CERN SUMMER STUDENT PROJECT REPORT

Heavy-flavour jet tagging studies

Author: Dagmar Bendová

Supervisors: Lorenzo Sestini
Oscar Augusto de Aguiar Francisco

August 24$^{\text{th}}$, 2018

1 Introduction

Jet studies are crucial for the observation of many QCD processes, e.g. inclusive $b\bar{b}$ production, production and decay of top quark or production of electroweak bosons accompanied by jets. Jets are in general collimated sprays of particles, which are produced in hard QCD processes. The cone-like structure of the jet is dominated by particles produced from one parton, i.e. quark or gluon. We can distinguish between the light-parton jets (or q-jets), which originate in hadronization of a light u,d or s quark or a gluon, and c-jets and b-jets (together denoted as heavy-flavour jets), which originate from the c and b quark, respectively. The identification of jets originating from the hadronization of quarks with different flavours is therefore crucial to differentiate between processes like $Z + (udsg)$-jet and $Z + b$-jet. The precise tracking and vertex reconstruction system of the LHCb experiment makes it well suited for identification of the jet flavours. Well established method for jet tagging, so called secondary vertex tagging, is the reconstruction of a secondary vertex (SV), originating from a decay of heavy hadron, inside the jet. After that a boosted decision tree method is applied to assure more precise discrimination between the flavours. Current tagging method has been validated and optimized for Run I data [1]. However in Run II, there is a need to develop more updated algorithm for jet identification at LHCb due to different conditions between the two data-taking periods. Another approach to jet tagging is to use more sophisticated deep neural networks, which use the information of all of the jets constituents as inputs and therefore, they are not dependent on the SV tagging process.

2 Secondary vertex tagging

Tagging of heavy-flavour jets is performed by requiring that a secondary vertex (SV) is found within the jet. The b-, c- and light-parton jets are defined as the two leading jets in the Monte Carlo simulated samples used in this work. The contamination of light-parton jets is reduced by this requirement on the existence of the SV and also by removing the tracks which are produced directly in primary vertex (PV) of the collision. In this process, all possible 2-body secondary vertices which satisfy a given vertex requirements are reconstructed inside the event. For each jet, all the 2-body SVs with tracks fulfilling the condition on the radial distance from the jet axis $\Delta R < 0.5$ are linked to form n-body SVs. After this, all the SVs in the jet, which share tracks, are merged. In case that multiple SVs are found in the jet, the one with the highest transverse momentum p_T is chosen.

The efficiency of the SV-tagging method can be calculated by dividing the number of jets with identified SV by the total number of jets. The efficiencies for all the three categories of MC simulated jets can be seen in Fig. 1 for the dependence on the jet p_T and in Fig. 2 for the dependence on the jet pseudorapidity η. The p_T and η bins were chosen to match the size of bins in efficiency plots of [1]. Also the cut on jet pseudorapidity $2.2 < \eta < 4.2$, in case of p_T-dependent efficiencies, and the cut on jet transverse momentum $p_T > 20$ GeV, in case of η-dependent efficiencies, were applied in accordance with [1]. As can be seen from the plots, the efficiency dependent on p_T for b-jets varies from 25.6% in the lowest p_T bin up to 55.3%. A drop of $\approx 5\%$ is observed for b-jets with $p_T > 50$ GeV when compared to the efficiency curve obtained for the Run I in [1]. This drop in efficiency was already observed in the analysis [2]. For c-jets we observe efficiency varying from approx. 8.7% to approx. 19% and again, the drop for high p_T is observed. The misidentification of light-parton jets varies from 0.86% to approx. 3.15%. The efficiency dependent on η varies from approximately 37.5% to 53.9% for b-jets. from approx. 12.6% to 19% for c-jets and the misidentification is from approx. 1.63% to 1.95%. The overall light-parton jets misidentification probability calculated in [1] is around 0.1% [1] for low-p_T jets and increases to approx. 1% for high-p_T jets, which is very close to the result of this work.

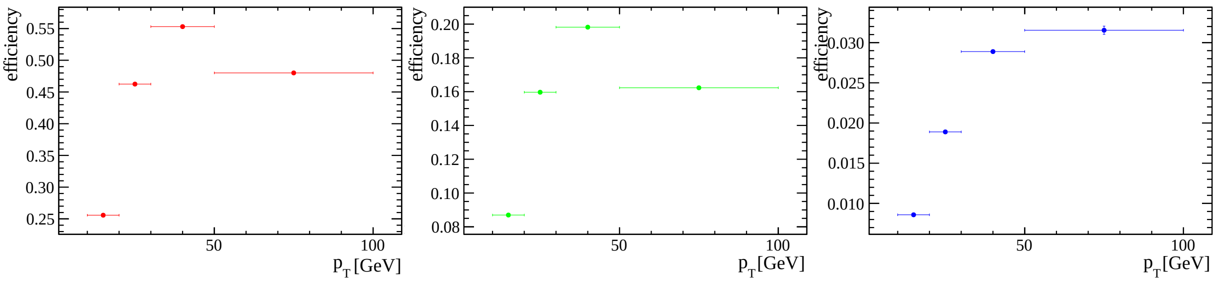

Fig. 1: The efficiency of the SV-tagging method dependent on the jet transverse momentum p_T [MeV]. In the most left plot, the efficiency for b-jets is depicted, in central plot, the efficiency for c-jets is shown and in the most right plot, the misidentification of light-parton jets is depicted.

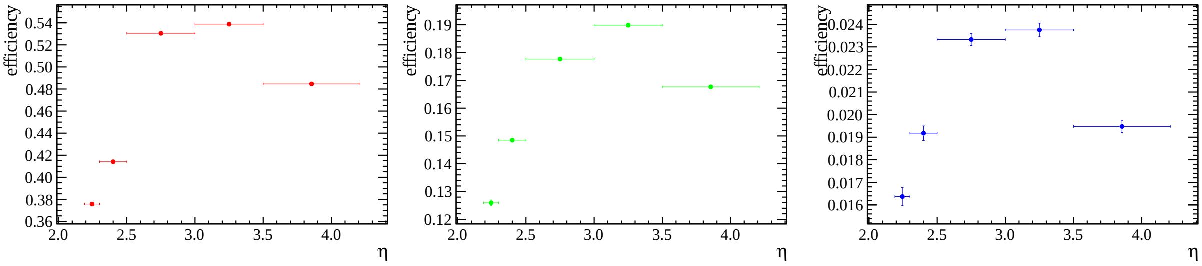

Fig. 2: The efficiency of the SV-tagging method dependent on the jet pseudorapidity η. In the most left plot, the efficiency for b-jets is depicted, in central plot, the efficiency for c-jets is shown and in the most right plot, the misidentification of light-parton jets is depicted.

3 Jet tagging using boosted decision trees

Although the requirement on the presence of the secondary vertex inside the jet highly suppresses the amount of background formed by light-parton jets, some of them may also contain a SV caused by the reconstruction of a fake heavy meson secondary vertex inside the light-parton jet and they can be subsequently be misidentified for the heavy-flavour jet. Therefore a boosted decision tree method is applied to further discriminate the light-parton jets and also to discriminate between the (b, c)-jets.

A decision tree is a consecutive set of questions (so called nodes) with only two possible answers which separates the signal from background in the sample after a given maximum number of nodes was reached. BDT method uses observables related to the secondary vertex inside the jet as an input, e.g. p_T of the SV, corrected mass of the SV, flight distance of the SV, etc.

In Fig. 3 we can see the comparison of two BDT discriminators in b-jets, c-jets and light-parton jets Monte Carlo samples. The BDT0 method (left plot) discriminates between the light-parton and heavy-flavour jets. As can be seen, the discrimination of light-parton jets is very poor against the heavy flavours. Although the light-parton jets were already reduced in the SV-tagging process, there is still some amount of them remaining and it is difficult to discriminate them from heavy-flavour jets. The lack of separation can also be a subsequence of the definition of b, c and light-parton jets as the two leading jets, which may result in the pollution from other the other types in the samples. The misidentification of light-parton jets as heavy-flavour jets is clearly present also in the BDT1 (right plot) result. The BDT1 method discriminates between the b-jets and c-jets, which appear to be satisfying, although there is some overlap between the two flavours caused by misidentification. This current BDT method was trained and tested on Run I samples. From the unsatisfying separation of BDT0, we conclude that there is a need to retrain the BDT method with Run II samples to take into account different conditions between the two runs.

Motivated by this result, I tried to perform the training of a new BDT discriminator. Training was

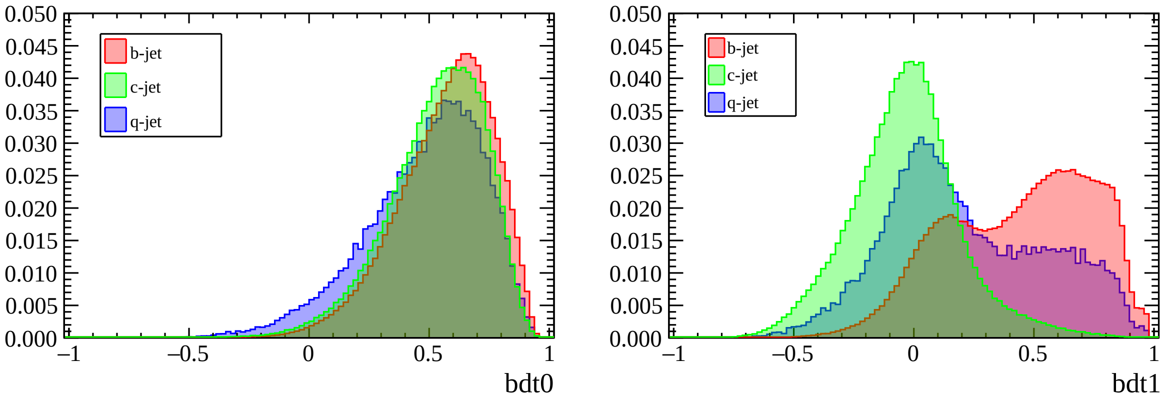

Fig. 3: The discrimination between the light-parton and heavy-flavour jets (left) and the discrimination between the *b*-jets and *c*-jets (right).

performed using the TMVA package [4] with Run II Monte Carlo samples using the same variables and same SV reconstruction procedure as in Run I [1]. For this purpose, the *b*-jets samples were taken as a set of signal events and light-parton jets samples were used as background. The presence of the SV was required in both signal and background, to obtain the discrimination between the heavy-flavour and misidentified light-parton jet. The list of variables used for the BDT training is following:

- the SV mass M
- the SV corrected mass M_{cor}
- the transverse flight distance of the two-track SV closest to the PV
- the fraction of the jet p_T carried by the SV
- ΔR between the SV flight direction and the jet
- the net charge of the tracks that form the SV
- the number of tracks in the SV
- the number of SV tracks with $\Delta R < 0.5$ relative to the jet axis
- the flight distance χ^2
- the sum of all SV tracks χ^2_{IP}

The distributions in the MC samples for these variables are shown in Appendix. The result of the training for several choices of BDT methods can be seen in Fig. 4 and Fig. 5. As can be seen, none of the available BDT algorithm provides satisfactory result with the upper set of conditions. This result is also apparent from a ROC curve depicted in Fig. 6. Although all the BDT methods show slightly improved signal efficiency vs background rejection when compared to simple rectangular cuts method, this preliminary result is still not satisfactory enough.

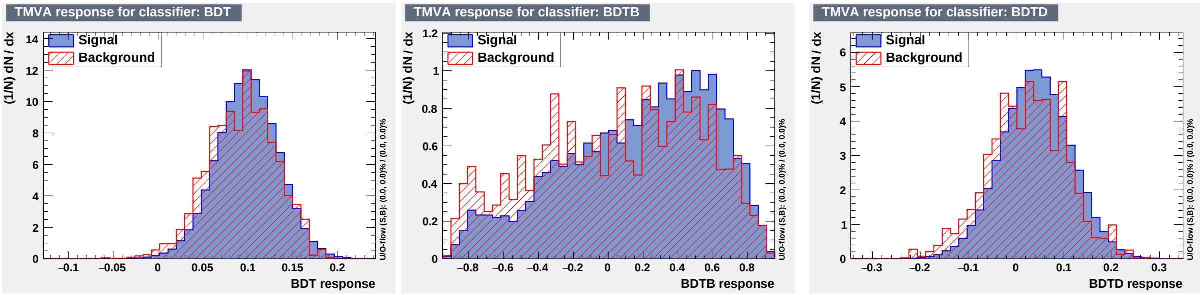

Fig. 4: The output of a new BDT discriminator training for various BDT methods – from left: Adaptive Boost, Bagging and decorrelation method combined with the Adaptive Boost [4].

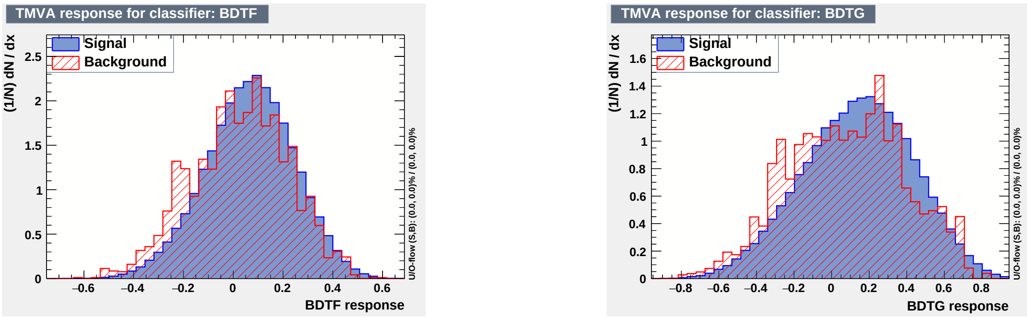

Fig. 5: The output of a new BDT discriminator training for two various BDT methods – the method with a Fisher discriminator allowed for mode splitting in the left plot and Gradient Boost in the right plot.

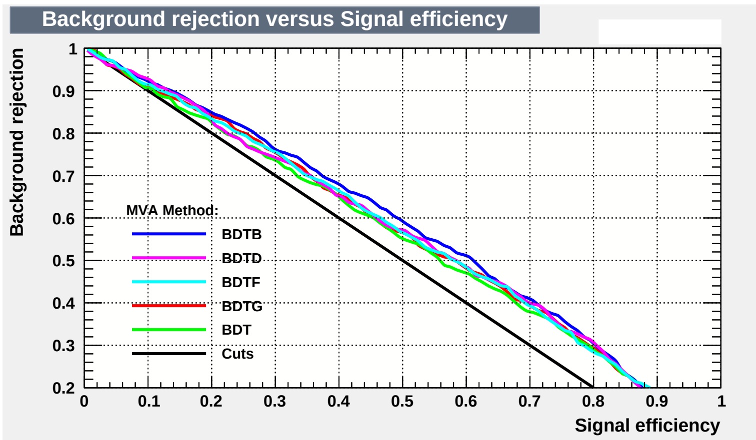

Fig. 6: ROC curve for various BDT methods compared to simple Rectangular cuts method.

4 Jet tagging using deep neural networks

Deep neural network is a multi-layer fully-connected neural net which consist of an input layer, several hidden layers and an output layer, each of the layers consists of a given number of nodes. The main difference when compared to the BDT method is that every node in each layer is connected to all of the nodes of the previous and also in the subsequent layer. DNN processes the given sample in the following way: a given node in the first hidden layer takes the weighted sum of it inputs and passes it through a non-linear activation function. This output of the node then becomes the input of the following second hidden layer. Final decision is obtained by performing these steps on all layers and the weights associated with all the connections between the nodes are learned in the process [5]. The deep neural network used for the purpose of jet tagging takes all the jet constituents like p_T of particles, charge of particles, etc. as inputs and it is trained on all MC jets, unlike BDT which is trained only on the jets with a tagged SV inside them.

The output of DNN method applied to MC jets is the probability of being a jet of a given flavour. These probabilities for b-jets, c-jets and light-parton jets can be seen in Fig. 7. The b-jets are clearly discriminated from other flavours. However c-jets and light-parton jets are mostly mixed and therefore we can not clearly discriminate between them with the current state of this method.

The efficiencies of the DNN-tagging method for individual jet flavours were calculated in such manner, so that they can be compared to SV-tagger efficiencies. The probability of being a light-parton jet P_q was used as a discriminative variable between the "selected" and "total" samples. The cut on $P_q < x$ where $x \in (0, 1)$ was applied in each of the p_T and η bins of the light-parton jet MC samples. By precisly tuning this parameter, the difference between the SV-tagger and DNN-tagging misidentification of light-parton jets is lower than 0.02 % in all of the p_T and η bins. The obtained DNN-tagging misidentification for light-jet sample can be seen in the rightmost plot of Fig. 8 and Fig. 9. It varies from approx. 0.86 % to 3.16 % for p_T-dependence and 1.63 % – 1.93 % for η-dependence. The efficiency for b-jets varies from approx. 25.9 % to 56.1 % for p_T dependence and approx. 37.8 % – 54.3 % for η-dependence. And for c-jets, the efficiencies vary from from approx. 8.1 % to 21.9 % for p_T dependence and approx. 12.3 % – 19.8 % for η-dependence. In accordance with the SV-tagger efficiencies in Fig. 1, there is apparent drop in the last p_T bins for both b-jets and c-jets as can be seen from Fig. 8.

Although the DNN approach could theoretically provide more precise results than the combined SV-tagging + BDT method, since it uses all the information about the jets, the current result shows that the performance of this method is very similar. However the better performance of the method for c-jets, where the efficiency is higher by 3% is very promising.

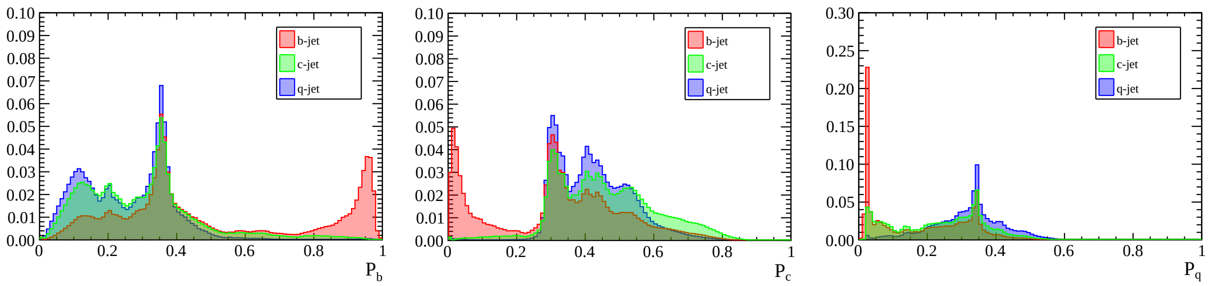

Fig. 7: The output of DNN-tagging method. From left – probability P_b for being a b-jets, probability P_c for being a c-jets and probability P_q for being a light-parton jet.

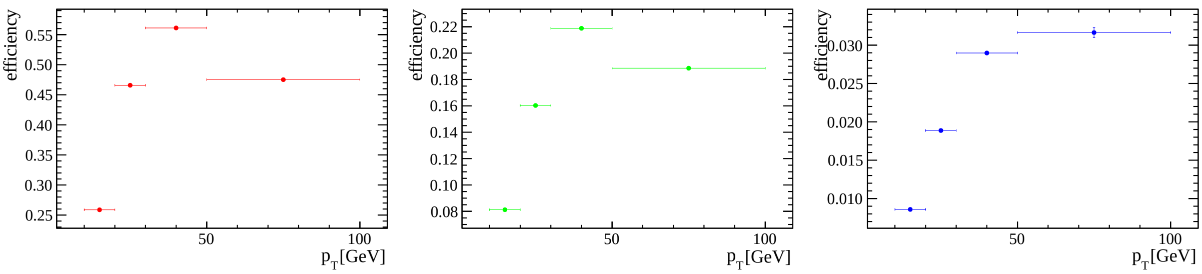

Fig. 8: The efficiency of the DNN-tagging method dependent on the jet transverse momentum p_T [MeV]. In the most left plot, the efficiency for b-jets is depicted, in central plot, the efficiency for c-jets is shown and in the most right plot, the misidentification of light-parton jets is depicted.

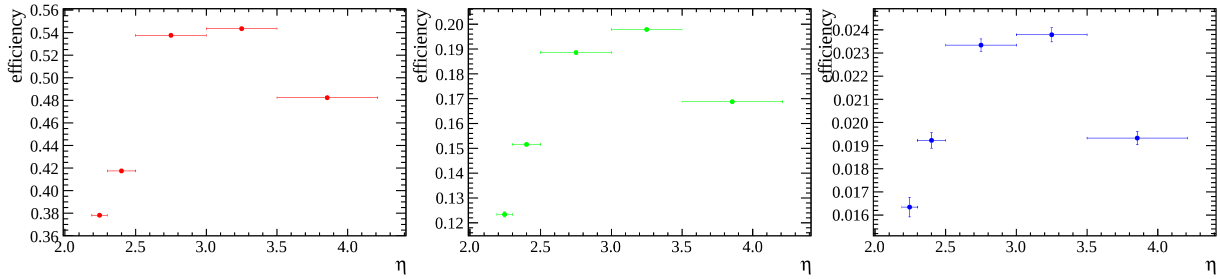

Fig. 9: The efficiency of the DNN-tagging method dependent on the jet pseudorapidity η. In the most left plot, the efficiency for b-jets is depicted, in central plot, the efficiency for c-jets is shown and in the most right plot, the misidentification of light-parton jets is depicted.

5 Conclusion

The SV-tagger represent a powerful approach to identify different jet flavours. The efficiency for high transverse momentum ($p_T > 20$ GeV) is around 50 % for b-jets, around 20 % for c-jets and with a misidentification of light-parton jets between 0.86 % – 3.15 %. Two methods are under investigation to improve the jet tagging performance. Deep neural network method performance is comparable to the SV tagging with a gain of few percents in b-jets and c-jets efficiency in some bins, however they can be considered comparable to the SV-tagger efficiency. The overall difference between the SV-tagger and DNN-tagging misidentification of light-parton jets is lower than 0.02 %. The difference between the two methods is lower than 1 % for b-jets and lower than 3 % for c-jets. The overall difference between the SV-tagger and DNN-tagging method can be seen in Fig. 10 – Fig. 11. BDT method shows a reduced discrimination power against light-parton jets compared with previous results from Run I [1]. This is the main motivation to retrain the BDT discriminators for Run II.

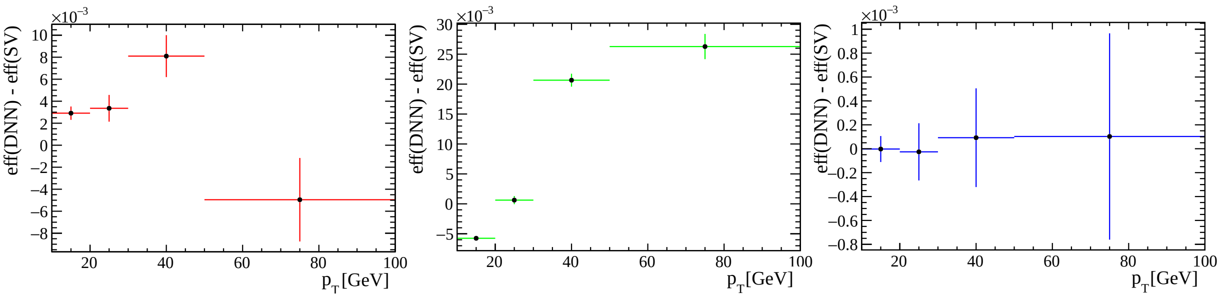

Fig. 10: The difference between the SV-tagging and DNN-tagging methods efficiencies dependent on jet transverse momentum p_T for b-jets (left plot), c-jets (middle plot) and for light-parton jets (right plot).

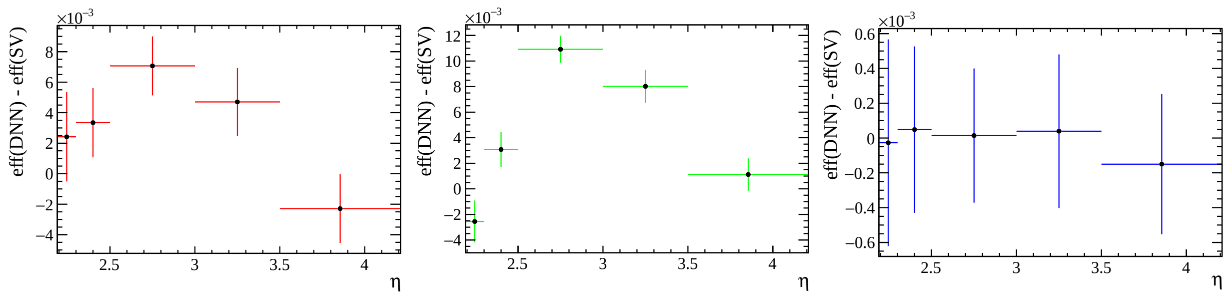

Fig. 11: The difference between the SV-tagging and DNN-tagging methods efficiencies dependent on jet pseudorapidity η for b-jets (left plot), c-jets (middle plot) and for light-parton jets (right plot).

References

[1] LHCb Collaboration, *Identification of beauty and charm quark jets at LHCb*, JINST 10 (2015) P06013.

[2] LHCb Collaboration, *Measurement of forward top pair production in the dilepton channel in pp collisions at* $\sqrt{s} = 13$ TeV.
Available at: arXiv:1803.05188

[3] C. Boser, S. Fink, S. Rocker, *Introduction to Boosted Decision Trees: A multivariate approach to classification problems*.
Available at: https://indico.scc.kit.edu/event/48/contributions/3410/attachments/1690/2312/BDT_KSETA_Freudenstadt.pdf

[4] A. Hoecker et al., *TMVA 4: Toolkit for Multivariate Data Analysis with ROOT - Users Guide*.
Available at: arXiv:physics/0703039

[5] A. Dertat, *Applied Deep Learning - Part 1: Artificial Neural Networks*.
Available at: https://towardsdatascience.com/applied-deep-learning-part-1-artificial-neural-networks-d7834f67a4f6

Appendix

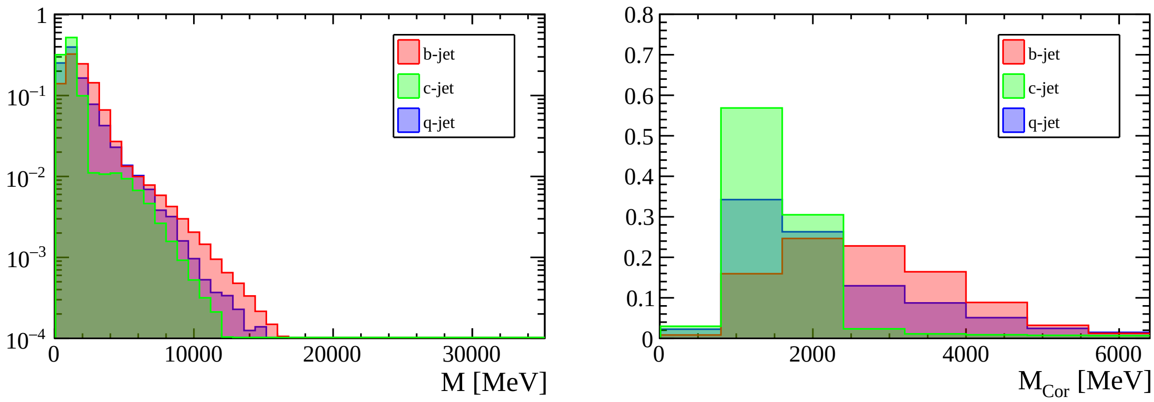

Fig. 12: Uncorrected mass M of the SV (left) and corrected mass M_cor of the SV (right).

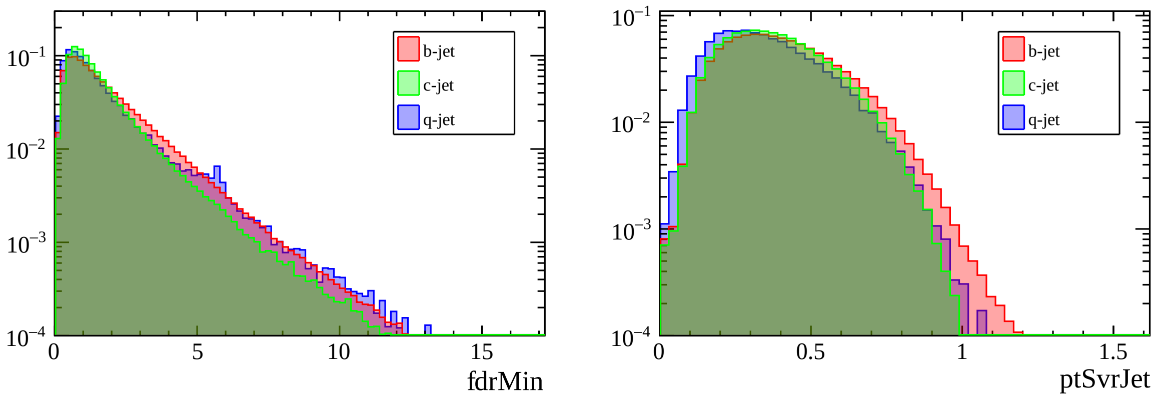

Fig. 13: The transverse flight distance of the two-track SV closest to the PV (left) and the fraction of the jet p_T carried by the SV (right).

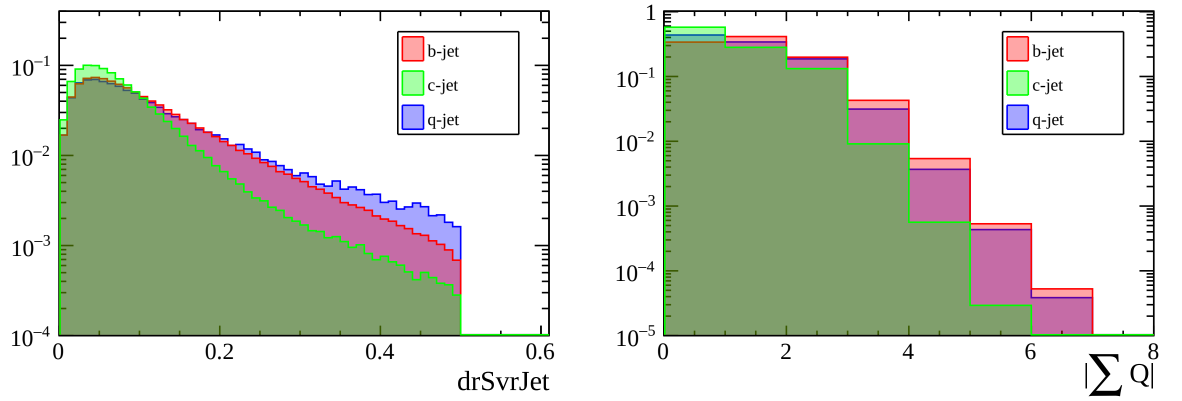

Fig. 14: Radial distance ΔR between the SV flight direction and the jet (left) and the net charge of the tracks that form the SV (right).

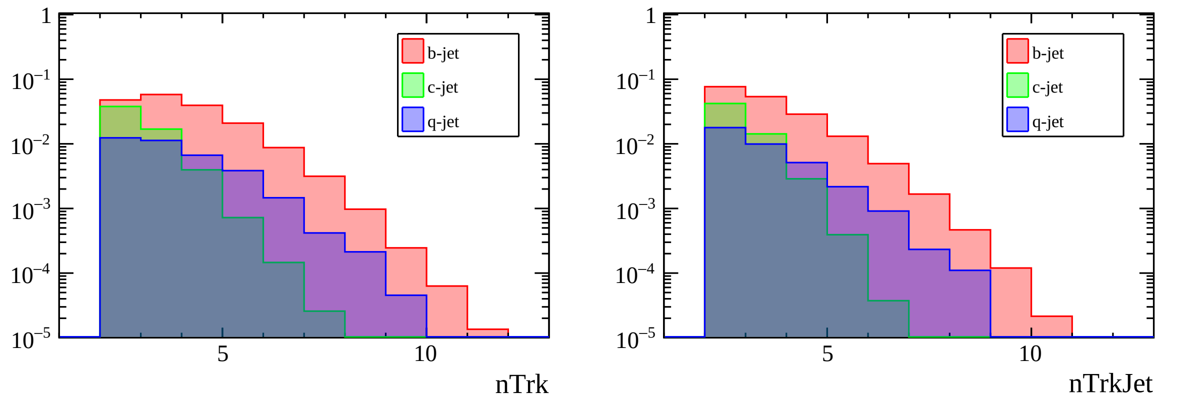

Fig. 15: The number of tracks in the SV (left) and the number of SV tracks with $\Delta R < 0.5$ relative to the jet axis (right).

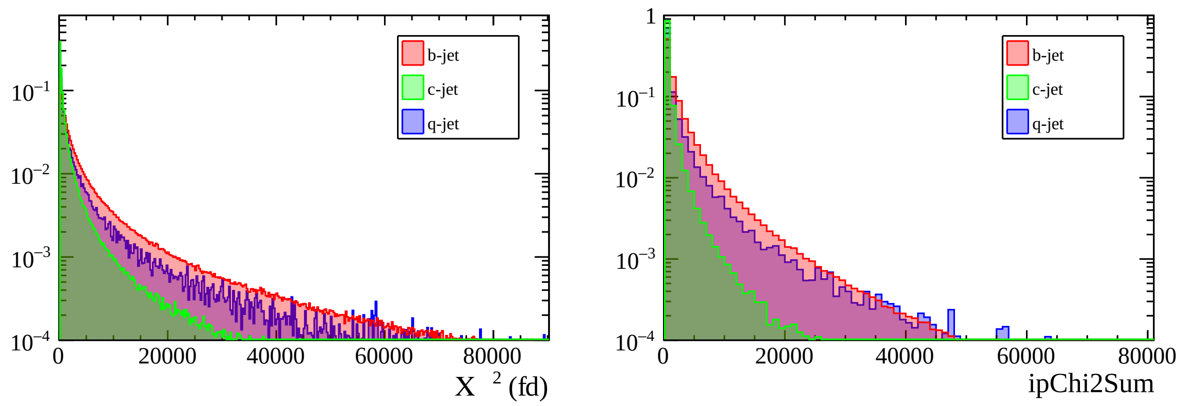

Fig. 16: The flight distance χ^2 (left) and the sum of all SV track χ^2_{IP}.