PROCEEDINGS OF SCIENCE

Reconstruction and identification of hadronic objects with CMS

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

The reconstruction and identification of hadronic objects has a wide variety of use-cases inside the physics program of the LHC, ranging from Standard Model measurements and Higgs searches up to the more exotic searches of boosted resonances. The CMS collaboration devoted a constant effort in improving such tools in order to extend the analyses reach.
A review of the current status of the identification algorithms for hadronic taus, heavy flavour jets, quark/gluon discrimination, and boosted resonances is presented together with their performance, either measured on data or expected from simulation.
Special attention is devoted to novel techniques recently introduced, and their impact on the final performance on the classification task.

EPS-HEP 2017, European Physical Society conference on High Energy Physics
5-12 July 2017
Venice, Italy

*Speaker.

PoS(EPS-HEP2017)526

1. Introduction

The hadronic object reconstruction in CMS [1] always starts with a jet being clustered and containing the decay products of the object of interest. The clustering algorithm of choice is the anti-k$_T$ algorithm [2] with radius $R = 0.4$ for normal jets and $R = 0.8$ for the identification of boosted object. After the jet clustering the identification can be divided into three main families: the identification of hadronically decaying taus, the jet flavour identification, comprising not only the heavy flavours, but quark/gluon discrimination as well, and the identification of boosted objects such as boosted decays of the top quark ($t \to bq\bar{q}$), of vector bosons ($W/Z \to q\bar{q}$) and generic resonances to a b-quark pair ($X \to b\bar{b}$).

The importance of these tools is shown by the great wealth of analyses exploiting these techniques. Such analyses had a large impact in the HEP community in the last few years and a constant struggle to improve and refine the tagging tools can have a positive impact on a large fraction of the CMS physics program.

2. Hadronic tau identification

Hadronic tau identification still relies on the Hadron Plus Strip (HPS) [3] algorithm which has been extremely successful during the Run I campaign, but few key changes have improved the performance for the Run II data taking. The algorithm still aims to reconstruct the full decay mode of the hadronic tau. The decay photons of the π^0 are clustered into strips in the electromagnetic calorimeter. The tau decay is classified into single prong ($\tau^\pm \to \pi^\pm \nu$), one prong plus strip (covering both $\tau^\pm \to \rho\nu \to \pi^\pm\pi^0\nu$ and $\tau^\pm \to a_1\nu \to \pi^\pm\pi^0\pi^0\nu$), and three prongs ($\tau^\pm \to a_1\nu \to \pi^\pm\pi^\pm\pi^\mp\nu$). The main change in the identification algorithm with respect to Run I is the size of the clustering ECAL strip which is now function of the e/γ p$_T$ to better follow the contour of real taus, as shown by Fig. 1 and detailed in Ref. [4].

The identification of hadronic taus decaying from heavy resonances still relies mainly on the isolation of the decay products. On top of the traditional cut-based approach a new multivariate method combines the isolation with strip shape and lifetime information into a boosted decision tree (BDT) to obtain a better performance as shown in Fig. 1.

The Z boson decays are used as standard candle to measure the efficiency and mis-identification rates with respect to electrons and muons.

3. Jet flavour identification

3.1 Heavy flavours

Several marking features help identifying jets originating from heavy-flavour quarks. As the open-charm and open-bottom hadrons contained in such jets have a significant lifetime, their decay products will create tracks displaced with respect to the main primary vertex (PV), and they will point to the same secondary vertex (SV). Moreover, the large branching fraction of such hadrons to semi-leptonic modes will enhance the presence of soft muons and electrons in the jet. Depending of which set of features is used to tag the jet, the classifiers are divided into *track-based taggers* (e.g. the Jet Probability tagger), using only the track displacement information, *combined taggers* (e.g.

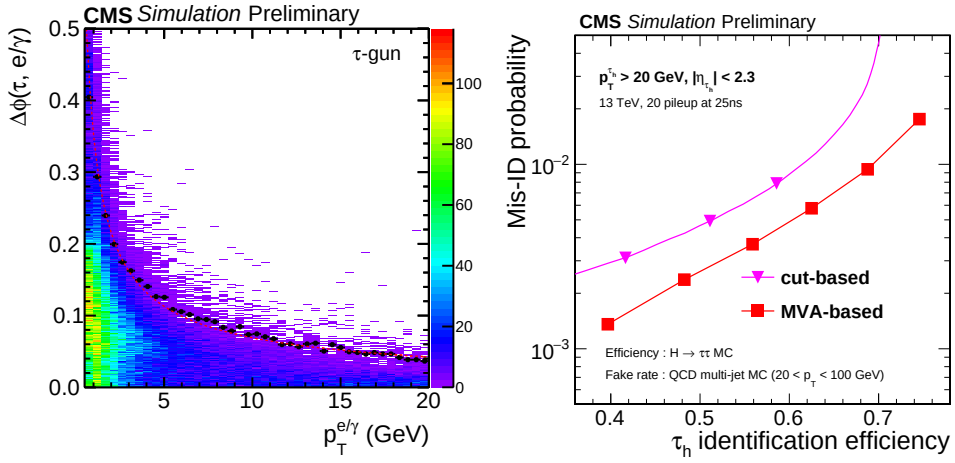

Figure 1: From Ref. [4]. Left: Distance in $\Delta\phi$ between the τ_h and e/γ, that are due to the tau decay products, as a function of e/γ p_T. A sample of simulated τ_h decays is used. The dotted point shows the 95% quantile for the given bin, and the dashed lines represent the fitted function used in the algorithm to determine the strip width. Right: Misidentification probability as a function of τ_h identification efficiency, evaluated using $H \to \tau\tau$ and QCD MC samples. The MVA-based discriminators are compared to that of the isolation sum discriminators. The points correspond to working points of the discriminators.

CSVv2 and DeepCSV), which use track displacement and SV information, and *super-combined taggers* (e.g. cMVAv2), which include the lepton information as well.

The large number of features used to achieve a competitive discrimination power requires the use of more advanced machine learning techniques such as shallow Neural Networks (CSVv2), deep Neural Networks (DeepCSV), and BDTs (cMVAv2). A more detailed description of the features and techniques used is available in Ref. [5] and Ref. [6]. The current best performing tagger is DeepCSV as shown in Fig. 2 by the ROC curves, computed on simulated $t\bar{t}$ events, and the data to simulation efficiency ratios.

DeepCSV uses four output nodes to properly classify jets containing one b-hadron, two b-hadrons, one c-hadron, and light jets. This multi-class approach enables the classifier to be used as a charm tagger as well, showing better performance than the charm tagger currently available in CMS and firstly documented in Ref. [7].

The big performance and flexibility gains obtained using DNNs, sparked a lot of interested on the subject leading to a detailed optimization study on the architecture of the DNN and on the size of the input features. A preliminary result of such optimization is the new DeepFlavour tagger, described in Ref. [8]. DeepFlavour takes as inputs a much larger set of input features (674 vs 68 of DeepCSV), divided into global features and features pertaining the charged candidates, neutral candidates, and secondary vertices. Each group of features, with the exception of the global ones, is passed though a set of convolutional filters that learn a compressed representation of the features that is then passed to a dense architecture similar to DeepCSV. The more advanced architecture and the large increase in input features results in a significant performance gain, even more pronounced at high jet p_T, as shown in Fig. 3.

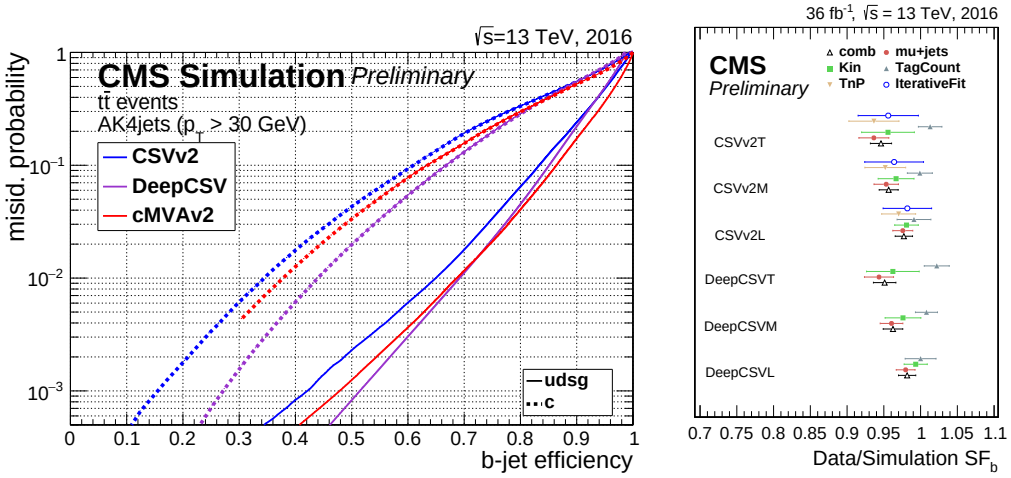

Figure 2: From Ref. [6]. Left: mis-identification probability of light jets (solid lines) and charm jets (dashed lines) as a function of b-jet tagging efficiency for the three best-performing alorithms. Right: data to MC simulation scale factors for b-jets for CSVv2 and DeepCSV, the SF compatibility among the homologous working points of the two different algorithms shows that the performance gain observed on simulation is retained in real data.

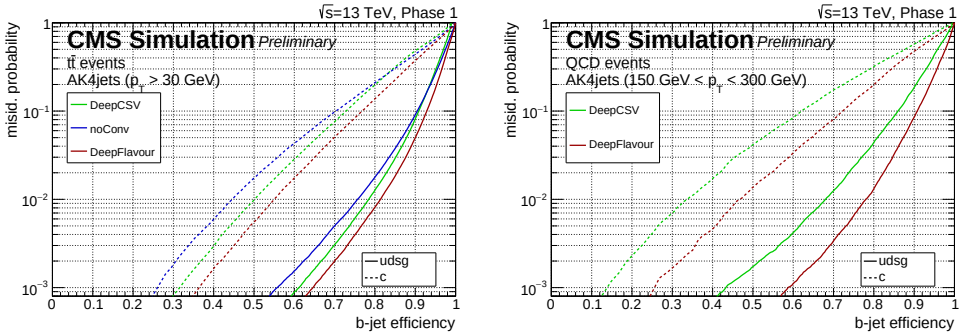

Figure 3: From Ref. [8]. Left: ROC curves for b vs. c jets (dashed) and b vs. light jets (solid) for DeepCSV, DeepFlavour and DeepFlavour without convolutional layers, the performance gain achieved by DeepFlavour cannot be only explained with the additional inputs, but requires the better architecture as well. Right: ROC curves for b vs. c jets (dashed) and b vs. light jets (solid) for DeepCSV (green) and DeepFlavour (red) for jets with 150 GeV $< p_T <$ 300 GeV.

3.2 Quark-gluon jets discrimination

The discrimination between jets stemming from light quarks and from gluons is achieved by combining three features sensitive to the jet particle multiplicity, shape and fragmentation function within a likelihood discriminant as described in Ref. [9]. The performance on data of the classifier is measured in dedicated Z+jets (quark-jets enriched) and dijet (gluon-jets enriched) regions. A comparison of the tagger output in such regions is presented in Fig. 4

4. Boosted objects

With the increasing Lorentz boost of a resonance, its decay products become more and more

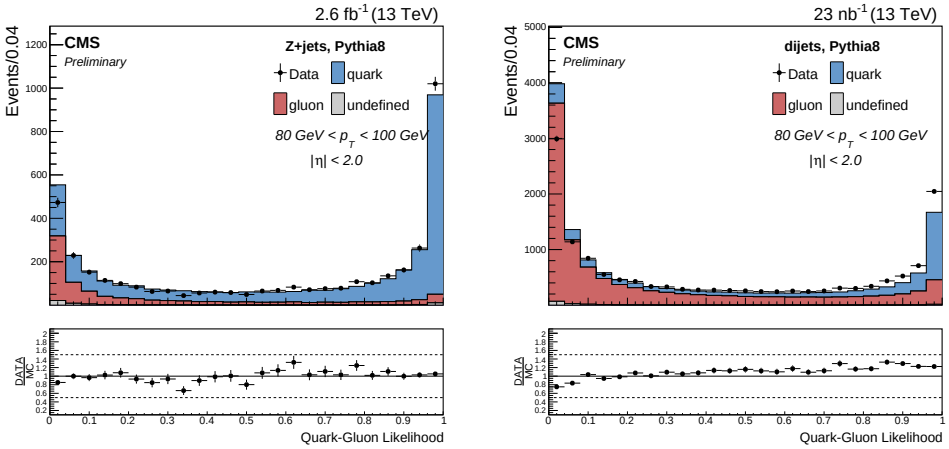

Figure 4: From Ref. [9]. Data-MC comparison for the quark-gluon discriminant in Z+jets (left) and dijet (right) events for jets in the central region with $80 < p_T < 100$ GeV. The data (black markers) are compared to the MADGRAPH/PYTHIA simulation, on which the different components are shown: quarks (blue), gluon (red) and unmatched/pileup (grey).

collimated, up until the point where, in case of hadronic decays, the two or more jets produced cannot be resolved. In such cases a jet with a larger cone is reconstructed, in the attempt of clustering the whole resonance decay.

The identification of boosted resonances relies heavily on techniques aimed at removing the soft collinear radiation like SoftDrop [10, 11] and jet pruning [12], as well as the sub-jettines, defined in Ref. [9]. The identification of boosted top decays additionally exploits the b-tagging of the sub-jets.

Several approaches, mostly cut-based, are explored. One notable approach decorrelates the tagger performance and the jet p_T using the DDT method [13]. This method achieves a flat rejection power as function of jet p_T, which is beneficial to the analyses background estimation. The performance of the different boosted W-tagging working points if shown in Fig. 5.

5. Conclusions

As the reconstruction of hadronic objects is of paramount importance for the collaboration physics program, more and more effort is put to refine and improve the performance and the impact on the analysis of such tools. In this perspective the introduction of modern machine learning techniques is showing a large impact potential in this field.

References

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

[2] M. Cacciari, G. P. Salam and G. Soyez, "The Anti-k(t) jet clustering algorithm", JHEP **0804**, 063 (2008) doi:10.1088/1126-6708/2008/04/063 [arXiv:0802.1189 [hep-ph]].

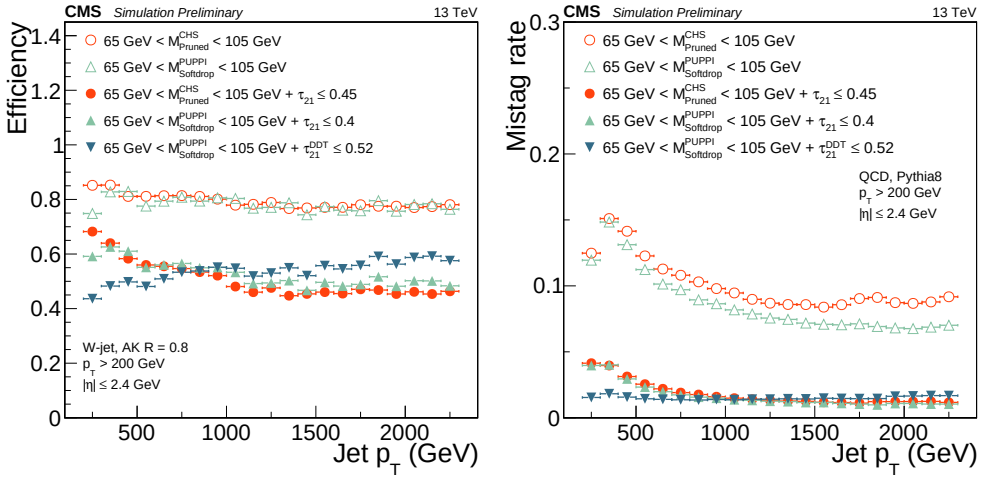

Figure 5: From Ref. [9]. Efficiency (left) and mis-identification rate (right) of the different boosted W-tagging working points. The legend summarizes the selection applied to achieve the classification. The last set of points (blue downward triangles) use the DDT method to obtain a flat rejection factor as a function of the jet p_T.

[3] S. Chatrchyan *et al.* [CMS Collaboration], "Performance of tau-lepton reconstruction and identification in CMS", JINST **7** (2012) P01001 doi:10.1088/1748-0221/7/01/P01001 [arXiv:1109.6034 [physics.ins-det]].

[4] CMS Collaboration, "Performance of reconstruction and identification of tau leptons in their decays to hadrons and tau neutrino in LHC Run-2", CMS-PAS-TAU-16-002.

[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, "Heavy flavor identification at CMS with deep neural networks", CMS-DP-2017-005.

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

[8] CMS Collaboration, "CMS Phase 1 heavy flavour identification performance and developments", CMS-DP-2017-013.

[9] CMS Collaboration, "Jet algorithms performance in 13 TeV data", CMS-PAS-JME-16-003.

[10] M. Dasgupta, A. Fregoso, S. Marzani and G. P. Salam, "Towards an understanding of jet substructure", JHEP **1309**, 029 (2013) doi:10.1007/JHEP09(2013)029 [arXiv:1307.0007 [hep-ph]].

[11] A. J. Larkoski, S. Marzani, G. Soyez and J. Thaler, "Soft Drop", JHEP **1405**, 146 (2014) doi:10.1007/JHEP05(2014)146 [arXiv:1402.2657 [hep-ph]].

[12] S. D. Ellis, C. K. Vermilion and J. R. Walsh, "Recombination Algorithms and Jet Substructure: Pruning as a Tool for Heavy Particle Searches", Phys. Rev. D **81**, 094023 (2010) doi:10.1103/PhysRevD.81.094023 [arXiv:0912.0033 [hep-ph]].

[13] J. Dolen, P. Harris, S. Marzani, S. Rappoccio and N. Tran, "Thinking outside the ROCs: Designing Decorrelated Taggers (DDT) for jet substructure", JHEP **1605**, 156 (2016) doi:10.1007/JHEP05(2016)156