ATLAS NOTE

ATL-PHYS-PUB-2017-010

8th June 2017

Variable Radius, Exclusive-k$_T$, and Center-of-Mass Subjet Reconstruction for Higgs($\rightarrow b\bar{b}$) Tagging in ATLAS

The ATLAS Collaboration

Abstract

Many physics searches in Run 2 of the Large Hadron Collider involve boosted Higgs bosons, which decay to two b-quarks with a large branching ratio. The Higgs boson is reconstructed as a large-R jet and the b-quarks are reconstructed as a pair of b-tagged subjets. This note documents alternative subjet techniques to reconstruct and identify the two b-jets from highly-boosted Higgs boson decays. New subjet tagging techniques are investigated, including the use of variable radius trackjets, exclusive-k_t subjets, and calorimeter subjets reconstructed in the center-of-mass frame of the Higgs jet. For Higgs jets with large transverse momenta (>1 TeV), these three new techniques significantly outperform the fixed radius trackjet tagging technique currently used as the standard method in ATLAS.

ATL-PHYS-PUB-2017-010
12 June 2017

1. Introduction

The ability to reconstruct boosted Higgs bosons[1] decaying to pairs of bottom quarks ($h \to b\bar{b}$) plays a very important role in the physics program at the Large Hadron Collider (LHC) [1]. Techniques to accomplish this were developed [2] and in the ATLAS experiment [3] were used throughout Run 1 and Run 2 [4–6]. In particular, the ability to identify high transverse momentum (p_{T}), or boosted $h \to b\bar{b}$ decays in which the two b-quarks are collimated, has played a central role in extending the sensitivity reach for Beyond the Standard Model (BSM) physics [7–9]. Conventionally, this is achieved by reconstructing the full hadronic $b\bar{b}$ system as a single large radius (large-R) jet and identifying the b-hadrons from the Higgs boson decay using fixed radius track jets which are then b-tagged. As the ATLAS dataset increases, it will become more important to extend these techniques to multi-TeV energy regimes to push the reach for new physics to higher mass scales and also to improve future measurements which will rely on multi-TeV $h \to b\bar{b}$ decays. This note presents three new subjet reconstruction techniques to accomplish this goal. They use variable radius track jets, exclusive-k_t subjets, or calorimeter subjets reconstructed in the rest frame of the Higgs jet to reconstruct the subjets originating from the b-quarks.

The main inputs to b-tagging algorithms required for the b-hadron reconstruction are the trajectories of charged particles (tracks) reconstructed in the inner detector (ID) [10, 11]. The ATLAS inner detector system, consisting of a silicon pixel detector, a silicon micro-strip detector (SCT), and the straw tubes of the transition radiation tracker, is used to measure the trajectories and momenta of charged particles in the region [2] $|\eta| < 2.5$. The ID surrounds the beam pipe, and is located inside a solenoid magnet that provides a 2 T axial magnetic field.

Section 2 summarizes the simulated samples used in this study. Section 3 outlines the physics object reconstruction and event selections applied in this note. The three subjet reconstruction techniques are presented in detail in Section 4 and comparison of their performance to the conventional method of subjet b-tagging using fixed radius track jets is discussed in Section 5.

2. Monte Carlo Simulation Samples

A sample of background jets initiated by light quarks and gluons is derived from a multi-jet process simulated using PYTHIA8 [12] with the NNPDF2.3 leading order (LO) parton distribution function (PDF) set [13] and the A14 [14] tuned modelling of showering and underlying event parameterisations. Jets from this sample are refered to as QCD jets. Fully hadronic top quark pair events are used for jets originating from hadronic top quark decays (top jets) and are generated using POWHEG [15, 16] interfaced to PYTHIA6 [17] with the PERUGIA 2012 [18] underlying event tune parameter set and the four flavor scheme of the CT10 PDF set [19]. The sample of top jets is reweighted on a jet-by-jet basis such that the large-R calorimeter jet kinematics in p_{T} and η matches that of the QCD jet sample.

[1] Throughout this work, the Higgs boson intended for tagging is specified by h and refers to the Standard Model Higgs boson with $m_h \sim 125$ GeV.

[2] ATLAS uses a right-handed coordinate system with its origin at the nominal interaction point (IP) in the center of the detector and the z-axis along the beam pipe. The x-axis points from the IP to the centre of the LHC ring, and the y-axis points upwards. Cylindrical coordinates (r, ϕ) are used in the transverse plane, ϕ being the azimuthal angle around the z-axis. The pseudorapidity is defined in terms of the polar angle θ as $\eta = -\ln\tan(\theta/2)$. Angular distance is measured in units of $\Delta R \equiv \sqrt{(\Delta\eta)^2 + (\Delta\phi)^2}$.

For the boosted Higgs jets, a sample of high p_{T} Higgs bosons is obtained from the BSM physics simulation of a Randall-Sundrum graviton (G^*) [20] decaying to a pair of Higgs bosons with both Higgs bosons subsequently decaying to $b\bar{b}$ pairs with the coupling[3] $\kappa/\bar{M}_P$=1 and the Higgs boson mass set to 125 GeV. This process is generated using MadGraph5 [21] interfaced with PYTHIA8 and with the ATLAS A14 tune, and NNPDF2.3 LO PDF set. The mass splitting between the massive graviton and the Higgs bosons provides a boost proportional to the mass of the G^*. Therefore, signal samples have been generated with graviton masses between 300 GeV and 6000 GeV to fully populate the kinematic region of interest. The samples of various masses are merged and reweighted on a jet-by-jet basis such that the large-R calorimeter jet kinematics in p_{T} and η in signal (defined in Section 3) matches that of the QCD jet sample. This reweighting is intended to mitigate the effects of any bias imposed on account of kinematic differences present between the unweighted samples. EvtGen [22] is used in all of the samples to model the decays of b- and c-flavoured hadrons.

The Monte Carlo samples are processed through the full ATLAS detector simulation [23] based on Geant4 [24]. Additional simulated proton–proton collisions generated using PYTHIA8 with the A2M tune [25] and MSTW2008LO PDF set [26] are overlaid to simulate the effects of additional collisions from the same and nearby bunch crossings (pile-up), with a mean number of 22 collisions per bunch crossing, the average in the combined 2015 and 2016 data taking period. All simulated events are then processed using the same reconstruction algorithms and analysis chain as would be used for real data.

3. Event and Object Reconstruction and Selections

The goal of this work is to improve the discrimination of Higgs jets from jets originating from multi-jet events and $t\bar{t}$ events. The baseline selection and identification procedure for jets follow closely those described in Ref. [6] and are summarised briefly in this section.

The reconstruction and identification of $h \to b\bar{b}$ decays involves a two step procedure. The first step is to reconstruct the Higgs boson kinematics by clustering its hadronic decay products within a single large radius (large-R) calorimeter jet. Calorimeter jets are initially reconstructed from noise suppressed topological clusters of calorimeter cells which are calibrated to the hadronic energy scale using the local calibration weighting method [27]. These form the set of constituents from which large-R calorimeter jets are reconstructed using the anti-k_t algorithm [28] with a radius parameter of $R = 1.0$ and further trimmed [29] to remove the effects of pileup and the underlying event. Trimming is a grooming technique in which the original constituents of a large-R jet are reclustered using the k_t algorithm [30] with a distance parameter R_{sub} in order to reconstruct a collection of subjets within the large-R jet. These subjets are then discarded if they carry less than a specific fraction (f_{cut}) of the p_{T} of the original large-R jet. The optimised values of the trimming parameters are $R_{\mathrm{sub}} = 0.2$ and $f_{cut} = 5\%$ [31]. The large-R calorimeter jet energy and mass scale are then calibrated to the particle level using correction factors derived from simulation [32, 33]. From this set of reconstructed large-R calorimeter jets, only those satisfying $p_{\mathrm{T}} > 250$ GeVand $|\eta| < 2.0$ are retained for further analysis. Where appropriate, the calorimeter jet mass is required to be between 76 GeV and 146 GeV which corresponds to a 90% signal selection efficiency to be consistent with the mass of the Higgs boson [6].

[3] The coupling $\kappa/\bar{M}_P$ is the free parameter of the Randall-Sundrum model where κ is a curvature parameter of the model and $\bar{M}_P$ is the Planck mass scale.

After the reconstruction of the Higgs boson kinematics, the next step is to identify the presence of two b-hadrons within the large-R calorimeter jet by reconstructing subjets containing the b-hadrons. The baseline subjet reconstruction algorithm in ATLAS, as presented in Ref. [6], is to use track jets clustered with the $R = 0.2$ anti-k_t jet algorithm. The inner detector tracks which are used as inputs to the jet algorithm are required to have $p_{\mathrm{T}} > 0.4$ GeV and $|\eta| < 2.5$. In addition, the tracks must have at least 7 hits in total in the pixel and SCT detectors, not more than one hit in the pixel detector shared by multiple tracks, not more than one missing hit in the pixel detector when it is expected, and not more than two missing hits in the SCT detector. The longitudinal impact parameter (z_0) of the tracks is required to be $|z_0 \cdot \sin\theta| < 3$ mm, where measurements are made with respect to the location of the primary vertex, determined as the reconstructed vertex with the highest scalar sum of the p_{T} of associated tracks. These requirements greatly reduce the number of fake tracks and tracks from pileup vertices, which ensures that the reconstructed track jets originated from the hard scatter vertex. The track jets are associated to the large-R calorimeter jets by using the ghost association method [34, 35], which provides a robust matching procedure that makes use of the catchment area of the untrimmed large-R calorimeter jet. Track-based subjets with $p_{\mathrm{T}} > 10$ GeV, $|\eta| < 2.5$ and having at least two track constituents are considered in this note.

As alternatives to the baseline subjet reconstruction algorithm above, we present three new algorithms to improve Higgs jet tagging performance at high p_{T}. The first algorithm extends the baseline algorithm by using the variable radius (VR) jet algorithm [36] to reconstruct the track jets, with the same input track selections and large-R calorimeter jet association as the baseline algorithm. The other two algorithms utilize the calorimeter cell cluster constituents of the trimmed large-R calorimeter jet to reconstruct the subjets, with one of the algorithms using the exclusive-k_t jet algorithm (ExKt) while the other reconstructs the subjets in the rest frame of the large-R calorimeter jet (the center-of-mass frame, CoM). For both algorithms, the calorimeter-based subjets are required to have $p_{\mathrm{T}} > 5$ GeV and $|\eta| < 2.5$. These algorithms, as well as the baseline algorithm, are described in greater detail in Section 4 and their performances are compared in Section 5.

After the reconstruction of the subjets, the identification of the flavour of the subjets is determined using a multivariate approach based on track and vertex information from the tracks associated with a given subjet [37]. For all the subjet algorithms, except for the CoM algorithm, tracks are associated exclusively to a subjet based on their angular separation, ΔR(track, subjet), and the ΔR selection decreases with increasing subjet p_{T} according to Equation 1.

$$\Delta R(\text{track, subjet}) < 0.239 + e^{-1.220 - 1.64\times10^{-5}\cdot p_T} \quad p_T[\text{MeV}] \tag{1}$$

For high p_T jets, the ΔR cut plateaus at a value of 0.239. The subjets from the CoM algorithm collect tracks in the center-of-mass frame of the large-R calorimeter jet within a fixed cone around the subjet axis. Details about the track association in the CoM algorithm can be found in Section 4.3. The tracks associated to the jet are selected from a different set of tracks than those used for track jet reconstruction and follow the selection described in Reference [38]. These tracks are then provided as inputs to b-jet identification algorithms [37] that individually exploit impact parameter information (IP2D, IP3D), secondary vertex information (SV1), and b- to c-hadron decay chain information (JetFitter) in a number of ways. The MV2c10 algorithm then combines information from these algorithms in a boosted decision tree which is trained to discriminate b-jets from a background sample composed of approximately 93% light flavour jets and 7% c-jets. This algorithm is described in more detail in Ref. [39].

The flavour-labelling of the large-R calorimeter jets and subjets is characterized by geometrically matching generator-level (truth) particles to the jets. A truth particle is associated to a large-R calorimeter jet if

it is within $\Delta R < 1.0$ of the large-R calorimeter jet axis and it is associated to a subjet if it is within $\Delta R < 0.3$ of the subjet axis. If the truth particle is within $\Delta R < 0.3$ of more than one subjet, it is associated to the subjet closest in ΔR. A *Higgs jet* is specifically defined as a reconstructed jet with $|\eta|<2.0$ and p_{T} >250 GeV that is matched to two b-hadrons (b_1, b_2), which are required to have $p_{\mathrm{T}} > 5$ GeV, and one Higgs boson. In Sections 4 and 5, an additional labelling technique is used to more closely mimic the large-R jet identification criteria using regions of interest defined by subjets while removing reconstruction deficiencies entering in the determination of the b-tagging discriminant. This labelling is referred to as *double subjet b-labelling* and defines a Higgs jet as a jet in which the two b-hadrons are exclusively matched to the two leading p_{T} subjets (j_1, j_2) with $\Delta R(b, j) < 0.3$.

When identifying Higgs jets based on the fully reconstructed flavour tagging information of the subjets, two cases are considered. The first, referred to as a *double tag*, defines a Higgs jet as a large-R jet with at least two subjets where the two leading p_{T} subjets are both b-tagged using the MV2c10 observable. The second, a *single tag*, defines a Higgs jet as a large-R jet with at least one subjet where at least one of the two leading p_{T} subjets is b-tagged using the MV2c10 observable.

4. Subjet Reconstruction Algorithms

The most common jet algorithms are iterative recombination algorithms which rely on two metrics to successively combine protojets[4] into jets. The metrics are defined to be the pairwise distance d_{ij} of two protojets i and j, and the distance d_{iB} between a protojet and the beam axis:

$$d_{ij} = min(p_{\mathrm{T},i}^{2n},\ p_{\mathrm{T},j}^{2n})\ \Delta R_{ij}^2 \tag{2}$$

$$d_{iB} = p_{\mathrm{T},i}^{2n}\ R^2 \tag{3}$$

where $\Delta R_{ij} = \sqrt{(\Delta y_{ij})^2 + (\Delta \phi_{ij})^2}$ denotes the distance between i and j with Δy_{ij} being the difference between the protojet rapidities. The radius parameter R sets the angular scale of the jet algorithm while n determines the type of jet algorithm. The case $n = 0$ corresponds to the Cambridge-Aachen (C/A) algorithm [40], the case $n = 1$ refers to the k_t algorithm [30] and $n = -1$ yields the anti-k_t algorithm [28].

The baseline subjet reconstruction algorithm used in ATLAS is the anti-k_t jet algorithm with the radius parameter fixed to be $R = 0.2$ using tracks as inputs. The fixed-radius track jet approach works well to identify two b-jets within the large-R calorimeter jet until the hadronisation products from the two b-quarks from the Higgs boson decay begin to overlap due to the high p_{T} of the Higgs boson. At this point, the jet algorithm is not able to resolve the two subjets [4]. To mitigate this effect of the fixed angular scale of the jet algoritm, three alternative methods are developed in the following sections.

Throughout this study, the two primary metrics that are used to benchmark the performance of a given algorithm are the signal efficiency, to identify jets seeded by a Higgs boson, and the background rejection,

[4] Protojets can be constructed from either stable truth particles, calorimeter cell clusters, or charged particle tracks. In this work, only the latter two are used.

both for jets seeded by a light quark or gluon (QCD jet) or a top quark. In the case of signal efficiency, this is defined both at truth level and at reconstruction level as

$$\text{Double Subjet B-Labelling Efficiency} = \epsilon_{\text{truth}}^{\text{Double subjet b-label}} = \frac{N(\text{Double subjet b-label} \,|\, \text{Higgs jet})}{N(\text{Higgs jet})} \tag{4}$$

$$\text{Double B-Tagging Efficiency} = \epsilon_{\text{reco}}^{\text{Double b-tag}} = \frac{N(\text{double tag} \,|\, \text{Higgs jet})}{N(\text{Higgs jet})} \tag{5}$$

$$\text{Single B-Tagging Efficiency} = \epsilon_{\text{reco}}^{\text{Single b-tag}} = \frac{N(\text{single tag} \,|\, \text{Higgs jet})}{N(\text{Higgs jet})} \tag{6}$$

where N represents the number of jets passing the specified requirements described in Section 3. In the case of background rejection for QCD and top jets, the rejection is defined as the inverse of the efficiency to accept a background jet, where the initial ensemble of QCD or top jets is taken as the inclusive set of jets passing the kinematic requirements imposed on large-R jet p_{T} and η.

4.1. Variable-R Track Jets

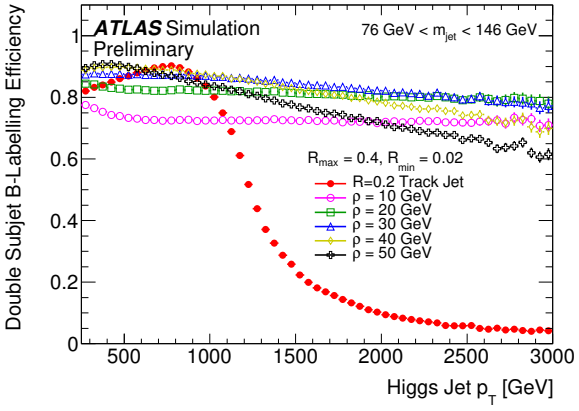
The first alternative to fixed radius track-based subjets is to apply the technique of using a "variable" radius jet algorithm. Initially described in Ref. [36], this algorithm modifies the conventional iterative recombination algorithm by making the radius parameter a function of the jet p_{T} as:

$$R \longrightarrow R_{\text{eff}}(p_{\text{T}}) = \frac{\rho}{p_{\text{T}}} \tag{7}$$

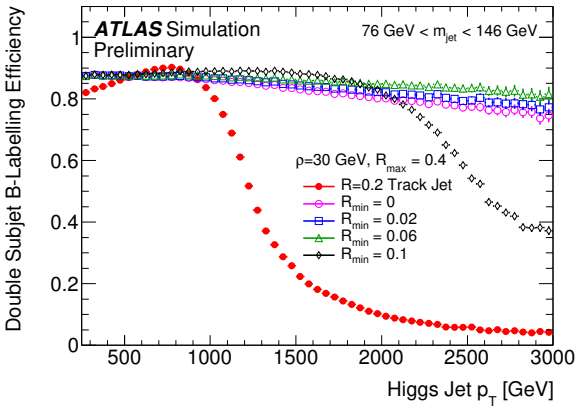
The new parameter ρ determines how fast the effective jet size decreases with the transverse momentum of the jet. In addition to ρ, the VR algorithm requires two additional parameters, R_{min} and R_{max}, to impose lower and upper cut-offs on the jet size, respectively. The additional parameters prevent the jets from becoming too large at low p_{T} and from shrinking below the detector resolution at high p_{T}. The effective jet size varies smoothly between R_{min} and R_{max}. In principle, VR formulations of C/A and k_t algorithms are also possible but are beyond the scope of this note.

The ρ, R_{min} and R_{max} parameters are scanned to find physically motivated and sensible values for the parameters of the VR jet algorithm that are to be used in reconstructing track jets from boosted $h \to b\bar{b}$ decays. The scan for each parameter is performed by examining the truth subjet double b-labelling efficiency of Higgs jets for different values of a given parameter while using fixed values for the other parameters. The optimal value for each parameter is then chosen to be the value which gives the highest truth subjet double b-labelling efficiency.

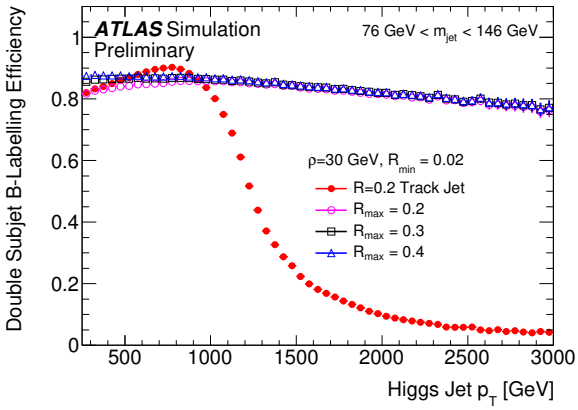
In Figure 1, the truth subjet double b-labelling efficiency is shown for $R = 0.2$ track jets as a function of the Higgs jet p_{T} as a benchmark for comparing multiple VR track jet reconstruction techniques with different combinations of VR algorithm parameters. As seen in Figure 1(a), VR track jets with $\rho = 30$ GeV have the best truth subjet double b-labelling efficiency over the largest Higgs jet p_{T} range. In Figure 1(b), the configuration for VR track jets with $\rho = 30$ GeV and $R_{\text{max}} = 0.4$, the $R_{\text{min}} = 0.06$ value is observed to be the optimal value because the truth subjet double b-labelling efficiency is the most stable as a function of Higgs jet p_{T} compared to other R_{min} values. The R_{max} value of 0.4 for VR track jets



(a)



(b)



(c)

Figure 1: Efficiency of subjet double b-labelling at the truth level of a Higgs jet as a function of the Higgs jet p_T. (a) The efficiency for VR track jets with R_{min} = 0.02 and R_{max} = 0.4 for several ρ values. (b) The efficiency for VR track jets with ρ = 30 GeV and R_{max} = 0.4 for different values of R_{min}. (c) The efficiency for VR track jets with ρ = 30 GeV and R_{min} = 0.02 for varying values of R_{max}. The efficiency for R = 0.2 track jets is also included in all of the plots. The error bars include statistical uncertainties only.

with ρ = 30 GeV and R_{min} = 0.02 gives the highest truth subjet double b-labelling efficiency across the whole Higgs jet p_{T} range as shown in Figure 1(c). Furthermore, it can be observed in Figure 1(c) that the truth subjet double b-labelling efficiency decreases at low Higgs jet p_{T} as the values of R_{max} is decreased, converging to that of the fixed radius track jets for R_{max}=0.2. This decrease in performance is expected for lower p_{T} Higgs jets because the jet is composed of more spread out, lower p_{T} constituents which, when forced to cluster with a smaller radius algorithm, tend to spawn a greater number of associated subjets, leading to an increased probability that the b-hadrons will not be associated to the leading two subjets. Allowing the subjets to grow beyond this bound, as is done with the VR algorithm, mitigates this effect. Lastly, for all VR parameter combinations, it can be seen that the truth subjet double b-labelling efficiency using VR track jets is substantially higher at high Higgs jet p_{T} compared to R = 0.2 track jets. For Higgs jets with 600 GeV $< p_T <$ 1000 GeV, the R = 0.2 track jet approach performs better than the VR track jet approach, and this is studied in more detail in Section 5.

4.2. Exclusive-k_t Subjets

Another alternative to fixed radius track jets is the exclusive-k_t algorithm, which is a variation of the k_t jet algorithm. While the exclusive-k_t algorithm uses exactly the same distance metric as the ordinary inclusive k_t jet algorithm, it will not stop clustering constituents until either all d_{ij} are above a certain threshold, or a certain fixed number of jets are obtained. For the purpose of $h \to b\bar{b}$ tagging, the fixed number of jets is set to 2, and three different sets of constituents have been tried as input to the ExKt algorithm in order to better understand its performance. These three ExKt variants are:

- Calorimeter subjets re-clustered from the trimmed large-R calorimeter jet constituents ("ExKt (Trimmed)"): For a large-R calorimeter jet, many calorimeter cell clusters are from pile-up and underlying event activities, which leads to incorrect subjet reconstruction. The trimming technique [29] is able to remove activity unrelated to the $h \to b\bar{b}$ hard process inside the large-R calorimeter jet before any splitting procedure is performed.
- Calorimeter subjets re-clustered from the untrimmed large-R calorimeter jet constituents ("ExKt (Untrimmed)"): It has been found that the trimming procedure might incorrectly remove decay products from soft b-hadrons from the $h \to b\bar{b}$ decay, especially when the boost direction of the Higgs boson is anti-parallel to the b-hadron direction in the center-of-mass frame of the Higgs boson [4]. Exclusive-k_t subjet finding would consequently fail to divide the large-R jet into two subjets corresponding to the two b-hadrons. Therefore, applying ExKt to the untrimmed large-R jet is also studied here.
- Subjets re-clustered from tracks ghost associated to the untrimmed large-R calorimeter jet ("ExKt Track Jets"): One technique that might overcome the problems from the two methods previously described is to use tracks ghost associated to the untrimmed large-R jet as inputs to the exclusive-k_t algorithm. Due to the excellent resolution of the tracking system, it is possible to see whether or not a track comes from pile-up by extrapolating it to the primary vertex. The selection of tracks is the same as that used for R = 0.2 and VR track jets, and is described in Section 3. However, there will still be some inefficiency when using tracks due to the loss of information from neutral particles.

A significant feature of the exclusive-k_t algorithm is that, unlike an inclusive algorithm, it does not have a distance parameter. This suggests that there is no intrinsic angular lower bound on the distance between

subjets. Furthermore, since exclusive-k_t with exactly two subjets is equivalent to reversing the jet clustering with the k_t algorithm for the last clustering step, exclusive-k_t effectively splits the large-R calorimeter jet into two parts, each of which is used to define a region of interest that is expected to contain one b-hadron from the $h \to b\bar{b}$ decay. This technique is particularly useful in the case of very high p_T Higgs jets because the jet is divided into two components by construction, thereby more closely resembling the $h \to b\bar{b}$ topology.

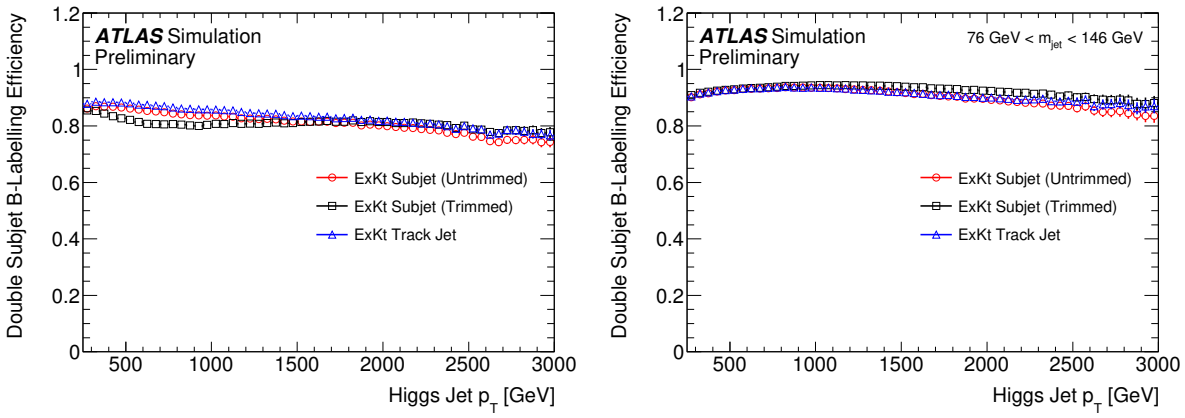


Figure 2: For the set of large-R jets before (left) and after (right) the Higgs boson mass cut $76 < m_H < 146$ GeV, the probability that both leading and subleading subjets have exactly one b-hadron matched for different exclusive-k_t variations, as function of Higgs jet p_{T}. The error bars include statistical uncertainties only.

In practice, the exclusive-k_t algorithm might not split the large-R calorimeter jet appropriately : the algorithm might incorrectly reconstruct the two subjets, with both b-hadrons matched to only one of the subjets. This would mostly be due to detector resolution, pile-up, underlying event and parton shower. Nevertheless, Figure 2 shows the truth subjet double b-labelling efficiency as a function of Higgs jet p_T for both the inclusive sample of Higgs jets described in Section 3 and only those that satisfy the Higgs boson mass cut $76 < m_H < 146$ GeV and shows that the exclusive-k_t algorithm can in general maintain a very high probability of correct splitting, even in the case of extremely high p_{T} Higgs jets.

Furthermore, a large difference on the truth labelling efficiency can be seen in Figure 2 before and after applying the Higgs mass window cut. Before the mass window cut, the truth labelling efficiency is around 80% when subjets are built from trimmed jet constituents. This increases to around 90% by using subjets associated to the untrimmed jet, due to efficiency recovery from soft b-hadrons that are trimmed out. Since jets with these soft b-hadrons trimmed out typically fall below the Higgs mass window, such an efficiency margin between two different variations of exclusive-k_t gets much smaller when the Higgs mass window cut is applied. Moreover, it is observed that in this kinematic region of interest near the Higgs mass, the expected improvement against increased resolution coming from using track jets is not greater than by grooming the calorimeter jet and the truth labelling efficiency becomes flat between $90 \sim 95\%$ across a wide range of large-R calorimeter jet p_{T} after the Higgs mass window cut is applied.

Figure 3 shows the distribution of ΔR between the subjet axis and the associated b-hadron flight direction. This can provide another view on how the b-hadron is reconstructed by the exclusive-k_t algorithm once the splitting is done. A good ΔR alignment is observed across various p_{T} ranges. Furthermore, it can be seen that the alignment of ΔR is much better when tracks are used as the input to the exclusive-k_t algorithm. It has been observed in previous ATLAS studies that track jets are able to obtain a better

ΔR resolution than calorimeter-based jets [4]. However, given the alignment of ΔR is already very good, with position of peak less than 0.02 in most cases, differences of ΔR alignment as observed here would not necessarily lead to improvements in the b-tagging performance, as the tracks used for b-tagging are gathered in a cone around the jet axis which shrinks as a function of jet p_T until it plateaus at a ΔR value slightly greater than 0.2, as described in Section 3.

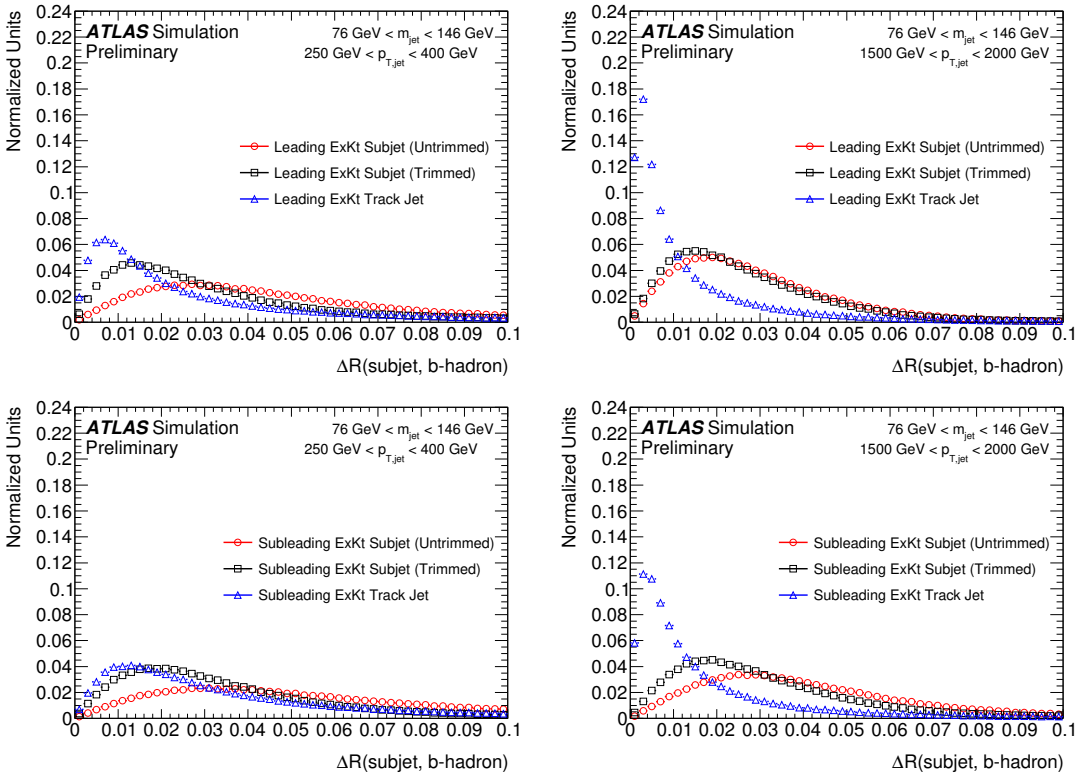


Figure 3: ΔR distribution between reconstructed subjet axis and associated b-hadron flight direction for low and high p_{T} regions for leading and subleading exclusive-k_t subjets. The error bars include statistical uncertainties only.

Figure 4 shows the MV2c10-based double b-tagging QCD jet rejection as a function of Higgs jet p_T for a fixed Higgs jet efficiency of 50% for all variations of the exclusive-k_t algorithm, both with and without a mass cut. Analogous plots for top jet rejection are shown in Figure 5. As seen in these figures, the performance is not optimal for a single choice of inputs for the subjet reconstruction when examining the background rejection. Instead, the optimal choice depends both on the type of background jet against which the identification technique is discriminating as well as the kinematic regime. For instance, when comparing the rejection power against QCD jets prior to the application of a Higgs jet mass window cut, at low p_{T} the "ExKt (Untrimmed)" algorithm is optimal whereas at higher p_{T} the "ExKt Track Jet" algorithm is optimal. However, after applying the Higgs jet mass window cut, the performance difference between the three algorithms decreases with the "ExKt (Trimmed)" subjet choice performing better at low p_{T} both in the case of QCD jets and top quark jets. Because of this observation, along with the fact that the "ExKt (Trimmed)" subjet collection is found to have the best truth subjet double b-labelling efficiency after the Higgs mass window cut is applied as in Figure 2, the "ExKt (Trimmed)" subjet collection is used for the rest of the studies presented in this note.

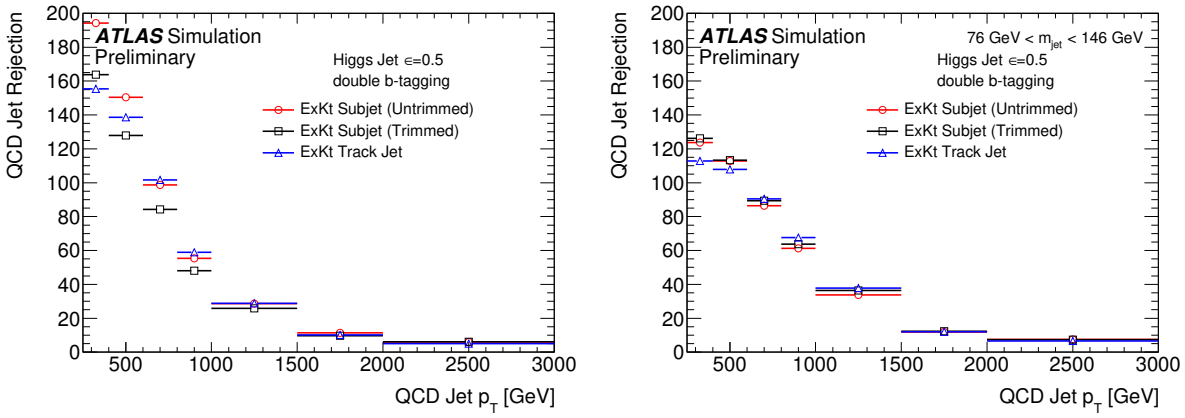


Figure 4: For the ExKt subjet b-tagging algorithm using three different sets of subjet inputs (calorimeter cell clusters from the ungroomed jet, calorimeter cell clusters from the trimmed jet, and charged tracks), the rejection against QCD jet background as function of Higgs jet p_T for a fixed Higgs jet double b-tagging efficiency of 50% both without (left) and with (right) a Higgs jet mass window requirement. The error bars include statistical uncertainties only.

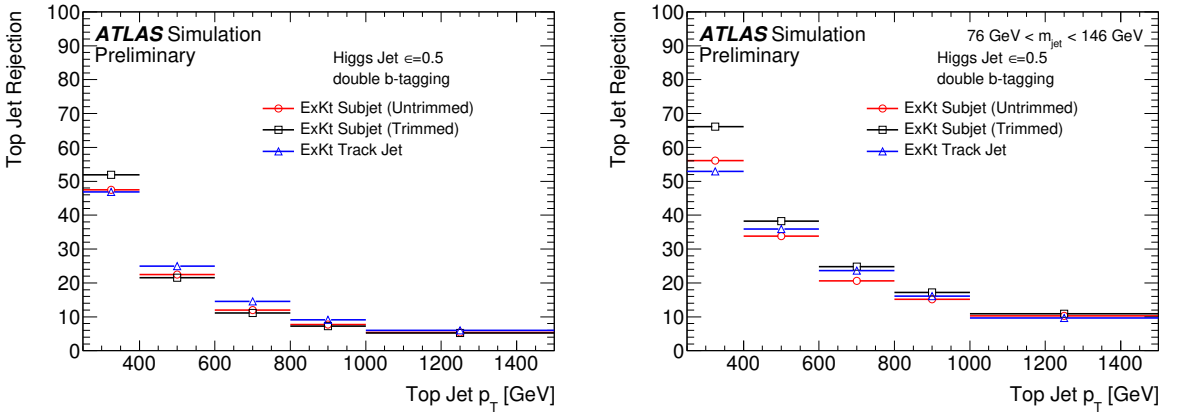


Figure 5: For the ExKt subjet b-tagging algorithm using three different sets of subjet inputs (calorimeter cell clusters from the ungroomed jet, calorimeter cell clusters from the trimmed jet, and charged tracks), the rejection against top jet background as function of Higgs jet p_T for a fixed Higgs jet efficiency of 50% double b-tagging both without (left) and with (right) a Higgs jet mass window requirement. The error bars include statistical uncertainties only.

4.3. Exclusive Center-of-Mass Calorimeter Subjets

Jet substructure information in the center-of-mass frame defined by the reconstructed jet four-momentum, in this case the large-R trimmed jet described in Section 3, has been shown to discriminate effectively between hadronically-decaying high p_T W or Z bosons and QCD jets [41, 42]. This method can also be applied to the identification of a high p_T Higgs boson decaying to a $b\bar{b}$ quark pair [43]. In this approach, the jet constituents belonging to the large-R calorimeter jet are boosted to the center-of-mass frame of the large-R calorimeter jet (*jet rest frame*). In this study, the center-of-mass frame is defined as the reference frame in which the four-momentum of the large-R calorimeter jet is equal to $p_\mu = (m_{\text{jet}}, 0, 0, 0)$ and m_{jet} is the invariant mass of the large-R calorimeter jet. In the jet rest frame, the constituents of the large-R calorimeter jet are reclustered using the EECambridge jet algorithm [44] to form subjets. The EECambridge jet algorithm sequentially combines calorimeter cell clusters with the smallest angular separation $y_{ij} = 2 \times (1 - \cos\theta_{ij})$, where θ_{ij} is the angle between the momenta of the i^{th} and j^{th} energy cluster in the jet rest frame. The algorithm stops either when y_{ij} exceeds the cut-off value $y_{\text{cut}}^{\text{subjet}}$ or a certain fixed number of subjets is reached, and in this case, as with the exclusive-k_t algorithm described in Section 4.2, the CoM clustering algorithm stops when exactly two subjets are found, one for each of the b-quarks from the Higgs boson decay.

The track-to-jet association also adopts the idea of boosting the constituents, in this case charged ID tracks, to the CoM of the large-R calorimeter jet. The following procedure is applied to each event:

- All tracks which pass the pre-selection, detailed in Section 3, are considered in turn. For the large-R calorimeter jet under study, those tracks satisfying ΔR(track, large-R calorimeter jet) < 1.0 are associated to the large-R calorimeter jet in the laboratory frame.
- The tracks matched to the large-R jet are boosted into the jet rest frame.
- The angular distance y_{ij} is calculated between each track and subjet in the CoM, where i represents the i^{th} track and j the j^{th} subjet. When $y_{ij} < y_{\text{cut}}^{\text{track}}$, the i^{th} track is associated to the j^{th} subjet. The parameter $y_{\text{cut}}^{\text{track}}$ is a cone size parameter analogous to the R parameter in the anti-k_t algorithm. Each track can only be associated to the subjet which is closest to it.
- The subjets and tracks are boosted back to the laboratory frame after association.

The tracks associated to each subjet after this association procedure are used in the calculation of the flavour tagging weight, MV2c10. The CoM algorithm, by definition, has a variable cone size for track association in the lab frame, as the cone size ($y_{\text{cut}}^{\text{track}}$) is fixed in the center-of-mass frame. When the subjets and tracks are boosted back to the CoM frame, the cone size depends on the four momentum of the Higgs candidate, the large-R jet, and the decay angle of the b-quarks from the Higgs direction. In this sense, the CoM algorithm can be viewed as an adaptive variable cone size algorithm in the lab frame in which the cone size depends on the jet momentum and Higgs decay topology.

To determine the value to use for the $y_{\text{cut}}^{\text{track}}$ parameter, the double b-tagging background rejection at fixed signal efficiencies of 50% and 70% are shown for a few $y_{\text{cut}}^{\text{track}}$ values (0.6, 0.7, 0.8 and 0.9) in Figure 6. No strong trend is seen in the performance when scanning this parameter. At a fixed signal efficiency of 50%, $y_{\text{cut}}^{\text{track}} = 0.6$ performs better in terms of QCD jet rejection for $p_{\text{T}} < 800$ GeV whereas there is little dependency on this parameter when evaluating the top jet rejection. For the signal efficiency working points tested, no $y_{\text{cut}}^{\text{track}}$ parameter is found to be optimal across the whole p_{T} region and $y_{\text{cut}}^{\text{track}} = 0.8$ is chosen as the configuration used for CoM subjet reconstruction.

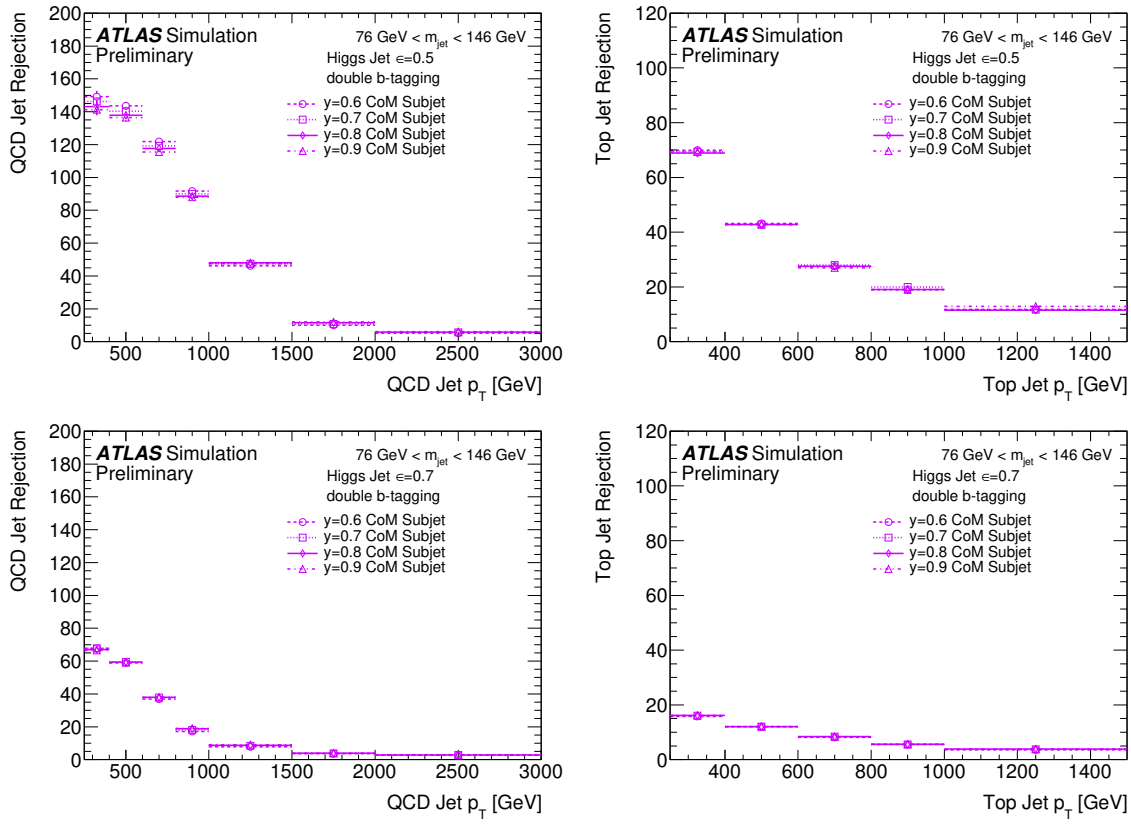


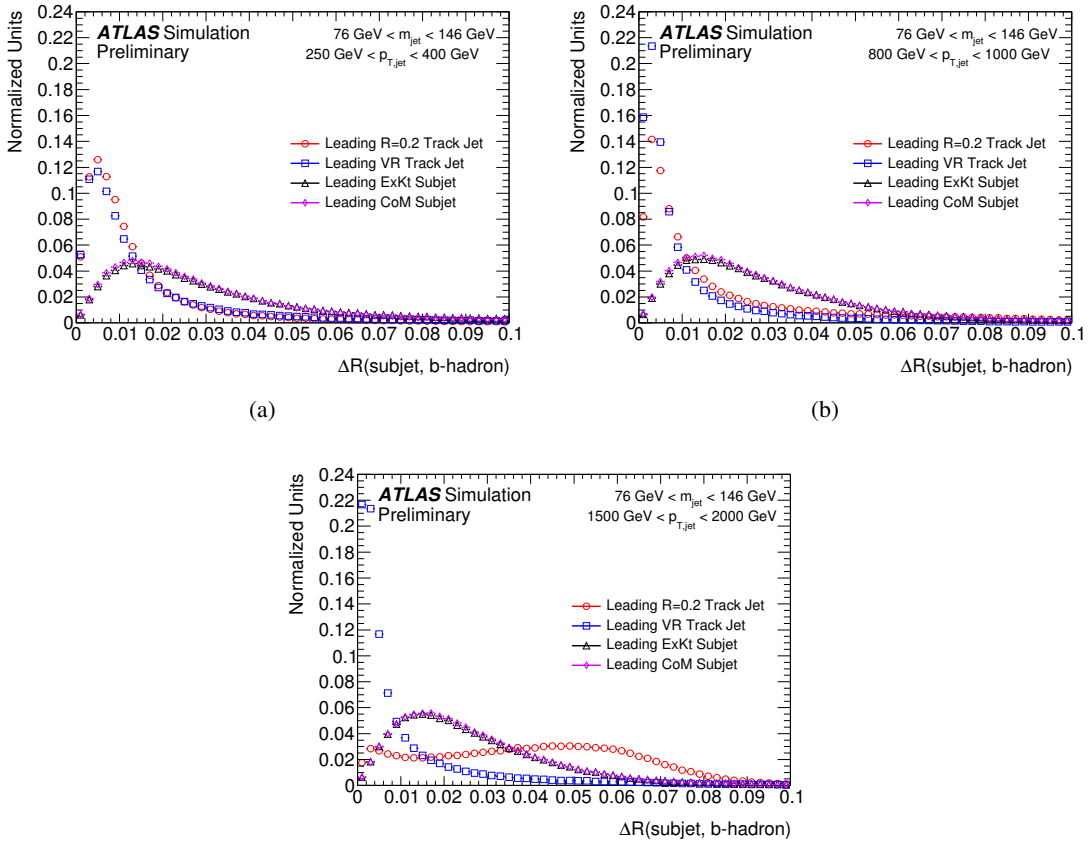
Figure 6: Double b-tagging rejection of QCD jets (left) and top jets (right) at fixed signal efficiencies of 50%(top) and 70%(bottom). The track-to-subjet association cone size parameter, $y_{\text{cut}}^{\text{track}}$, in the CoM method is studied. Values of $y_{\text{cut}}^{\text{track}}$ are varied from 0.6 to 0.9 with a step of 0.1. At the signal efficiency of 50% $y_{\text{cut}}^{\text{track}} = 0.6$ performs better in terms of background rejection at $p_{\text{T}} < 800$ GeV. The error bars include statistical uncertainties only.

5. Results

To benchmark the performance of the VR, exclusive-k_T, and CoM algorithms, it is useful to compare performance metrics both with b-tagging and without b-tagging. Without directly applying the MV2c10 b-tagging discriminant it is still possible to study how well the alternative algorithms reconstruct the b-hadrons from Higgs boson decays using the b-hadron truth information. These comparisons are useful because the MV2c10 b-tagging performance itself drops at high p_T and additionally the MV2c10 training, which is taken from the standard training for anti-k$_T$ R=0.4 jets built from calorimeter cell clusters, is expected to be suboptimal when applied to both the standard $R = 0.2$ track jets and the subjets from the alternative algorithms. This is because the training includes kinematic observables which behave differently for different types of jet collections; these differences may propagate into the final performance obtained for the different subjet types. However, it is still important to apply MV2c10 to the alternative algorithms in order to benchmark their performance in as close a way as possible to what would be done in a physics analysis.

5.1. Comparisons without *b*-tagging

One metric of *b*-hadron reconstruction performance is the ΔR separation between a subjet axis and the nearest matched truth *b*-hadron. For Higgs jets with low p_T (250 GeV $< p_T <$ 400 GeV), $R = 0.2$ and VR track jets reconstruct axes closer to the truth *b*-hadrons than the calorimeter-based exclusive-k_T and center-of-mass subjets, as can be seen in Figures 7(a) and 8(a). As mentioned previously in the discussion of ExKt in Section 4, it is a well-known feature that track jets tend to have a better ΔR resolution than calorimeter-based jets. For Higgs jets with high p_T (1500 GeV $< p_T <$ 2000 GeV), the leading $R = 0.2$ track jet's ΔR separation shows a pronounced shoulder as in Figure 7(c), which corresponds to cases where the two *b*-hadrons become reconstructed as one track jet which is off-axis from both of the *b*-hadrons. However, the VR, exclusive-k_T, and CoM ΔR distributions exhibit the same qualitative features from low to high large-R jet p_T. For these same high p_T Higgs jets, the subleading $R = 0.2$ track jet retains a small ΔR separation to the truth *b*-hadron, however reconstructing two $R = 0.2$ track jets for Higgs jets in this p_T region is rare, as is indicated by the small area under the normalized distributions in Figure 8(c).



(a) (b)

(c)

Figure 7: Distributions of the ΔR between leading subjets and matched truth *b*-hadrons for three different Higgs jet p_T bins. The error bars include statistical uncertainties only. All algorithms have been normalized to an area corresponding to the fraction of signal jets which contain a leading subjet.

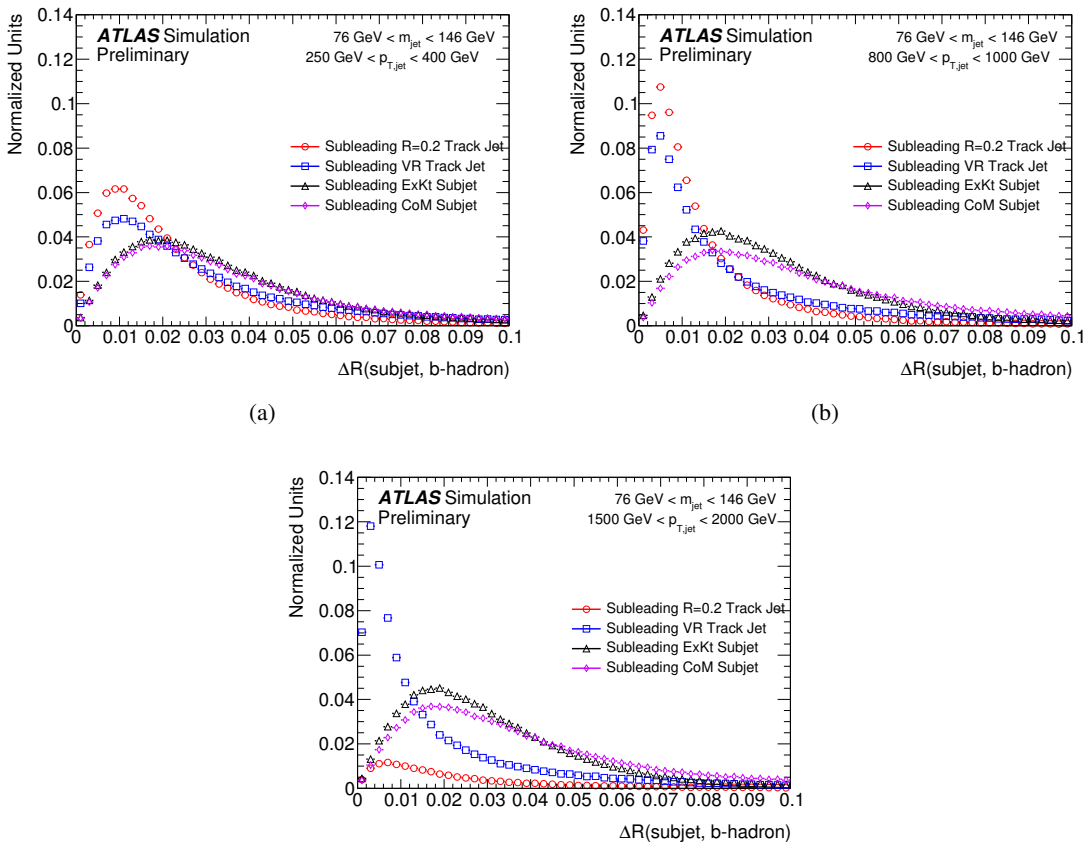


Figure 8: Distributions of the ΔR between subleading subjets and matched truth b-hadrons for three different Higgs jet p_T bins. The error bars include statistical uncertainties only. All algorithms have been normalized to an area corresponding to the fraction of signal jets which contain a subleading subjet.

Another useful metric is the ΔR separation between the subjet axes themselves. The ΔR between the truth b-hadrons decreases with increasing Higgs boson p_T, and this behavior should be apparent in the reconstructed subjets. As can be seen from Figure 9, for Higgs jets with 250 GeV $< p_T <$ 1000 GeV, the ΔR of the leading subjets from all techniques match the behavior of the truth b-hadrons well. However, for Higgs jets with $p_T >$ 1000 GeV, only the leading subjets from the alternative subjet finding techniques are able to match the behavior of the truth b-hadrons. This indicates that, in addition to having a low efficiency to reconstruct two subjets in this high p_T regime, when the $R = 0.2$ track jet technique does reconstruct two subjets, the behavior of their ΔR separation does not match that of the truth b-hadrons. Instead, the leading track jet is found in the core of the large-R jet and contains both b-hadrons, as is seen from the poor ability to reconstruct the b-hadron direction in Figure 7(c), whereas the subleading track jet is reconstructed from additional radiation within the jet but far from its core and not coming from a b-hadron.

Perhaps the most direct performance comparison metric between the different subjet reconstruction algorithms which does not require b-tagging is the efficiency for a Higgs jet to have its two leading associated subjets matched to truth b-hadrons. This truth efficiency ($\epsilon_{\text{truth}}^{\text{Double subjet } b\text{-label}}$) metric is plotted as

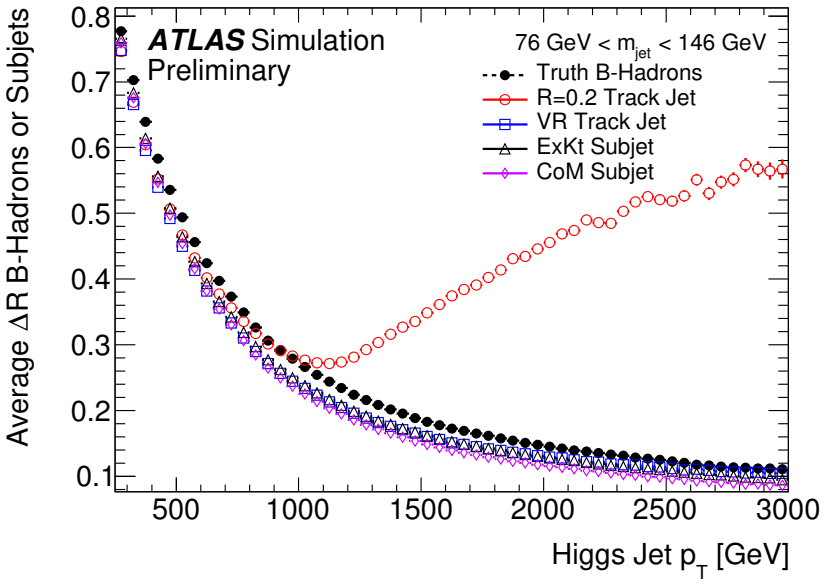


Figure 9: The ΔR between the two leading truth b-hadrons or subjets associated to Higgs jets as a function of Higgs jet p_T. The error bars include statistical uncertainties only.

a function of Higgs jet p_T in Figure 10. For Higgs jets with $p_T < 1000$ GeV, all subjet reconstruction techniques have high truth double b-tagging efficiencies. However, for Higgs jets with $p_T > 1000$ GeV, the efficiency for $R = 0.2$ track jets decreases rapidly, while the other alternative techniques retain a high efficiency up to 3000 GeV.

To better understand why the truth double b-tagging performance of the $R = 0.2$ track jet tagger is better than that of the VR track jet tagger for Higgs jets with 600 GeV $< p_T <$ 1000 GeV, a slight variation of truth double b-tagging is studied. In Figure 11, it is seen that for a large fraction of large-R jets (30% to 40%), there are three VR track jets, as opposed to the expected number of two. In these cases, only considering the two leading p_{T} subjets leads to an inefficiency in correctly labelling the jet as a Higgs jet when this additional track jet is the one that should have been queried for b-labelling. To mitigate this effect, the variation that is implemented considers the three leading subjets, and requires that two out of the three be truth matched to a b-hadron. This truth double b-tagging variant efficiency is plotted vs Higgs jet p_T in Figure 12. The $R = 0.2$ and VR track jet efficiencies are higher when the third leading subjet is considered, and the VR track jet efficiency is larger than the $R = 0.2$ track jet efficiency. This indicates that VR track jets reconstruct one of the b-hadrons as the third leading subjet more frequently than $R = 0.2$ track jets.

In general the alternative subjet reconstruction techniques outperform the $R = 0.2$ track jet technique in terms of b-hadron axis reconstruction for high p_T Higgs jets.

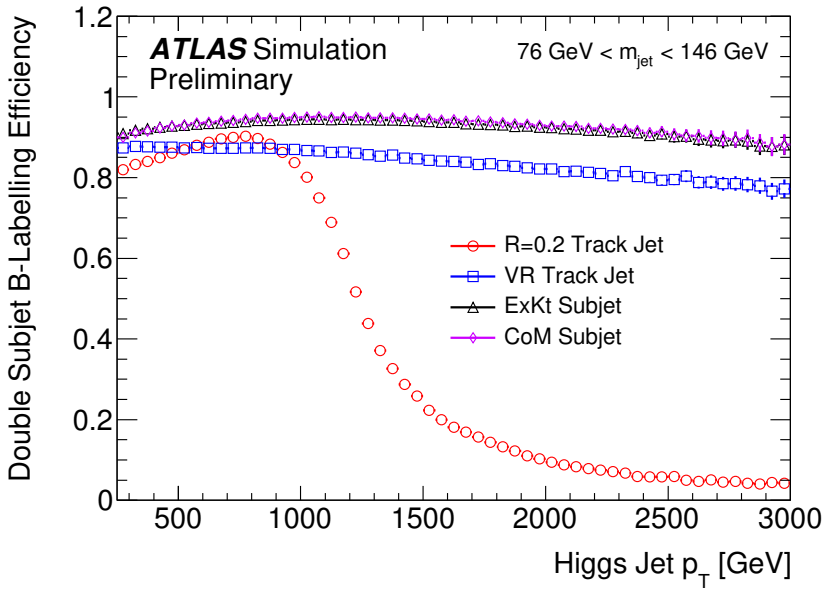


Figure 10: The efficiency for a Higgs jet to have its two leading associated subjets matched to truth b-hadrons vs Higgs jet p_T. The error bars include statistical uncertainties only.

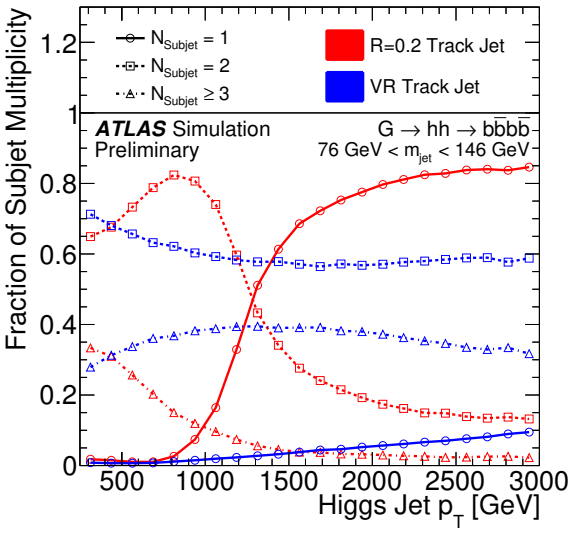


Figure 11: The subjet multiplicity fractions as a function of the Higgs jet p_{T} for Higgs jets with exactly 1 subjet (solid line), 2 subjets (dashed line) and at least 3 subjets (dotted line). The red line refers to fixed radius track jets (anti-k_t R = 0.2) while the blue line refers to variable radius track jets (ρ = 30 GeV and R_{max} = 0.4, R_{min} = 0.02). In this figure, a Higgs jet mass window cut of 76 GeV < m_{jet} < 146 GeV is applied.

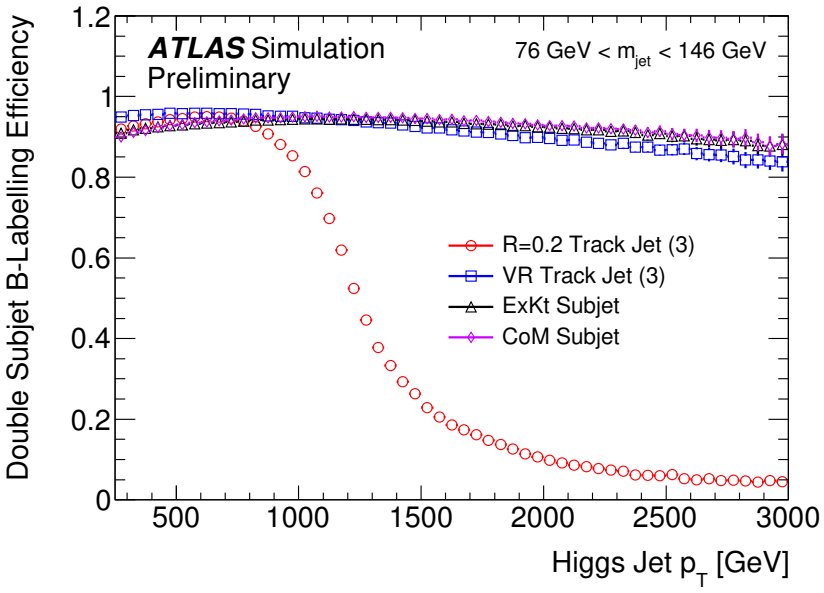


Figure 12: The efficiency for a Higgs jet to have two of the leading three associated subjets matched to truth b-hadrons vs Higgs jet p_T. The error bars include statistical uncertainties only. The results for ExKt and CoM shown here are identical to those in Figure 10 but are presented for comparison.

5.2. Comparisons with b-tagging

Further insight into the performance of these techniques can be gained by applying MV2c10-based b-tagging instead of truth b-labelling, however it must be stressed that the b-tagging MVA is trained on $R = 0.4$ calorimeter jets, so it will be suboptimal on all subjet algorithms considered here.

Based on the comparison results from the previous subsection, it is expected that all subjet techniques will result in similar b-tagging performance for low p_T Higgs jets, while the double b-tagging performance of the $R = 0.2$ track jet technique should degrade significantly more than that of the alternative techniques for high p_T Higgs jets. This can be seen in the receiver operating characteristic (ROC) curves shown in Figures 13 and 14, which benchmark the double b-tagging performance of the algorithms by plotting the QCD jet and top jet rejection as a function of Higgs jet efficiency, respectively. The new taggers outperform the standard $R = 0.2$ track jet tagger increasingly more at higher p_{T}.

The double b-tagging ROC curve for QCD jets in the p_T region 800 GeV $< p_T <$ 1000 GeV confirms that in this region $R = 0.2$ track jets largely outperform VR track jets. However, in the 800 GeV $< p_T <$ 1000 GeV region, $R = 0.2$ track jets also outperform exclusive-k_T subjets. In this p_{T} region, the two b-hadrons from Higgs boson decay start to get closer. This makes it more difficult for a ΔR based track-to-jet matching in the laboratory frame, which happens to VR track jets and exclusive-k_T subjets. With the advantage of the track-to-jet association in the CoM frame of the Higgs boson, where the two b-hadrons are back to back, the performance of the CoM algorithm is better than that of the $R = 0.2$ track jets.

Another interesting note is that from Figures 7 and 8, it is expected that the $R = 0.2$ and VR track jets reconstruct the b-hadron axes better than the calorimeter-based exclusive-k_T and CoM subjets and

should therefore have an improved determination of the MV2c10 b-tagging discriminant on account of improved track-to-subjet association as described in Section 3 as well as determination of fundamental tagger discriminants, like JetFitter and IP3D, which depend on the axis direction [37]. However, because the tracks used by the b-tagging algorithms are collected within a radius around each jet axis which is greater than or equal to 0.239, this improvement in reconstruction will not lead to an improved association and an improved b-tagging discriminant reconstruction. In fact, when considering the performance of the various subjet reconstruction techniques using fully reconstructed b-tagging as in Figures 13 or 14, CoM subjets are found to perform better than VR track jets indicating that this improvement in angular resolution is smaller than the level at which the simple taggers are sensitive.

The single b-tagging performance of the algorithms is benchmarked for QCD and top jet backgrounds in Figures 15 and 16, respectively. Since all algorithms are typically able to reconstruct at least one subjet, the single b-tagging performance of the algorithms are quite similar, though the new taggers do start to show an improvement over the $R = 0.2$ track jet tagger for Higgs jets with high p_{T}.

Another useful way to represent the information in the ROC curves is to pick a fixed Higgs jet efficiency and plot the QCD and top jet rejection as a function of p_T holding the signal efficiency fixed by varying the selection on MV2c10 for the subjets dynamically as a function of the large-R jet p_{T}. Evaluating the QCD and top jet rejection for a fixed Higgs jet efficiency of 50% from the ROC curve data results in Figures 17 and 18 for double and single b-tagging, respectively. In this evaluation of the performance, the selection value of MV2c10 is tuned as a function of p_{T} separately in each case to provide the fixed signal efficiency and is not the same in each case. In comparing Figures 17 and 18 in the case of QCD background jets, the performance of double b-tagging is improved when using the advanced algorithms. However, at high p_{T} (>1500 GeV) it is preferable to apply a single b-tagging selection as this allows for a tighter selection of MV2c10 on either of the two leading subjets, the reconstruction of which has itself been improved by the alternative subjet finding techniques as examined in Section 5.1. However, in the case of background top jets, applying a double b-tagging selection is better than single b-tagging. Therefore, when applying these techniques, it is preferable to have the knowledge of the composition of the background that is faking the Higgs jet signal.

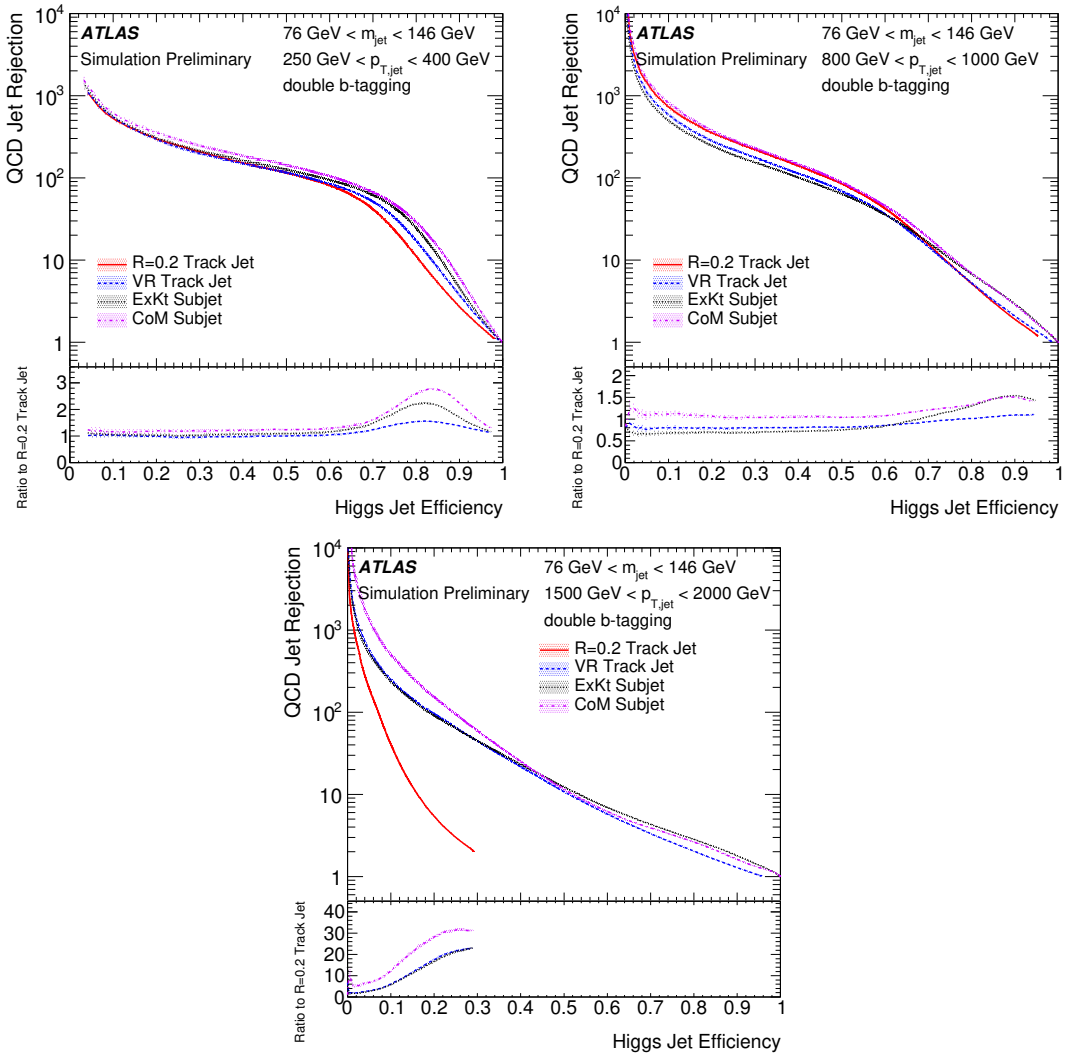


Figure 13: QCD jet rejection as function of $h \to b\bar{b}$ jet efficiency when applying double b-tagging on subjets found by the $R = 0.2$ track jet, VR track jet, exclusive-k_T subjet, and CoM subjet algorithms in different p_T regions. The error bars include statistical uncertainties only.

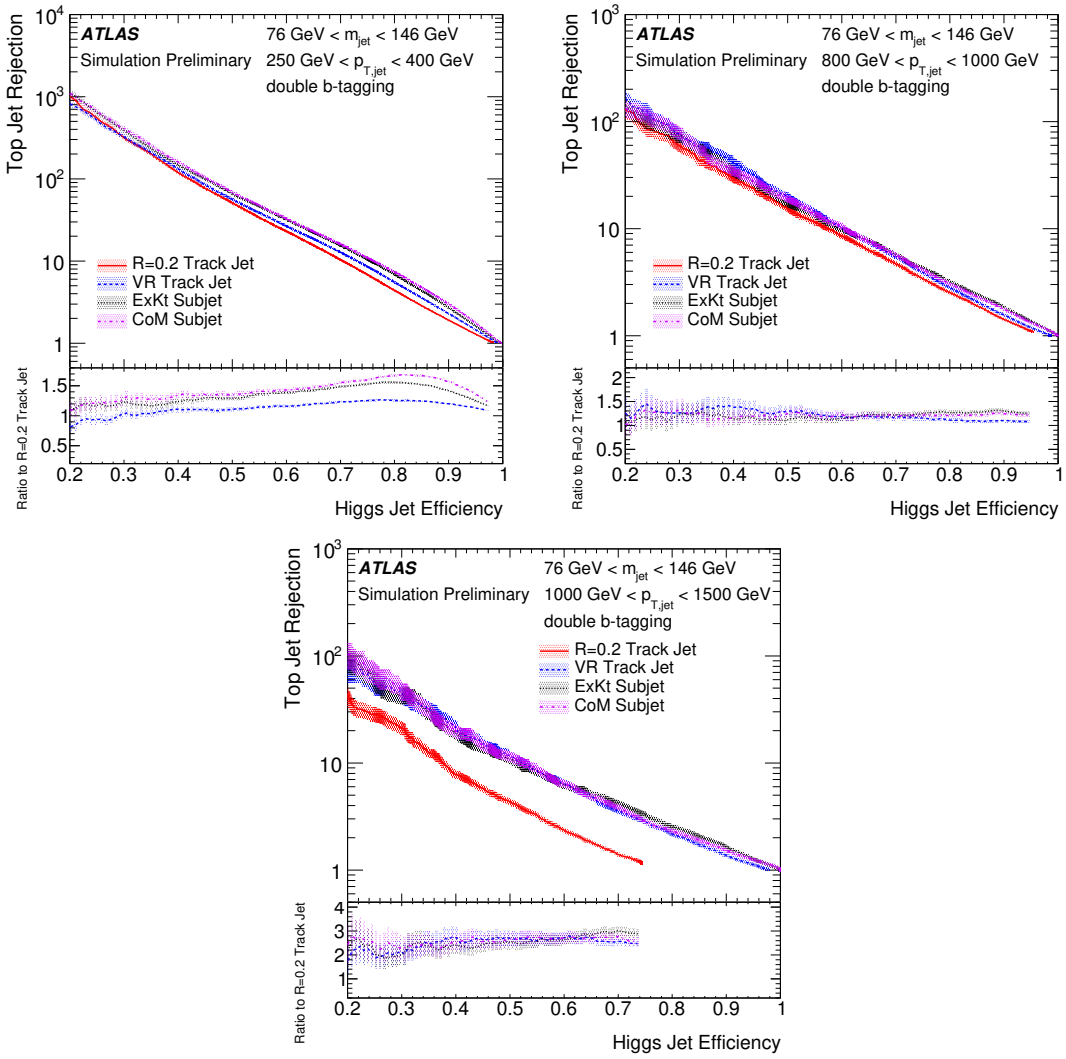


Figure 14: Top jet rejection as function of $h \to b\bar{b}$ jet efficiency when applying double b-tagging on subjets found by the $R = 0.2$ track jet, VR track jet, exclusive-k_T subjet, and CoM subjet algorithms in different p_T regions. The error bars include statistical uncertainties only.

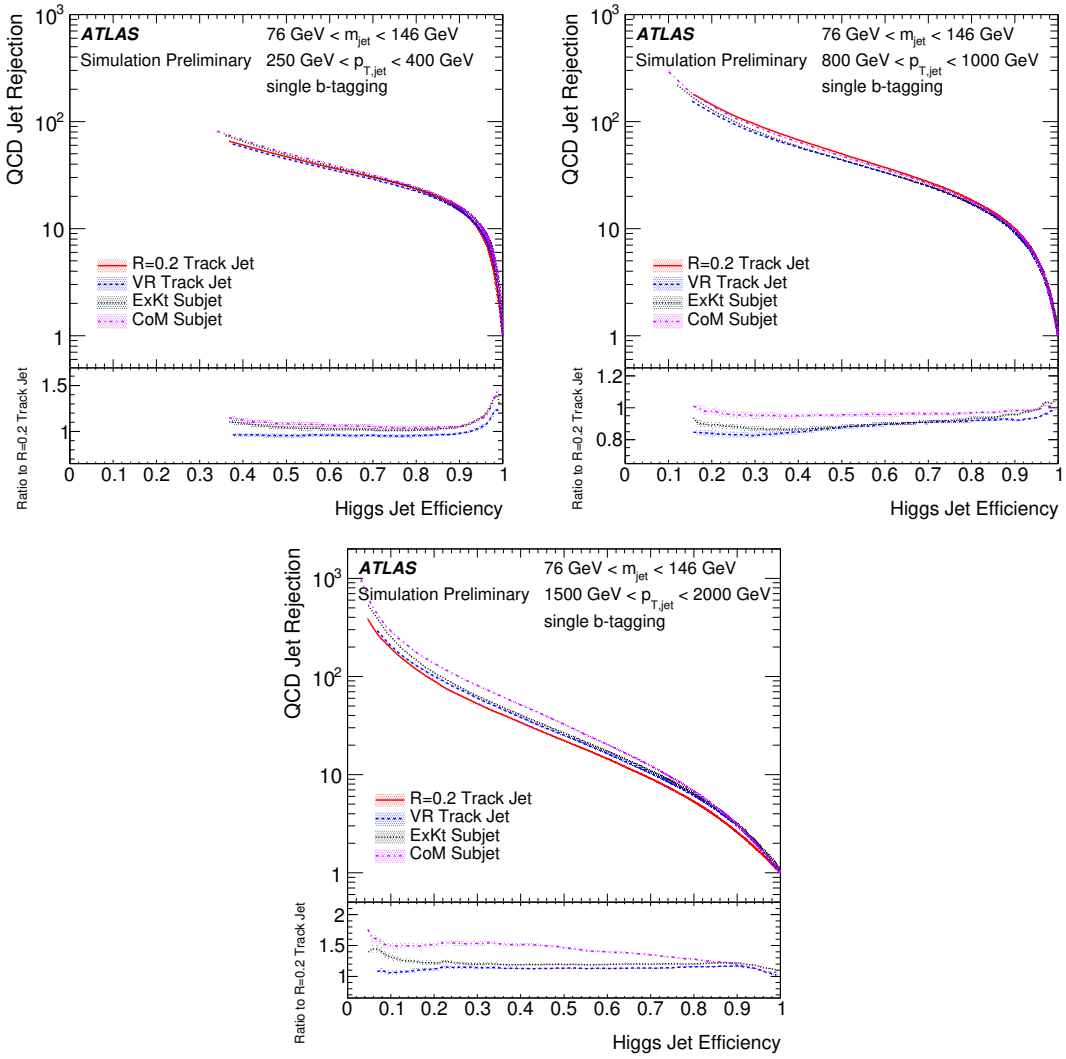


Figure 15: QCD jet rejection as function of $h \to b\bar{b}$ jet efficiency when applying single b-tagging on subjets found by the $R = 0.2$ track jet, VR track jet, exclusive-k_T subjet, and CoM subjet algorithms in different p_T regions. The error bars include statistical uncertainties only.

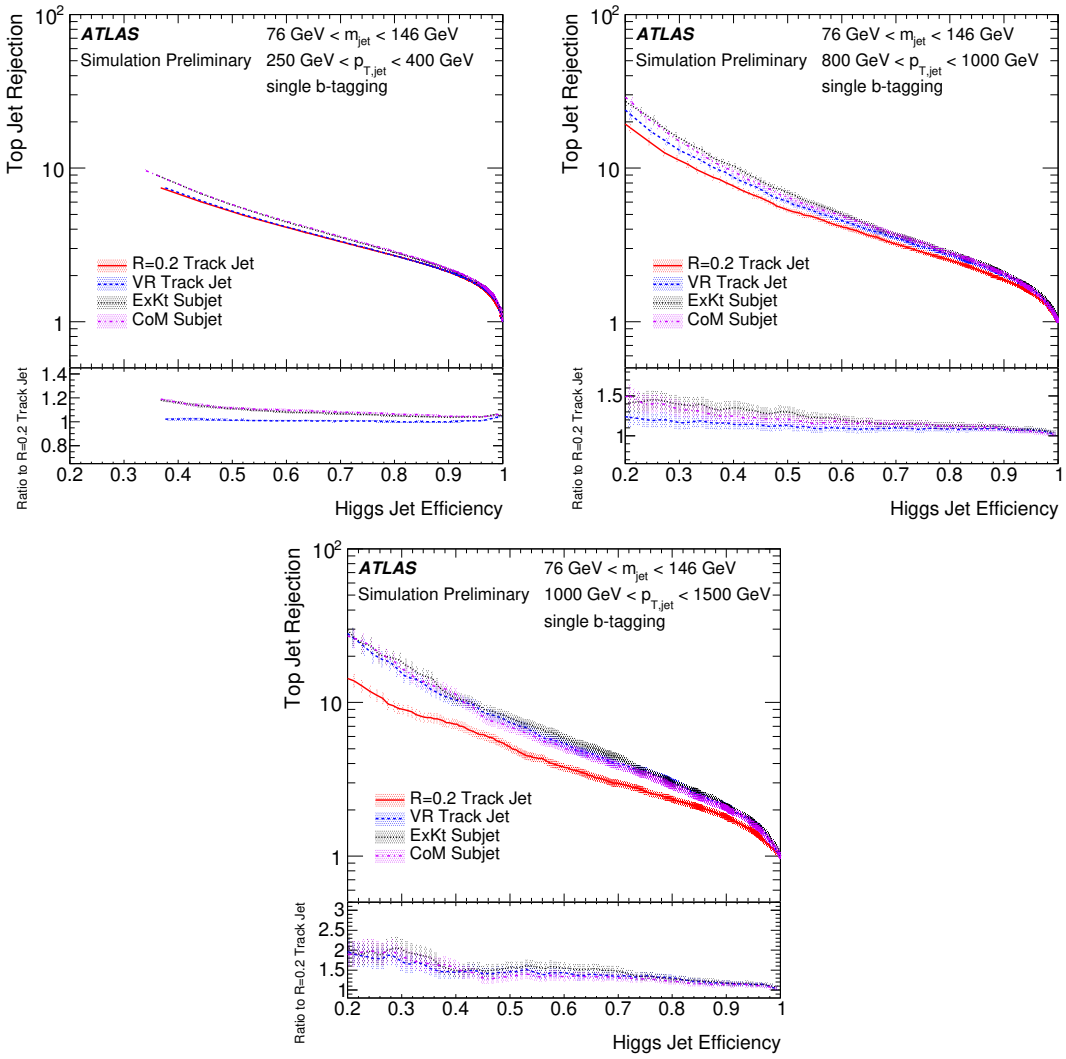


Figure 16: Top jet rejection as function of $h \to b\bar{b}$ jet efficiency when applying single b-tagging on subjets found by the $R = 0.2$ track jet, VR track jet, exclusive-k_T subjet, and CoM subjet algorithms in different p_T regions. The error bars include statistical uncertainties only.

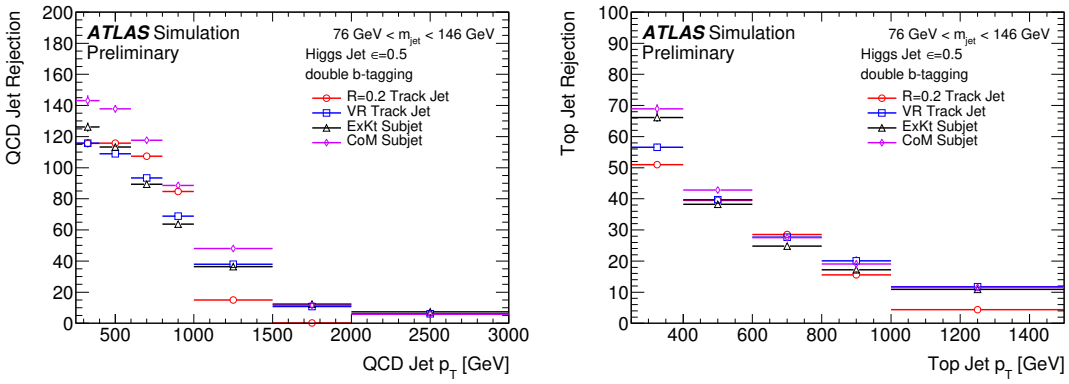


Figure 17: QCD and top jet double b-tagging rejection as a function of p_T for a fixed Higgs jet efficiency of 50%. The error bars include statistical uncertainties only.

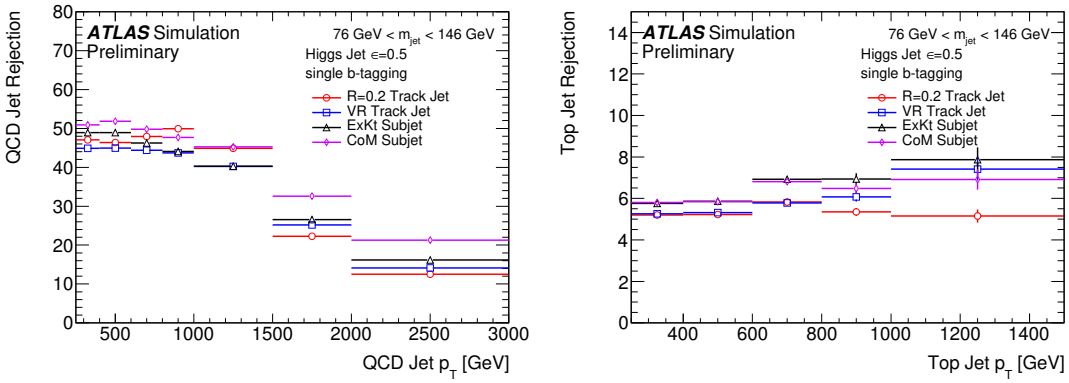


Figure 18: QCD and top jet single b-tagging rejection as a function of p_T for a fixed Higgs jet efficiency of 50%. The error bars include statistical uncertainties only.

6. Conclusion

Three new Higgs tagging techniques using subjets have been developed which show strong performance improvements over the nominal $R = 0.2$ track jet technique for identifying $h \to b\bar{b}$ decays with $p_T > 1000$ GeV. The variable radius track jet, exclusive-k_T subjet, and center-of-mass subjet techniques outperform the $R = 0.2$ track jet technique in this p_T region in both b-tagging metrics and performance metrics which do not involve b-tagging (e.g. b-hadron axis reconstruction). Across the jet p_{T} range studied, the CoM algorithm has the best performances among the Higgs tagging techniques under study. In addition to the subjet reconstruction optimizations investigated here a number of additional identification methods can further be optimized in the future. These include alternative track-to-subjet association prior to the determination of the flavor tagging discriminant as well as the optimization of the multivariate-based discriminant itself. These investigations as well as those concerning data modelling and systematic uncertainties are outside the scope of this note and will provide an important next step for future work.

Appendix

A. Subjet Reconstruction Cartoons

Figures 19-22 show illustrative cartoons to conceptually understand the working principles of the various subjet reconstruction techniques.

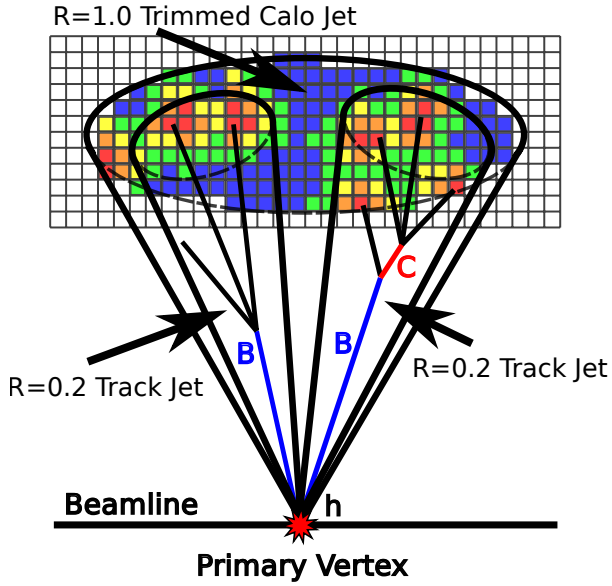


Figure 19: Cartoon illustrating subjet reconstruction using fixed radius R=0.2 track jets.

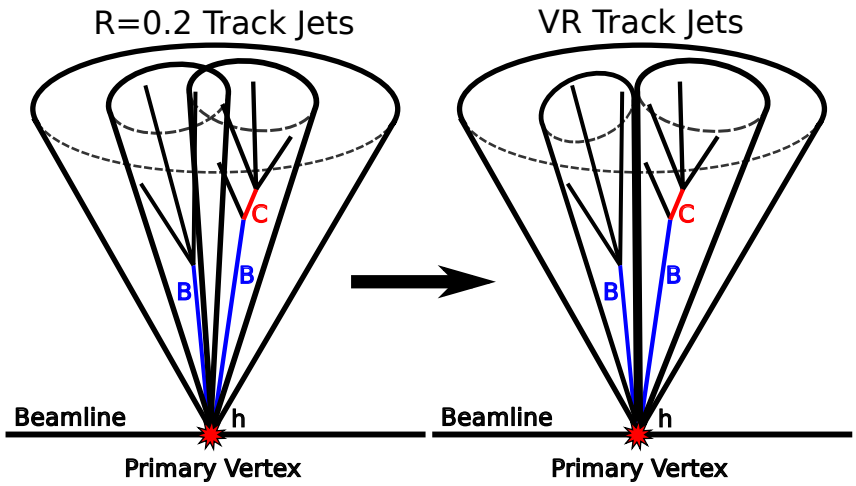


Figure 20: Cartoon illustrating subjet reconstruction using variable radius track jets.

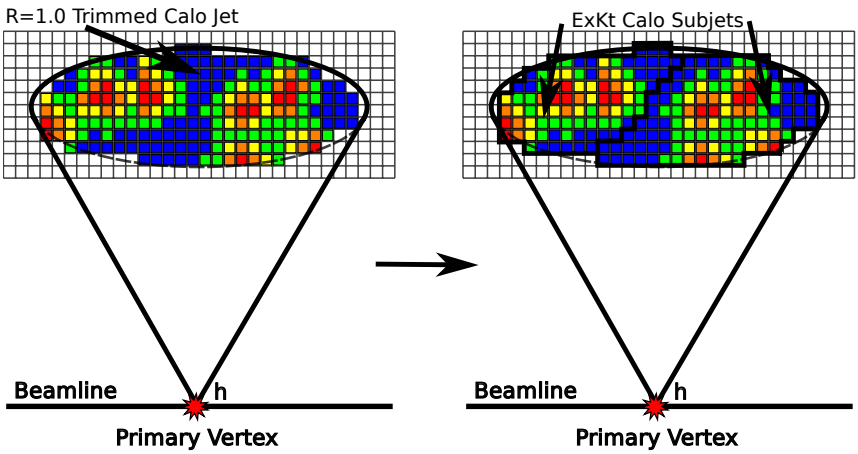


Figure 21: Cartoon illustrating subjet reconstruction using exclusive-k_T subjets.

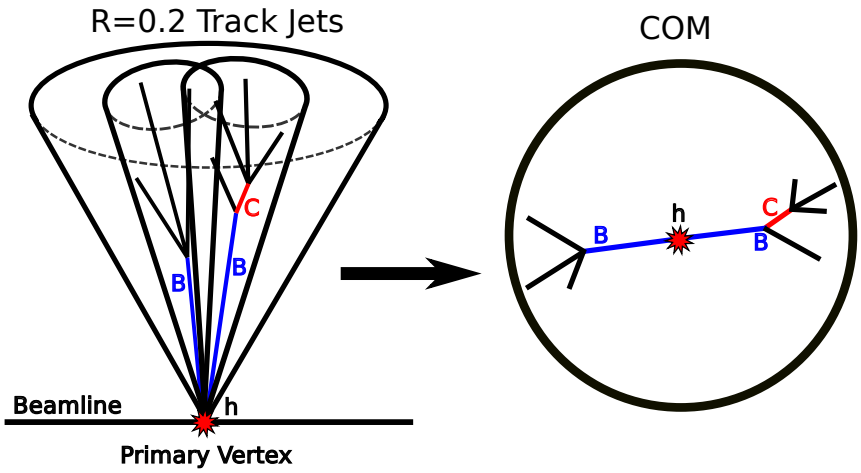


Figure 22: Cartoon illustrating subjet reconstruction using CoM subjets.

B. Sample Event Counts

For completeness, Table 1 provides a summary of the information for the event samples used in this analysis and shown in Figure 23 is the large-R calorimeter jet p_{T} spectrum of the multijet sample, to which the signal is reweighted as described in Section 2. This precise information is presented because the analysis is divided into regions of Higgs jet p_{T}, each of which can be populated by ensembles of events originating from a number of Graviton samples of different masses. It is found that the signal efficiency for a fixed selection is dependent on the $G*$ mass sample from which they are derived. Therefore, for reproducibility, these details are presented.

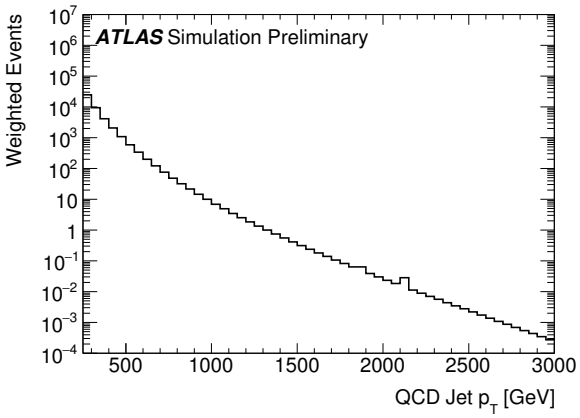


Figure 23: The large-R jet p_{T} spectrum from the combined sample of jets from the multijet samples listed in Table 1 with each sample weighted according to its theoretical cross section. When evaluating the performance of the tagging methods throughout this note, the Higgs jet and sample of jets from $t\bar{t}$ were both weighted on a jet-by-jet basis such that the resulting p_{T} spectra of each sample matches that shown here.

Sample Type	Specifier	Value	N_{events}	Jets Used	Jet-by-Jet Weighting
$G* \to hh$, $h \to b\bar{b}$	$G*$ Mass	300	79800	*"Higgs Jets"* (Section 3) : $\Delta R(J, x)$ <1.0 x=(h,b_1,b_2) p_{T} >250 GeV $\lvert\eta\rvert$<2.0	Reweight to multijet in large-R jet(p_{T},η)
		400	99800		
		500	94400		
		600	99800		
		700	54800		
		800	70000		
		900	83000		
		1000	10000		
		1100	99800		
		1200	99800		
		1300	19800		
		1400	99600		
		1500	99400		
		1600	99800		
		1800	15000		
		2000	89800		
		2250	99800		
		2500	60000		
		2750	59600		
		3000	78000		
		4000	100000		
		4500	99000		
		5000	99000		
		6000	99000		
$t\bar{t}$	$m_{t\bar{t}}$ sliced	[1100-1300)	550000	Inclusive p_{T} >250 GeV $\lvert\eta\rvert$<2.0	Reweight to multijet in large-R jet(p_{T},η)
		[1300-1500)	230000		
		[1500-1700)	100000		
		[1700-2000)	75000		
		[2000-14000)	45000		
Multijet	Truth jet p_{T} sliced	[160-400)	10^6	Inclusive p_{T} >250 GeV $\lvert\eta\rvert$<2.0	Theory cross section
		[400-800)	10^6		
		[800-1300)	10^6		
		[1300-1800)	10^6		
		[1800-2500)	10^6		

Table 1: The graviton mass value (m_{G^*}) and the number of simulated events (N_{events}) for each MC signal graviton sample. Summarized in the "Jets Used" column is the set of jets from the respective sample used in the evaluation of performance of the tagging methods. Summarized in the "Jet-by-Jet Reweighting" column is the weight applied per jet when evaluating the performance.

C. Truth level based subjet b-tagging efficiency

The goal of tagging a Higgs boson decaying to $b\bar{b}$ is to identify the presence of the two b-hadrons via the use of regions of interest defined by subjets within the large-R jet. However, if these regions of interest do not accurately reconstruct the directions of the b-hadrons, then the ability to make a positive identification of a Higgs jet when it is seeded by a Higgs boson, is hindered. In some cases, this may be due to the presence of additional radiation in a subjet which biases its direction away from the b-hadron (e.g. in ExKt or CoM) or due to the algorithm reconstructing more than two subjets (in the case of the fixed and variable radius track jets), which causes combinatorial difficulties when determining which of the subjets to use to identify the jet as a Higgs jet or not. Knowing the precise composition of these cases can provide added insight into the failures and merits of each technique.

Figures 24 and 25 show (for the p_{T} bins used in the main body of the note and for a set of more finely balanced p_{T} bins, respectively) the decomposition of the labelling of subjets within the sample of signal Higgs jets for different Higgs jet p_{T} intervals. On the x-axis of each plot is the subjet reconstruction algorithm, with variable radius track jets on the left, exclusive-k_t and CoM subjets in the center, and fixed radius track jets on the right. On the y-axis are the various categories of how a large-R jet can be decomposed into subjets, showing separately the categories for one, two, three, four, and five or greater subjets as determined by the respective algorithm. Each category is further subdivided into the manner in which the set of subjets has the pair of b-hadrons distributed among them. For example, in the case of two subjets, the "(20/02)" label means that either both b-hadrons were matched to the leading subjet ("20") or both were matched to the subleading subjet ("02"). A more precise description of all the categories is as follows :

- **1 subjet:** Higgs jet has exactly 1 subjet.
  - **(0):** The only subjet has no associated b-hadron.
  - **(1):** The only subjet has exactly 1 associated b-hadron.
  - **(2):** The only subjet has exactly 2 associated b-hadrons.
- **2 subjets:** Higgs jet has exactly 2 subjets.
  - **(00):** Neither subjet has any associated b-hadrons.
  - **(10/01):** Either the leading subjet or the sub-leading subjet has exactly 1 associated b-hadron.
  - **(11):** Both subjets have exactly one associated b-hadron
  - **(20/02):** Either the leading subjet or the sub-leading subjet has exactly 2 associated b-hadrons.
- **3 subjets:** Higgs jet has exactly 3 subjets.
  - **(000):** None of the subjets have any associated b-hadrons.
  - **(100/010):** Either the leading subjet or the sub-leading subjet has exactly 1 associated b-hadron.
  - **(110):** The leading and sub-leading subjets each have exactly 1 associated b-hadron.
  - **(101/011):** The third leading subjet and one of either the leading or subleading subjets have exactly 1 associated b-hadron.

  - **(200/020/002):** Either the leading subjet or the sub-leading subjet or the sub-sub-leading subjet has exactly 2 associated b-hadron.
  - **(001):** Only the third leading subjet has exactly one associated b-hadron.
- **4 subjets:** Higgs jet has exactly 4 subjets.
  - **(1100):** The leading and sub-leading subjets have exactly 1 associated b-hadron.
  - **(Others):** Does not fall into the subjet category above.
- **More than 4 subjets:** Higgs jet has more than 4 subjets.

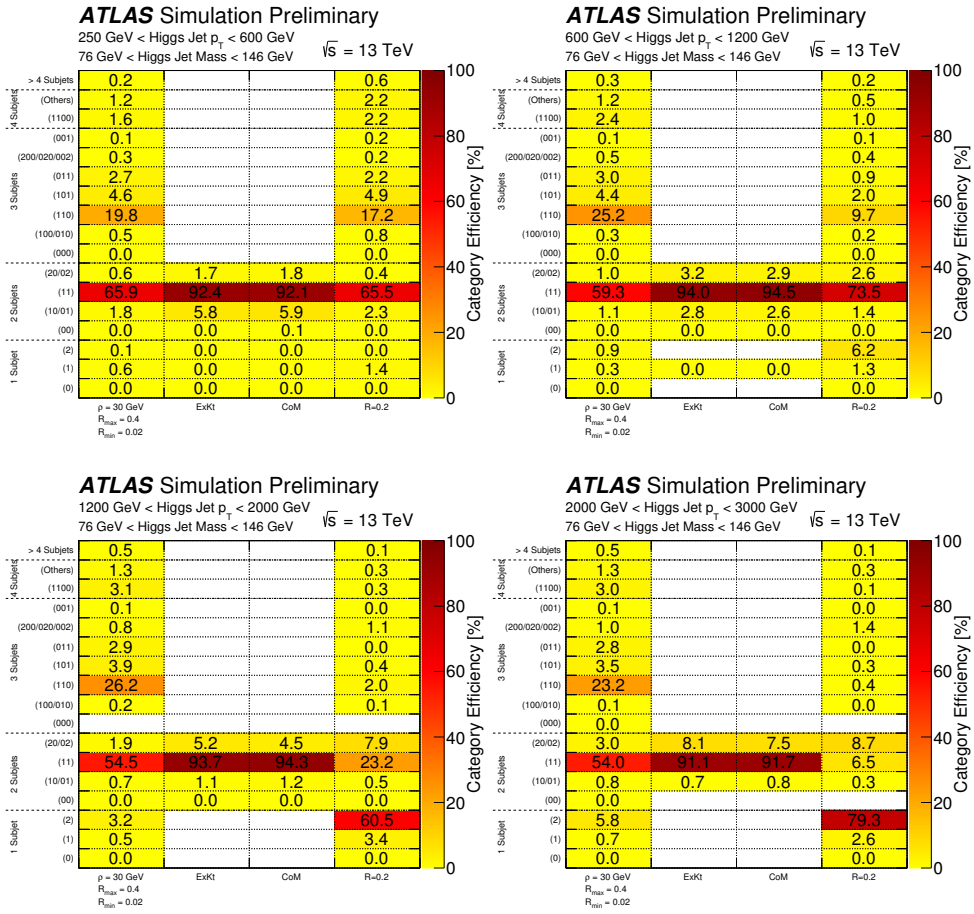


Figure 24: Category efficiency matrix for different subjet collections in various Higgs jet p_{T} regimes. In this figure, a Higgs jet mass window cut of 76 GeV $< m_{jet} <$ 146 GeV is applied.

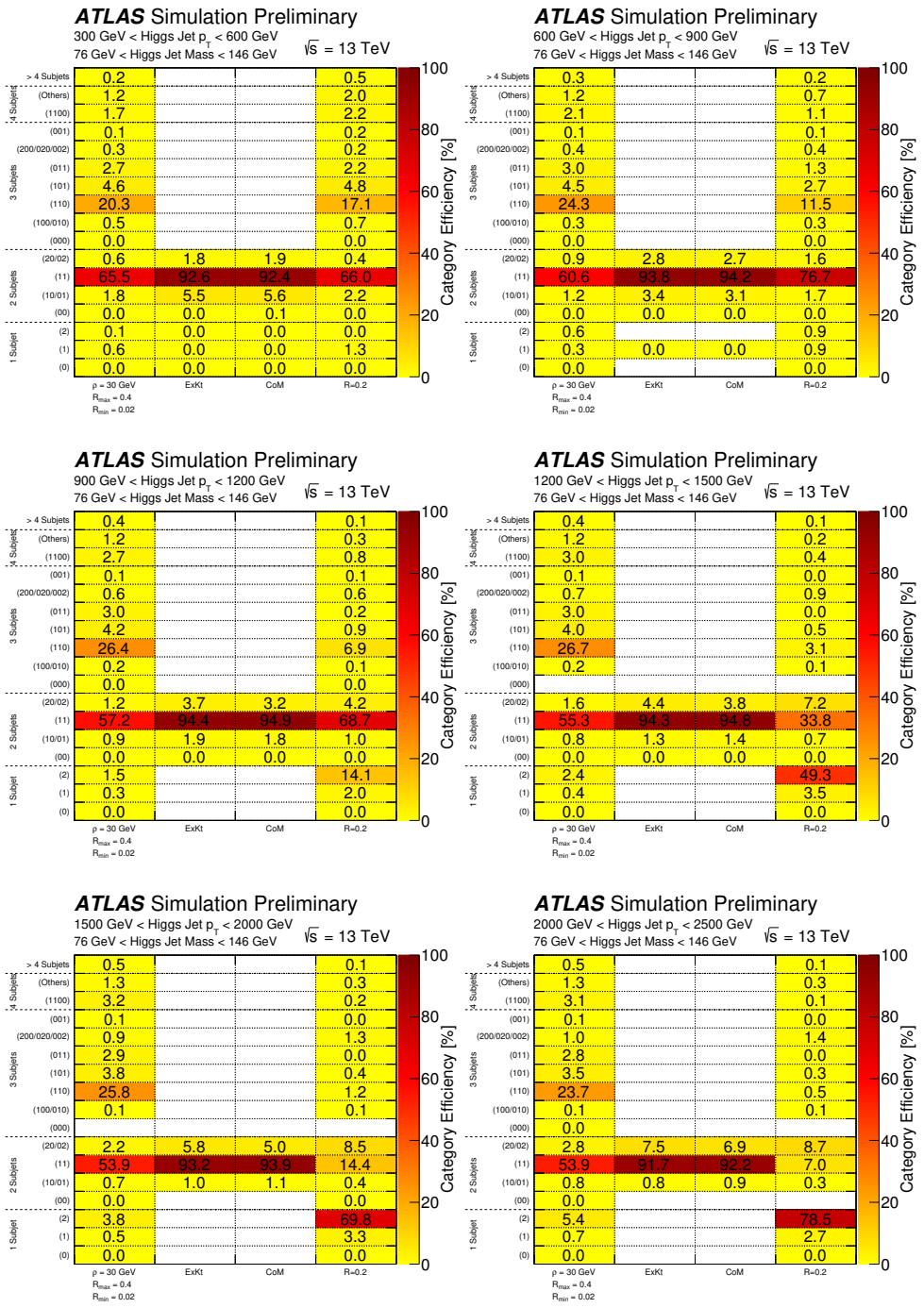


Figure 25: Category efficiency matrix for different subjet collections in various fine Higgs jet p_{T} bins. In this figure, a Higgs jet mass window cut of 76 GeV $< m_{jet} <$ 146 GeV is applied.

References

[1] L. Evans and P. Bryant, *LHC Machine*, JINST **3** (2008) S08001, ed. by L. Evans.

[2] J. M. Butterworth et al., *Jet substructure as a new Higgs search channel at the LHC*, Phys. Rev. Lett. **100** (2008) 242001, arXiv: 0802.2470 [hep-ph].

[3] ATLAS Collaboration, *The ATLAS Experiment at the CERN Large Hadron Collider*, JINST **3** (2008) S08003.

[4] ATLAS Collaboration, *Flavor Tagging with Track Jets in Boosted Topologies with the ATLAS Detector*, ATL-PHYS-PUB-2014-013, 2014, URL: https://cds.cern.ch/record/1750681.

[5] ATLAS Collaboration, *Expected Performance of Boosted Higgs ($\to b\bar{b}$) Boson Identification with the ATLAS Detector at $\sqrt{s} = 13$ TeV*, ATL-PHYS-PUB-2015-035, 2015, URL: https://cds.cern.ch/record/2042155.

[6] ATLAS Collaboration, *Boosted Higgs ($\to b\bar{b}$) Boson Identification with the ATLAS Detector at $\sqrt{s} = 13$ TeV*, ATLAS-CONF-2016-039, 2016, URL: https://cds.cern.ch/record/2206038.

[7] ATLAS Collaboration, *Search for dark matter in association with a Higgs boson decaying to b-quarks in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector*, Phys. Lett. **B765** (2017) 11–31, arXiv: 1609.04572 [hep-ex].

[8] ATLAS Collaboration, *Search for pair production of Higgs bosons in the $b\bar{b}b\bar{b}$ final state using proton–proton collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector*, ATLAS-CONF-2016-049, 2016, URL: https://cds.cern.ch/record/2206131.

[9] ATLAS Collaboration, *A Search for Resonances Decaying to a W or Z Boson and a Higgs Boson in the $q\bar{q}^{(\prime)}b\bar{b}$ Final State*, ATLAS-CONF-2016-083, 2016, URL: https://cds.cern.ch/record/2206276.

[10] ATLAS Collaboration, *The ATLAS Inner Detector commissioning and calibration*, Eur. Phys. J. **C70** (2010) 787–821, arXiv: 1004.5293 [physics.ins-det].

[11] ATLAS Collaboration, *ATLAS Insertable B-Layer Technical Design Report*, CERN-LHCC-2010-013, ATLAS-TDR-19 (2010), URL: https://cds.cern.ch/record/1291633.

[12] T. Sjöstrand, S. Mrenna and P. Z. Skands, *A Brief Introduction to PYTHIA 8.1*, Comput. Phys. Commun. **178** (2008) 852–867, arXiv: 0710.3820 [hep-ph].

[13] R. D. Ball et al., *Parton distributions with LHC data*, Nucl. Phys. **B867** (2013) 244–289, arXiv: 1207.1303 [hep-ph].

[14] ATLAS Collaboration, *ATLAS Pythia 8 tunes to 7 TeV data*, ATL-PHYS-PUB-2014-021, 2014, URL: http://cdsweb.cern.ch/record/1966419.

[15] P. Nason, *A New Method for Combining NLO QCD with Shower Monte Carlo Algorithms*, JHEP **11** (2004) 040, arXiv: hep-ph/0409146 [hep-ph].

[16] S. Frixione, P. Nason and C. Oleari, *Matching NLO QCD computations with Parton Shower simulations: the POWHEG method*, JHEP **11** (2007) 070, arXiv: 0709.2092 [hep-ph].

[17] T. Sjöstrand, S. Mrenna and P. Skands, *PYTHIA 6.4 Physics and Manual*, JHEP **0605** (2006) 026, arXiv: 0603175 [hep-ph].

[18] P. Z. Skands, *Tuning Monte Carlo generators: The Perugia tunes*, Phys. Rev. **D 82** (2010) 074018, arXiv: 1005.3457 [hep-ph].

[19] J. Gao et al., *CT10 next-to-next-to-leading order global analysis of QCD*, Phys. Rev. D **89** (2014) 033009, arXiv: 1302.6246 [hep-ph].

[20] L. Randall and R. Sundrum, *A Large Mass Hierarchy from a Small Extra Dimension*, Phys. Rev. Lett. **83** (1999) 3370–3373, arXiv: hep-ph/9905221.

[21] J. Alwall et al., *The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations*, JHEP **07** (2014) 079, arXiv: 1405.0301.

[22] D.J. Lange, *The EvtGen particle decay simulation package*, Nucl. Instr. Meth. **A462** (2001) 152.

[23] ATLAS Collaboration, *The ATLAS Simulation Infrastructure*, Eur. Phys. J. **C70** (2010) 823–874, arXiv: 1005.4568 [physics.ins-det].

[24] S. Agostinelli et al., *GEANT4: A simulation toolkit*, Nucl. Instrum. Meth. **A506** (2003) 250–303.

[25] ATLAS Collaboration, *2012 Summary of ATLAS Pythia 8 tunes*, ATL-PHYS-PUB-2012-003, 2012, URL: http://cdsweb.cern.ch/record/1474107.

[26] A. D. Martin et al., *Parton distributions for the LHC*, Eur. Phys. J. **C63** (2009) 189–285, arXiv: 0901.0002 [hep-ph].

[27] ATLAS Collaboration, *Topological cell clustering in the ATLAS calorimeters and its performance in LHC Run 1* (2016), arXiv: 1603.02934 [hep-ex].

[28] M. Cacciari, G. P. Salam and G. Soyez, *The Anti-k(t) jet clustering algorithm*, JHEP **04** (2008) 063, arXiv: 0802.1189 [hep-ph].

[29] D. Krohn, J. Thaler and L.-T. Wang, *Jet Trimming*, JHEP **02** (2010) 084, arXiv: 0912.1342 [hep-ph].

[30] S. Catani et al., *Longitudinally invariant K_t clustering algorithms for hadron hadron collisions*, Nucl. Phys. **B406** (1993) 187–224.

[31] ATLAS Collaboration, *Identification of boosted, hadronically decaying W bosons and comparisons with ATLAS data taken at $\sqrt{s} = 8$ TeV*, Eur. Phys. J. **C76**.3 (2016) 154, arXiv: 1510.05821 [hep-ex].

[32] ATLAS Collaboration, *Performance of jet substructure techniques for large-R jets in proton–proton collisions at $\sqrt{s} = 7$ TeV using the ATLAS detector*, JHEP **1309** (2013) 076, arXiv: 1306.4945 [hep-ex].

[33] ATLAS Collaboration, *Jet mass reconstruction with the ATLAS Detector in early Run 2 data*, ATLAS-CONF-2016-035, 2016, URL: https://cds.cern.ch/record/2200211.

[34] M. Cacciari, G. P. Salam and G. Soyez, *The Catchment Area of Jets*, JHEP **04** (2008) 005, arXiv: 0802.1188 [hep-ph].

[35] M. Cacciari and G. P. Salam, *Pileup subtraction using jet areas*, Phys. Lett. **B 659** (2008) 119, arXiv: 0707.1378 [hep-ph].

[36] D. Krohn, J. Thaler and L.-T. Wang, *Jets with Variable R*, JHEP **06** (2009) 059, arXiv: 0903.0392 [hep-ph].

[37] ATLAS Collaboration, *Performance of b-Jet Identification in the ATLAS Experiment*, JINST **11** (2016) P04008, arXiv: 1512.01094 [hep-ex].

[38] ATLAS Collaboration, *Expected performance of the ATLAS b-tagging algorithms in Run-2*, ATL-PHYS-PUB-2015-022, 2015, URL: https://cds.cern.ch/record/2037697.

[39] ATLAS Collaboration, *Optimisation of the ATLAS b-tagging performance for the 2016 LHC Run*, ATL-PHYS-PUB-2016-012, 2016, URL: https://cds.cern.ch/record/2160731.

[40] M. Wobisch and T. Wengler, *Hadronization corrections to jet cross-sections in deep inelastic scattering* (1998) 270–279, arXiv: hep-ph/9907280 [hep-ph].

[41] C. Chen, *New approach to identifying boosted hadronically-decaying particle using jet substructure in its center-of-mass frame*, Phys. Rev. **D85** (2012) 034007, arXiv: 1112.2567 [hep-ph].

[42] ATLAS Collaboration, *Measurement of the cross-section of high transverse momentum vector bosons reconstructed as single jets and studies of jet substructure in pp collisions at $\sqrt{s}$ = 7 TeV with the ATLAS detector*, New J. Phys. **16**.11 (2014) 113013, arXiv: 1407.0800 [hep-ex].

[43] C. Chen, *Identification of a bottom quark-antiquark pair in a single jet with high transverse momentum and its application*, Phys. Rev. **D92**.9 (2015) 093010, arXiv: 1507.06913 [hep-ph].

[44] Y. L. Dokshitzer et al., *Better jet clustering algorithms*, JHEP **08** (1997) 001, arXiv: hep-ph/9707323 [hep-ph].