charm quarks*, Tech. Rep. CMS-PAS-HIG-18-031, CERN, Geneva, 2019.

[8] CMS collaboration, *Performance of Deep Tagging Algorithms for Boosted Double Quark Jet Topology in Proton-Proton Collisions at 13 TeV with the Phase-0 CMS Detector*, CMS-DP-2018-046, Jul, 2018.