PROCEEDINGS OF SCIENCE

Heavy flavor jet tagging in CMS

Soureek Mitra,
on behalf of the CMS collaboration

Institut für Experimentelle Teilchenphysik (ETP), Karlsruhe Institute für Technologie (KIT),
Wolfgang-Gaede-Straße 1, 76131, Karlsruhe, Germany

E-mail: soureek.mitra@kit.edu

Identification of hadronic jets originating from heavy flavor quarks in the final state is extremely important to study the properties of the top quark and the Higgs boson, along with various searches for signatures of new physics beyond the standard model. The latest developments in the identification algorithms based on deep learning methods make it an interesting topic also from a technical perspective. In this article, a summary of various identification algorithms along with their performance in simulation and pp collision data, in boosted and resolved topologies, will be presented.

PoS(ICHEP2022)945

41st International Conference on High Energy physics - ICHEP2022
6-13 July, 2022
Bologna, Italy

© Copyright owned by the author(s) under the terms of the Creative Commons
Attribution-NonCommercial-NoDerivatives 4.0 International License (CC BY-NC-ND 4.0).

1. Introduction

Hadronic jets originating from heavy flavor (b/c) quarks arise often in the studies involving the top quark and the Higgs boson. Identification of these jets is therefore extremely important to determine the properties and interactions of the Higgs bosons and the top quarks with ultimate precision and compare them with the predictions obtained from the standard model (SM) of elementary particle physics. In addition, they also play a crucial role to search for signatures of new physics beyond the SM at the LHC. Heavy flavor (HF) jets have the following distinct features that are exploited for their identification.

- Contains secondary vertices (SV) significantly away from the primary vertex (PV) due to decay of b(c)-hadrons with large mass ≈ 5.3 (1.9) GeV and long lifetime ≈ 1.5 (1.0) ps in their rest frame.

- High track multiplicity within jets with high impact paramater (IP) relative to the PV.

- Presence of soft leptons within the jets due to significant semileptonic decay probabilities of the b(c)-hadrons, e.g., $\mathcal{B}(B^- \to \mu^- X) \approx 20\%$.

The identification of HF jets usually depends on the tracking information, such as track IPs, multiplicity etc.; the SV information, such as the SV invariant mass, flight distances relative to the PV etc.; and the properties of the charged and neutral hadron and soft lepton candidates reconstructed using particle-flow (PF) algorithm [1]. In addition, some combination of the these are also used.

2. Tagging algorithms

- **DeepCSV** is a multi-classification deep-neural-network (DNN) algorithm [2] to distinctly identify b-, c- or light (udsg) jets, based on secondary vertex information obtained with the Inclusive Vertex Finder (IVF) algorithm and track-based lifetime information, that are then fed into a fully-connected DNN with 5 hidden layers (i.e. 7 layers altogether) of a width of 100 nodes each. It has 4 output nodes, namely, b, bb, c and light (udsg); that assign a probability to a jet for each category.

- **DeepJet** is a multi-classification DNN algorithm [3] with a more complex architecture compared to DeepCSV, replacing the track-based lifetime information used in DeepCSV with more general (low-level) properties of several charged and neutral PF jet constituents, supplemented also with properties of secondary vertices (using the IVF algorithm) associated with a jet. For each collection of charged and neutral particles and vertices, separate 1×1 convolutional layers are trained with different levels of filters acting on each particle (charged and neutral) or vertex individually. The output of the convolutional layers is further propagated to a collection of recurrent layers (LSTMs). The outputs of the LSTMs are merged with global jet properties and propagated through one dense layer with 200 nodes, followed by 7 hidden dense layers with 100 nodes each. It has 6 output nodes, namely, *b, bb, leptonic b, c, uds,* and *g*, that assign a probability for each jet to belong to any of these categories.

PoS(ICHEP2022)945

Truth labels are obtained from hadron flavor definition. Jets from pileup vertices are excluded during training that is performed on a mixture of simulated QCD multijet and top pair ($t\bar{t}$) events. The performances are demonstrated via receiver-operator-characteristic (ROC) curves evaluated for reconstructed jets using simulated $t\bar{t}$ events in Figure 1. Three working points are defined based on the light (udsg) misidentification rates; namely, *Loose* (10%), *Medium* (1%), and *Tight* (0.1%). The efficiency (ε) of a jet of flavor f in Monte Carlo (MC) simulation and data are defined as:

$$\varepsilon_f^{MC} = \frac{N_f^{Tagged}}{N_f^{Total}}, \ \varepsilon_f^{Data} = SF_f \times \varepsilon_f^{MC} \tag{1}$$

where, N_f^{Tagged}, N_f^{Total}, and SF_f represent the number of tagged jets, the number of total jets, and the calibration scale factor for the jet flavor f, respectively.

3. Performance

3.1 Resolved jets

The calibration of the b-, c-, and light-jets with cone size 0.4 are performed using various methods based on QCD multijet, $t\bar{t}$, Drell-Yan (DY), and W + c events [4]. Calibration SFs are estimated at different working points as well as for the entire discriminant shape. Calibration for c-tagging discriminant [5] is performed simultaneously in c-enriched (W + c), b-enriched ($t\bar{t}\,\ell$ + jets), and light-enriched (DY, QCD multijet) regions. Figure 2 shows the calibration with data using different methods.

3.2 Trigger

For Run3, a dedicated training is performed online, i.e., with the raw inputs for the high-level triggers (HLTs) [7]. The online training model shows better performance in simulated events relative to the offline training model evaluated on the HLT-level inputs as shown in Figure 1. This online training model has been deployed for recording data during Run3.

3.3 Boosted jets

A heavy resonance, X, when lorentz-boosted ($\mathrm{p_T^X} \gg \mathrm{m}_X$), can decay into a pair of HF quarks with a very small opening angle ($\approx \frac{2\mathrm{m_X}}{\mathrm{p_T^X}}$). In such a case, the two daughter quarks cannot be reconstructed separately into small-radius jets and hence are usually merged into a single large-radius jet due to the heavy resonance X. The state-of-the-art tagging algorithms, namely, the **double-b** and the **DeepDoubleX** taggers [6] utilize several properties of the boosted jet as inputs, such as correlations between the flight directions of the b-quarks, N-subjettiness etc. A comparison of their performance is shown in Figure 3 by studying the ROC curves obtained from a combined sample of QCD multijet and $\mathrm{H} \rightarrow \mathrm{b\bar{b}(c\bar{c})}$ events.

4. Summary

This report contains a summary of the state-of-the-art algorithms for HF jet tagging and their performance in simulation as well as in data in various event topologies. These algorithms are trained online with HLT-level inputs in simulated events and deployed for data recording during Run3. Further developments towards higher performance in simulation as well as in data with advanced neural-network architectures are in full swing.

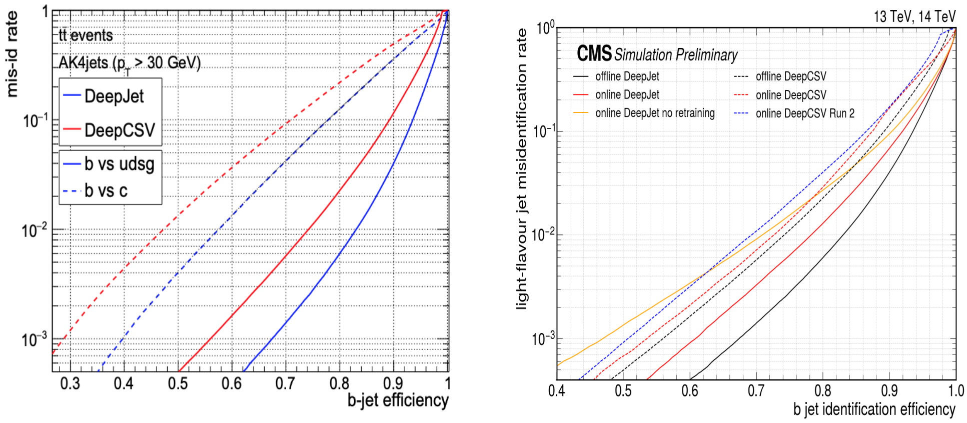

Figure 1: The performances of DeepCSV and DeepJet algorithms in simulated $t\bar{t}$ events is shown for jets with cone size 0.4 reconstructed offline (left) and at HLT (right).

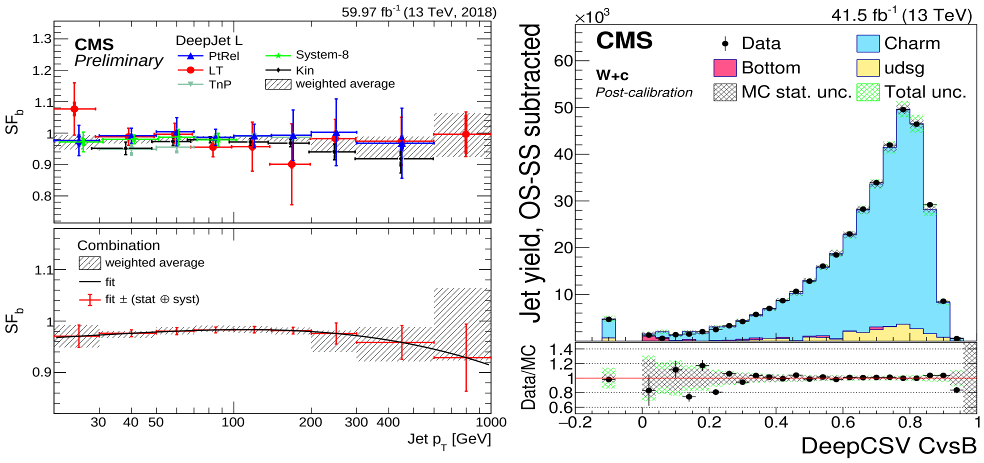

Figure 2: The calibration of the DeepJet b- and c-tagging discriminants in data events are shown for jets with cone size 0.4. The SF_b is evaluated at the *Loose* working point (left) using various methods in QCD multijet and $t\bar{t}$ events. The entire shape of the *CvsB* discriminant is calibrated to data (right) in the W + c events.

PoS(ICHEP2022)945

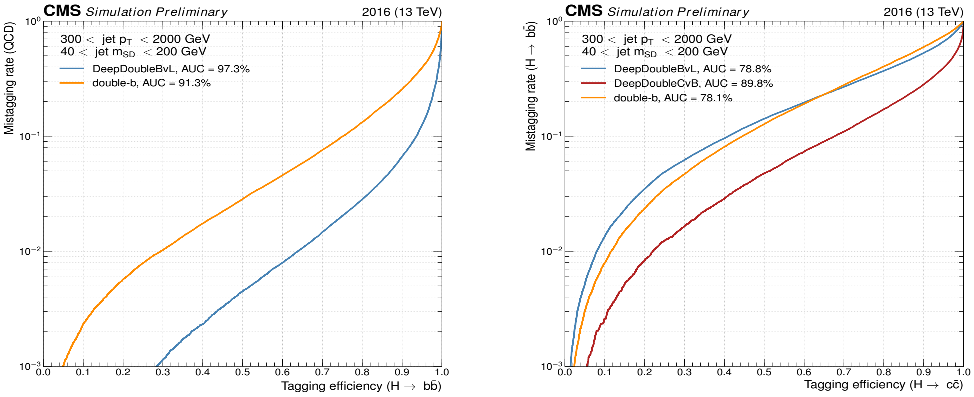

Figure 3: The performances of double-b and DeepDoubleX taggers are shown in simulated events for large-radius jets with cone size 0.8.

References

[1] CMS Collaboration, *Particle-flow reconstruction and global event description with the CMS detector*, JINST 12 (2017) P10003.

[2] CMS Collaboration, *Identification of heavy-flavour jets with the CMS detector in pp collisions at 13 TeV*, JINST 13 (2018) P05011.

[3] Emil Bols et. al., *Jet Flavour Classification Using DeepJet*, JINST 15 (2020) P12012.

[4] CMS Collaboration, *B-tagging performance of the CMS Legacy dataset 2018*, CMS-DP-2021-004.

[5] CMS Collaboration, *A new calibration method for charm jet identification validated with proton-proton collision events at 13 TeV*, JINST 17 (2022) P03014.

[6] 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.

[7] CMS Collaboration, *Expected Performance of Run-3 HLT b-quark jet identification*, CMS-DP-2022-030.

PoS(ICHEP2022)945