Hit multiplicity approach to b-tagging in FCC-hh

E. Perez\textsuperscript{1)}*, P. Roloff\textsuperscript{*}

\textsuperscript{*} CERN, Switzerland

Abstract

The FCC-hh project aims to study proton-proton collisions at $\sqrt{s}=100$ TeV. The ability to tag heavy flavour jets of high energy produced in those collisions will be crucial for many beyond the Standard Model searches such as $Z'$ production. Tagging highly energetic jets at FCC-hh is challenging not only due to the presence of up to 1000 pile-up interactions per bunch crossing, but also due to the dense environments inside the boosted jets and the very displaced secondary vertices arising from the B- and C-hadron decays. In particular, the difficulties of reconstructing tracks from B-hadron decay products in such environments have a dramatic impact on the b-tagging performance. A new approach for tagging high energy b-jets without reconstructing tracks and secondary vertices has been suggested. The so-called "hit multiplicity jump" tagger is based on the number of hits in the different detector layers. It makes use of the fact that a long-lived particle decaying between two layers will cause the hit multiplicity to increase between those layers. This document describes the adaptation of this approach to the FCC-hh detector and its test on central jets with transverse momentum of up to 5 TeV for the first time. The approach of combining the information on the number of hits in each detector layer in a multivariate discriminator achieves a performance that is not only comparable but potentially superior to the discrimination achieved by the traditional approach based on track and secondary vertex reconstruction.

This work was carried out in the framework of the FCC Collaboration

\textsuperscript{1}estel.perez.codina@cern.ch
1. Introduction

The Future Circular Collider (FCC) is a project for building a 100 km perimeter collider near CERN. The main facility planned is a hadron collider (FCC-hh) that aims to provide proton-proton collisions at $\sqrt{s} = 100\text{ TeV}$ and a luminosity of up to $30 \times 10^{34}\text{ cm}^{-2}\text{s}^{-1}$. The pile-up conditions are expected to range from an average of 200 pile-up events per bunch crossing in Phase I to 1000 pile-up interactions per bunch crossing in Phase II [1].

Previous work in the context of the LHC experiments has shown the difficulties of reconstructing tracks in dense environments [2] and also the resulting flavour tagging performance decrease for increasing jet transverse momentum ($p_T$) [3].

The flavour tagging performance decreases at high jet energy due to the significant flight distance of the B-hadrons, which can traverse several detector layers before decaying. This leaves fewer measurement layers available to reconstruct their decay products. On the other hand, the small angular separation between daughters challenges their track reconstruction due to hit sharing and mis-association. The tracking performance for such tracks decreases especially for jets with energies beyond 1 TeV.

The hit multiplicity jump technique described in [4, 5] exploits the fact that when a B-hadron decays between two detector layers, the number of hits in the second layer increases with respect to the first one, the so-called "hit multiplicity jump". This "jump" can be quantified by defining the variable: $f_j = (N_{j+1} - N_j)/N_j$, where $N_j$ is the number of hits in the $j^{th}$ innermost detector layer. The rejection power of this variable is evaluated standalone as well as in combination with other variables in an Artificial Neural Network discriminator.

This approach is revised and adapted to the FCC-hh detector, exploiting further physics properties of the decay. The main improvements investigated are:

- The use of information from various concentric cones around the jet direction to exploit the differences in jet structure between heavy- and light-flavour jets.
- All vertex and tracker detector layers are taken into consideration, to make use of all the information in the detector potentially providing discrimination power for the different jet energies.

This note focuses on jets of $p_T = 5\text{ TeV}$ since the performance of traditional taggers for such high $p_T$ jets is limited (background mis-identification is worse than for $p_T = 500\text{ GeV}$ by more than one order of magnitude [6]) and can therefore benefit the most from this complementary approach. Flavour tagging of such high $p_T$ jets will be essential for many Beyond the Standard Model (BSM) searches such as $Z'$ bosons decaying to $t\bar{t}$ or $b\bar{b}$ [7]. Additional results for lower jet momenta can be found in the Appendix A. The studies are performed on central jets, measured only in the barrel part of the vertex and tracker detectors, since the simple detector geometry in the barrel makes it straightforward to apply this hit-counting approach. In addition, since the depth of the tracker in the radial direction is one order of magnitude smaller than in the $z$ direction, central jets are the most challenging benchmark in terms of reconstructing extremely displaced secondary vertices from multi-TeV jets. In this preliminary study the simulation of pile-up events is not included.

This note is structured as follows: In Section 2 the MC samples, geometry and simulation software are introduced. The basic variables are described and optimised in Section 3. In the first part of the study (Section 4.1) the combination of parameters that gives the best rejection power for a single-variable selection is identified, and in the second part (Section 4.2) a sensible set of variables is used as input to train a Boosted Decision Tree (BDT) discriminator. These studies use simulated hits, therefore the effect of detector granularity on the performance is estimated ad-hoc in Section 4.3. The conclusions and outlook are summarised in Sections 5 and 6.
2. Simulation

2.1. Event samples

Dijet events in p-p collisions at $\sqrt{s} = 100\text{ TeV}$ are generated using Madgraph5 [8]. Two sets of Monte Carlo (MC) samples are used: heavy flavour samples, $pp \rightarrow b\bar{b}$, and light flavour (LF) samples, $pp \rightarrow q\bar{q}$, where $q$ is either an u-, d-, s-quark or a gluon. In both cases, the partons are produced very centrally, with a pseudorapidity of $|\eta| < 0.05$, and with a transverse momentum within a narrow range. Four samples with different $p_T$ ranges are produced for each of the b$\bar{b}$ or LF final states:

- "$p_T = 500\text{ GeV}$": $490\text{ GeV} < p_T (b \text{ or } q) < 510\text{ GeV},$
- "$p_T = 1000\text{ GeV}$": $990\text{ GeV} < p_T (b \text{ or } q) < 1010\text{ GeV},$
- "$p_T = 2000\text{ GeV}$": $1950\text{ GeV} < p_T (b \text{ or } q) < 2050\text{ GeV},$
- "$p_T = 5000\text{ GeV}$": $4950\text{ GeV} < p_T (b \text{ or } q) < 5050\text{ GeV}.$

Fragmentation and hadronisation are modelled using Pythia6 [9], including both Initial and Final State Radiation (ISR/FSR). As a result, the generated quark/gluon $p_T$ distribution is smeared with respect to the ranges listed above. The MC simulation does not include Multiple Interactions or pile-up events.

During hadronisation of high-$p_T$ quarks and gluons, pairs of b$\bar{b}$ and c$\bar{c}$ quarks can be produced. In order to obtain a pure sample of light flavour jets, the events containing either a b- or c- quark or a B- or C-hadron are removed from the LF sample.

2.2. Detector geometry

The FCC-hh tracker geometry version v3.03 [10] is used to perform this study. In the barrel region, the FCC-hh detector consists of four pixel layers with average radii of 25 mm, 60 mm, 100 mm and 150 mm respectively, as depicted in Figures 1 and 2. The material budget is $X/X_0 = 1\%$ per layer, which includes an estimation for the non-sensitive material like cables and support structures. Surrounding the pixel layers, there are four layers of macro-pixels, with radii of 270 mm, 400 mm, 530 mm and 742 mm and material budget of $X/X_0 = 2\%$ per layer. Finally, in the outermost part of the tracker there are four striplet layers of radii 0.9 m, 1.1 m, 1.3 m and 1.5 m, respectively, and $X/X_0 = 2.5\%$ of material budget per layer. The sensor pitch in $R - \phi$ and $Z$ is $25 \times 50 \mu m^2$, $33 \times 400 \mu m^2$ and $33 \mu m \times 50\text{ mm}$ for pixels, macropixels and striplets, respectively. The tracker is immersed in a 4 T solenoidal magnetic field.

2.3. Properties of B-hadron decays in the FCC-hh detector

This section describes briefly the kinematic properties of B-hadrons in high $p_T$ jets decaying in the FCC-hh detector. Figure 3 shows the B-hadron decay vertex position in the radial direction for the different jet $p_T$ values, and Figure 4 presents the corresponding B-hadron $p_T$ spectra.

Table 1 shows the fraction of B-hadrons in the $p_T = 5\text{ TeV}$ sample which decay before each of the detector layers. The denominator considers all B-hadrons that have a non-zero flight distance, including decays happening beyond the last tracker layer. More than half of the B-hadrons decay after the third detection layer.

---

2The quark/gluon $p_T$ before ISR and FSR is sometimes referred to as "jet $p_T$" in this document, to denote the samples described above.
Figure 1: FCC-hh vertex and tracking detectors. The different colours represent different detector granularities.

Figure 2: Vertex sub-detector (in red) and tracker layers (in dark green). The different colours represent different detector granularities.

Figure 3: B-hadron decay vertex position in the radial direction for different b-quark $p_T$ values.

Figure 4: B-hadron $p_T$ distributions for different b-quark $p_T$ samples.
3. Basic variables for the hit-based approach

Table 1: Fraction of B-hadrons decaying before each detector layer for the $p_T = 5$ TeV sample.

<table>
<thead>
<tr>
<th>Layer</th>
<th>layer R [mm]</th>
<th>Fraction of B-hadrons [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beampipe</td>
<td>20</td>
<td>19</td>
</tr>
<tr>
<td>1</td>
<td>25</td>
<td>22</td>
</tr>
<tr>
<td>2</td>
<td>60</td>
<td>35</td>
</tr>
<tr>
<td>3</td>
<td>100</td>
<td>45</td>
</tr>
<tr>
<td>4</td>
<td>150</td>
<td>56</td>
</tr>
<tr>
<td>5</td>
<td>270</td>
<td>72</td>
</tr>
<tr>
<td>6</td>
<td>400</td>
<td>82</td>
</tr>
</tbody>
</table>

2.4. Software

The tracker geometry is defined using tkLayout [11], and translated to DD4hep [12] format. The particles interaction with the detector is simulated using DDSim [13], which is based on GEANT4 [14–16]. The simulated hits are not digitised.

3. Basic variables for the hit-based approach

The hit multiplicity jump approach to flavour tagging is based on counting the number of hits in each detector layer that are contained within a certain angular region $\Delta R$ around the direction of the jet ($\Delta R_{jet-hit}$). Contrary to [4, 5], this study is performed using GEANT4 simulated hits.

3.1. Jet direction

Jets are defined using stable MC particles (those with generator status = 1, excluding neutrinos) as input to the anti-$k_t$ jet clustering algorithm [17]. Only the jet with highest $p_T$ in the event is considered in this study in order to reduce the impact of high energy gluon radiation on the jet $p_T$.

Figure 5 shows the angular distance between the MC jet direction and the position of the simulated hits produced by B-hadron daughter particles, produced by all particles in b-jets and produced by all particles in LF jets. The B-hadron daughters are more collimated around the jet direction than the other particles in the b-jet or LF-jet. Therefore by considering only particles within a certain $\Delta R_{jet-hit}$ distance from the jet direction, a significant fraction of particles not originating from the B-hadron decay can be rejected.

Figure 6 presents the $\Delta R_{jet-hit}$ distance distribution for hits from B-hadron daughters for different values of the jet R parameter and compares it to the distribution of the hit angular distance with respect to the actual B-hadron direction ($\Delta R_{Bhad-hit}$). For the $p_T = 5$ TeV sample, a jet R parameter of R=0.006 is used as default since the jet direction gives a good estimate of the B-hadron direction. A more accurate estimation of the B-hadron direction, allows for a smaller $\Delta R_{jet-hit}$ distance to be used, which provides larger rejection of particles not coming from the B-hadron decay. The default R value is compatible with the FCC-hh jet angular resolution of 0.001 in $\eta$ and 1 mrad in $\phi$ for $p_T = 5$ TeV jets [18].

The R-parameter used for jet clustering was optimised for each jet $p_T$ sample, as shown in Figure 7, to contain most of the particles coming from the B-hadron decay chain and provide a good estimate of the B-hadron direction. The default jet algorithm R-parameter values used are also quoted in Table 2.

3.2. Additional $\Delta R_{jet-hit}$ cones

By default, hits within a $\Delta R_{jet-hit}$ cone of the same size as the jet algorithm R parameter are considered. However, to further exploit the b-jet structure properties, three additional concentric cones are defined,

\[ \Delta R \equiv \sqrt{\Delta \eta^2 + \Delta \phi^2} \]
3. Basic variables for the hit-based approach

3.3. Number of hits per layer

Figure 8 shows the average number of hits per jet inside the default cone (see Table 2) in each of the detector layers for the $b\bar{b}$ sample and the LF sample, and also the number of hits from B-hadron daughters in $b\bar{b}$ events. The number of hits from the B-hadron daughters increases with the layer number up to layer number nine. Comparing this distribution for other jet energies (shown in Figure 9) suggests that the energy (and therefore the lifetime) of the B-hadrons causes this increase, which reaches a plateau at lower radii for lower energies. In Figure 9 the cone $\Delta R_{\text{jet-hit}}$ considered is 0.2, and is the same for all $p_T$ samples so that the effect of hits falling out of the $\Delta R_{\text{jet-hit}}$ cone is minimised and therefore the different samples can be compared.

For LF jets, the number of hits decreases with the layer number. The decrease is mainly due to particles at relatively low $p_T$ with sizeable curvature falling out of the defined $\Delta R_{\text{jet-hit}}$ cone. Figure 10 shows a breakdown of the origin of the particles producing the hits. The number of particles created in simulation is significant and increases with the layer number. Particles created in the simulation are due to the interaction with the detector material and can also lead to a hit multiplicity increase. They are usually of low $p_T$ and hence they likely fall out of the $\Delta R_{\text{jet-hit}}$ cone in subsequent layers. The

Table 2: Default jet R parameter used for the different $p_T$ samples.

<table>
<thead>
<tr>
<th>Sample $p_T$ [GeV]</th>
<th>jet R parameter</th>
</tr>
</thead>
<tbody>
<tr>
<td>500</td>
<td>0.06</td>
</tr>
<tr>
<td>1000</td>
<td>0.03</td>
</tr>
<tr>
<td>2000</td>
<td>0.015</td>
</tr>
<tr>
<td>5000</td>
<td>0.006</td>
</tr>
</tbody>
</table>

with $\Delta R_{\text{jet-hit}} = 1/2$, 1/4 and 1/8 of the default size, respectively.

Figure 5: $\Delta R_{\text{jet-hit}}$ for anti-$k_t$ jets with $R=0.006$ in the $p_T = 5$ TeV sample. Simulated hits produced by B-hadron daughter particles (in red), all particles in b-jets (in black) and all particles in LF jets (in blue) are shown separately.

Figure 6: $\Delta R_{\text{jet-hit}}$ for anti-$k_t$ jets of different R parameters in the $p_T = 5$ TeV $b\bar{b}$ sample showing simulated hits produced by B-hadron daughter particles only (solid lines in shades of red), compared to $\Delta R_{\text{Bhad-hit}}$ (dashed black line).
3. Basic variables for the hit-based approach

3.4. The variable \( f_j \)

Instead of using the average number of hits, the hit multiplicity jump is calculated event by event. Figure 11 compares the number of hits per event in b-jets and in LF jets, and shows that up to the 6\(^{th}\) layer the average number of hits is significantly smaller for the b\(\bar{b}\) sample than for the LF sample.

The hit multiplicity jump approach is based on the difference in the number of hits per event between one layer and the next. Figure 12 shows the number of hits in layer \( j + 1 \) versus the number of hits in layer \( j \) in b\(\bar{b}\) events considering only hits from the daughters of the B-hadrons. Each figure shows the distribution for one of the 11 layer pairs. The hit multiplicity jump is expected to appear as a higher number of entries in the bins above the diagonal, which can be seen in Figure 12 especially in the first (innermost) layers. If the jump is due to a particle decaying between two consecutive layers, their hit multiplicity difference should be a multiple of two due to charge conservation. In the outer layers, the entries that appear in bins below the diagonal are probably due to tracks with sizeable curvature and falling out of the defined \( \Delta R_{\text{jet-hit}} \) cone, as already mentioned. Figure 13, shows the same set of figures considering all the particles in the b-jets, and Figure 14 presents similar figures for the LF sample. For the b\(\bar{b}\) sample the hit multiplicity jump due to B-hadron decays is visible, especially in the six innermost layers, while for the LF sample such even jumps above the diagonal are not prominent. The spread around the diagonal is significantly larger for the outer layers. Therefore those layers are expected to be less powerful in the discrimination between light and heavy flavour jets.

The variable \( f_j = (N_{j+1} - N_j)/N_j \) is calculated for each of the layers and shown in Figure 15 for b\(\bar{b}\) events and for LF events. This variable reaches higher values in the b\(\bar{b}\) sample and has a dependence on the layer used, especially in the LF sample.
3 Basic variables for the hit-based approach

Figure 9: Number of hits per jet inside inside a cone of $\Delta R_{\text{jet-hit}} = 0.2$ for each layer, for various $p_T$ samples, considering simulated hits produced by B-hadron daughter particles only.

Figure 10: Number of hits per jet inside the default cone for each layer in the $p_T = 5\text{TeV}$ LF sample. Breakdown of the origin of the particles producing the hits as explained in the text.

Figure 11: Number of hits in the default $\Delta R_{\text{jet-hit}}$ cone per event for the $p_T = 5\text{TeV}$ b$b$ sample (black) and the LF sample (blue). Each figure shows the number of hits in one of the layers. All distributions are normalised to unit area.
3.5. The effect of pile-up interactions

The effect of pile-up events is not simulated in the present study, and is left as open topic for future studies. In \cite{4, 5} it was found that pile up did not to have an important effect in the method for the LHC. In the FCC-hh case, the number of pile-up interactions to be considered is higher, but also the $\Delta R_{\text{jet-hit}}$ cone considered for high energy jets is smaller. Preliminary estimates (see Appendix 4) suggest that 1000 pile-up interactions would add less than one hit per layer in a cone of $\Delta R_{\text{jet-hit}}=0.006$ (default for $p_T=5$ TeV jets), and therefore the effect of pile-up interactions on this method is not expected to be dominant for central jets of such energy.

4. Flavour tagging using hit information

4.1. Single variable selection

The goal of this first part of the study is to explore different definitions of "hit multiplicity jump" and rank the most discriminating variables. For that scope, a cut-and-count selection using a single stand-alone variable is performed. In the following, results for the jet $p_T=5$ TeV sample are discussed.

There are two parameters to be optimised: the size of the cone used ($\Delta R_{\text{jet-hit}}$), and the number of layers that are considered in the multiplicity jump. All combinations of the four $\Delta R_{\text{jet-hit}}$ cone sizes and the six innermost layers are considered. In addition, other definitions of the hit multiplicity jump requirement (by default $f_j \geq 1$ in any of the layers) are also studied for all the combinations:

- $f_j \geq 1$ and even multiplicity difference ($N_{j+1} - N_j$) (to exploit charge conservation)
4 Flavour tagging using hit information

Figure 13: Number of hits in layer $j+1$ vs number of hits in layer $j$ in $b\bar{b}$ events for the $p_T=5$ TeV sample, considering hits from all particles. One figure per layer pair.

Figure 14: Number of hits in layer $j+1$ vs Number of hits in layer $j$ in LF events for the $p_T=5$ TeV sample, considering hits from all particles. One figure per layer pair.
Flavour tagging using hit information

Figure 15: The variable $f_j$ vs layer number. For $b\bar{b}$ events (left) and LF events (right) in the $p_T = 5$ TeV samples.

Figure 16: Distribution of the $f_l$ variable calculated in the default $\Delta R_{jet-hit}$ cone and using the sixth innermost layer. The variable is calculated using the simulated hits produced by the B-hadron decay products (in red), by all particles in $b$-jets (in black) and by all particles in LF jets (in blue).

- $f_j \geq 1.5$ (tighter cut on $f_j$)
- $f_l \geq 1$: fractional difference $((N_l - N_1)/N_1)$ between the first layer and a given layer $l$ (as suggested in referenced paper)

The requirement giving the best standalone rejection is $f_l \geq 1$ calculated in the default $\Delta R_{jet-hit}$ cone and using the sixth innermost layer, therefore: $(N_6 - N_1)/N_1 \geq 1$; giving an efficiency of 38% for the $b$-jet signal and a 2.1% efficiency for the LF jet background, meaning an efficiency ratio of about 18. This variable is shown in Figure 16 comparing the distribution for simulated hits produced by the B-hadron decay products, by all particles in $b$-jets and by all particles in LF jets.

The other five most discriminating requirements are variations of the best one mentioned above: using half the cone instead of the default $\Delta R_{jet-hit}$ cone, setting the threshold at 1.5 instead of at 1, using layer
This study on the discrimination of standalone variables guides the choice of variables to include in the multivariate analysis, presented in Section 4.2.

### 4.2. Multivariate analysis using simulated hits

A discriminant based on a single standalone variable does not exploit the full potential of this tagging approach. Therefore a multivariate analysis is performed, based on a BDT classifier and making use of the TMVA [19] package. The use of variables like \(f_j\) or \(f_l\) in a BDT is found not to be as powerful as providing as input to the BDT classifier the lower level information: the number of hits in each layer. Hence, only the hit multiplicity information is used as input to the BDT.

The \(b\bar{b}\) and LF samples for \(p_T=5\) TeV jets are used to train the discriminator as signal and background, respectively. Figure 17 shows the discrimination (b-tagging efficiency vs LF mis-identification efficiency) obtained by combining the hit multiplicity in each of the 12 layers in the default \(\Delta R_{\text{jet-hit}}\) cone size. The performance using different cone sizes (1/8, 1/4, 1/2 of the jet R parameter) is compared to the default. Using the \(\Delta R_{\text{jet-hit}}\) cone of half the R-parameter value, leads to the highest rejection. If the information on the hit multiplicity in all of the four \(\Delta R_{\text{jet-hit}}\) cone sizes is included in the BDT, the rejection significantly increases by up to 50%, as shown in Figure 17.

Other options such as using the mean, variance, or skewness of the hit position with respect to the jet direction instead of using different \(\Delta R_{\text{jet-hit}}\) cones, were tested but did not lead to any improvement.

Figure 18 compares the discrimination obtained by using the hit multiplicity information from different numbers of detector layers. For such high energetic jets, using only the four innermost layers does not exploit the full discrimination power of the method. Considering the information from six, eight or even ten layers, provides a significant improvement in discrimination power. Adding the two outermost layers does not have such a significant impact on the performance. As expected, providing the BDT discriminator with the information from all layers leads to the highest rejection and therefore has been used as default in this study.

### 4.3. Multivariate analysis considering detector granularity

The results presented in the previous sections consider only simulated hits. Therefore, the number of particles going through the sensors are effectively counted. In this section, the impact of detector granularity is partially estimated by defining pseudo-reconstructed hits ("reco" hits) as described next. The detector sensitive area is binned in \(R\phi\) and \(Z\) according to the sensor pitch. If two or more simulated hits fall in the same bin, a single "reco" hit is counted, i.e. it is assumed that the two simulated hits cannot be distinguished by the detector readout.

The position of the simulated hits is not smeared according to the detector resolution. This is expected to have a small effect on the hit multiplicity for the innermost layers since the sensor pitch is small with respect to the \(\Delta R_{\text{jet-hit}}\) cone size. However for the four outermost layers the strip pitch in the \(z\) direction is larger than the \(\Delta R_{\text{jet-hit}}\) cone size. This is an aspect that needs to be addressed in future studies, and could potentially motivate an improvement of the spatial resolution in the \(z\) direction for the outermost layers.

Figure 19 shows the decrease in performance when "reco" hits are considered, which ranges from about 20 to 50% depending on the b-tagging efficiency.

Several variations on the detector granularity were studied, such as improving the granularity of the four innermost layers (pixel layers) to \(20 \times 20\) \(\mu\text{m}^2\), which lead to a recovery of about 20-30% rejection, as extracted from Figure 19.

If the sensors would be able to distinguish between 0, 1, 2 or 3-or-more particles going through the same pixel (through analog readout as in ATLAS [2]), it would be possible to recover the performance achieved by using simulated hits, as can be seen in Figure 19.
4.4. Multivariate analysis for different jet energies

The multivariate analysis is performed similarly for the different jet $p_T$ samples, and the results are presented in Figure 20. In contrast to the track-based tagging approach, in this case the performance improves with increasing jet energy, since displaced decays taking place inside the tracker volume are an advantage in this method. The performance for the two highest jet $p_T$ samples is similar since the $\Delta R_{\text{jet-hit}}$ cone has been optimised for each sample.

5. Conclusions

A hit-based approach to flavour tagging has been adapted to FCC-hh. This document studies the discrimination between b-jets and LF jets of $p_T = 5$ TeV, which is a challenging benchmark for the traditional approach based on tracking and secondary vertex reconstruction.

The outer detector layers are most affected by the particle interactions with the detector material, as
Figure 19: LF mis-identification rate vs b-tagging efficiency. Comparison of the BDT classifier rejection power when considering simulated hits (independent of the detector granularity in red), and "reco" hits (as defined in the text) with default granularity (blue), improved granularity (violet) or default granularity with analog readout (green). The background efficiency ratio is calculated with respect to the performance using simulated hits.

Figure 20: LF mis-identification rate vs b-tagging efficiency for various $p_T$ samples, considering "reco" hits. The background efficiency ratio is calculated with respect to the performance for $p_T = 5$ TeV jets.

well as the particle hits falling out of the considered $\Delta R_{\text{jet-hit}}$ cone. Therefore when used standalone to identify a hit multiplicity jump, those layers have less rejection power.

Training a BDT discriminator based uniquely on the number of hits is the best of the considered approaches. The simultaneous use of various $\Delta R_{\text{jet-hit}}$ cones of different sizes improves the rejection significantly (up to a 50% increase) as this exploits the differences in jet structure between b-jets and LF jets.

When used in the BDT, the outer detector layers provide significant additional rejection power for very high $p_T$ jets due to the longer lifetimes and the differentiated jet structure.

The effect of detector granularity was partially studied, and it was found that improving the granularity of the four innermost layers from $25 \times 50 \mu m^2$ to $20 \times 20 \mu m^2$, improves the rejection power moderately (20%-30%). Using an analog readout able to distinguish up to three particles traversing the detector would also improve the rejection such that it would be possible to recover the performance achieved by using simulated hits.
A Results for various jet energies

This b-tagging method performs best for very high jet energies (jet \( p_T > 1 \text{ TeV} \)) since a sizeable fraction of B-hadrons decay after traversing several detector layers. Also because the optimal \( \Delta R_{\text{jet-hit}} \) cone used is smaller and therefore less affected by background hits from the same event and potentially also from other pile-up interactions. This highlights the complementarity of this method with the track-based flavour tagging methods. Hence, the greatest benefit would potentially be obtained by combining the two approaches.

The hit multiplicity approach seems to be very promising for FCC-hh, where tracking and secondary vertex reconstruction at high energies will be very challenging due to the large boost of the decaying particles. Nevertheless, this method should be further studied, in particular the effect of the expected pile-up interactions on the tagging performance. A list of considerations for future studies can be found in Section 6.

6. Outlook

This section summarises the remaining aspects and extensions to be addressed in future studies.

Detector granularity  As mentioned in Section 4.3 the effect of the hit spatial resolution on the hit multiplicity needs to be studied in detail, including the effect of the granularity on the pile-up estimates.

Extension to higher \(|\eta|\)  The present study is performed in the detector barrel region, which is where the track-based tagging is most affected by the very large jet boost. The extension of the method to higher pseudorapidity regions (transition region between barrel and endcaps) is non trivial since the number of layers traversed by the particle and the distance of the layers from the interaction point depend strongly in \( \eta \) and cannot be assumed to be constant as in the current study. Hence the BDT discriminator should be dependent on the jet pseudorapidity. In the endcap region the background due to pile up interactions is larger, therefore its impact on the method should be taken into consideration.

Pile-up  The effect of hits from tracks in pile-up events need to be studied by including those events in the simulation, in order to correctly estimate the effect of correlated hits in the various layers and \( \Delta R_{\text{jet-hit}} \) cones.

Fragmentation and hadronisation modelling  The jet structure is sensitive on the hadronisation and fragmentation parameters. The uncertainty coming from the modelling of such processes should be investigated.

Combination with track-based approach  Jets with \( 500 \text{ GeV} < p_T < 1 \text{ TeV} \) are expected to benefit most from the combination of the two approaches.

Acknowledgments

The authors would like to acknowledge Andrea Coccaro for the generation of the input samples.

A. Results for various jet energies

A.1. Properties of B-hadron decays

Table 3 shows the fraction of B-hadrons decaying before a certain layer, for various jet energies.
Table 3: Fraction of B-hadrons decaying before each detector layer, for the different $p_T$ samples.

<table>
<thead>
<tr>
<th>Layer</th>
<th>layer R[mm]</th>
<th>$p_T = 500$ GeV</th>
<th>$p_T = 1$ TeV</th>
<th>$p_T = 2$ TeV</th>
<th>$p_T = 5$ TeV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beampipe</td>
<td>20</td>
<td>0.58</td>
<td>0.43</td>
<td>0.28</td>
<td>0.19</td>
</tr>
<tr>
<td>1</td>
<td>25</td>
<td>0.66</td>
<td>0.51</td>
<td>0.32</td>
<td>0.22</td>
</tr>
<tr>
<td>2</td>
<td>60</td>
<td>0.89</td>
<td>0.75</td>
<td>0.53</td>
<td>0.35</td>
</tr>
<tr>
<td>3</td>
<td>100</td>
<td>0.97</td>
<td>0.88</td>
<td>0.67</td>
<td>0.45</td>
</tr>
<tr>
<td>4</td>
<td>150</td>
<td>0.99</td>
<td>0.95</td>
<td>0.79</td>
<td>0.56</td>
</tr>
<tr>
<td>5</td>
<td>270</td>
<td>0.999</td>
<td>0.99</td>
<td>0.92</td>
<td>0.72</td>
</tr>
<tr>
<td>6</td>
<td>400</td>
<td>1</td>
<td>0.999</td>
<td>0.97</td>
<td>0.82</td>
</tr>
</tbody>
</table>

Figure 21: $\Delta R_{\text{jet-hit}}$ for anti-$k_t$ jets in various $p_T$ samples. Considering simulated hits produced by B-hadron daughter particles (in red), all particles in b-jets (in black) and all particles in LF jets (in blue). Normalised to the same number of hits from all particles.

Figure 22: Average number of hits per jet inside the default cone in each layer for various $p_T$ samples. For the b$b$ sample (black) and the LF sample (blue) and hits from B-hadron daughters (red). Normalised to the same number of jets.

A.2. $\Delta R_{\text{jet-hit}}$ and hit multiplicity distributions

Figure 21 shows the $\Delta R_{\text{jet-hit}}$ for anti-$k_t$ jets in various $p_T$ samples. Figure 22 shows the average number of hits per jet inside the default cone per layer in various $p_T$ samples.

Figures 23, 24 and 25 show the hit multiplicity distribution in the default $\Delta R$ cone of each layer, for the $p_T = 500$ GeV, $p_T = 1$ TeV and $p_T = 2$ TeV, comparing the b$b$ and LF samples.

Figures 26, 27, 28, and 29 show the average number of hits per jet in each detector layer, including a breakdown for the in various $\Delta R_{\text{jet-hit}}$ cones, for the $p_T = 5$ TeV, $p_T = 2$ TeV, $p_T = 1$ TeV and $p_T = 500$ GeV samples respectively. They show the distributions for hits from particles in the LF jets, from particles in the b-jets, from the particles in the b-jets coming from the B-hadron decay, and from the particles in the b-jets not coming from the B-hadron decay.
Figure 23: Number of hits in the default $\Delta R$ cone per event, for the $p_T=500 \text{ GeV}$ $b\bar{b}$ sample (black) and LF sample (blue). Each figure shows the number of hits in one of the layers respectively. Distributions are normalised to unit area.

Figure 24: Number of hits in the default $\Delta R$ cone per event, for the $p_T=1 \text{ TeV}$ $b\bar{b}$ sample (black) and LF sample (blue). Each figure shows the number of hits in one of the layers respectively. Distributions are normalised to unit area.
B. Pile-up effect calculation

Figure 25: Number of hits in the default $\Delta R_{\text{jet-hit}}$ cone per event, for the $p_T = 2 \text{ TeV} \ b\bar{b}$ sample (black) and LF sample (blue). Each figure shows the number of hits in one of the layers respectively. Distributions are normalised to unit area.

Figure 26: Average number of hits per jet in each detector layer in various $\Delta R_{\text{jet-hit}}$ cones. Figure (a) shows the distributions considering all particles in LF jets. Figure (b) considers all particles in $b$-jets. Figures (c) and (d) consider particles in $b$-jets coming from the B-hadron decay, and not coming from the B-hadron decay, respectively. All distributions are for jets with $p_T = 5 \text{ TeV}$.

A.3. Dependence on the $\Delta R_{\text{jet-hit}}$ cone and number of layers used

Figure 30 shows for different $p_T$ samples the dependence of the $b$-tagging performance on the $\Delta R_{\text{jet-hit}}$ cone used, compared to using simultaneously the information in all of the four cones.

Figure 31 shows the effect of including the hit multiplicity information from a different number of layers, for various jet energies.

B. Pile-up effect calculation

To estimate the effect of pile-up interactions on the hit multiplicity tagger, the charged particles fluence map is used to estimate the maximum sensor occupancy in each layer [20]. Given the layer radius,
Figure 27: Average number of hits per jet in each detector layer in various $\Delta R_{\text{jet-hit}}$ cones. Figure (a) shows the distributions considering all particles in LF jets. Figure (b) considers all particles in b-jets. Figures (c) and (d) consider particles in b-jets coming from the B-hadron decay, and not coming from the B-hadron decay, respectively. All distributions are for jets with $p_T = 2$ TeV.

Figure 28: Average number of hits per jet in each detector layer in various $\Delta R_{\text{jet-hit}}$ cones. Figure (a) shows the distributions considering all particles in LF jets. Figure (b) considers all particles in b-jets. Figures (c) and (d) consider particles in b-jets coming from the B-hadron decay, and not coming from the B-hadron decay, respectively. All distributions are for jets with $p_T = 1$ TeV.

Figure 29: Average number of hits per jet in each detector layer in various $\Delta R_{\text{jet-hit}}$ cones. Figure (a) shows the distributions considering all particles in LF jets. Figure (b) considers all particles in b-jets. Figures (c) and (d) consider particles in b-jets coming from the B-hadron decay, and not coming from the B-hadron decay, respectively. All distributions are for jets with $p_T = 500$ GeV.
Figure 30: LF mis-identification probability vs b-tagging efficiency for the different $p_T$ samples. Comparison of the BDT classifier rejection power when using the hit multiplicity information in a single cone for different cone sizes (in black and in grey) or using simultaneously the information in all of the four different cones (in red). The background efficiency ratio is calculated with respect to the performance using a single $\Delta R$ cone with the default size.

Figure 31: LF mis-identification probability vs b-tagging efficiency for various $p_T$ samples. Comparison of the BDT classifier rejection power when using the hit multiplicity information in different number of innermost layers. The background efficiency ratio is calculated with respect to the performance using 12 layers.

Granularity and the $\Delta R_{\text{jet-hit}}$ of the considered cone, the number of pixels inside the cone is estimated. The estimated number of pile-up hits entering the considered $\Delta R_{\text{jet-hit}}$ cone in each layer is detailed in Table 4, showing that in the $p_T = 5$ TeV case, the effect of pile-up translates into less than one additional hit in each layer. For lower energies the number of hits scales with the squared of the $\tan(\Delta R_{\text{jet-hit}})$, i.e. 25 hits in the first layer for $p_T$ 500 GeV jets, using a $\Delta R_{\text{jet-hit}} = 0.06$ cone.
Table 4: Input used for the estimation of the number of hits coming from 1000 pile-up interactions inside the default $\Delta R_{\text{jet-hit}} = 0.006$ (for 5 TeV jets) in each layer.

<table>
<thead>
<tr>
<th>Barrel layer</th>
<th>Layer radius [mm]</th>
<th>Cone area [mm$^2$]</th>
<th>Pixel area [mm$^2$] in cone</th>
<th># pixels</th>
<th>max. occupancy [%]</th>
<th># pileup hits</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>25</td>
<td>0.0707</td>
<td>0.00125</td>
<td>56.6</td>
<td>0.45</td>
<td>0.25</td>
</tr>
<tr>
<td>2</td>
<td>60</td>
<td>0.407</td>
<td>0.00125</td>
<td>326</td>
<td>0.11</td>
<td>0.36</td>
</tr>
<tr>
<td>3</td>
<td>100</td>
<td>1.13</td>
<td>0.00125</td>
<td>905</td>
<td>0.05</td>
<td>0.45</td>
</tr>
<tr>
<td>4</td>
<td>150</td>
<td>2.54</td>
<td>0.00125</td>
<td>2036</td>
<td>0.02</td>
<td>0.41</td>
</tr>
<tr>
<td>5</td>
<td>270</td>
<td>8.24</td>
<td>0.0132</td>
<td>625</td>
<td>0.08</td>
<td>0.50</td>
</tr>
<tr>
<td>6</td>
<td>400</td>
<td>18.1</td>
<td>0.0132</td>
<td>1371</td>
<td>0.04</td>
<td>0.55</td>
</tr>
<tr>
<td>7</td>
<td>530</td>
<td>31.8</td>
<td>0.0132</td>
<td>2407</td>
<td>0.02</td>
<td>0.48</td>
</tr>
<tr>
<td>8</td>
<td>742.4</td>
<td>62.3</td>
<td>0.0132</td>
<td>4722</td>
<td>0.01</td>
<td>0.47</td>
</tr>
<tr>
<td>9</td>
<td>937.2</td>
<td>99.3</td>
<td>1.65</td>
<td>60.2</td>
<td>0.75</td>
<td>0.45</td>
</tr>
<tr>
<td>10</td>
<td>1132</td>
<td>145</td>
<td>1.65</td>
<td>87.8</td>
<td>0.43</td>
<td>0.38</td>
</tr>
<tr>
<td>11</td>
<td>1326.7</td>
<td>199</td>
<td>1.65</td>
<td>121</td>
<td>0.27</td>
<td>0.33</td>
</tr>
<tr>
<td>12</td>
<td>1539.5</td>
<td>268</td>
<td>1.65</td>
<td>162</td>
<td>0.21</td>
<td>0.34</td>
</tr>
</tbody>
</table>
References


