Identification of double b-hadron jets from gluon-splitting with the ATLAS Detector

María Laura González Silva

Doctoral Thesis in Physics

Physics Department

University of Buenos Aires

November 2012

UNIVERSIDAD DE BUENOS AIRES

Facultad de Ciencias Exactas y Naturales

Departamento de Física

Identificación de jets con hadrones b producidos por desdoblamiento de gluones con el detector ATLAS.

Trabajo de Tesis para optar por el título de
Doctor de la Universidad de Buenos Aires en el área Ciencias Físicas

por **María Laura González Silva**

Director de Tesis: Dr. Ricardo Piegaia
Lugar de Trabajo: Departamento de Física

Buenos Aires, Noviembre 2012

Agradecimientos

Quiero agradecer a mi director, Ricardo Piegaia, por darme la oportunidad de trabajar en el proyecto ATLAS, por su dedicación y su enseñanza constante; y a mis compañeros de grupo, Gastón Romeo, Gustavo Otero y Garzón, Hernán Reisin y Sabrina Sacerdoti por el trabajo compartido y por brindarme su amistad a lo largo de estos años. Quiero agradecer a Ariel Schwartzman por darnos este análisis, por su caudal inagotable de ideas y por su generosidad y la de todo su equipo. Agradezco al Laboratorio CERN, al Experimento ATLAS, a los programas HELEN y e-Planet, al CONICET y al Fundación Exactas por hacer posible la realización de esta tesis.

Quiero agradecer el apoyo de mis compañeros de la carrera, especialmente a mis amigos Cecilia, Tomás y Leandro. Quiero agradecer también a mis compañeros de grupo y oficina, Javier, Yann, Pablo, y Orel por estar siempre dispuestos a darme una mano. Quiero agradecer a mis colegas y amigos de la Universidad de La Plata, Fernando, Martín y Xabier por todos los momentos compartidos; y a los amigos que hice a lo largo de estos años en mis visitas al Laboratorio CERN, Dodo, Laura, Lucile, Bárbara, Teresa, Manouk, Alex, Olivier y Haris, por ser mi familia en la distancia.

Agradezco profundamente a mis amigos y a toda mi familia por su apoyo y aliento; y de manera especial a mamá y a Juan, por comprenderme y acompañarme en todo. A ellos les dedico esta tesis.

Identificación de jets con hadrones b producidos por desdoblamiento de gluones con el detector ATLAS.

Resumen

En esta tesis se presenta un estudio de la subestructura de jets que contienen hadrones b con el propósito de distinguir entre jets-b genuinos, donde el quark b se origina a nivel de elemento de matriz (por ejemplo, en decaimientos de top, W, o Higgs) y jets-b producidos en la lluvia partónica de QCD, por el desdoblamiento de un gluón en un quark y un antiquark b cercanos entre sí. La posibilidad de rechazar jets-b producidos por gluones es importante para reducir el fondo de QCD en análisis de física dentro del Modelo Estándar, y en la búsqueda de canales de nueva física que involucran quarks b en el estado final. A tal efecto, se diseñó una técnica de separación que explota las diferencias cinemáticas y topológicas entre ambos tipos de jets-b. Esta se basa en observables sensibles a la estructura interna de los jets, construídos a partir de trazas asociadas a éstos y combinados en un análisis de multivariable. En eventos simulados, el algoritmo rechaza 95% (50%) de jets con dos hadrones b mientras que retiene el 50% (90%) de los jets-b genuinos, aunque los valores exactos dependen de p_T, el momento transverso del jet. El método desarrollado se aplica para medir la fracción de jets con dos hadrones b en función del p_T del jet, con 4,7 fb^1 de datos de colisiones pp a $\sqrt{s} = 7$ TeV, recogidos por el experimento ATLAS en el Gran Colisionador de Hadrones en 2011.

Palabras clave: Experimento ATLAS, Jets, Subestructura de Jets, QCD, Producción de jets b, Etiquetado de Jets b.

Identification of double b-hadron jets from gluon-splitting with the ATLAS Detector.

Abstract

This thesis presents a study of the substructure of jets containing b-hadrons with the purpose of distinguishing between "single" b-jets, where the b-quark originates at the matrix-element level of a physical process (e.g. top, W or Higgs decay) and "merged" b-jets, produced in the parton shower QCD splitting of a gluon into a collimated b quark-antiquark pair. The ability to reject b-jets from gluon splitting is important to reduce the QCD background in Standard Model analyses and in new physics searches that rely on b-quarks in the final state. A separation technique has been designed that exploits the kinematic and topological differences between both kinds of b-jets using track-based jet shape and jet substructure variables combined in a multivariate likelihood analysis. In simulated events, the algorithm rejects 95% (50%) of merged b-jets while retaining 50% (90%) of the single b-jets, although the exact values depend on p_T, the jet transverse momentum. The method developed is applied to measure the fraction of double b-hadron jets as a function of jet p_T, using 4.7 fb^{-1} of pp collision data at $\sqrt{s} = 7$ TeV collected by the ATLAS experiment at the Large Hadron Collider in 2011.

Keywords: ATLAS Experiment, Jets, Jet Substructure, b-jet Production, QCD, Gluon Splitting, b-tagging.

Contents

Chapter 1

Event reconstruction and b-Tagging

The event reconstruction packages, which in ATLAS are implemented in the software framework ATHENA [1], process the events, starting from the raw data obtained from the various sub-detectors (energy deposits and hits), through different stages to finally interpreting them as a set of charged tracks, electrons, photons, jets, muons and, in general, of possible kinds of final state objects with related four momenta. In this chapter the reconstruction of these objects is briefly described together with the algorithms for the identification of b-quark jets. These algorithms are mainly based on the reconstruction of the primary interaction vertex, on the reconstruction of charged particles in the Inner Detector and on the reconstruction of jets in the calorimeter.

1.1 Jet reconstruction and calibration

Hadronic jets used for ATLAS analyses are reconstructed by a jet algorithm, starting from the energy depositions of electromagnetic and hadronic showers

in the calorimeters. The ATLAS performance group, addressing the calibration of jets and the missing transverse energy (Jet/Etmiss), has made the decision to adopt the anti-k_t algorithm (Chapter **??**) as its default jet algorithm. This choice was driven by multiple requirements ranging from physics performance to those intimately involved with the computing, trigger and detector: the anti-k_t algorithm is fast and its memory consumption is low, it is well adapted to algorithms used in the trigger, and it has the best jet reconstruction efficiency at low p_T. Moreover, this algorithm exhibits the smallest fluctuations of the jet area showing good stability under pile-up [2].

Two different size parameters are used: $R = 0.4$, for narrow jets, more adecuate to describe the event substructure and associate matrix element partons to jets in multiparton final states; and $R = 0.6$, for wider jets, with very little out of cone radiation, more suitable for QCD studies.

The input to calorimeter jet reconstruction can be calorimeter towers or topological cell clusters. Charged particle tracks reconstructed in the Inner Detectors are also used to define jets. The latter have the further advantage of being insensitive to pile-up and they provide a stable reference for systematic studies. Both towers and topological clusters are combined as massless four-momentum objects. In the case of track-jets, the track four-momentum is constructed assuming the π meson mass for each track. The final four-momentum of the jet is obtained from summing the four-momenta of its constituents in the so called "four-vector recombination scheme". This scheme conserves energy and momentum and allows a meaningful definition for the jet mass. In Monte Carlo simulation, reference jets ("truth jets") are formed from simulated stable particles using the same jet algorithm as for the calorimeter jets.

Calorimeter towers are static, $\Delta\eta \times \Delta\phi = 0.1 \times 0.1$, grid elements built

directly from calorimeter cells. There are two types of calorimeter towers: with or without noise supression. The latter are called "noise-suppressed", and use only the cells with energies above a certain noise threshold. The noise of a calorimeter cell is measured by recording calorimeter signals in periods where no beam is present in the acelerator. The standard deviation σ around the mean no-beam energy is interpreted as the noise of the cell, and it depends on the sampling layer in which the cell resides and the position in η.

The results presented in this thesis use jets built from noise-suppressed topological clusters, also known as "topo-clusters" [3]. Topological clusters are groups of calorimeter cells that are designed to follow the shower development taking advantage of the fine segmentation of the ATLAS calorimeters. The topological cluster formation starts from a seed cell with $|E_{cell}| > 4\sigma$ above the noise. In a second step, neighbor cells that have an energy at least 2σ above their mean noise are added to the cluster. Finally, all nearest-neighbor cells surrounding the clustered cells are added to the cluster, regardless of the signal-to-noise ratio[1]. The position of the cluster is assigned as the energy-weighted centroid of all constituent cells (the weight used is the absolute cell energy).

Jet calibration

The baseline EM energy scale of the calorimeters is the result of the calibration of the electronics signal to the energy deposited in the calorimeter by electromagnetic showers (see Chapter **??**). The purpose of the jet energy calibration, or jet energy scale (JES), is to correct the measured EM scale energy

[1]Noise-supressed towers also make use of the topological clusters algorithm [3] to select cells, i.e. only calorimeter cells that are included in topo-clusters are used.

to the energy of the stable particles within a jet. The jet energy calibration must account then for the calorimeter non-compensation; the energy lost in inactive regions of the detector, such as the cryostat walls or cabling; energy that escapes the calorimeters, such as that of highly-energetic particles that "punch-through" to the muon system; energy of cells that are not included in clusters, due to inefficiencies in the noise-suppression scheme; and energy of clusters not included in the final reconstructed jet, due to inefficiencies in the jet reconstruction algorithm. The muons and neutrinos that may be present within the jet are not expected to interact within the calorimeters, and are not included in this energy calibration. Due to the varying calorimeter coverage, detector technology, and amount of upstream inactive material, the calibration that must be applied to each jet to bring it to the hadronic scale varies with its η position within the detector.

A number of complex calibration schemes, taking into account these effects, have been developed in ATLAS. The simplest procedure used for 2011 data, referred to as "EM+JES" calibration, utilizes an energy and η-dependent calibration scheme that is primarily based on Monte Carlo simulation with some direct in-situ measurements. This is the calibration used in this thesis. It consists of three subsequent steps:

- Pile-up correction: An offset correction is applied in order to substract the additional average energy measured in the calorimeter due to multiple proton-proton interactions. This correction is derived from minimum bias data as a function of NPV, the jet pseudorapidity and the bunch spacing. This additional energy is substracted before the hadronic energy scale is restored such that the derivation of the jet energy scale calibration is factorized and does not depend on the number of interactions in the event.

- Vertex correction: The jet four momentum is corrected such that the jet originates from the primary vertex of the interaction instead of the geometrical centre of the detector.

- Jet energy and direction correction: The jet energy and direction are corrected using constants derived from the comparison of the kinematic observables of reconstructed jets and those from truth jets in the simulation.

In the final step the calibration is derived in terms of the energy response of the jet, or the ratio of the reconstructed jet energy to that of a "truth" jet built of all truth stable interacting particles in the Monte Carlo. This response, written as

$$\mathcal{R} = E_{reco}/E_{truth} \tag{1.1}$$

may be defined at any energy scale. In Equation 1.1, E_{truth} is the energy of the closest isolated truth jet, within $\Delta R < 0.3$[2]. The isolation requirement is applied in order to factorize the effects due to close-by jets from those due to purely detector effects such as dead material and non-compensation. The isolation criterion requires that no other jet with a $p_T > 7$ GeV be within $\Delta R < 2.5R$, where R is the distance parameter of the jet algorithm.

The jet energy response is binned in truth jet energy and the calorimeter jet η. For each (E_{truth}, η)-bin, the averaged jet energy response is defined as the peak position of a Gaussian fit to the E_{reco}/E_{truth} distribution. The jet p_T response, which will be used later, uses the ${p_T}^{reco}/{p_T}^{truth}$ distribution.

The EM+JES calibration constants consist in the inverse of the response: $\mathcal{C}({p_T}^{EM}) = \mathcal{R}^{-1}_{reco}({p_T}^{EM})$, where $\mathcal{C}$ is the calibration constant and $\mathcal{R}_{reco}$ is

[2]This value was chosen because it results in a reconstructed-to-truth jet match more than 99% of the times.

the response calculated as a function of reconstructed jet p_T. They are derived as a function of ${p_T}^{truth}$, to remove the impact of the underlying p_T spectrum on the response. The jet response determined as a function of ${p_T}^{truth}$, $\mathcal{R}_{truth}$, is used to apply the constants as a function of ${p_T}^{EM}$, that is $\mathcal{R}_{reco}({p_T}^{EM}) = \mathcal{R}_{truth}(\mathcal{R}_{truth} \cdot {p_T}^{truth})$. This relationship is valid in ATLAS due to the linearity of the jet response as a function of p_T. The correct energy scale is obtained by multiplying the EM scale energy of a jet by the calibration constant

$$E^{EM+JES} = \mathcal{C} \cdot E^{EM}. \tag{1.2}$$

Other calibrations schemes are the global calorimeter cell weighting (GCW) calibration and the local cluster weighting (LCW) calibration. The GCW scheme exploits the observation that electromagnetic showers in the calorimeter leave more compact energy depositions than hadronic showers with the same energy. Energy corrections are derived for each cell within a jet. The cell corrections account for all energy losses of a jet in the detector. Since these corrections are only applicable to jets and not to energy depositions, they are called "global" corrections.

The LCW calibration method first classifies topo-clusters as either electromagnetic of hadronic, based on the measured energy density. Energy corrections are derived according to this classification from single charged and neutral pion Monte Carlo simulations. Dedicated corrections are derived for the effects of non-compensation, signal losses due to noise threshold effects, and energy lost in non-instrumented regions. Since the energy corrections are applied without reference to a jet definition they are called "local" corrections. Jets are then built from these calibrated clusters using a jet algorithm.

A further jet calibration scheme called global sequential (GS) calibration,

starts from jets calibrated with the EM+JES calibration and corrects the energy jet-by-jet, without changing the average response. This scheme exploits the topology of the energy deposits in the calorimeter to characterize fluctuations in the jet particle content of the hadronic shower development. Correcting for such fluctuations can improve the jet energy resolution. The correction uses several jet properties, and each correction is applied sequentially.

For the 2011 data the recomended calibration schemes were the EM+JES and the LCW calibrations. The simple EM+JES calibration does not provide the best resolution performance, but allows in the central detector region the most direct evaluation of the systematic uncertainties from the calorimeter response to single isolated hadrons measured *in situ* and in test-beams and from systematic variations in the Monte Carlo simulation. For the LCW calibration scheme the JES uncertainty is determined from *in situ* techniques. For all calibration schemes, the JES uncertainty in the forward regions is derived from the uncertainty in the central region using the transverse momentum balance in events where only two jets are produced.

Jet energy scale uncertainties for the EM+JES scheme

For many physics analyses, the uncertainty on the JES constitutes the dominant systematic uncertainty because of its tendency to shift jets in and out of analysis selections due to the steeply falling jet p_T spectrum. The uncertainty on the EM+JES scale is determined primarily by six factors: varying the physics models for hadronization and parameters of the Monte Carlo generators, evaluating the baseline calorimeter response to single particles, comparing multiple models for the detector simulation of hadronic showers, assesing the calibration scales as a function of pseudorapidity, and by ad-

justing the JES calibration methods itself. The final JES uncertainty in the central region, $|\eta| < 0.8$, is determined from the maximum deviation in response observed with respect to the response in the nominal sample. For the more forward region, the so called "η-intercalibration" contribution is estimated. This is a procedure that uses direct di-jet balance measurements in two-jet events to measure the relative energy scale of jets in the more forward regions compared to jets in a reference region. The technique exploits the fact that these jets are expected to have equal p_T due to transverse momentum conservation. Figure 1.1 shows the final fractional jet energy scale uncertainty and its individual contributions as a function of p_T for a central η region. The JES uncertainty for anti-k_t jets with $R = 0.4$ is between $\approx$4% (8%, 14%) at low jet p_T and $\approx$2.5%-3% (2.5%-3.5%, 5%) for jets with $p_T > 60$ GeV in the central (endcap, forward) region.

In addition to the tests above, *in situ* tests of the JES using direct γ-jet balance, multi-jet balance, and track-jets indicate that the uncertainties in Fig. 1.1 reflect accurately the true uncertainties in the JES.

In the case of jets induced by bottom quarks (b-jets), the calorimeter response uncertainties are also evaluated using single hadron response measurements *in situ* and in test beams [4]. For jets within $|\eta| < 0.8$ and $20 \leq p_T < 250$ GeV the expected difference in the calorimeter response uncertainty of identified b-jets with respect to the one of inclusive jets is less than 0.5%. It is assumed that this uncertainty extends up to $|\eta| < 2.5$.

The JES uncertainty arising from the modelling of the b-quark fragmentation can be determined from systematics variations of the Monte Carlo simulation. The fragmentation function is used to estimate the momentum carried by the b-hadron with respect to that of the b-quark after quark fragmentation. The fragmentation function included in PYTHIA originates from

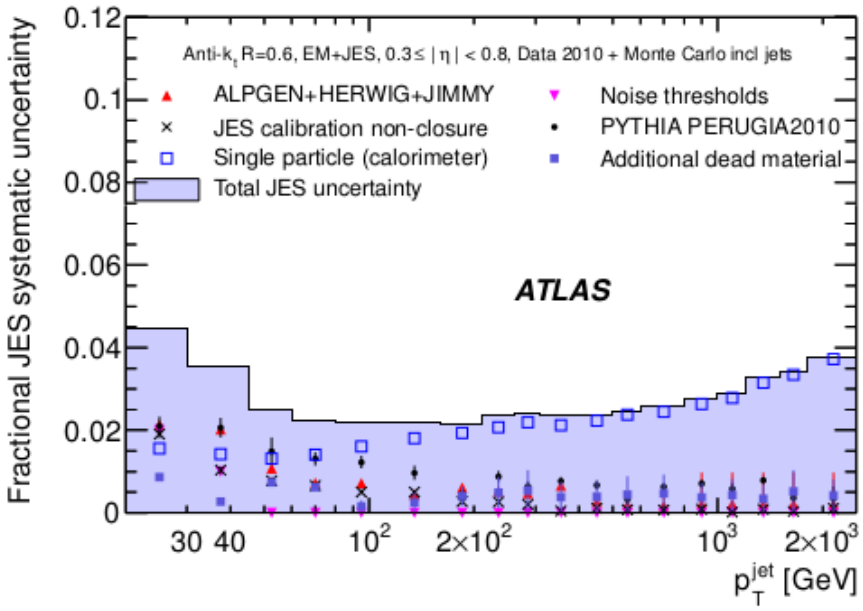

Figure 1.1: Fractional jet energy scale uncertainty as a function of jet p_T for jets in the pseudorapidity region $0.3 < |\eta| < 0.8$ in the calorimeter barrel. The total uncertainty is shown as the solid tight blue area. The individual sources are also shown.

a detailed study of the b-quark fragmentation function in comparison with OPAL [5] and SLD [6] data. To assess the impact of the b-fragmentation, the nominal parameters of the PYTHIA fragmentation function are replaced by the values from a tune using the Professor framework [7]. In addition, the nominal fragmentation function is replaced by the modified Bowler-Lund fragmentation function [8]. The b-jet response uncertainty is evaluated from the ratio between the response of b-jets in the varied Monte Carlo samples to the nominal PYTHIA. The response variations are well within 2%.

The b-jet JES uncertainty is obtained adding the calorimeter response uncertainty and the uncertainties from the systematic Monte Carlo variations in quadrature. The resulting additional JES uncertainty for b-jets is shown in Fig. 1.2. It is about 2% up to $p_T \approx 100$ GeV and below 1% for higher p_T. To obtain the overall b-jet uncertainty this uncertainty is added in quadrature to the JES uncertainty for inclusive jets.

1.2 Reconstruction of charged particle tracks

The Inner Detector layout and the characteristics of its main sub-detectors were presented in Section ?? of Chapter ??. The algorithm used for track reconstruction is based on a modular software framework, which is described in more detail in Ref. [9]. The main steps are the following:

- Firstly, the raw data from the pixel and SCT detectors are converted into clusters, while the TRT raw timing information is turned into calibrated drift circles. The SCT clusters need to be further transformed into space-points, by combining the clusters information from opposite sides of the SCT module (stereo strip layers).

- In a second stage, the track-finding is performed, in which the pattern

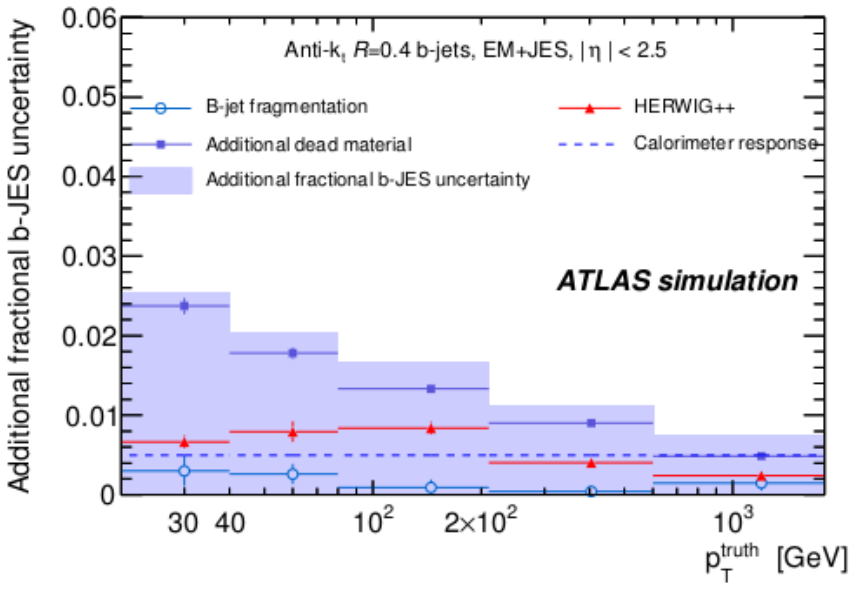

Figure 1.2: Additional fractional b-jet JES uncertainty as a function of the truth jet transverse momentum for anti-k_t jets with $R = 0.4$ calibrated with the EM+JES scheme for $|\eta| < 2.5$. Shown are systematic Monte Carlo variations using different modelling of the b-quark fragmentation and physics effects as well as variations in the detector geometry and the uncertainty in the calorimeter response to b-jets as evaluated from single hadron response measurements. Uncertainties in the individual points are statistical only.

recognition and a global χ^2 minimization procedure is implemented as a default.

In the track-finding stage, track seeds are found in the first three pixel layers and in the first SCT layer. These are extended throughout the SCT to form track candidates and a first track fit is performed. Afterwards, ambiguities in the track candidates found in the silicon detectors are resolved, and tracks are extended into the TRT (which covers up to $|\eta| < 2.$, while Pixel and SCT cover up to 2.5). The final track candidate is refitted with the full information from the three tracking subdetectors. The baseline algorithm is designed for the efficient reconstruction of primary charged particles. Primary particles are defined as particles with a meanlife of greater than 3×10^{-11} s directly produced in a proton-proton interaction, or from the subsequent decays or interactions of particles with lifetime shorter than 3×10^{-11} s. The tracks reconstructed in this stage are required to have $p_T > 400$ MeV.

In a complementary stage, a track search starts from segments reconstructed in the TRT and extends them inwards by adding silicon hits, which is referred to as "back-tracking". This recovers tracks for which the first hits in the pixel layers are missing, e.g. because they originate from secondaries, which are produced in decays or the interaction of primaries.

The final reconstructed track trajectory is usually specified at its closest point to the interaction region on the transverse plane by its impact parameters in the transverse plane and in the longitudinal direction, respectively called d_0 and z_0[3], and by its momentum, typically expressed in azimuthal angle ϕ, polar angle θ and inverse momentum $1/p$.

[3]Stricktly speaking the impact parameter is $|z_0|sin\theta$, where θ is the polar angle of the track.

The track reconstruction efficiency is defined as the fraction of primary particles with $p_T > 400$ MeV and $|\eta| < 2.5$ matched to a reconstructed track. The reconstruction efficiency for primary tracks with transverse momentum above 1 GeV and central η is above 80%, going down to values below 70% for tracks at the edge of the Inner Detector acceptance [10]. The dense environment of a jet decreases the track reconstruction efficiency and increases the fake rate. This is caused by the ocurrence of shared hits between different tracks, which makes the pattern recognition and track fitting tasks more difficult.

The relative transverse momentum scale and resolution of tracks is defined as the Gaussian mean and width of

$$ {p_T}^{MC} \times (1/{p_T}^{MC} - 1/{p_T}^{reco}) = 1 - \frac{{p_T}^{MC}}{{p_T}^{reco}} \tag{1.3} $$

where ${p_T}^{MC}$ (${p_T}^{reco}$), refers to the track's transverse momentum given by simulation truth (MC) or by reconstruction (reco). It should be noted that the $(1/p_T)$ resolution is used instead of $\sigma(p_T)$ as the Inner Detector measures the sagitta and not directly the transverse momentum[4]. However, the resolution obtained from the equation above is the relative transverse momentum resolution, $\sigma(p_T)/p_T$. At low p_T the multiple scattering dominates the resolution, and at high momenta, the resolution is limited by the bending power of the solenoid field and by the instrinsic detector resolution. For a central track with p_T =5 GeV the transverse momentum resolution is around 75 MeV and the transverse impact parameter resolution is about 35 μm.

[4]The relation between sagitta s and transverse momentum (p_T) is given by $s \sim 1/p_T$.

1.3 Vertex reconstruction

Primary vertices are reconstructed using an iterative vertex finding algorithm [11]. In a first step, a dedicated vertex finding algorithm associates tracks to vertex candidates. Vertex seeds are obtained by looking for the global maximum in the distribution of the z coordinates of the tracks. In a second stage, an iterative χ^2 fit is made using the seed and nearby tracks. Each track carries a weight which is a measure of its compatibility with the fitted vertex depending on the χ^2 of the fit. Tracks displaced by more than 7σ from the vertex are used to seed a new vertex and the procedure is repeated until no additional verteces can be found. The parameters of the beam spot are used both during the finding to preselect compatible tracks and during the fitting step to constrain the vertex fit.

The knowledge of the position of the primary interaction point (primary vertex) of the proton-proton collision is important for b-quark jets identification since it defines the reference point with respect to which impact parameters and vertex displacements are measured. The typical vertexing resolution in z is $\mathcal{O}(100\mu\text{m})$.

To ensure a good resolution on the vertex position, the primary vertex must be reconstructed from at least five tracks. The choice of the primary vertex is less trivial in the presence of minimum-bias events from pile-up: the primary vertex from a pile-up event may be mistakenly used as the signal vertex, or a fake primary vertex built from tracks from two different vertices may be reconstructed. The current strategy is to choose the primary vertex candidate that maximizes $\sum_{tracks} {p_T}^2$.

1.4 b-jet Tagging

The ability to identify jets originating from *bottom*-quarks (denoted as b-tagging in the following) is important for the high-p_T physics program of a general-purpose experiment at the LHC such as ATLAS since many interesting physics processes contain b-quarks in the final state, while the most abundant backgrounds contain mostly up, down and strange quark or gluon jets or, in a smaller fraction of cases, charm quark jets. The aim of b-tagging is therefore to identify the b-quark jets with high efficiency, while rejecting most of the background contamination from jets originating from the fragmentation of light (u, d, and s) quarks, gluons and c-quarks.

A b-quark, once produced, fragments necessarily into a b-flavoured hadron, b-hadron in the following. In most of the cases ($\approx$87%), first an excited b-hadron is produced, like a B^* or a B^{**}, which decays inmediately, strongly or electromagnetically, into a ground state b-hadron plus one or more further particles; while in the remaining cases, a ground state b-hadron is produced directly. One is only interested in the transition from a b-quark into the final state b-hadron, since the typical timescale for electromagnetic or strong interactions is so small that the B^*, B^{**} decay vertices are not significantly displaced with respect to the primary vertex. In most of the cases ($\approx$ 91%) a b-meson is produced out of the fragmentation of an original b-quark (40% B^+, 40% B^0 and 11% B_s^0). The rest are b-baryons.

Due to the b-quark fragmentation function being very hard, most of the original b-quark energy is transmitted to the final b-hadron. This fraction is for example 70% for b-quarks with a momentum of $\approx$45 GeV. This property can be exploited during b-tagging, since the fragmentation for light quarks into light hadrons or c-quarks into c-hadrons is softer.

Any of the finally produced b-hadrons decay through weak interactions

and therefore have a significant lifetime, which is on average, for all b-hadrons considered, $(1.568 \pm 0.009) \times 10^{-12}$ s. The effective distance travelled in the detector by the b-hadron before decaying depends on the b-hadron momentum, which enters the relativistic boost factor $\beta\gamma$. A b-quark with momentum of 50 GeV will travel around 3 mm, which is a visible flight length in the detector. Due to the combination of the b-hadron lifetime and relatively high mass ($m_B \approx 5.28$ GeV), which results in a non-negligible decay angle of the b-hadron decay products with respect to the b-hadron flight direction, the charged particles produced at the decay vertex will be on average significantly displaced with respect to the primary vertex position.

This is the main signature which is exploited by the *lifetime* based b-tagging algorithms, which depend either on the presence of significantly displaced tracks, as in impact parameter based b-tagging algorithms, or on the explicit reconstruction of the b-hadron decay vertex, as in secondary vertex based b-tagging algorithms.

b-hadrons decay preferably into a c-hadron plus additional particles[5]. The lifetime of a c-hadron is not much lower than for b-hadrons, but in general the momentum of the c-hadron will be lower than the original b-hadron momentum. However, the c-hadron can still travel for a significant path in the detector and form with its decay producs a visible *tertiary* vertex.

Another property which is usually exploited by b-tagging is the fraction of b- and c-hadron decays into leptons: a lepton from the semi-leptonic decay of a b-hadron ($b \to l$) or from the subsequent c-hadron decay ($b \to c \to l$) is produced in ≈21% of the cases. This is valid both in case the lepton is a an electron or a muon, which brings the overall fraction of b-quarks ending up into final state containing a lepton to ≈42%. Due to the b- or c-hadron

[5]Weak decays are governed by the CKM matrix mechanism, and $|V_{cb}|^2 >> |V_{ub}|^2$.

mass, the lepton will be emitted with an average transverse momentum comparable with m_{b-had} or m_{c-had}. By identifying either an electron or a muon originating from a jet and by requiring it to have sufficiently high p_T with respect to the jet axis, it is possible to identify b-jets.

Association of tracks to jets

The b-tagging performance relies critically on the accurate reconstruction of the charged tracks in the ATLAS Inner Detector. The actual tagging is performed on the sub-set of tracks in the event that are associated to jets. The b-tagging algorithm takes as input the three-momenta of the jets, reconstructed by a jet algorithm, and uses the jet direction to associate the charged particles tracks to the jet. Since the 2 Tesla solenoidal magnetic field of the ATLAS Inner Detector bends charged particles in the transverse plane, in particular in the case of low p_T tracks, the tracks are best matched to the jet by using the direction of their momenta at the point of closest approach to the interaction region. The criterion for associating charged particle tracks to jets is simply:

$$\Delta R(jet, track) < \Delta R_{cut} \tag{1.4}$$

where usually the value of $\Delta R_{cut} = R$ is used; with R, the distance parameter of the jet algorithm used for jet reconstruction.

After the tracks are associated to the jets, they are filtered in order to remove tracks with bad quality or which can easily be erroneously identified as secondary tracks from b-decays. These include tracks originating from decays of even longer lived particles, like K_s^0 ($c\tau \approx$2.69 cm) and Λ baryons ($c\tau \approx$7.89 cm); from electromagnetic interactions in the detector material, like conversions in electron-positron pairs ($\gamma \rightarrow e^+e^-$); or from hadronic interactions with the detector material, which result in two or more tracks

with high impact parameter. In order to reject badly reconstructed tracks, quality cuts are applied. Requirements are imposed on the number of silicon hits, the track fit quality, the track momentum, and the transverse and longitudinal impact parameters. The track selection needs to be particularly tight in the case of the impact parameter based b-tagging algorithms, since in that case the explicit presence of a vertex is not required, so that the influence of badly reconstructed tracks or tracks from long lived particles does directly limit the performance. The minimum track p_T required is of 1 GeV in the case of the impact parameter based algorithms and of 400-500 MeV otherwise. The transverse and longitudinal impact parameters must fulfill $|d_0|$ <1 mm (3.5 mm) and $|z_0|\sin\theta < 1.5$ mm (no cut on z_0) in the case of the algorithms relying on the impact parameters of tracks (on the reconstruction of secondary vertices). The minimum number of precision hits required is typically 7, for both approaches.

1.4.1 b-tagging algorithms

For the 2011 data-taking a set of lifetime taggers were commissioned and calibrated. In this section a brief description of the main features of these algorithms will be given.

Impact parameter based b-tagging algorithms

The charged particle tracks originating from b-hadrons are expected to have significantly higher transverse and longitudinal impact parameters compared to prompt tracks originating directly from fragmentation. If the effect of long lived particles, conversions and hadronic interactions can be reduced, the best discrimination between prompt tracks and displaced tracks from b- and c-hadron decays can be obtained using the impact parameter significance,

both in the transverse and longitudinal plane. With

$$IP_{r\phi} = d_0 \text{ and } IP_z = z_0 \sin\theta, \tag{1.5}$$

the transverse and longitudinal impact parameter significances are obtained by dividing $IP_{r\phi}$ and IP_z by their respective errors,

$$IP_{r\phi}/\sigma(IP_{r\phi}) \text{ and } IP_z/\sigma(IP_z). \tag{1.6}$$

On the basis that the decay point of the b-hadron must lie along its flight path, and in order to increase the discriminating power of the impact parameter significance, a lifetime sign is assigned to these variables (replacing the sign of the geometrical definition of the impact parameter). The sign is positive if the track extrapolation crosses the jet direction in front of the primary vertex (i.e. is more compatible with having its origin in a secondary decay vertex in the direction of flight expected for the b-hadron) or negative if the track is more likely to intersect the flight axis behind the primary vertex, opposite to the jet direction. Both cases are illustrated in Fig. 1.3.

The lifetime sign can be defined in three-dimensions, acording to the variables $\vec{p_{Tjet}}$, $\vec{p_{Ttrk}}$ and $\vec{\Delta r}_{IP} = \vec{r}_{IP} = \vec{r}_{PV}$, the three-dimensional impact parameter of the track with respect to the primary vertex:

$$\text{sign}_{3D} = \text{sign}([\vec{p}_{trk} \times \vec{p}_{jet}] \cdot [\vec{p}_{trk} \times \vec{\Delta r}_{IP}]). \tag{1.7}$$

The computation of the lifetime sign assumes that the jet direction reproduces, up to a good approximation, the b-hadron direction. Under this assumption and up to resolution effects both on the jet direction and on the impact parameter and momentum of the track, the lifetime sign of tracks originating from b-hadron decays is positive.

The lifetime sign can also be defined on the transverse plane $(x-y)$ or on the longitudinal plane $(r\phi - z)$ by considering respectively the transverse

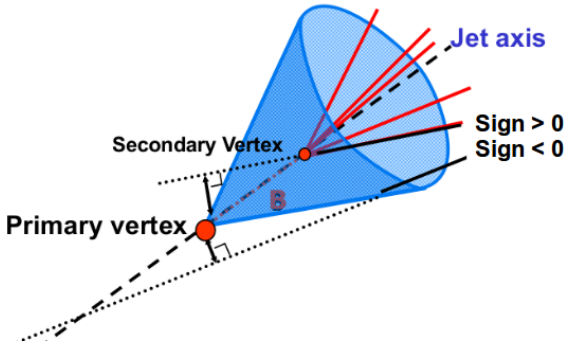

Figure 1.3: Lifetime sign of tracks. A positive and a negative lifetime signed track is shown.

and longitudinal impact parameters (the projections of the three-dimensional impact parameter on the respective planes):

$$\text{sign}_{r\phi} = \text{sign}(\sin(\phi_{jet} - \phi_{trk}) \cdot d_{0,trk}); \text{ and } \text{sign}_z = \text{sign}((\eta_{jet} - \eta_{trk}) \cdot z_{0,trk}). \tag{1.8}$$

Distributions of the signed transverse impact parameter and signed transverse impact parameter significance for light, c-, and b-jets, are shown in Fig. 1.4 for experimental data and for simulation; the sign is defined by "$\text{sign}_{r\phi}$". Tracks from the fragmentation in light-jets tend to have a signed impact parameter distribution which is symmetric around 0, since they have no correlation with the jet direction. Tracks from b- and c-hadron decays, as expected, have an asymmetric distribution, with the most significant contribution at positive significances; however a negative tail extending beyond the pure fragmentation contribution is also seen, corresponding to resolution effects and to an eventual mismatch between the b-jet and the b-hadron directions.

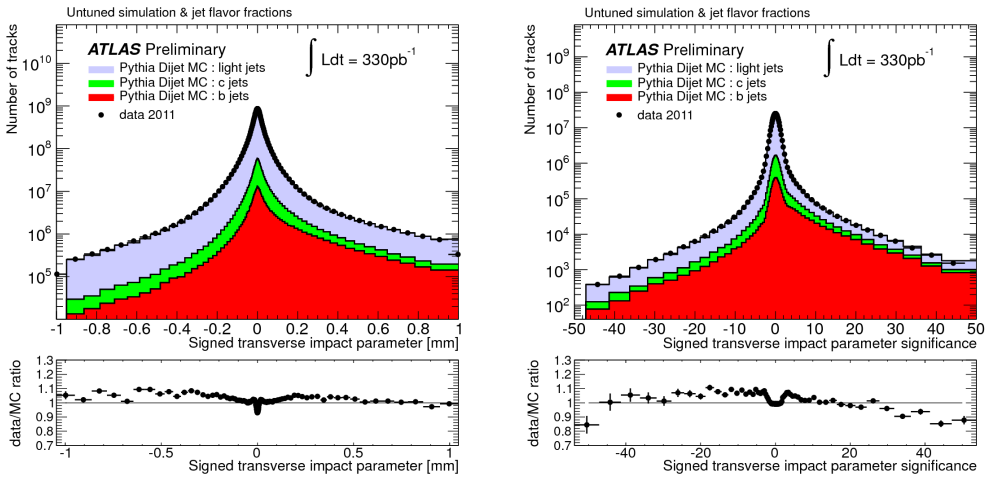

Figure 1.4: Distribution of the signed transverse impact parameter (left) and signed transverse impact parameter significance with respect to primary vertex for tracks associated to jets, for experimental data (solid black points) and for simulated data (filled histograms for the various flavors). The ratio data over simulation is shown at the bottom of the plot.

The significance, which gives more weight to tracks measured precisely, is the main ingredient of the tagging algorithms based on impact parameters. Now, the impact parameter significance of all N tracks associated to the jet to tag need to be combined into a single discriminating variable. It is assumed that tracks are uncorrelated, so their probability density functions (PDF), defined based on the transverse and/or longitudinal impact parameter significance distributions for the different hipothesis, are uniquely defined as a function of the jet flavour. Using a likelihood function defined according to the product of these PDFs, under the hypothesis of uncorrelated tracks, the following likelihood ratio provides the optimal separation, according to Neyman-Person lemma [12]:

$$\mathrm{LR}(IP_1, IP_2, ..., IP_N) = \frac{\prod_{i=1}^{N} \mathrm{PDF}_b(IP_i)}{\prod_{i=1}^{N} \mathrm{PDF}_l(IP_i)} \tag{1.9}$$

For convention, the discriminant variable used for b-tagging is then defined as:

$$\mathrm{weight}(IP_1, IP_2, ..., IP_N) = \log(\mathrm{LR}(IP_1, IP_2, ..., IP_N)) \tag{1.10}$$

Using such a formalism, two impact parameter based b-tagging algorithms are constructed, based on the definition of $\mathrm{PDF}(IP_i)$:

1. IP2D: $\mathrm{PDF}(IP_i) = \mathrm{PDF}(IP_{i,r\phi})$
2. IP3D: $\mathrm{PDF}(IP_i) = \mathrm{PDF}(IP_{i,r\phi}, IP_{i,z})$

In the first case the track PDF is one-dimensional, based on the transverse impact parameter significance. In the second case the PDF is based on a two-dimensional histogram of the transverse and longitudinal impact parameter significance.

The **IP3D** is one of the high-performance tagging algorithms supported for the 2011 data-taking, in which input variables are compared to pre-defined

smooth Monte Carlo PDFs for both b-jet and light jet hypotheses [13]. Prior to the use of these advanced tagger, a simpler tagging algorithm, the **JetProb**, combining the impact parameter significances of all tracks associated to the jet was devised to be used for early data, being extensively used during 2010 [14].

The impact parameter based algorithm permits to obtain a very good b-tagging performance, as will be shown at the end of this chapter. This performance can be improved by using some information from the secondary vertex based algorithms in two aspects: tracks associated to long lived particle vertices can be removed from the tracks considered for the impact parameter based algorithms; and, the direction between the secondary and the primary vertex positions can be used to improve the reliability of the lifetime sign, subtituting $\vec{p}_{jet}$ with $\vec{r}_{SV} - \vec{r}_{PV}$. The latter improves significantly the estimation of the b-hadron direction. Both kinds of information improve the performance of the impact parameter based b-tagging algorithms.

Secondary vertex based b-tagging algorithms

The typical topology of particle decays in a b-jet is a decay chain with two vertices, one stemming from the b-hadron decay and at least one from c-hadron decays. The reconstruction of these secondary vertices is done in an inclusive way, where the number of charged particle tracks originating from the b- and c-hadron decays is not known a-priori. An exclusive reconstruction of the huge number of different possible b-decay modes cannot be performed, the set of selection cuts needed to reconstruct all of them would severely limit the reconstruction efficiency.

Two strategies to detect a secondary decay vertex in b-jets are available in ATLAS. The first one is based on the fit of a single geometrical vertex. Even

if this hypothesis is not correct, this approximation works well for a large fraction of cases. The second algorithm is based on a kinematic approach, which assumes that the primary event vertex and the b- and the c-hadron decay vertices lie approximately on the same line, the flight path of the b-hadron.

The inclusive fit of a single displaced vertex in b-jets is based on the VKalVrt [15] reconstruction package. The main idea of the algorithm is to maximise the b/c-hadron vertex detection efficiency, keeping at the same time the probability to find a vertex inside a light jet low.

The algorithm begins with all tracks associated to the jet and passing a loose cut selection. The vertex search starts with looking for all track pairs and trying to form a two-track vertex. Each track of the pair must have a three-dimensional impact parameter significance with respect to the primary vertex larger than 2σ and the sum of these two significances must be larger than 6σ. To reduce the influence of badly measured tracks, the two-tracks vertices are required to be produced in the direction of flight of the b-quark, by requiring the scalar product of $(\vec{r}_{2-track}-\vec{r}_{PV})\cdot\vec{p}_{jet}$ to be positive. Charged particles coming from long lived particles and conversions are not considered. All the tracks corresponding to the accepted two-track vertices are used to determine a single secondary vertex. If the resulting vertex has a very small vertex probability, the track with the highest contribution to the vertex χ^2 is removed and the vertex fit is repeated until the χ^2 of the fit is good. The result of this procedure is the (eventual) presence of a vertex, its position, and the list of associated tracks.

The **SV1** secondary vertex algorithm uses this procedure to reconstruct inclusive secondary vertices. This advanced tagger takes advantage of three of the reconstructed vertex properties: the invariant mass of all tracks as-

sociated to the vertex, the ratio of the sum of the energies of the tracks in the vertex to the sum of the energies of all tracks in the jet, and the number of two-track vertices. These variables are combined using a likelihood ratio technique. SV1 relies on a two-dimensional distribution of the two first variables and a one-dimensional distribution of the number of two-track vertices. In addition the distance ΔR between the jet axis and the line joining the primary vertex to the secondary one is used.

The three-dimensional decay length significance alone, signed with respect to the jet direction can be used as a discriminating variable between b-jets and light jets: this is the principle of the early data **SV0** tagger, extensively used as well with the 2010 and 2011 data [16].

As opposed to the algorithm described above, in which the displaced tracks are selected and an inclusive single vertex is obtained, a second algorithm, called **JetFitter**, is based on a different hypothesis. It assumes that the b- and the c-hadron decay vertices lie on the same line defined through the b-hadron flight path. All charged particle tracks stemming from either decay intersect this b-hadron flight axis. This method has the advantage of reconstructing incomplete topologies, with, for instance, a single track from the b-hadron and a single track from the c-hadron decay. The fit in this case evaluates the compatibility of the given set of tracks with a b-c-hadron like cascade topology, increasing the discrimination power against light quark jets. The transversal displacement of the c-hadron decay vertex with respect to the b-hadron flight path is small enough not to violate significantly the basic assumption within the typical resolutions of the tracking detector. The discrimination between b-, c- and light jets is based on a likelihood using similar variables as in the SV1 tagging algorithm above, and additional variables such as the flight length significances of the vertices.

Algorithm combinations and performance

The IP3D and SV1 tagging algorithms both use the likelihood ratio method, and due to this they can be easily combined: the weights of the individual tagging algorithms are simply summed up.

The combination of the JetFitter and the IP3D algorithms can be performed using an artificial neural network technique with Monte Carlo simulated training samples and additional variables describing the topology of the decay chain.

Figure 1.5 compares the performance for the various ATLAS b-tagging algorithms described in a simulated sample of $t\bar{t}$ events. It can be seen that by combining the vertexing techniques and the impact parameter information, the IP3D+SV1 and IP3D+JetFitter algorithms can reach very high tagging efficiencies.

The performance of a b-tagging algorithm is usally measured in terms of the *light-jet rejection* obtained for a given *b-jet tagging efficiency*. Curves are obtained by varying continuously the *operating point* of each tagger, i.e. the cut on its output discriminating variable (weight). The b-jet tagging efficiency, ϵ_b, is the fraction of jets labeled as b-jets that are properly tagged while the light-jet rejection, defined as $1/\epsilon_{light}$, is the reciprocal of the fraction of jets that are labeled as light jets and are actually incorrectly tagged by the algorithm.

The labeling procedure used for b-tagging is based on the flavor of true quarks: a jet is labeled as a b-quark jet if a b-quark is found in a cone of size $\Delta R = 0.3$ around the jet direction. The various labeling hypotheses are tried in this order: b quark, c quark and τ lepton. When none of these hypotheses are satisfied, the jet is labeled as a light jet. No attempt is made to distinguish light jets originating from gluons from those originating from

quarks at this stage.

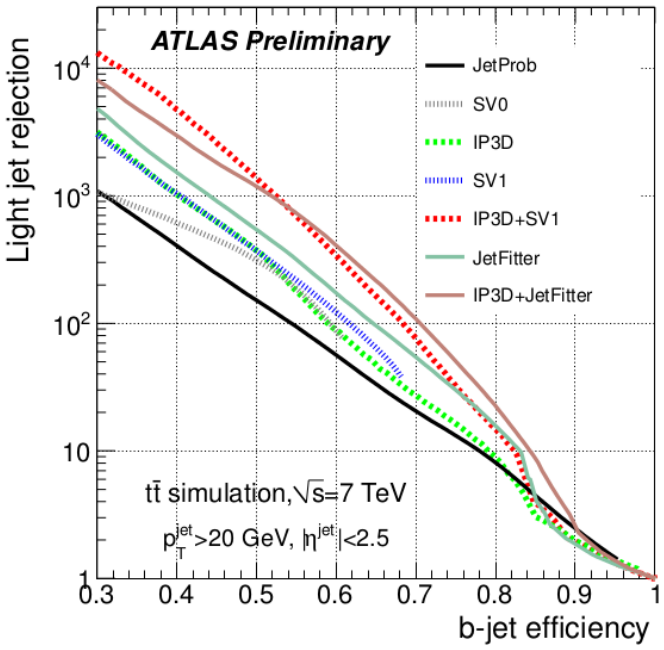

Figure 1.5: Light-jet rejection as a function of the b-jet tagging efficiency for the early tagging algorithms (JetProb and SV0) and for the high-performance algorithms, based on simulated $t\bar{t}$ events.

The MV1 tagging algorithm

The **MV1** b-tagging algorithm is a combined algorithm based on a neural network using the output weights of the IP3D and SV1 algorithms and the JetFitter+IP3D combination as input. Being the best performing algorithm (better light rejection for a given signal efficiency) it is the recomended tagger for 2011 and 2012 analyses. This is the b-tagging algorithm used in this thesis.

1.4.2 b-tagging calibration

In order for b-tagging to be used in physics analyses, the efficiency with which a jet originating from a b-quark is tagged needs to be measured in data. Moreover, an appropriate description of the b-tagging efficiencies based on measurements with data is essential for correctly modelling the measurements in Monte Carlo simulation . A second necessary piece of information is the probability of mistakenly tagging a jet originating from a light-flavour (u-, d-, s-quark or gluon) jet as a b-jet, referred to as the mistag rate. The b-tagging "calibration" includes both the measurement of the mistag rates and b-tagging efficiency.

The measurements of the b-tag efficiency and mistag rate are provided in the form of jet p_T- and η-dependent scale factors that correct the b-tagging performance in simulation to that observed in data. The scale factors are defined as the ratio of the b-tag efficiency or mistag rate in data and simulation:

$$\kappa_{\epsilon_b}^{data/sim} = \frac{\epsilon_b^{data}}{\epsilon_b^{sim}}, \quad \kappa_{\epsilon_l}^{data/sim} = \frac{\epsilon_l^{data}}{\epsilon_l^{sim}}, \tag{1.11}$$

where ϵ_b^{sim} and ϵ_l^{sim} are the fractions of b- and light-flavour jets which are tagged in simulated events, with the jet flavour defined by matching to generator level partons as defined in the previous section.

In physics analyses, these p_T-dependent scale factors are then applied as weights to the jets in Monte Carlo simulation, to reproduce the b-tagging performance in data.

The main b-tagging efficiency calibration methods, the so called *system8* and $p_{T_{rel}}$ methods, are described in detail in ref [17].These measurements are based on a sample of jets with muons inside, where the muons are serving as a reference b-tagging algorithm to obtain a b-jet sample on which the calibrations can be performed. At the LHC, the large $t\bar{t}$ production cross section of

$\sigma_{t\bar{t}} = 177 \pm 3(\text{stat.}) \, ^{+8}_{-7}(\text{syst.}) \pm 7(\text{lum.})$ pb [18] offers an alternative source of events enriched in b-jets. Calibrations using samples of $t\bar{t}$ events have been obtained for SV0, IP3D+SV1, JetFitter and MV1 b-tagging algorithms [19]. All these algorithms provide an output weight w, discriminating between b-jets and non-b-jets. Lower values of w are assigned to c- and light-flavour jets, whereas the purity of b-jets increases with w. For each b-tagging algorithm a set of operating points, corresponding to a certain w cut value, are defined and calibrated:

- SV0: $\epsilon_b^{sim} = 50\%$
- IP3D+SV1: $\epsilon_b^{sim} = 60\%$, $\epsilon_b^{sim} = 70\%$, $\epsilon_b^{sim} = 80\%$
- JetFitter: $\epsilon_b^{sim} = 57\%$, $\epsilon_b^{sim} = 60\%$, $\epsilon_b^{sim} = 70\%$, $\epsilon_b^{sim} = 80\%$
- MV1: $\epsilon_b^{sim} = 60\%$, $\epsilon_b^{sim} = 70\%$, $\epsilon_b^{sim} = 75\%$, $\epsilon_b^{sim} = 85\%$

where ϵ_b^{sim} is the nominal b-tagging efficiency derived from an inclusive sample of simulated $t\bar{t}$ events.

The mistag rate is measured in data using two methods, both based on an inclusive sample of jets, referred to as the *negativetag* and *sv0mass* methods [20]. The first method uses the invariant mass spectrum of tracks associated with reconstructed secondary vertices to separate light and heavy-flavour jets, and the other is based on the rate at which secondary vertices with negative decay length, or tracks with negative impact parameter, are present in the data.

Currently, there is no explicit measurement of the c-tag efficiency available in ATLAS. As both the b- and c-tag efficiencies are dominated by decays of long-lived heavy flavour hadrons, they are expected to show a similar behaviour. In general, for physics analyses, it is thus assumed that the scale

factor is the same for b- and c-jets. However, to take into account possible deviations from this assumption, the systematic uncertainty for the c-tag efficiency scale factor is inflated by a factor of two, which is considered to be a conservative choice based on simulation studies. In the future, the c-tag efficiency is expected to be measured in dedicated analyses.

Chapter 2

Double b-hadron jet identification

In this chapter we focus on the understanding of the internal structure of b-jets containing two b-hadrons by investigating the differences between these and single b-quark jets. These differences are expected to arise from the two-subjet structure of double b-hadron or "merged" jets, which would tend to be wider and with a larger number of constituents. Based on these envisaged characteristics, simulated QCD samples of b-tagged jets were used to explore properties with potential discrimination power. The Monte Carlo distributions were compared to data from the 2011 run for validation. We present results from these studies and discuss the choice of the observables selected to build the multivariable tool presented in Chapter **??**.

2.1 Data sample

The tagging technique presented in this thesis relies on Monte Carlo predictions for the signal (single b) or background (merged b) hypotheses. The accuracy of the simulation is validated with data by comparing the distributions of the different variables studied.

The data samples employed correspond to proton-proton collisions at $\sqrt{s} = 7$ TeV delivered by the LHC and recorded by ATLAS between May and November 2011, with the LHC running with 50 ns bunch spacing, and bunches organized in bunch trains. Only data collected during stable beam periods in which all sub-detectors were fully operational are used. After the application of the data quality selection, the surviving data corresponds to an integrated luminosity of 4.7 fb^{-1}. The LHC instantaneous luminosity steadily increased during 2011. As a result, the average number of minimum-bias pile-up events, originating from collisions of additional protons in the same bunch as the signal collision, grew from from 3 to 20 (see Fig.**??**). This fact will be of importance when discussing the selection of discriminating variables.

The events were collected using the ATLAS single jet triggers which select events with at least one jet with transverse energy above a given threshold. At the hardware Level 1 and local software Level 2 (see Section **??**), cluster-based jet triggers are used to select events with high-p_T jets. The Event Filter, in turn, runs the offline anti-k_t jet finding algorithm with $R = 0.4$ on topological clusters over the complete calorimeter. At this stage, the transverse energy thresholds, expressed in GeV, are: 20, 30, 40, 55, 75, 100, 135, 180. These triggers reach an efficiency of 99% for events having the leading jet with an offline energy higher than the corresponding trigger thresholds by a factor ranging between 1.5 and 2. The jet triggers with the lowest p_T thresholds were prescaled by up to five orders of magnitude.

2.2 Monte Carlo sample

The Monte Carlo samples employed were produced with the event generators discussed in Section **??**. Samples of dijet events from proton-proton collision processes were simulated with PYTHIA version 6.423 [21], used both for the simulation of the hard $2 \rightarrow 2$ process as well as for the parton shower, underlying event, and hadronization models. The ATLAS AMBT2 tune of the soft model parameters was used [22].

In order to have sufficient statistics over the entire p_T spectrum, seven samples were generated with different thresholds of the hard-scattering partonic transverse momentum $\hat{p}_T$: 8-17 GeV, 17-35 GeV, 35-70 GeV, 70-140 GeV, 140-280 GeV, 280-560 GeV and 560-1120 GeV. For the Monte Carlo p_T distribution (or the distribution of any other observable), to be compared to that in experimental data, events from the different samples need to be weighted by their respective production cross sections. The unweighted distribution is shown in Fig. 2.1. The p_T spectrum obtained after performing this procedure is displayed in Fig. 2.2.

The simulated data sample used for the analysis gives an accurate description of the pile-up content and detector conditions for the full 2011 data-taking period.

2.2.1 Event and jet selection

The data sample in the analysis is selected online using a set of single jet triggers as described in Section 2.1. In the case of the Monte Carlo, a trigger simulator is used. In this way both the simulated and real data from the detector can then be run through the same ATLAS trigger packages [23].

The offline event selection comprises an additional set of cuts on the

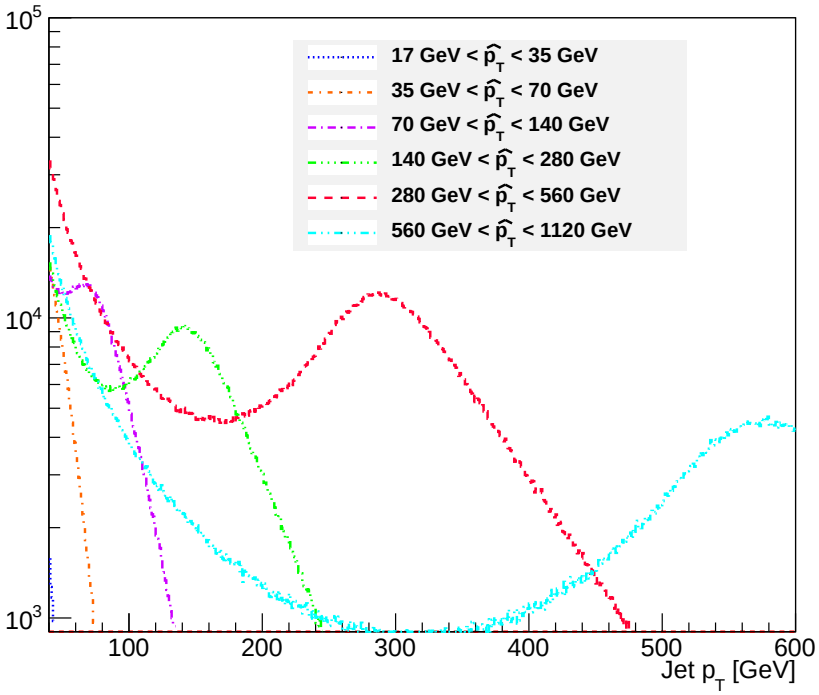

Figure 2.1: Calibrated jet p_T distribution for anti-k_t jets in a dijet Monte Carlo sample composed of different sub-samples generated with increasing thresholds of the hard-scattering partonic transverse momentum, $\hat{p}_T$.

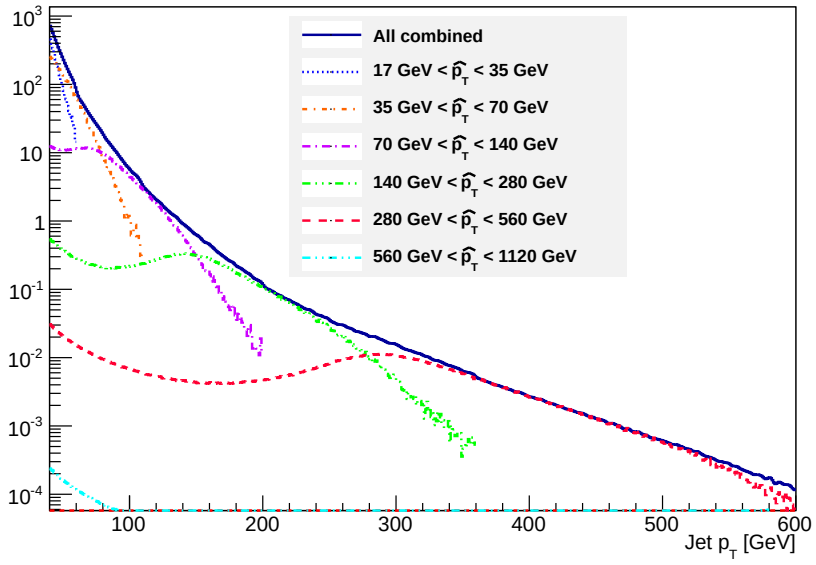

Figure 2.2: Calibrated jet p_T distribution for anti-k_t jets in a dijet Monte Carlo sample composed of different sub-samples generated with increasing thresholds of the hard-scattering partonic transverse momentum, $\hat{p}_T$. In order to obtain the falling p_T spectrum observed in data, the different samples were weighted by their respective production cross sections.

reconstructed objects, including jet kinematic and jet-specific data quality cuts. A vertex cut is also included, requiring at least one primary vertex with five or more associated tracks in the event. This cut serves as a first rejection for events originating from cosmic rays and particles produced in interactions of the beam with particles in the beam tunnel ("beam halo" and "beam gas"). No requirements are placed on the longitudinal position (along the beam line) of the vertex as the beam spot is used as a constraint when fitting the vertex.

The jet algorithm selected for the analysis was the ATLAS default anti-k_t algorithm (Section **??**), with a distance parameter $R = 0.4$, using calorimeter topological clusters as input (Section 1.1). All jets were calibrated using the EM+JES scheme (Section 1.1). A high cut on the minimum jet p_T is implemented to select jets in the region where the triggers used in the analysis are most efficient. Jets are required to have a minimum p_T of 40 GeV. Jets with transverse momentum above this threshold were also required to be in a region with full tracking coverage, $|\eta_{jet}| < 2.1$. Although the Pixel and SCT detectors cover up to $|\eta| < 2.5$, a lower pseudorapidity cut is used in order to account for the size of the calorimeter jets, $R = 0.4$. Jets passing this selection were classified in eight p_T bins chosen such as to match the jet trigger 99% efficiency thresholds (in GeV): 40, 60, 80, 110, 150, 200, 270, 360. An event is used if it satisfies the highest threshold trigger that is 99% efficient for the p_T bin that corresponds to the p_T of its leading jet. The upper limit of our highest p_T bin was set to 480 GeV; beyond this energy the b-tagging efficiency becomes very poor.

Several quality criteria are applied to jets to eliminate "fakes" that are caused by noise bursts in the calorimeters and energy depositions belonging to a previous bunch crossing. A detailed description of these quality cuts can

be found in reference [24].

In addition to these kinematic and quality cuts, two more cuts are imposed to jets:

- b-**tagging**. Jets are only accepted if they are tagged as b-jets using the MV1 b-tagging algorithm, at its 60% efficiency working point.

- **Isolation**. Jets are only accepted if they are isolated. The isolation criterion requires that no other jet with a $p_T > 7$ GeV be within $\Delta R < 2R$, where R is the distance parameter of the jet algorithm.

Finally, in the case of MC, the reconstructed b-tagged jets were further classified into single and merged b-jets based on truth Monte Carlo information. A b-hadron is considered to be associated to a jet if the ΔR distance in $\eta - \phi$ space between the direction of the hadron and the jet axis is smaller than 0.4. Jets were labeled as merged (single) b-jets if they contained two (only one) b-hadron:

$$\text{single } b\text{-jets: } \Delta R(j, B_i) < 0.4 \;\&\; \Delta R(j, B_j) > 0.4 \;\; for\; i \neq j \tag{2.1}$$

$$\text{merged } b\text{-jets: } \Delta R(j, B_i) < 0.4 \;\&\; \Delta R(j, B_j) < 0.4 \;\; for\; i \neq j \tag{2.2}$$

where j is a jet in the event and $B_{i(j)}$ are the b-hadrons in the event. In the case another size parameter is used for jet finding, the definitions in equations 2.1 and 2.2 change accordingly.

2.2.2 Track selection

The tracking system provides a very precise tool for understanding the structure of jets and for mitigating the pile-up background. Charged particle jet constituents that leave tracks in the inner detector provide 3-dimensional

information on the jet origin and direction as a result of the vertexing provided by the tracks. The combination of tracking and calorimetry therefore greatly enhance the identification and selection of hadronic jets from primary interactions that do typically have associated charged tracks.

In the study of the internal structure of jets containing b-hadrons, the tracking information will be used to define jet variables with potential discriminating power between single and merged b-jets. For this reason the selection of genuine tracks belonging to jets is of great importance.

The jet direction is used to associate the charged particles reconstructed as tracks in the inner detector to the jet. A simple $\Delta R < 0.4$ matching criterion is used, where the matching is performed using the track coordinates at the point of closest approach to the primary vertex.

Tracks are required to fulfill cuts on their transverse momentum, number of hits and transverse and longitudinal impact parameters, similar to those applied by b-tagging algorithms (see Section 1.4). Cuts on ${p_T}^{trk} > 1.0$ GeV and the χ^2 of the track fit, $\chi^2/ndf < 3$, are applied. The effect of a lower cut on the track transverse momentum, ${p_T}^{trk} > 0.5$ GeV, is discussed in the next section. In addition, tracks are required to have a total of at least seven precision hits (pixel or micro-strip) in order to guarantee at least 3 z-measurements. As cutting on impact parameter (IP) might be detrimental for b-jets, where large IP values are expected, relaxed cuts were used, $|d_0| <$ 2 mm, and $|z_0 \sin\theta| < 2$ mm, with θ being the polar angle measured with respect to the beam axis. The track quality cuts are summarized in table 2.1.

Track parameter	Selection
p_T	> 1 GeV
d_0^{PV}	< 2 mm
$z_0^{PV} \sin\theta$	< 2 mm
$\chi^2/ndof$	< 3
Number of Pixel hits	≥ 2
Number of SCT hits	≥ 4
Number of Pixel+SCT hits	≥ 7

Table 2.1: Track selection criteria used for tracks associated to b-jets, where d_0^{PV} and z_0^{PV} denote the transverse and longitudinal impact parameters derived with respect to the primary vertex. The $\chi^2/ndof$ is that of the track fit.

2.3 Kinematic differences between single and double b-hadron jets

The differences between genuine b-quark jets and double b-hadron jets, that in QCD originate mainly from gluon splitting, are expected to arise from the two-subjet structure of merged jets. In this section we present the study of a set of jet shape and substructure variables for the discrimination between single and merged b-jets. These variables are built from jet constituents either at calorimeter level (topological clusters) or tracks associated to the jet.

Jet track multiplicity

The jet track multiplicity is a variable simple to calculate that carries important information of the jet inner structure. It is defined as the number of tracks with p_T above 1 GeV, satisfying the quality cuts described in section 2.2.2, and contained within a cone of radius $R = 0.4$ around the jet axis. Figure 2.3 shows its distribution for two p_T bins, representative of the range covered in this study. It is observed that merged b-jets contain on average around two more tracks than single b-jets at low jet p_T, with a larger difference at higher p_T values.

The effect of the minimum track p_T requirement was examined by lowering the selection cut to $p_T > 0.5$ GeV. On the one hand this could lead to an improvement in discrimination if it captured more information about the fragmentation process; on the other hand, a lower minimum track p_T can make the method more sensitive to pile-up with the addition of soft tracks incorrectly associated to the jets. It was observed that reducing the p_T cut of the tracks degrades the discrimination because it widens the distributions without increasing the separation between single and merged jets.

We also considered the possibility of restricting ourselves to using tracks significantly displaced from the PV ($|d_0|/\sigma(d_0) > 2.5$), which are more likely to originate from the b-hadrons decays. In order to evaluate the effect of this particular selection, a preliminary study was done with a sample of dijet events generated with PYTHIA and with no detector simulation (denoted as "standalone" PYTHIA in the following). For this study jets were reconstructed using all stable particles in the event, clustered with the anti-k_t algorithm. The association of charged particles, the equivalent of tracks at the level of event generation, was done in the same way as with the full ATLAS simulation. Distributions of the track multiplicity built using all charged

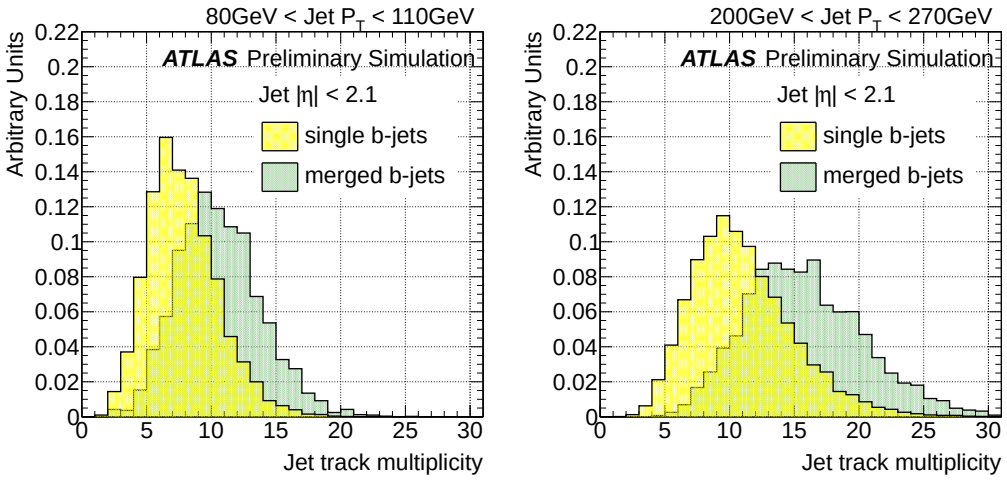

Figure 2.3: Distribution of the track multiplicity for single and merged b-jets from 80 GeV to 110 GeV (left) and 200 GeV to 270 GeV (right).

particles and using only charged particles coming from the b-hadron decay ("b-tracks") are illustrated in Fig. 2.4. A better discrimination between single and merged b-jets, measured in terms of the significance, $s = \Delta \mathrm{n}_{trk}/\sigma(\Delta \mathrm{n}_{trk})$ with n_{trk} the mean jet track multiplicity, is observed when using b-tracks only: $s = 5.9 \cdot 10^{-1}$ compared to $s = 4.4 \cdot 10^{-1}$ when using all charged particles. The result obtained with standalone PYTHIA suggests that a potential improvement in single-merged separation can be achieved by circumscribing the track selection, in the full simulation, to tracks with large impact parameter significance. A comparison of track multiplicity distributions using all tracks and distributions built with displaced tracks only is shown in Fig. 2.5. No improvement is obtained by using displaced tracks. The potential sensitivity achieved by enriching the sample in tracks associated to the b-hadron is counterbalanced by the lower number of associated tracks.

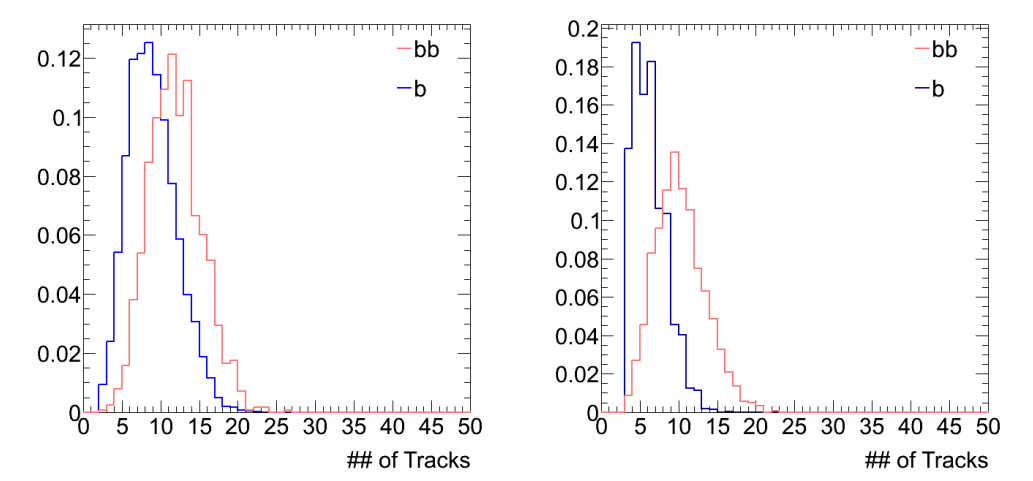

Figure 2.4: Distribution of the charged particle multiplicity for single (b) and merged (bb) jets from 80 GeV to 120 GeV in a sample of dijet events generated with PYTHIA and no detector simulation. Distributions are shown using all charged particles (left) and using only charged particles coming from b-hadron decay (right). A better discrimination between single and merged b-jets is obtained when using tracks from b-decay only.

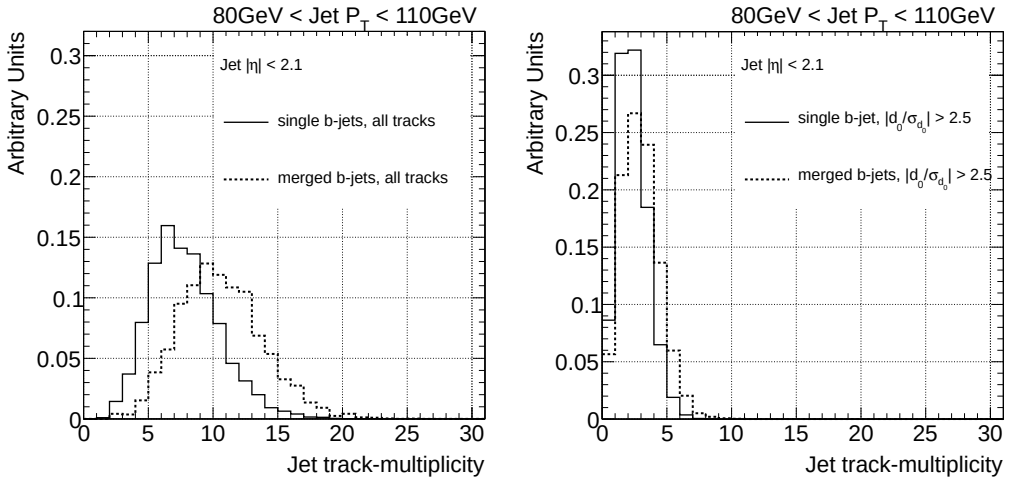

Figure 2.5: Distribution of the jet track multiplicity for single and merged b-jets from 80 GeV to 110 GeV, for all (left) and displaced tracks only (right). No improvement is obtained by using displaced tracks.

Jet width

The jet width is part of a set of continuous variables, like geometric moments, that are sensitive to the distribution of the constituents within a jet. This particular combination is a linear moment which sums the distances between the jet constituents and its axis, weighted by the constituents p_T. Its definition is,

$$Jet\ width = \frac{\sum_{i=1}^{N} {p_T}^{const_i}\, \Delta R(const_i, jet)}{\sum_{i=1}^{N} {p_T}^{const_i}} \tag{2.3}$$

where N is the total number of calorimeter, track or particle constituents.

This observable has also found use in the discrimination between gluon initiated and light quark initiated jets, see for instance [25] and [26]. Gluon jets are seen to be broader than quark jets. In the case of jets originating from b-quarks, these resemble gluon jets more closely than quarks jets [27]: due to the mass diference between b-hadrons and light-quark hadrons the

angular spread is larger for a b-jet than a light-quark jet.

In order to explore how merged jets, originating from a gluon splitting into a $b\bar{b}$ pair, compare to single b-quark jets and pure gluon jets, a standalone PYTHIA analysis was performed. Figure 2.6 illustrates the result. b-jets containing two b-hadrons present a greater angular width relative to single b-jets and gluon initiated jets. The latter, in turn, look broader than single b-jets. This behavior is somehow expected in the LHC's higher p_T jets because the QCD shower produces more particles resulting in broader gluon jets, with more jet-to-jet fluctuations, while the particle multiplicity is relatively fixed in the b-hadron decay.

The distribution of the track-jet width for the full ATLAS simulation is shown in Fig. 2.7. In this case the sum in equation 2.3 runs over the N tracks associated to the jet, using the same criteria as for the jet track multiplicity. As expected, merged b-jets are wider than single b-jets.

PYTHIA standalone samples were also used to evaluate the potential gain in discrimination obtained by utilising all stable particles in the event to build the observable, as opposed to using the charged particles only. A 10% improvement in merged b-jet rejection (for a 50% efficiency in selecting single b-jets) was achieved.

In full simulation, the jet width can be measured in terms of calorimeter variables by replacing tracks by topological clusters in the sum (this is somehow the equivalent in full simulation of switching from charged to all particles). Although it offers good separation, this variable is more sensitive to the amount of pile-up in the event than its track-based counterpart. This is illustrated in Fig. 2.8, which shows the distribution of calorimeter width and track-jet width for single b-jets in events with low and high number of primary vertices (NPV) in a low p_T region where the effect of pile-up is more

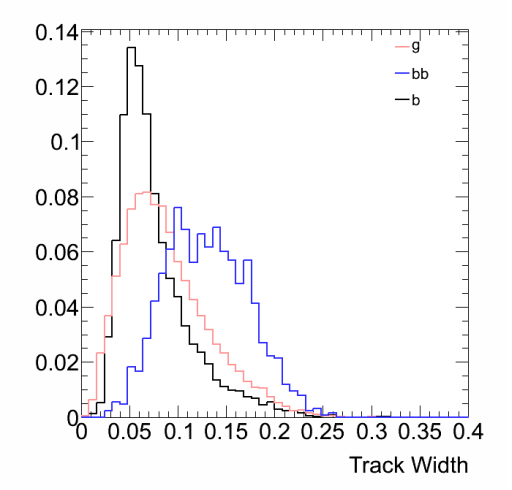

Figure 2.6: Distribution of track-jet width for gluon-initiated (g), single (b) and merged (bb) jets from 80 GeV to 120 GeV in a sample of dijet events generated with PYTHIA and no detector simulation.

important.

In general, all the studied calorimeter-based jet variables show similar dependences with NPV. For this reason the track-based versions are preferred as more robust discriminators.

Jet Mass

The reconstructed jets, built from massless topological clusters, obtain mass in the recombination process. The single-jet mass is defined as

$$Jet\ mass = E_{jet}^2 - \ \mathbf{p}^2 = (\sum_i E_i)^2 - (\sum_i \mathbf{p}_i)^2 \tag{2.4}$$

with E_i and $\mathbf{p}_i$, the energy and momentum of the jet constituent i. This observable is highly correlated to the jet width.

The jet mass, like the linear radial moment, depends on the radiation pattern of the event. It is the most basic observable for distinguishing massive boosted objects from jets originating from quarks or gluons [28].

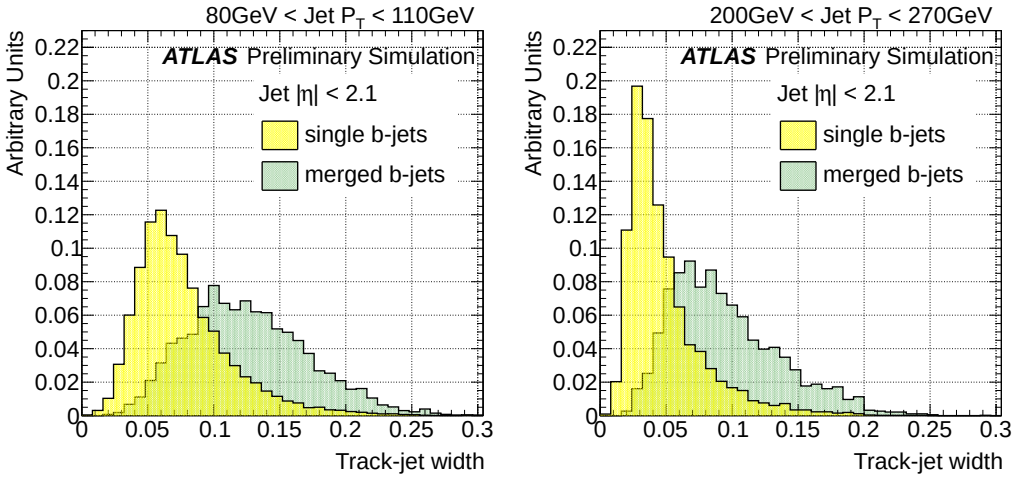

Figure 2.7: Distribution of track-jet width for single and merged b-jets from 80 GeV to 110 GeV (left) and 200 GeV to 270 GeV (right).

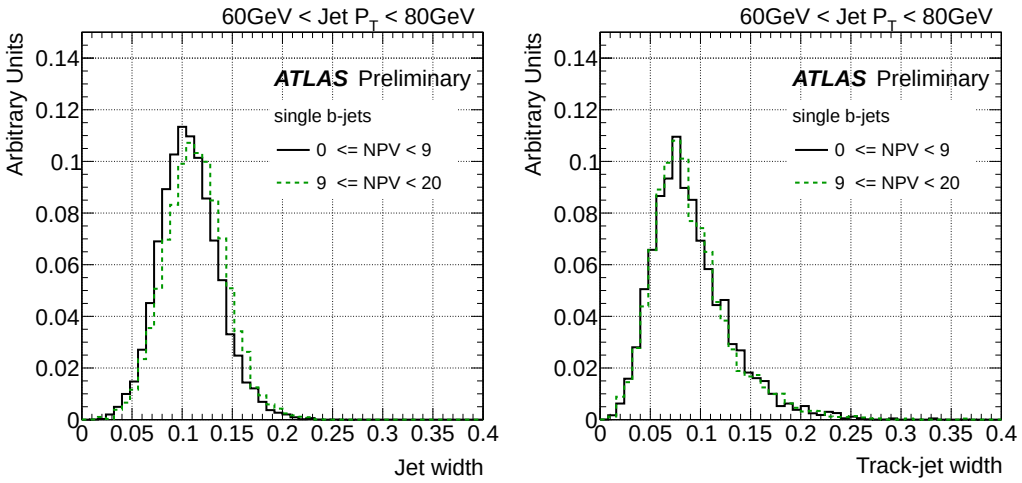

Figure 2.8: Distribution of jet width using topological clusters (left) and tracks (right) for single b-jets in two bins of number of primary vertices (NPV) for jets from 60 GeV to 80 GeV.

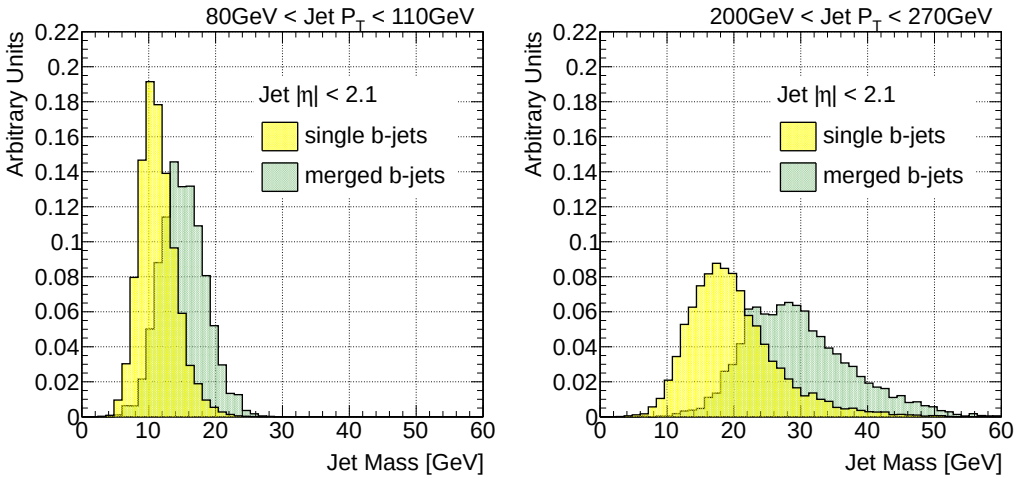

Figure 2.9: Distribution of jet mass in GeV for single and merged b-jets from 80 GeV to 110 GeV (left) and 200 GeV to 270 GeV (right).

Detector level jet mass distributions for jets selected to have $80 < p_T < 110$ GeV and $200 < p_T < 270$ GeV are shown in Fig. 2.9, both for single and merged b-jets. Merged jets tend to have higher masses than single b-jets for the same p_T bin. Although it shows good separation, this calorimeter based variable can be significantly affected by the amount of pile-up in the event as even a single soft wide angle deposition will have an effect on the jet mass, shifting the distribution to higher values[1].

ΔR between leading tracks

An alternative approach to measuring the width is to use the angular separation of the two hardest constituents inside jets. This has the advantage of removing any dependence on the shower development within the calorimeter

[1] In the ATLAS analysis of 35 pb^{-1} of 2010 data, the sensitivity of individual jet mass to pile-up is directly tested (for jets with at least 300 GeV). The mean jet mass is observed to increase linearly with NPV [29].

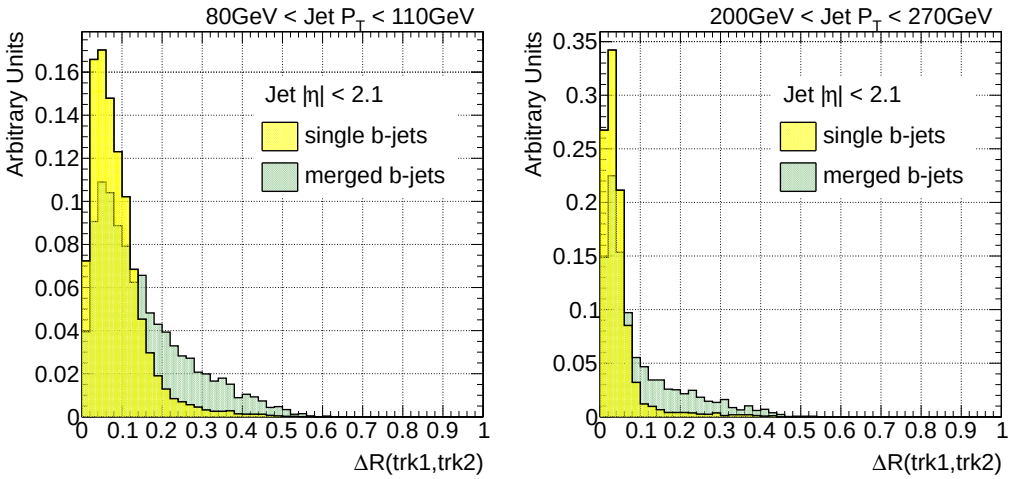

Figure 2.10: Distribution of ΔR between leading tracks for single and merged b-jets from 80 GeV to 110 GeV (left) and 200 GeV to 270 GeV (right).

and focuses on the hard components of the jet.

Figure 2.10 shows the distribution of the ΔR between leading tracks in the jet for single and merged b-jets. The merged b-jet distributions are slightly broader than single b-jet distributions for medium jet p_T. The effect diminishes as we go to higher transverse momentum values, offering very poor discrimination.

Maximum ΔR between track pairs

Several other variables, besides the jet width, were investigated to expose the expected two-subjet substructure of merged b-jets. The maximum ΔR separation between pairs of tracks associated to the jet ($\max\{\Delta R(trk, trk)\}$) is one example. Its distribution is shown in Fig. 2.11, for single and double b-hadron jets. The latter show significantly higher values over a broad range of jet p_T. The distinct characteristic of this variable is that the separation between single b-jets and merged does not depend on jet p_T.

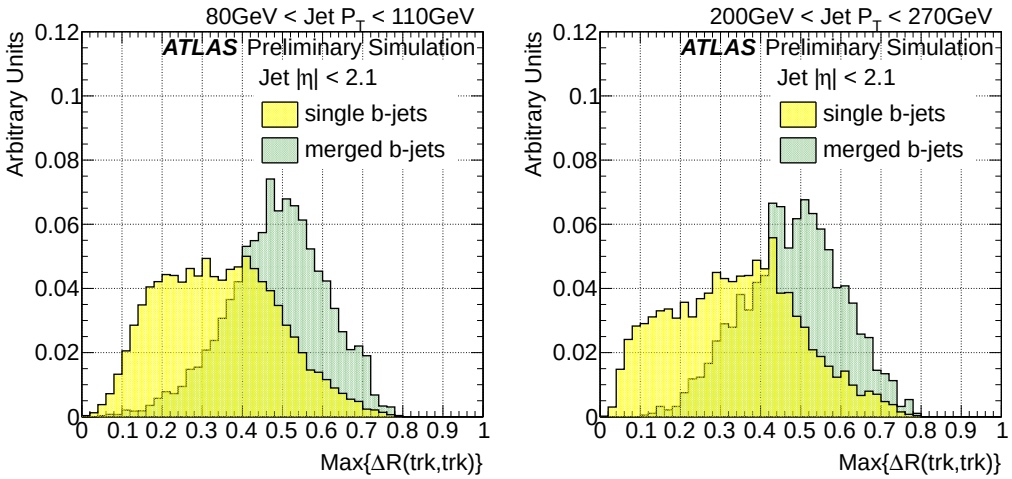

Figure 2.11: Distribution of the maximum ΔR between pairs of tracks for single and merged b-jets from 80 GeV to 110 GeV (left) and 200 GeV to 270 GeV (right).

In spite of its good discrimination power, alternative characterising variables are desirable as $\max\{\Delta R(trk, trk)\}$ is not infrared safe as it is affected by soft radiation. Furthermore it is sensitive to soft tracks originating from pile-up.

Subjet multiplicity

Subjet reconstruction has a similar approach as jet reconstruction but, rather than looking at all clusters (for topocluster jets) in an event, the subjet analysis is limited to objects only within a jet. The subjet multiplicity is the number of the reconstructed subjets within a jet and it provides information on the distribution of energy and multiplicity of particles within a jet. A measurement of this observable for quark and gluon jets indicates that gluon-initiated jets tend to have on average higher subjet multiplicity [30]. This result is consistent with the QCD prediction that gluons radiate more than

quarks.

The subjets were resolved by use of the inclusive k_t jet algorithm on the jet constituents with a fixed distance parameter. The k_t algorithm is the only jet algorithm that correctly identifies the resulting substructure as physical objects and therefore is the algorithm used for substructure analysis. As an alternative to fixed distance parameter subjets, it is also possible to undo the last step in the recombination sequence in order to identify the decay products of an object. This corresponds conceptually to undoing the first step in the fragmentation process that leads from interacting partons to jets. This approach is used in several jet grooming procedures[2], see for instance [32].

Figure 2.12 shows the distribution of the number of subjets for single and merged b-jets. The subjets in this case were built using the associated tracks as constituents, clustered by the inclusive k_t algorithm with a fixed distance parameter of $R = 0.2$. Merged jets tend to have on average one more subjet than single b-jets. The discrimination power of this variable is very poor and has the problem of being discrete with small numbers.

ΔR between the axes of two k_t subjets

The ΔR between k_t subjets is obtained by applying the k_t algorithm to the tracks associated to the jet using a large k_t distance parameter in order to ensure that all tracks get combined. The clustering is stopped once it reaches exactly two jets. This is done in Fastjet (see Section **??**) by the so called "exclusive" k_t algorithm. The exclusive k_t subjets correspond to reversing one step the process of clusterization, obtaining thus the two objects that,

[2]Jet grooming comprises dedicated techniques to remove uncorrelated radiation within a jet. A review of these procedures can be found in [31].

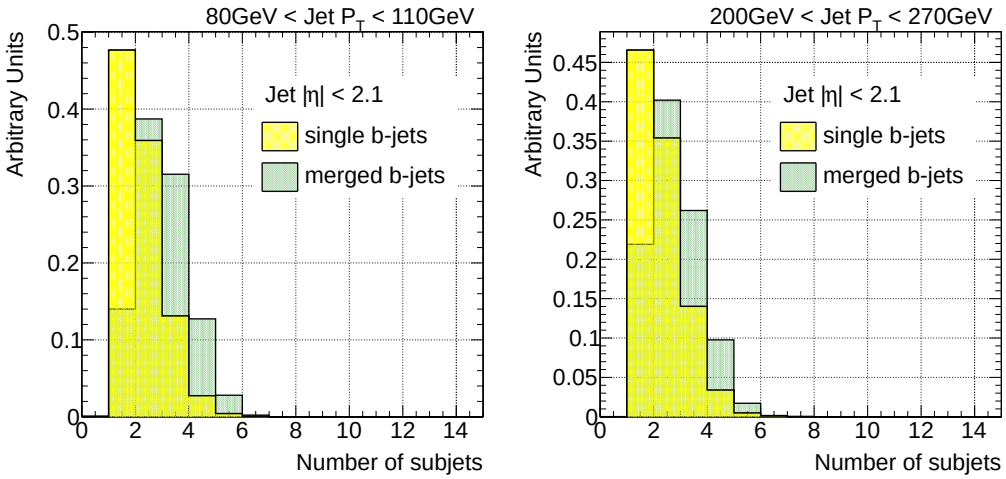

Figure 2.12: Distribution of the number of k_t sub-track-jets for single and merged b-jets from 80 GeV to 110 GeV (left) and 200 GeV to 270 GeV (right).

upon merging, give rise to the final jet. This cannot be done with anti-k_t.

The ΔR between the axes of the two exclusive subjets is shown in Fig. 2.13. As expected, it is larger for merged than for single jets. We observe that this variable provides very good separation, with the advantage of infrared safety and insensitivity to pile-up as opposed to $\max\{\Delta R(trk, trk)\}$.

In order to illustrate what this variable represents, an event display of a merged b-jet with a large (> 0.3) ΔR value is shown in Fig. 2.14. The plot illustrates in a 0.1×0.1 grid the area covered by the jet (in blue) and the position of the clusters associated to each jet, with the color indicating the value of their energy. The area of the jet (the section of the $\eta - \phi$ space belonging to it) is obtained by means of the "jet active area" concept, proposed in Ref. [33]. The event display indicates how the high energy cells in the jet with two b-hadrons (in red) are grouped around the b-hadrons directions, leading to the two-subjet substructure of merged jets.

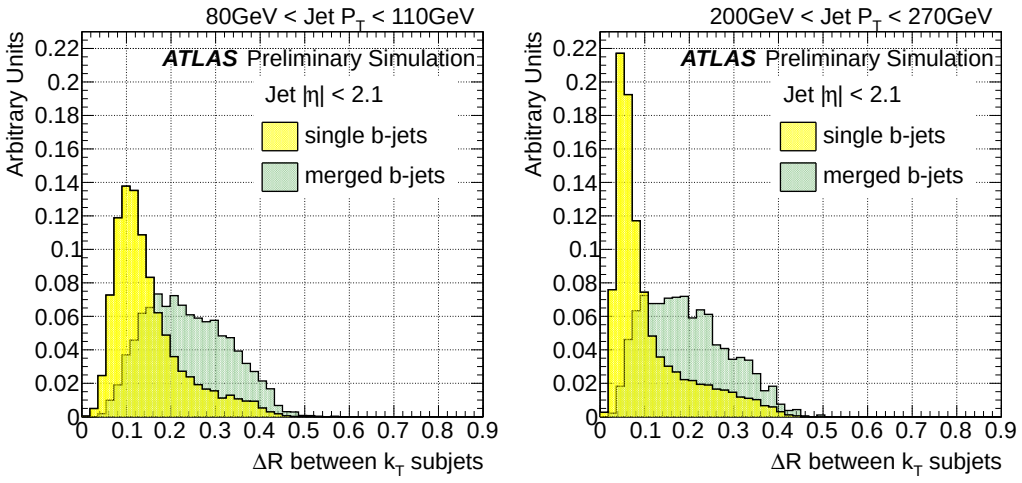

Figure 2.13: Distribution of the ΔR between the axes of the two k_t subjets in the jet for single and merged b-jets from 80 GeV to 110 GeV (left) and 200 GeV to 270 GeV (right).

N-subjettiness variables

It is possible to extend the use of individual subjets in conjunction with more sophisticated jet shape variables. Using these tools, an inclusive jet shape based on the substructure topology of a single jet, "N-subjettiness" has been recently proposed [34]. This variable describes the energy flow within a jet, quantifying the degree to which radiation is aligned along N subjet axes. That is, it characterizes how consistent a jet is with an N-subjet substructure. This jet shape was adapted from the event shape N-jettiness [35].

Given candidate subjets directions determined by an external algorithm such as the exclusive k_t procedure, the variable is defined as,

$$\tau_N^{(\beta)} = \frac{1}{\sum_k p_{Tk} (R_0)^\beta} \sum_k p_{Tk} (\min\{\Delta R_{j1,k}, \Delta R_{j2,k}, ..., \Delta R_{jN,k}\})^\beta. \quad (2.5)$$

The sum runs over the k constituents in a given jet where $p_{T,k}$ are their transverse momenta, and $\Delta R_{j1,k}$ is the distance between the candidate subjet

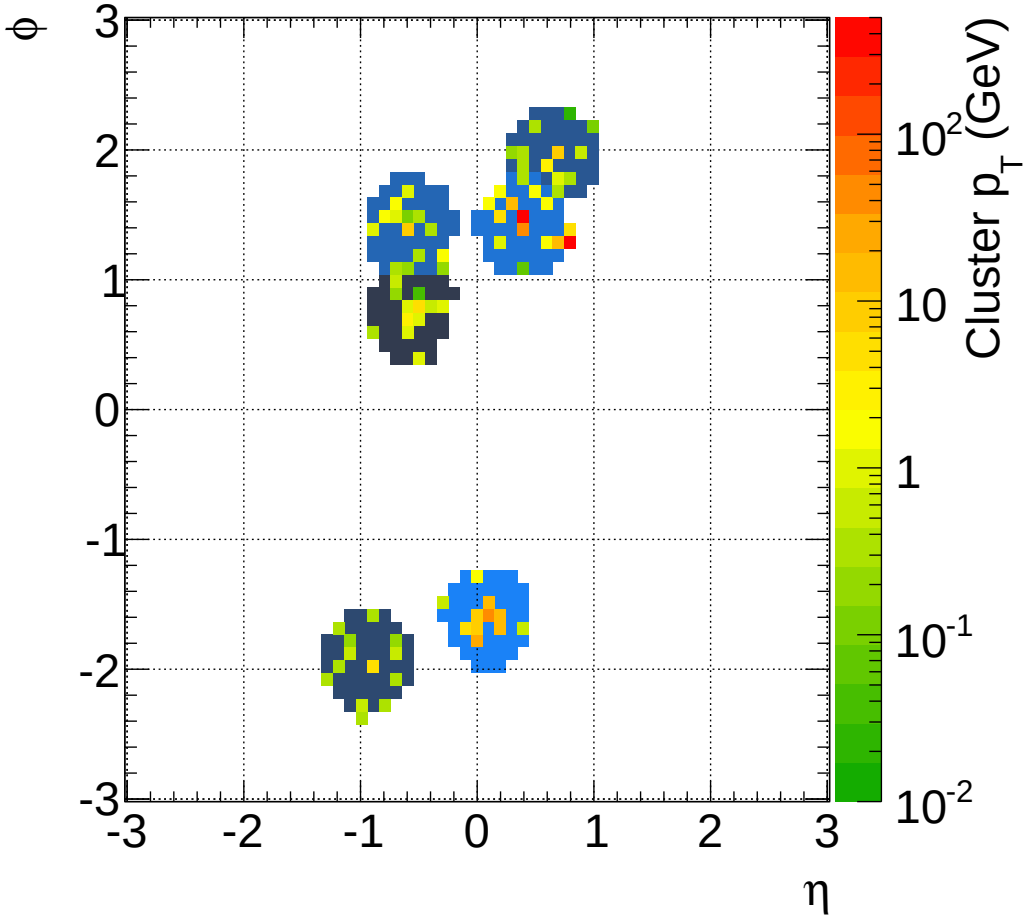

Figure 2.14: Event display of a merged b-jet in $(\eta, \phi) = (0.46, 1.41)$ and $p_T = 110$ GeV. The two b-hadrons are indicated as two red squares. The area of the jet is shown in blue and the topoclusters belonging to the jet are shown in different colors, from green to orange, depending on their transverse momentum. The ΔR between the axes of the two k_t subjets in the jet is larger than 0.3. The two-subjet structure of the merged jet is displayed.

$j1$ and a constituent particle k. R_0 is the characteristic jet radius used in the original jet clustering algorithm. The exponential weight, β, can optionally be applied to the angular distance computed between the subjets and the jet constituents. Since eq. 2.5 is linear in the p_T of the constituent particle, this variable is an infrared-safe observable.

This jet shape was designed to separate boosted hadronic objects, like electroweak bosons and top quarks decaying into collimated showers of hadrons which a standard jet algorithm would reconstruct as single jets. A simple cut on the ratio τ_N/τ_{N-1} provides excellent discrimination power for N-prong hadronic objects [34] . In particular, τ_2/τ_1 can identify boosted W/Z and Higgs bosons, with the angular weighting exponent $\beta = 1$ providing the best discrimination.

The definition of N-subjettiness is not unique, and different choices can be used to give different weights to the emissions within a jet. The initial step of choosing candidate subjet axes is in fact unnecessary; the quantity in equation 2.5 can be minimised over the candidate subjet directions, further improving boosted object discrimination.

To avoid dependence on pile-up we consider track-based N-subjettiness, where the sum is over the tracks in the b-tagged jet. As seen for massive boosted objects, a jet with a two pronged structure, with all tracks clustered along two directions, is expected to have a smaller τ_2 value than a jet with a more uniform track distribution. The distributions of τ_2, shown in Fig. 2.15, display good separation between single and merged jets, but with the latter showing larger values than single. This behavior can be traced to the level of correlation between τ_2 and track-jet width, displayed in Fig. 2.16a, to be compared to the much lower correlation presented, for instance, between track-jet width and jet track multiplicity, shown in Fig. 2.16b.

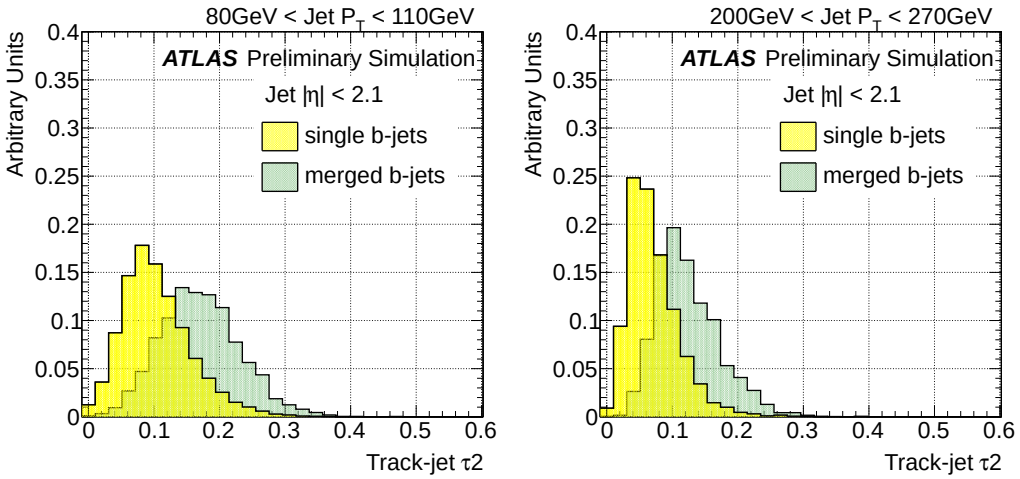

Figure 2.15: Distribution of τ_2 for single and merged b-jets from 80 GeV to 110 GeV (left) and 200 GeV to 270 GeV (right).

The correlation observed suggests to switch from an absolute to a width-normalized τ_2, and evaluate the ratio τ_2/τ_1, as shown in Fig. 2.17. Somewhat larger values are obtained for single than for merged b-jets, specially at high p_T, as expected. However, the difference is small, producing only a marginal discrimination, indicating that gluon splitting jets do not present a marked 2-subjet structure as boosted Z or H fat jets.

Jet eccentricity

In defining a jet moment there are several ways to weight the momentum and define the center of the jet. We have described the jet width as the first moment of the transverse energy with respect to the jet axis. But it is also natural to look at higher moments, such as those contained in the 2×2 matrix,

$$\begin{bmatrix} \sum p_{Ti}\delta\eta_i^2 & -\sum p_{Ti}\delta\eta_i\delta\phi_i \\ -\sum p_{Ti}\delta\eta_i\delta\phi_i & \sum p_{Ti}\delta\phi_i^2 \end{bmatrix} \tag{2.6}$$

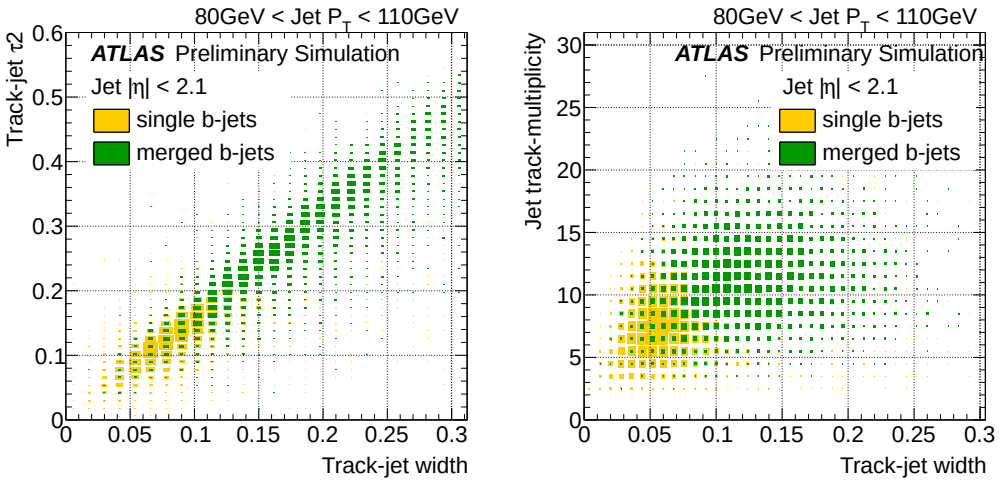

Figure 2.16: Correlation between τ_2 and track-jet width (left) and jet track multiplicity and track-jet width (right) for single and merged b-jets from 80 GeV to 110 GeV.

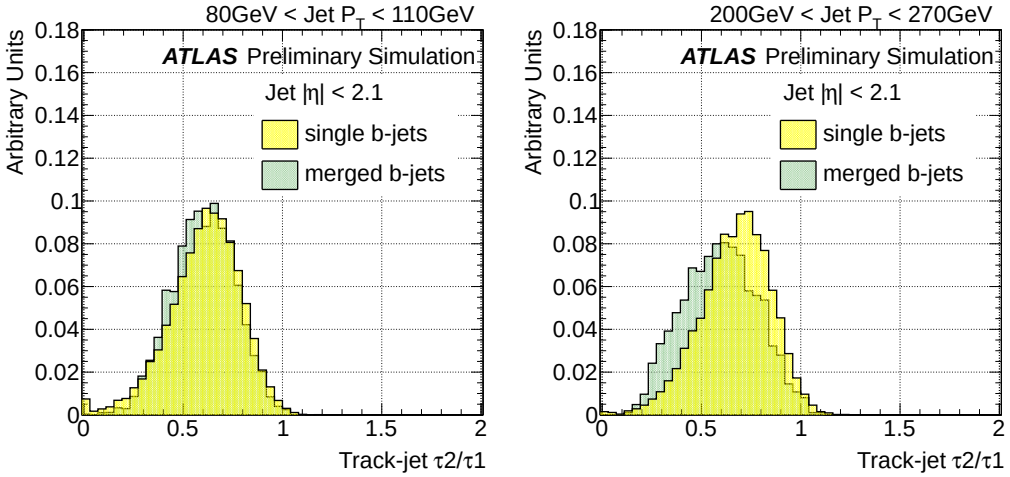

Figure 2.17: Distribution of τ_2/τ_1 for single and merged b-jets from 80 GeV to 110 GeV (left) and 200 GeV to 270 GeV (right).

Here, $(p_{Ti}, \delta\eta_i, \delta\phi_i)$ are the jet constituent transverse momentum and its pseudorapidity and azimuthal angle measured with respect to the jet axis, respectively. The eigenvalues $\lambda_m \leq \lambda_p$ of this tensor are associated to the semiminor and semimajor axes of an elliptical approximation to the jet shape in the $\eta - \phi$ plane. The jet eccentricity, defined below, is a combination of these eigenvalues, and it is a measure of how elongated the area of a jet is,

$$e = \sqrt{1 - r^2} \tag{2.7}$$

where the parameter r is defined as the ratio of the eigenvalues,

$$r = \frac{\lambda_m}{\lambda_p} = \frac{\sum p_{Ti}\delta\eta_i^2 + \sum p_{Ti}\delta\phi_i^2 - \sqrt{(\sum p_{Ti}\delta\eta^2 - \sum p_{Ti}\delta\phi_i^2)^2 + 4(\sum p_{Ti}\delta\eta_i\delta\phi_i)^2}}{\sum p_{Ti}\delta\eta_i^2 + \sum p_{Ti}\delta\phi_i^2 + \sqrt{(\sum p_{Ti}\delta\eta^2 - \sum p_{Ti}\delta\phi_i^2)^2 + 4(\sum p_{Ti}\delta\eta_i\delta\phi_i)^2}}. \tag{2.8}$$

Figure 2.18 shows the distribution of the jet eccentricity, built using track constituents. Although merged jets tend to be less spherical than single jets the difference is only marginal and essentially nonexistent for high p_T jets. The definition of the track-eccentricity, in Equation 2.7, weights the angular distances by the associated tracks p_T. Therefore, any pair of tracks with transverse momentum much higher than the rest will lead to a jet eccecntricity $\sim$1.

2.4 Validation of the jet variables in data

In order to study the extent to which the simulation reproduces the distributions observed in data for the different variables explored a comprehensive programme of data and MC comparisons was carried on. A few examples are presented in this section. Figures 2.19 to 2.23 show distributions of jet track multiplicity, track-jet width, ΔR between the axes of the two k_t subjets, $\max\{\Delta R(trk, trk)\}$ and τ_2 in two different p_T bins for b-tagged jets in dijet

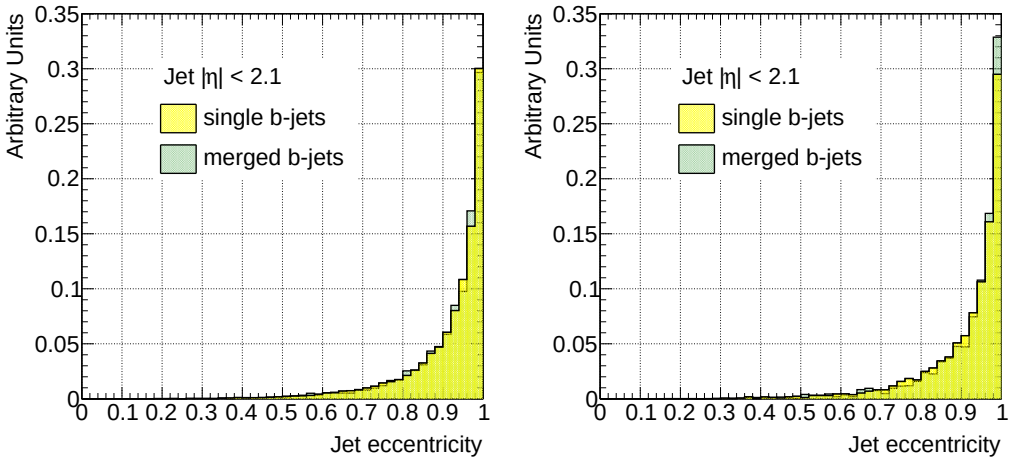

Figure 2.18: Distribution of the jet eccentricity for single and merged b-jets from 80 GeV to 110 GeV (left) and 200 GeV to 270 GeV (right).

Monte Carlo and data events passing the selection described in Section 2.2.1. The distributions are normalized to unit area to allow for shape comparisons. There is a very good agreement between data and simulation in all cases.

It should be remarked that the observed agreement is actually not a direct validation of the description in the MC of the relevant variables, but its convolution with the simulated relative fractions of light-, c-, b- and bb-jets in the b-tagged generated jet sample. To some extent, there could be some level of compensation between these two effects. To study this posibility, the agreement between data and simulation was evaluated in b-jets selected with a looser cut of MV1 tagger (70% efficiency working point) as well as with another b-tagging algorithm, the JetFitter. The result is shown for the jet track multiplicity in Figures 2.24 and 2.25.The agreement is still very good, suggesting that this compensation is not likely to occur in samples sufficiently enriched in b-jets.

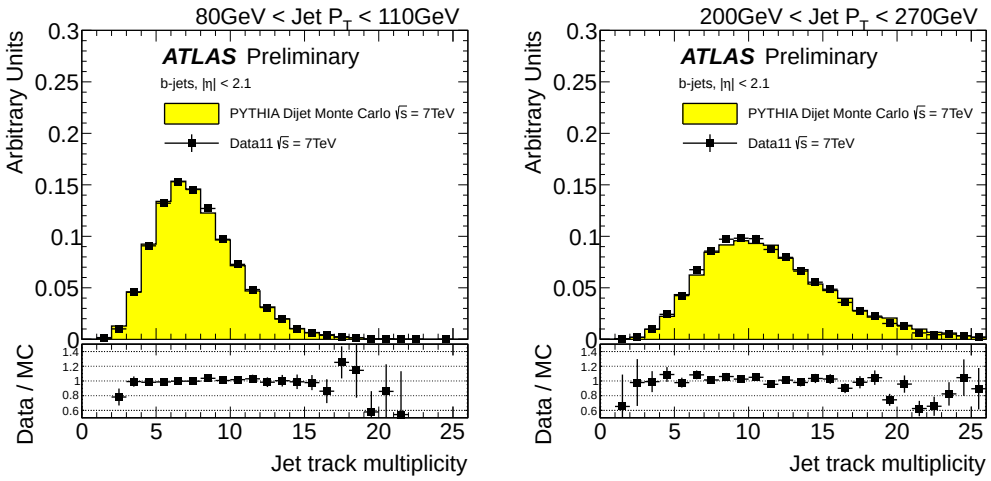

Figure 2.19: Distribution of the jet track multiplicity in 2 different jet p_T bins, for experimental data collected by ATLAS during 2011 (solid black points), and simulated data (filled histograms). The ratio data over simulation is shown at the bottom of each plot.

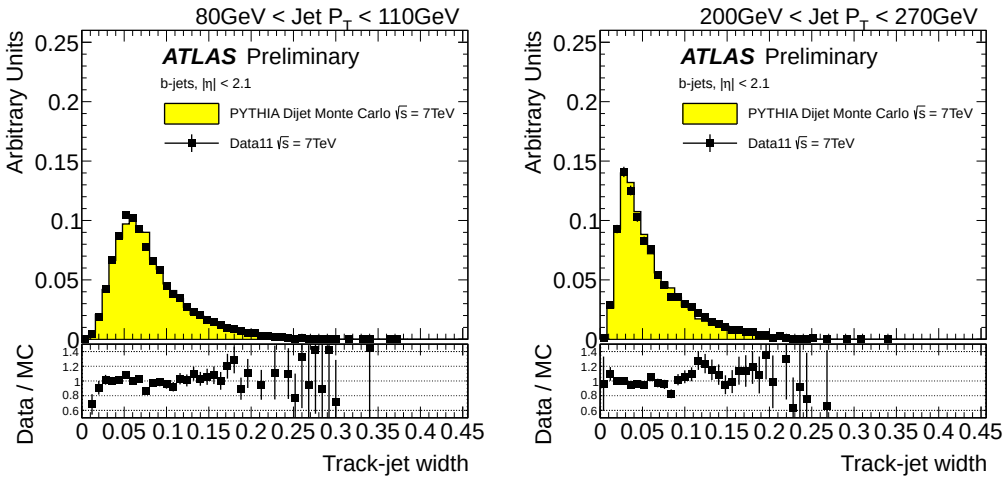

Figure 2.20: Distribution of the track-jet width in 2 different jet p_T bins, for experimental data collected by ATLAS during 2011 (solid black points), and simulated data (filled histograms). The ratio data over simulation is shown at the bottom of each plot.

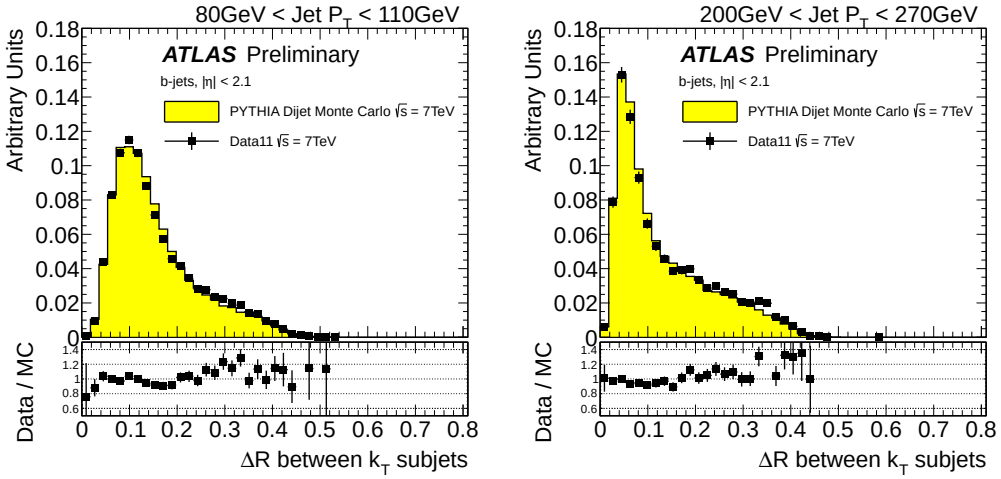

Figure 2.21: Distribution of the ΔR between the axes of the two k_t subjets in the jet in 2 different jet p_T bins, for experimental data collected by ATLAS during 2011 (solid black points), and simulated data (filled histograms). The ratio data over simulation is shown at the bottom of each plot.

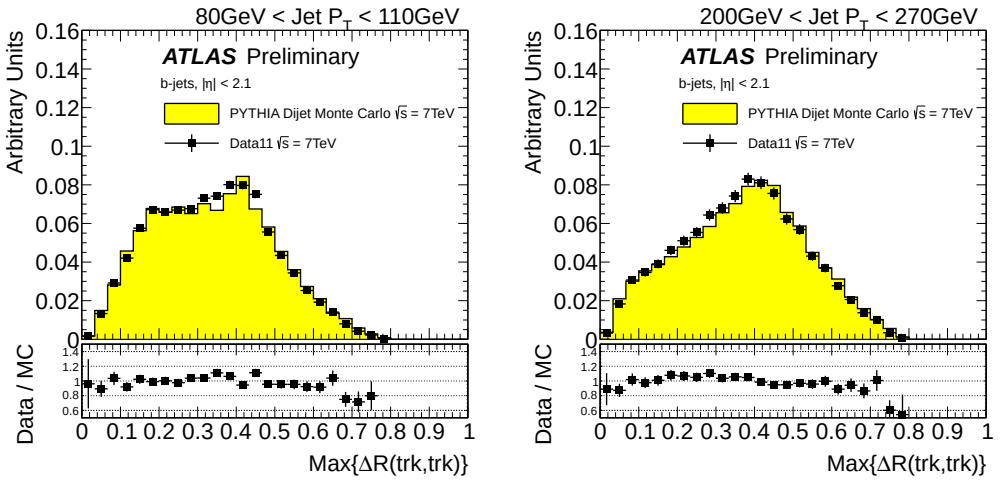

Figure 2.22: Distribution of maximum ΔR between pairs of tracks in two different jet p_T bins, for experimental data collected by ATLAS during 2011 (solid black points), and simulated data (filled histograms). The ratio data over simulation is shown at the bottom of each plot.

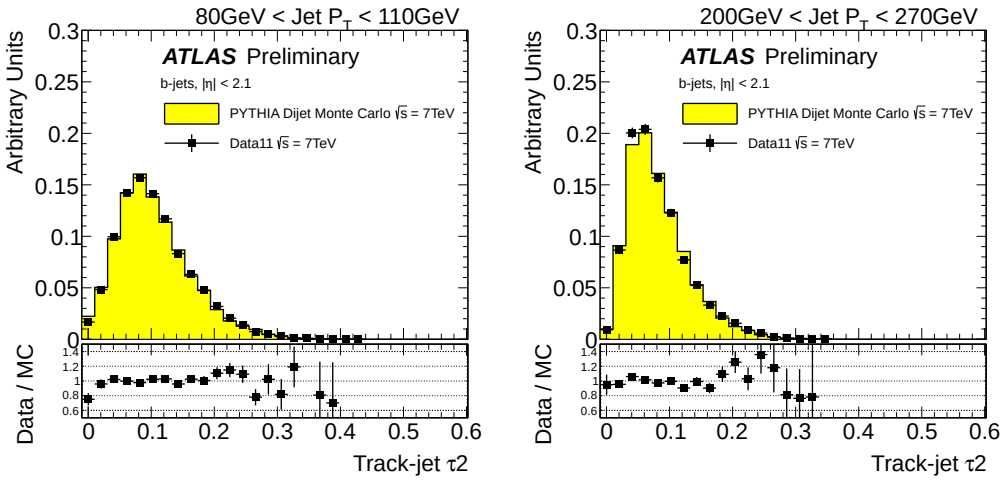

Figure 2.23: Distribution of τ_2 in two different jet p_T bins, for experimental data collected by ATLAS during 2011 (solid black points), and simulated data (filled histograms). The ratio data over simulation is shown at the bottom of each plot.

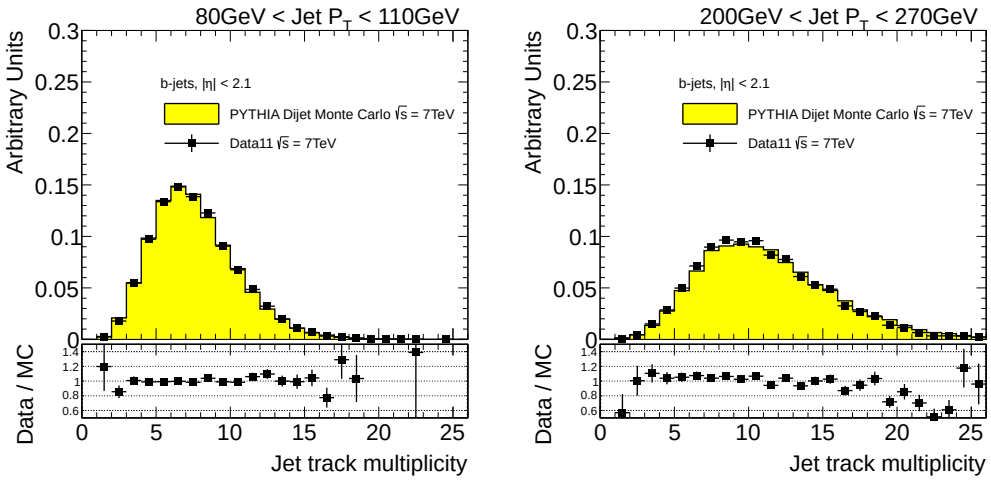

Figure 2.24: Distribution of the jet track multiplicity in 2 different jet p_T bins, for experimental data collected by ATLAS during 2011 (solid black points), and simulated data (filled histograms). Jets were selected using MV1 tagger at its 70% b-jet efficiency working point. The ratio data over simulation is shown at the bottom of each plot.The agreement is very good.

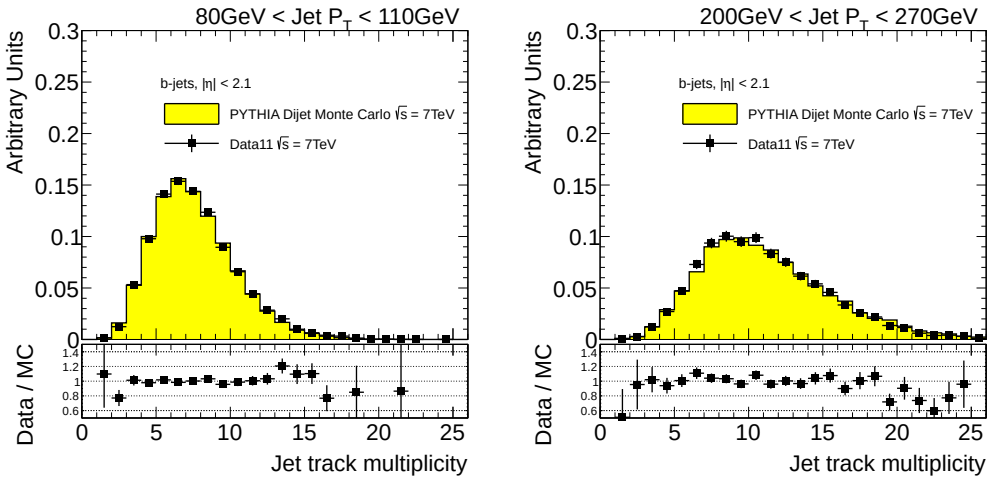

Figure 2.25: Distribution of the jet track multiplicity in 2 different jet p_T bins, for experimental data collected by ATLAS during 2011 (solid black points), and simulated data (filled histograms). Jets were selected using JetFitter tagger at its 60% b-jet efficiency working point. The ratio data over simulation is shown at the bottom of each plot.The agreement is very good.

Bibliography

[1] Calafiura, P. and Lavrijsen, W. and Leggett, C. and Marino, M. and Quarrie, D. The athena control framework in production, new developments and lessons learned. pages 456–458, 2005.

[2] ATLAS Collaboration. Performance of Jet Algorithms in the ATLAS Detector. *ATL-PHYS-INT-2010-129*, 2010.

[3] W. Lampl et al. Calorimeter Clustering Algorithms: Description and Performance. *ATL-LARG-PUB-2008-002. ATL-COM-LARG-2008-003*, Apr 2008.

[4] ATLAS Collaboration. ATLAS Calorimeter Response to Single Isolated Hadrons and Estimation of the Calorimeter Jet Scale Uncertainty. *ATLAS-CONF-2011-028*, 2011.

[5] G. Abbiendi et al. Inclusive analysis of the b quark fragmentation function in Z decays at LEP. *Eur.Phys.J.*, C29:463–478, 2003.

[6] Koya Abe et al. Measurement of the b quark fragmentation function in Z0 decays. *Phys.Rev.*, D65:092006, 2002.

[7] Andy Buckley, Hendrik Hoeth, Heiko Lacker, Holger Schulz, and Jan Eike von Seggern. Systematic event generator tuning for the LHC. *Eur. Phys. J. C*, 65:331–357, 2010.

[8] Bowler, M. G. Production of heavy quarks in the string model. *Zeitschrift fur Physik C Particles and Fields*, 11:169–174, 1981. 10.1007/BF01574001.

[9] T Cornelissen, M Elsing, S Fleischmann, W Liebig, E Moyse, and A Salzburger. Concepts, Design and Implementation of the ATLAS New Tracking (NEWT). *ATL-SOFT-PUB-2007-007. ATL-COM-SOFT-2007-002*, 2007.

[10] G. Aad et al. Charged-particle multiplicities in pp interactions measured with the ATLAS detector at the LHC. *New J.Phys.*, 13:053033, 2011.

[11] ATLAS Collaboration. Performance of primary vertex reconstruction in proton-proton collisions at $\sqrt{s}$ =7 TeV in the ATLAS experiment. *ATLAS-CONF-2010-069*, 2010.

[12] J. Neyman and E.S. Pearson. On the Problem of the Most Efficient Tests of Statistical Hypotheses. *Royal Society of London Philosophical Transactions Series A*, 231:289–337, 1933.

[13] ATLAS Collaboration. Commissioning of the ATLAS high-performance b-tagging algorithms in the 7 TeV collision data. *ATLAS-CONF-2011-102*, 2011.

[14] ATLAS Collaboration. Performance of Impact Parameter-Based b-tagging Algorithms with the ATLAS Detector using Proton-Proton Collisions at $\sqrt{s} = 7$ TeV. *ATLAS-CONF-2010-091*, 2010.

[15] V Kostyukhin. Vkalvrt - package for vertex reconstruction in atlas. *ATL-PHYS-2003-031*, Aug 2003.

[16] ATLAS Collaboration. Performance of the ATLAS Secondary Vertex b-tagging Algorithm in 7 TeV Collision Data. *ATLAS-CONF-2010-042*, 2010.

[17] ATLAS Collaboration. Calibrating the b-Tag Efficiency and Mistag Rate in 35 pb^{-1} of Data with the ATLAS Detector. *ATLAS-CONF-2011-089*, 2011.

[18] ATLAS Collaboration. Statistical combination of top quark pair production cross-section measurements using dilepton, single-lepton, and all-hadronic final states at $\sqrt{s} = 7$ TeV with the ATLAS detector. *ATLAS-CONF-2012-024*, Mar 2012.

[19] ATLAS Collaboration. Measuring the b-tag efficiency in a top-pair sample with 4.7 fb-1 data from the ATLAS detector. *ATLAS-CONF-2012-097*, 2012.

[20] ATLAS Collaboration. Measurement of the Mistag Rate with 5 fb^1 of Data Collected by the ATLAS Detector. *ATLAS-CONF-2012-040*, 2012.

[21] Torbjorn Sjostrand, Stephen Mrenna, and Peter Skands. PYTHIA 6.4 Physics and Manual. *JHEP*, 05:026, 2006.

[22] Atlas tunes of pythia 6 and pythia 8 for mc11. Technical Report ATL-PHYS-PUB-2011-009, CERN, Geneva, Jul 2011.

[23] G. Aad et al. The atlas simulation infrastructure. *The European Physical Journal C*, 70:823–874, 2010.

[24] ATLAS Collaboration. Selection of jets produced in proton-proton collisions with the ATLAS detector using 2011 data. *ATLAS-CONF-2012-020*, 2012.

[25] Jason Gallicchio and Matthew D. Schwartz. Quark and gluon tagging at the lhc. *Phys. Rev. Lett.*, 107:172001, Oct 2011.

[26] ATLAS Collaboration. Light-quark and gluon jets in atlas. *ATLAS-CONF-2011-053*, 2011.

[27] D. Buskulic et al. Quark and gluon jet properties in symmetric three-jet events. *Physics Letters B*, 384(1-4):353–364, 1996.

[28] Leandro G. Almeida, Seung J. Lee, Gilad Perez, Ilmo Sung, and Joseph Virzi. Top quark jets at the lhc. *Phys. Rev. D*, 79:074012, Apr 2009.

[29] ATLAS Collaboration. Measurement of Jet Mass and Substructure for Inclusive Jetsin $\sqrt{s} = 7$ TeV pp Collisions with the ATLAS Experiment. *ATLAS-CONF-2011-073*, May 2011.

[30] R. Snihur. Subjet multiplicity in quark and gluon jets at d0. *Nuclear Physics B - Proceedings Supplements*, 79(1âĂŞ3):494 – 496, 1999. Proceedings of the 7th International Workshop on Deep Inelastic Scattering and QCD.

[31] A. Abdesselam, E. Bergeaas Kuutmann, U. Bitenc, G. Brooijmans, J. Butterworth, et al. Boosted objects: A Probe of beyond the Standard Model physics. *Eur.Phys.J.*, C71:1661, 2011.

[32] Stephen D. Ellis, Christopher K. Vermilion, and Jonathan R. Walsh. Techniques for improved heavy particle searches with jet substructure. *Phys. Rev. D*, 80:051501, Sep 2009.

[33] G.P. Salam M. Cacciari and Gregory Soyez. The Catchment Area of Jets. *JHEP*, 0804:42, 2008.

[34] Jesse Thaler and Ken Van Tilburg. Identifying Boosted Objects with N-subjettiness. *JHEP*, 1103:015:026, 2011.

[35] Iain W. Stewart, Frank J. Tackmann, and Wouter J. Waalewijn. n jettiness: An inclusive event shape to veto jets. *Phys. Rev. Lett.*, 105:092002, Aug 2010.