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, Haris y Patricia, 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

Theoretical framework

In this chapter a short overview of the theory of elementary particles and fundamental interactions is presented, with emphasis on the strong interactions and the description of the hadronic final state in hadron collisions.

1.1 The Standard Model

The Standard Model (SM) is a quantum field theory that describes the behavior of all experimentally-observed particles under the influence of the electromagnetic, weak and strong forces[1]. In this model, all forces of nature are the result of particle exchange. The force mediators interact on the particles of matter, and, in some cases, due to the non-Abelian character of the theory[2], with each other.

Elementary particles are categorized into two classes of particles: *bosons*

[1] In principle gravitational forces should also be included in the list of fundamental interactions but their impact is fortunately negligible at the distance and energy scales usually considered in particle physics experiments.

[2] The transformations of the symmetry group do not commute in the case of the QCD and weak groups.

and *fermions*. Bosons have integer spin and obey the Bose-Einstein statistics, whereas fermions have half-integer spin and follow Fermi-Dirac statistics. Each elementary particle has a corresponding anti-particle, whose quantum numbers are opposite in sign.

The fundamental building blocks of matter predicted by the SM are fermions with spin 1/2:

- six leptons (and their antiparticles), organized in three families,

$$\begin{pmatrix} \nu_e \\ e \end{pmatrix} \begin{pmatrix} \nu_\mu \\ \mu \end{pmatrix} \begin{pmatrix} \nu_\tau \\ \tau \end{pmatrix}$$

- and six quarks (and their antiparticles), organized in three families,

$$\begin{pmatrix} u \\ d \end{pmatrix} \begin{pmatrix} c \\ s \end{pmatrix} \begin{pmatrix} t \\ b \end{pmatrix}$$

.

These particles are considered point-like, as there is no evidence of any internal structure of leptons or quarks to date. The six types of quarks are also known as the six quark flavors. Collectively, the u (up), d (down), and s (strange) quarks are frequently referred to as the light quarks. The heaviest quark of the Standard Model, the quark t (top), was the last to be found [1, 2]. The electric charge[3] Q of quarks adopts fractional values, i.e. $+2/3$ for quarks u, c and t and $-1/3$ for quarks d, s and b; yet they are only observed as the integer charge combinations of three quarks (baryons) or a quark and an antiquark (mesons).

[3]The electric charge is given in units of the elementary charge, e, which is the charge carried by the positron.

In addition, the model contains the vector bosons which are the carriers of the fundamental forces:

- a gauge boson for the electromagnetic interactions, the photon γ;
- three gauge bosons for the weak interactions, $W^{\pm}$ and Z^0;
- eight gauge bosons for the strong interactions, called gluons.

The Standard Model is based on a symmetry group of the kind $SU(3)_C \times SU(2)_L \times U(1)_Y$, where $SU(3)_C$ describes de *colour* symmetry of strong interactions, $SU(2)_L$ describes the *weak isospin* for the unified electroweak interactions and $U(1)_Y$, the invariance under *hypercharge* Y transformations. The twelve gauge bosons are associated with the generators of the symmetry groups of the theory. The exact symmetry of the SM predicts massless particles, one possible mechanism for breaking this symmetry is the existence of a massive scalar Higgs field that has non-zero vacuum expectation value [3]. Very recently, a Higgs-like particle was discovered by ATLAS and CMS experiments at the LHC [4]. This scalar boson completes the table of Standard Model particles.

Quantum electrodynamics (QED) is the relativistic quantum field theory based on the symmetry group $U(1)$ that describes the interaction of charged particles via the exchange of one (or more) photon. The coupling of charged fermion fields ψ to the photon field A^{μ} is described by the QED Lagrangian density, which is given by

$$\mathcal{L}_{QED} = \bar{\psi}(i\gamma^{\mu}D_{\mu} - m)\psi - \frac{1}{4}F_{\mu\nu}F^{\mu\nu}. \tag{1.1}$$

The covariant derivative D_{μ} and the field strength tensor $F_{\mu\nu}$ are given by

$$D_{\mu} = \partial_{\mu} - ieA_{\mu} \tag{1.2}$$

$$F^{\mu\nu} = \partial^{\mu}A^{\nu} - \partial^{\nu}A^{\mu} \tag{1.3}$$

such that the Lagrangian is invariant under local $U(1)$ gauge transformations. The γ^μ are the Dirac matrices, which satisfy $\{\gamma^\mu, \gamma^\nu\} = 2g^{\mu\nu}$. The strength of the interaction is characterized by the coupling $\alpha = e^2/4\pi$.

The full theory of QED was developed by Feynman, Schwinger and Tomonaga throughout the 1940s [5]. The structure of the SM is, in a sense, a generalisation of this theory, extending the gauge invariance of electrodynamics to a larger set of conserved currents and charges.

In addition to electromagnetic interactions, fermions are subject to weak interactions. Both are manifestations of the unified electroweak theory, which is described by the gauge symmetry $SU(2)_L \times U(1)_Y$. The fermion fields are expressed by Dirac spinors which can be decomposed into a left- and a right-handed component. The matrix operator $\gamma^5 = i\gamma^0\gamma^1\gamma^2\gamma^3$ has eigenvalues -1 for left-handed fermions and $+1$ for right-handed fermions. Consequently, the left- and right-handed projections are obtained by applying the chirality operators

$$P_L = \frac{1-\gamma^5}{2} \quad P_R = \frac{1+\gamma^5}{2} \tag{1.4}$$

respectively. The left-handed fermion fields $\psi_i = \binom{\nu_i}{l_i}_L$ and $\binom{u_i}{d_i'}_L$ of the i^{th} generation transform as doublets under the $SU(2)_L$ symmetry group. The conserved quantum number under $SU(2)_L$ transformations is the third component of the weak isospin, I_3, which is equal to $+1/2$ for the upper component in each doublet and $-1/2$ for its isospin partner. The right-handed fermion fields are invariant under $SU(2)_L$. The violation of parity in weak interactions is thus incorporated in the Standard Model.

The weak eigenstates of the quark fields are not identical to their mass eigenstates. Instead, they are linear combinations parametrized by the CKM (Cabibbo-Kobayashi-Maskawa) matrix V_{ij} [6], such that $d' = \sum_j V_{ij} d_j$. The coupling between fermions from different generations is thus proportional to

the (very small) off-diagonal elements of the CKM matrix.

Glashow, Weinberg and Salam proposed the unified description of the electromagnetic and weak interactions by introducing the $SU(2)_L \times U(1)_Y$ electroweak theory [7, 8, 9]. The gauge fields corresponding to the generators of the gauge symmetry are W^i_μ with $i = 1, 2, 3$, for $SU(2)_L$, and B_μ for $U(1)_Y$. The respective coupling strengths are denoted g and g' and the field strength tensors are given by

$$W^i_{\mu\nu} = \partial_\mu W^i_\nu - \partial_\nu W^i_\mu + g\epsilon_{ijk} W^j_\mu W^k_\nu \tag{1.5}$$

$$B_{\mu\nu} = \partial_\mu B_\nu - \partial_\nu B_\mu. \tag{1.6}$$

Analogous to $\mathcal{L}_{QED}$, the interactions between the gauge fields and fermions are described by the Lagrangian density

$$\mathcal{L}_{EW} = i \sum_f \bar{\psi}_f \gamma^\mu D_\mu \psi_f - \frac{1}{4} W^i_{\mu\nu} W^{i\mu\nu} - \frac{1}{4} B_{\mu\nu} B^{\mu\nu}, \tag{1.7}$$

which is invariant under local $SU(2)_L \times U(1)_Y$ gauge transformations when the covariant derivative is given by

$$D_\mu = \partial_\mu + \frac{1}{2} i g \tau^i W^i_\mu - \frac{1}{2} i g' Y B_\mu \tag{1.8}$$

The generators associated with the $SU(2)$ symmetry group are the Pauli matrices τ_i and the generator of the $U(1)_Y$ symmetry is the hypercharge Y, which is defined via

$$Q = Y + I_3 \tag{1.9}$$

Initially, the proposed unification failed because it predicted massless gauge fields associated to the generators of the $SU(2)_L$ symmetry group, analogous to the photon in QED, which were not observed. Instead there was indirect evidence for the massive charged $W^\pm$ and neutral Z^0 bosons, which have masses close to 80 and 90 GeV, respectively [10, 11]. A mechanism was

required for the weak bosons to acquire mass. The proposed solution involves spontaneous symmetry breaking through the Higgs mechanism.

The current theory of the strong interactions began with the identification of the elementary fermions that make up the hadrons (baryons and mesons). In 1963, Gell-Mann and Zweig proposed the quark model [12, 13, 14], which asserts that hadrons are in fact composites of smaller constituents. The quark model was formalized into the theory of Quantum Chromodynamics (QCD) with quarks carrying an additional quantum number called color. Without color charge, it would seem that the quarks inside some hadrons exist in symmetric quantum states, in violation of the Pauli exclusion principle (this was indeed the problem of the quark model as proposed by Gell-Mann and Zweig). The color theory extends the electroweak Lagrangian to be symmetric under $SU(3)_C$ transformations, which introduces eight new physical gauge fields, the gluons.

In this new picture a hadron is actually a complex composite object. A "core" set of *valence* quarks, as well as a *sea* of virtual quarks and gluons that are constantly being emitted and absorbed, comprise each hadron. Both valence quarks and sea quarks, along with the gluons, share the total momentum of the hadron.

The Quantum Chromodynamics (QCD) Lagrangian density is given by

$$\mathcal{L}_{QCD} = \sum_q \bar{\psi}_{q,a}(i\gamma^\mu(D_\mu)_{ab} - m_q\delta_{ab})\psi_{q,b} - \frac{1}{4}W^A_{\mu\nu}W^{A\mu\nu}. \tag{1.10}$$

The $\psi_{q,a}$ are the quark fields for flavor q and carry a color index a, which runs from 1 to $N_c = 3$. The covariant derivative D_μ and the gluon field strength tensor $G^A_{\mu\nu}$ are given by

$$D_\mu = \partial_\mu + ig_s t^A \mathcal{A}^A_\mu, \tag{1.11}$$

$$G^A_{\mu\nu} = \partial_\mu \mathcal{A}^A_\nu - \partial_\nu \mathcal{A}^A_\mu + g f^{ABC} \mathcal{A}^B_\mu \mathcal{A}^C_\nu, \tag{1.12}$$

where $\mathcal{A}_\mu^A$ are the gluon fields with index A, B, C running from 1 to $N_c^2 - 1 = 8$. The 3×3 matrices t^A are the generators of the $SU(3)$ group and satisfy $[t^A, t^B] = if^{ABC}t^C$. The strong coupling strength g_s is usually replaced by $\alpha_s = g_s^2/(4\pi)$. The QCD Feynman rules that follow the Lagrangian are the quark and gluon propagators and the vertices $q\bar{q}g$, ggg, and $gggg$.

1.2 Perturbative QCD

As described in section 1.1, the fundamental actors of the theory of the strong interactions are quarks and gluons or, collectively, partons [15]. Partons are confined in hadrons, but, act free at sufficiently small scales. This behaviour is called asymptotic freedom. The essence of asymptotic freedom is that the strong force couples particles together more strongly as the distance between them increases. This explains why quarks and gluons are only observed, at low energies, trapped together into color-neutral hadrons, in an effect known as confinement[4]. A quantitative representation of the decreasing power of the strong force with increasing energy is given by the negative β-function of QCD [16, 17], which describes how the coupling constant of the force changes with energy. The variation of the strength of the coupling with the energy is referred to as the "running" of the coupling constant.

The experimental consequence of asymptotic freedom is that quarks and gluons require interactions with high energy probes to be ejected from nucleons, and that its presence can only be inferred indirectly. Measurements of deep inelastic lepton-hadron scattering provided some of the first indications of the presence of quarks. The momentum transfer, Q^2, between the

[4]In very-high energy environments, such as the universe shortly after the Big Bang, quarks and gluons are only weakly linked by the strong force, forming what is called a quark-gluon plasma.

probe particles (leptons) and the target hadron is related to the distance scale within the hadron being measured.

The low value of the strong coupling constant at high-energies permits the use of perturbative techniques to calculate physical processes. Each higher order of the perturbative expansion contains an additional factor of the coupling constant, α_s. Since the value of α_s varies with energy, it must be evaluated at some energy scale, close to the energy scale of the interaction. At energies of $\sim M_Z$, α_s is ~ 0.117, and the higher order terms can be ignored to yield an approximate solution. Thus from an expansion of an infinite number of terms, only a few need to be computed. The complexity of the process determines the precision of the calculation that can be performed. For inclusive parton production, calculations are typically performed at next-to-leading order (NLO). To help organize the computation of the multitude of terms in perturbative calculations, the tool of Feynman diagrams is frequently used. Feynman diagrams are graphical representations of the terms of the perturbative expansion. The outer lines of the diagram correspond to incoming and outgoing particles, the inner lines correspond to virtual particles, and the vertices correspond to particle interactions. To each of these components of the diagram, a mathematical expression or operation is assigned. Each vertex contributes with $\sqrt{\alpha_s}$ to the matrix element, with the exception of the 4-gluon vertex wich contributes with α_s. Each increasing order in α_s of the perturbative expansion simply corresponds to a set of diagrams with the correct combination of vertices. By drawing all possible Feynman diagrams for a given order of perturbation theory, all the terms in the calculation can be read off. In this context, leading-order diagrams are also known as "tree-level" diagrams (with no internal loops).

Using this formalism, the cross section for two partons to interact can

be computed up to some fixed-order in perturbation theory, but there is a further complication. Colliders such as the LHC do not produce simple parton-parton interactions, but instead collisions of hadrons that consist of multiple partons.

The factorization theorem [18] allows the perturbative calculations for parton interactions to be extended to proton-proton collisions. This theorem states that the total cross section for two hadrons to interact can be obtained by weighting and combining the cross sections for two particular partons to interact. This weighting is done using parton distribution functions (PDFs), which state the probability for a certain parton to carry a momentum fraction x of the total hadron momentum. Thus the total cross section, at some momentum Q that characterizes the interaction, can be written as:

$$\sigma(P_1, P_2) = \sum_{i,j} \int dx_1 dx_2 f_i(x_1, \mu_f^2) f_j(x_2, \mu_f^2) \hat{\sigma}_{ij}(p_1, p_2, \alpha_s(\mu_r^2), Q^2/\mu_r^2, Q^2/\mu_f^2). \tag{1.13}$$

Here, P_1 and P_2 are the momenta of the two incoming hadrons, x_1 and x_2 are the momentum fractions carried by the two interacting partons, and $p_1 = x_1 P_1$ and $p_2 = x_2 P_2$ are the momenta of the two interacting partons. The partonic cross section $\hat{\sigma}_{ij}$, corresponding to the interaction of partons i and j, is calculated at a fixed order in α_s, which is evaluated at some renormalization scale, μ_r. The renormalization scale is the scale at which the natural divergences in the cross sections are canceled by counter-terms in the Lagrangian [19, 20]. The total cross section is obtained by summing over all possible parton flavors and integrating over all possible momentum fractions.

The parton distribution functions, f_i and f_j, are evaluated at a factorization scale, μ_f, which can be thought of as the scale that separates short-distance, perturbative physics, from long-distance, non-perturbative physics.

In a perturbative expansion carried at all orders, the cross section $\sigma(P_1, P_2)$ in Equation 1.13 would be independent of μ_F and μ_R. In actual finite order calculations this is not true. They are chosen at a typical scale of the process, in order to minimize the contribution of (uncalculated) higher order terms.

Often μ_f and μ_r are identified with one another and written $\mu_f = \mu_r = \mu$. Both scales will appear in ratios within the cross section integral and thus as logarithms when expanded order-by-order in perturbation theory. The presence of logarithms due to multiple scales is most apparent in the example of the dependence of α_s on μ_r and Λ_{QCD}, the ultraviolet cutoff scale used in QCD [21],

$$\alpha_s(\mu^2) \sim \frac{1}{\ln(\mu^2/\Lambda^2_{QCD})}. \tag{1.14}$$

Equation 1.14 reveals the order of magnitude of the scale at which α_s becomes large enough to destroy the assumption that perturbative expansion is valid ($\Lambda_{QCD} \approx 200$ MeV). Fundamentally this equation represents an ultraviolet divergence that results in a scaling logarithm in the process of renormalization. Infrared divergences like those due to soft gluon emission result in similar collinear logarithms of the form $\alpha_s^n \ln^{2n}(\mathrm{p}_T/m)$ and $\alpha_s^n \ln^{2n-1}(\mathrm{p}_T/m)$. One convention [22] is to refer to these terms as "leading logarithmic" (LL) and "next-to-leading logarithmic" (NLL), respectively. These terms may be exponentiated, leading to a sum of ratios of the relevant scales (here, p_T/m) to all orders.

Any differences in the calculated cross sections due to different choices of these scales can therefore be interpreted as an uncertainty due to the unknown higher-order corrections in the cross section calculation.

The fact that the cross-section of a process should be independent of the factorization scale μ_f led to the DGLAP equations, published separately in the 1970s by Yuri Dokshitzer, Vladimir Gribov and Lev Lipatov, and Guido

Altarelli and Giorgo Parisi [23]. These equations determine the evolution of the PDFs with Q. The dependence on x, on the other hand, must be obtained by fitting possible cross section predictions to data from hard scattering experiments.

1.3 Monte Carlo tools

Knowing QCD predictions is crucial in the design of methods to search for new physics, as well as for extracting meaning from data. Different techniques can be used to make QCD predictions at hadron colliders, and in particular at the LHC. The so called Matrix Element Monte Carlos use direct perturbative calculations of the cross-section matrix elements for each relevant partonic subprocesses. LO and NLO calculations are available for many processes. These "fixed-order predictions" include the first terms in the QCD perturbative expansion for a given cross-section; as more terms are involved in the expanstion, an improvement in the accuracy of the prediction is expected. The complexity of the calculations increases significantly with the number of outgoing legs.

An alternative approach is applied by the so called Monte Carlo parton shower programs. These simulation programs use LO perturbative calculations of matrix elements for $2 \to 2$ processes, relying on the parton shower to produce the equivalent of multi-parton final state. PYTHIA [24] and HERWIG++ [25] are the most commonly used parton shower Monte Carlos.

The Monte Carlo generators must account for and correctly model the showering of partons. To approximate the energy-evolution of the shower, the DGLAP equations that describe the evolution of the PDFs with changing

energy scale can be used. The separation of radiation into initial- (before the hard scattering process takes place) and final-state showers is arbitrary, but sometimes convenient. In both initial- and final-state showers, the structure is given in terms of branchings $a \to bc$: $q \to qg$, $q \to q\gamma$, $g \to gg$ and $g \to q\bar{q}$. Parton b carries a fraction z of the energy of the mother energy and parton c carries the remaining $1-z$ (the term "partons" includes the radiated photons). In turn, daughters b and c may also branch, and so on. Each parton is characterized by some evolution scale, which gives an approximate sense of time ordering to the cascade. In the initial-state shower, the evolution scale values are gradually increasing as the hard scattering is approached, while these values decrease in the final-state showers. The evolution variable of the cascade in the case of PYTHIA, Q^2, has traditionally been associated with the m^2 of the branching partons[5]. In the recent version of PYTHIA a $p_\perp$-ordered shower algorithm, with $Q^2 = p_\perp^2$ is available, and the shower evolution is cut off at some lower scale Q_0 typically around 1 GeV for QCD branchings. HERWIG++ provides a shower model which is angular-ordered.

There are two leading models for the description of the non-perturbative process of hadronization, after parton showering. PYTHIA uses the Lund string model of hadronization to form particles [26]. This model involves stretching a colour "string" across quarks and gluons and breaking it up into hadrons. HERWIG++ utilizes the cluster model of hadronization. In this model each gluon is split into a $q\bar{q}$ pair and then quarks and anti-quarks are grouped into colourless "clusters", which then give the hadrons.

Hadronization models involve a number of "non-perturbative" parameters. The parton-shower itself involves the non-perturbative cut-off Q_0^2 . These

[5]The final-state partons have $m^2 > 0$. For initial-state showers the evolution variable is $Q^2 = -m^2$, which is required to be strictly increasing along the shower.

different parameters are usually tuned to data from the LEP experiments.

In addition to the hard interaction that is generated by the Monte Carlo simulation, it is also necessary to account for the interactions between the incoming proton remnants. This is usually modelled through multiple extra $2 \to 2$ scattering, occurring at a scale of a few GeV. This effect is known as multiple parton interactions (MPIs). In addition, these partons may radiate some of their energy, either before or after the hard interaction. All the additional parton interactions, which are not involved in the hard scattering process, are grouped together in the term underlying event. The modelling of the underlying event is crucial in order to give an accurate reproduction of the (quite noisy) energy flow that accompanies hard scatterings in hadron-collider events.

It should be stressed that these multiple parton interactions are a completely separate effect from the multiple proton interactions that may occur in each bunch collision event in the LHC. These multiple proton collisions are referred to as pileup, and are not included in the definition of the underlying event.

No precise model exists to reproduce the underlying event activity. These are tuned to Tevatron and early LHC data. A specific set of chosen parameters for a generator is referred to as a "tune".

The two Monte Carlo generators used in this analysis are summarized below, indicating the particular versions and tunes that were implemented.

Pythia

The PYTHIA event generator has been used extensively for e^+e^-, ep, $pp/p\bar{p}$ at LEP, HERA, and Tevatron, and during the last 20 years has probably been the most used generator for LHC physics studies. PYTHIA contains an

extensive list of hardcoded subprocesses, over 200, that can be switched on individually. These are mainly 2→1 and 2→2, some 2→3, but no multiplicities higher than that. Consecutive resonance decays may of course lead to more final-state particles, as will parton showers.

As mentioned above, in this MC generator showers are ordered in transverse momentum [27] both for ISR and for FSR. Also MPIs are ordered in p_T [28]. Hadronization is based solely on the Lund string fragmentation framework.

For the results presented in this thesis simulated samples of dijet (see Section 1.4) events from proton-proton collision processes were generated with PYTHIA 6.423 [24]. The ATLAS AMBT2 tune of the soft model parameters was used [29]. This tune attemps to reproduce the ATLAS minimum bias charged particle multiplicity and angular distribution measurements and the ATLAS measurements of charged particle and p_T density observed collinear and transverse to the high-energy activity.

For systematic comparisons, a set of additional tunes, called the Perugia tunes [30] were also used. These utilize the minimum bias and p_T density measurements of CDF to model the underlying event, hadronic Z^0 decays from LEP to model the hadronization and final state radiation, and Drell Yann measurements from CDF and $D0$ to model the initial state radiation. In particular, the Perugia 2011, which is a retune of Perugia 2010 [31] includes 7 TeV data from 2011 data taking.

Herwig++

HERWIG++ [25] is based on the event generator HERWIG (Hadron Emission Reactions With Interfering Gluons), which was first published in 1986 and was developed throughout the LEP era. HERWIG was written in Fortran, and

the new generator, Herwig++ developped in C++. Some distinctive features of Herwig++ are: angular ordered parton showers and cluster hadronization, and hard and soft multiple partonic interactions to model the underlying event and soft inclusive interactions [32].

This MC generator was used for systematic uncertainties studies. The version utilized was 2.4.2 released in 2009.

Detector simulation

In order to use events produced by Monte Carlo generators to model events that one might observe with the detector, the output of these generators is passed through a detector simulation model. ATLAS uses the GEANT4 [33] toolkit. GEANT4 is an extensive particle simulation toolkit that governs all aspects of the propagation of particles through detectors, based on a description of the geometry of the detector components and the magnetic field. The physics processes include ionization, Bremsstrahlung, photon conversions, multiple scattering, scintillation, absorption and transition radiation.

The detector is described in terms of almost 30 millon volumes with properties, which in case of the ATLAS detector are constructed based on two databases: the geometry database and the conditions database. The former contains all basic constants, e.g. dimensions, positions and material properties of each volume. The latter is updated according to the circumstances at a given time and contains for instance dead channels, temperatures and misalignments. As a result, several layouts of the detector are available. Test beam data taken with components of the ATLAS detector before completion have aided the validation and further improvement of the detector simulation.

Due to the detailed and complicated geometry of ATLAS and the diversity and complexity of the physics processes involved, the consumed computing

time per event is large ($\mathcal{O}$(1hour)). This has been a motivation for the development of fast simulation alternatives. The standard GEANT4 simulation that exploits the full potential is referred to as *full simulation*. The majority of the events studied in this thesis are produced with full simulation.

1.4 Jet physics

Due to confinement quarks and gluons emerge from the interaction as constituents of final state "coloreless" hadrons[6]. This packet of particles produced tends to travel collinearly with the direction of the initiator quark or gluon. The result is a collimated "spray" of hadrons (also photons and leptons) entering the detector in place of the original parton; these clusters of objects are what we define as jets and are the experimental signature of the partons produced in the high energy interaction. The first evidence for jet production was observed in e^+e^- collisions at the SPEAR storage ring at SLAC in 1975 [34].

The evolution from a single parton to an ensemble of hadrons occurs through the processes of parton showering and hadronization. Since the strong coupling constant grows with increasing distance between color charges, a strong color potential forms as the parton from the "hard" (high Q^2) scattering process separates from the original hadron. This large potential causes quark/antiquark pairs ($q\bar{q}$) to be created, each carrying some of the energy and momentum of the original partons. As these new partons move away from one another, yet more color potentials are formed, and the process repeats. This process is perturbatively described as a parton shower, where quarks radiate gluons which in turn give rise, via pair production to $q\bar{q}$, in

[6]We use "colorless" to mean a singlet representation of the color group.

a process similar to the electromagnetic shower produced by a high energy electron or photon. The shower of partons travels basically along the same direction as the original. This process continues until there is no longer enough energy for the shower to develop. and instead the remaining partons combine to form stable hadrons. Since this progression involves successively lower energies and lower momentum transfers, perturbative QCD cannot describe the full process. The full parton shower and hadronization process then cannot be calculated from first principles, but has to be modelled.

1.4.1 Jet algorithms

As described above, quarks and gluons cannot be directly observed. Quarks and gluons hadronise, leading to a collimated spray of energetic hadrons, a jet. By measuring the jet energy and direction one can get close to the idea of the original parton. But one parton may form multiple experimentally observed jets, for example due to a hard gluon emission plus soft and collinear showering. Then, in comparing data to theory and MC programs predictions a set of rules for how to group particles into jets is needed. A jet algorithm, together with a set of parameters and a recombination scheme (how to assign a momentum to the combination of two particles) forms a jet definition.

By using a jet definition a computer can take a list of particle momenta for an event, be they quarks and gluons, or hadrons, or calorimeter depositions, and return a list of parton, particle or calorimeter jets, respectively. One important point to remark is that the result of applying a jet definition should be insensitive to the most common effects of showering and hadronization, namely soft and collinear emissions. This is illustrated in Fig. 1.1.

Traditionally, jet algorithms have been classified into two categories: cone and sequential recombination algorithms.

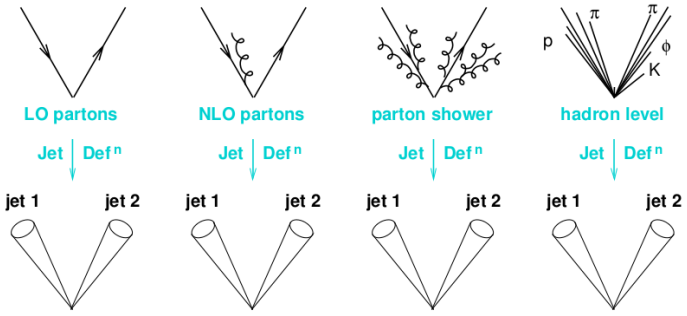

Figure 1.1: The application of a jet definition to a variety of events that differ just through soft/collinear branching and hadronization should give identical jets in all cases [35].

Fixed cone jet finder in ATLAS

Cone-like algorithms are based on the collinear nature of gluon radiation and the parton shower described above. The decay products of quarks and gluons and their emissions will tend to form a cone of particles in the $\eta - \phi$ plane[7] as they propagate outwards. The design of cone-like algorithms attempts to maximize the amount of energy present in a stable cone of fixed radius.

In ATLAS the standard jet algorithm for a long time was an iterative fixed-cone jet finder. First, it sorts all particles in the event according to their momentum, and identifies the one with largest p_T. This is referred to as a seed particle. Then a cone of radius R_{cone} in $\eta - \phi$ is drawn around

[7] In the ATLAS Coordinate System the azimuthal angle ϕ is measured around the beam axis, and the polar angle θ is the angle from the beam axis. The pseudorapidity is defined as $\eta = \ln(\tan(\frac{\theta}{2}))$. The transverse momentum p_T is defined in the plane transverse to the beam motion. See section **??**. The distance ΔR in the pseudorapidity-azimuthal angle space is defined as $\Delta R = \sqrt{\Delta\eta^2 + \Delta\phi^2}$. In collider physics p_T, η and ϕ are used instead of p_i, θ, and ϕ, since the former set is z-boost invariant and each partonic collision has a random boost in the pp center-of-mass frame.

the seed and all objects within a cone of $\Delta R < R_{cone}$ are combined with it. The direction of the sum of the momenta of those particles is identified and if it doesn't coincide with the seed direction then the sum is used as a new seed direction, and it iterates until the direction of the cone is stable (i.e, the direction of the sum of the cone contents coincides with the previous seed). The resulting cone is called a jet. The process is restarted with the highest p_T particle not yet associated to a cone. This type of algorithm is called "iterative" since it iterates the cone direction. The jets found in this way can share part of their constituents. Jets with common constituents are merged if their shared p_T is larger than 50% of the p_T of the softer jet. Otherwise, the overlapping part divided according to some algorithm between the two overlapping jets.

A difficulty and major drawback of this procedure is the use of the transverse momentum of the particle to select the first seed. This definition is collinear unsafe, i.e. a splitting of the hardest particle into a nearly collinear pair can have the consequence that another, less hard particle, pointing in a different direction suddenly becomes the hardest in the event, leading to a different final set of jets[8]. There are many other variants of cone algorithms, and nearly all suffer from problems of either collinear safety, or infrared safety (an extra soft particle creates a new seed, which can lead to an extra stable cone being found). With a seedless algorithm, the addition of one or more soft particles does not lead to new hard stable cones being found, therefore the algorithm is infrared safe at all orders.

[8]From the theoretical point of view, the splitting and merging procedures make this algorithm partially infrared safe, but the algorithm remains well defined only up to leading order of perturbation theory.

Sequential recombination algorithms

Recombination algorithms are both collinear and infrared safe. For this reason, they can be used in calculations to any order in perturbation theory. The term recombination is used since they attempt to follow the parton shower branchings which become progressively softer as the shower evolves. The resulting jet can be thought of as the final stage of this process and the algorithm is the device used to retrace the tree of sequential branchings. In general, recombination algorithms operate by successively combining pairs of particles using a distance metric, d_{ij}. At hadron colliders, due to the fact that one of the incoming partons may continue along the beam, for every pair of particles this metric is compared to a so-called "beam distance", d_{iB}, and only when $d_{ij} < d_{iB}$ the particle pair is combined and considered for subsequent clustering steps.

The k_t algorithm. The most common sequential recombination algorithm is the inclusive k_t algorithm. If was first implemented in the analysis of multi-jet events at e^+e^- colliders [36] and subsequently extended for use at hadron colliders [37, 38]. It is instructive to compare both the original algorithm as well as the ultimate definition of the modern k_t algorithm in order to identify relevant features of this algorithm. The distance measure in the original version is defined as:

$$d_{ij} = \frac{2E_iE_j(1-\cos\theta_{ij})}{Q^2}, \tag{1.15}$$

where Q is the total energy in the event, E_i is the energy of particle i and θ_{ij} the angle between particles i and j. In the collinear limit, d_{ij} is related to the relative transverse momentum between particles i and j (hence the name k_t algorithm), normalized to the total visible energy. The particles are combined if the minimum d_{ij}, d_{min}, is below a certain threshold, y_{cut}. The jet

multiplicity depends on the value of y_{cut}, as a lower value will result in more soft or collinear emissions surviving as jets. This is thus the first definition of an "event shape", this threshold marks the transition between two-jet events and three-jet events.

For a jet algorithm at a hadron collider, the notion of a beam distance is added. A distance scale, $\Delta R = \sqrt{\Delta y^2 + \Delta\phi^2}$, is introduced to define the typical radius for a jet, effectively replacing y_{cut}. In this case the particle distance metric becomes,

$$d_{ij} = \min(p_{ti}^2, p_{tj}^2)\frac{\Delta R_{ij}^2}{R^2} \tag{1.16}$$

and the beam distance,

$$d_{iB} = p_{ti}^2. \tag{1.17}$$

such that when no particle j is found such that $\Delta R_{ij} < R$ then i is promoted to the status of a jet.

The formulation of the modern inclusive k_t algorithm is formulated as follows:

1. Utilize the particle distance metric d_{ij} defined in Eq. 1.16.
2. Compute the minimum d_{ij}, $d_{min} = \min(d_{ij})$, among all particles.
3. If $d_{min} < d_{iB}, d_{jB}$, then combine particles i and j and repeat from step 1.
4. If $d_{ij} > d_{iB}$, then identify i as a jet and remove it from the list.
5. Continue until all particles are considered jets or have been clustered with other particles.

Jets built with this algorithm have quite irregular shapes, and particles with $\Delta R_{ij} > R$ can still be clustered within the jet. This is a problem

when, for example, an irregularly shaped jet happens to extend into poorly instrumented detector regions.

As defined, the k_t algorithm clusters first objects that are either very close in angle or have very low transverse momentum. The fact that soft particles are clustered first is a another drawback of this definition since it has the potential to introduce complications when the detector noise of energy density fluctuations are large.

A feature of the k_t algorithm that is attractive is that it does not only produce jets but it also assigns a clustering sequence to the particles within the jet. It is possible then to undo the clustering and to look back at the shower development history. This has been exploited in a range of QCD studies, and also in searches of hadronic decays of boosted massive particles and it will be used here for the search of two-pronged jets in gluon splitting.

The k_t algorithm can be generalized by introducing the following particle-particle and particle-beam distance measures:

$$d_{ij} = \min(p_{ti}^{2n}, p_{tj}^{2n}) \frac{\Delta R_{ij}^2}{R^2} \tag{1.18}$$

$$d_{iB} = p_{ti}^{2p}. \tag{1.19}$$

where p is a parameter which is 1 for the k_t algorithm. Two different algorithms can be obtained from this: The Cambridge-Aachen (C/A) algorithm [39], with $p = 0$, and the anti-k_t algorithm [40], with $p = -1$.

The Cambridge-Aachen algorithm. The C/A algorithm is obtained by choosing a value $p = 0$ in Equations 1.18 and 1.19. This algorithm recombines objects close in ΔR iteratively and reflects the angular ordering of the QCD radiation. It is ideally suited to reconstruct and decompose the various decay components of heavy objects like Higgs bossons or top quarks using subjet structure.

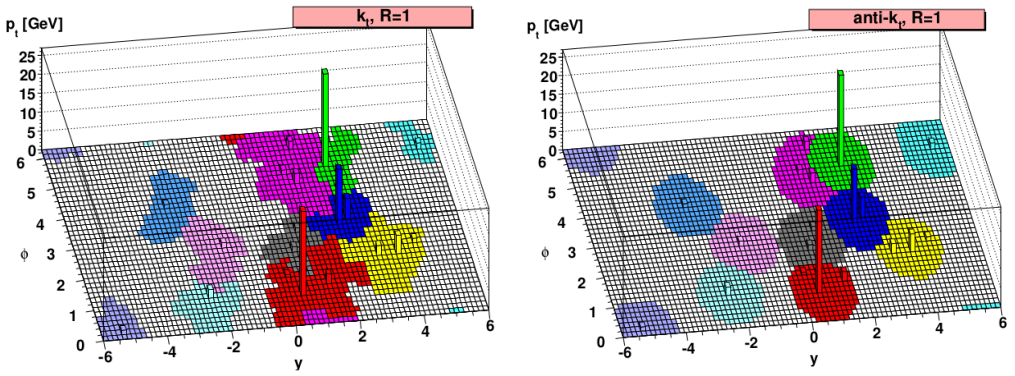

Figure 1.2: A sample parton-level event, generated with HERWIG, clustered with the k_t and anti-k_t algorithms, illustrating the active area of the resulting jets [41].

The anti-k_t algorithm. Contrary to the k_t algorithm, the anti-k_t algorithm, so named because of the inverted power law in the particle and beam distance metrics in Equations 1.18 and 1.19, first clusters hard objects together which results in more regular jets with respect to the k_t and C/A algorithms. This characteristic is illustrated for the k_t and anti-k_t algorithms in Fig. 1.2.

For this reason and the fact that this algorithm is less sensitive to soft emissions (see Chapter **??**) the anti-k_t algorithm was chosen as the default jet algorithm for ATLAS analyses.

Note that the anti-k_t algorithm does not provide useful information on jet substructure if a jet contains two hard cores, then the k_t (or C/A) algorithms first reconstruct those hard cores and merge the resulting two subjets. The anti-k_t will often first cluster the harder of the two cores and then gradually aglomerate the contents of the second hard core.

These algorithms, and more, are implemented in FASTJET [42] software package for jet-finding.

1.4.2 Jet substructure

The first evidence of jet structure resulted from the study of the spacial distribution and multiplicity of particles in the event phase space in hadron production in e^+e^- collisions [34]. Generally, all final hadronic states in $pp/p\bar{p}/e^+e^-$ collisions can be explored in terms of the structure and shape of the event energy flow by means of the so called "event shape" variables. This family of variables attempts to extract information about the global geometry of an event, usually distinguishing between di-jet events and multijet final states. Such variables have been successfully utilized in many SM measurements and BSM searches, see for example [43, 44].

Although very useful, event shape variables are not sensitive to the detailed structure and distribution of energy inside a particular jet. In SM and new physics searches, tools for the identification of individual objects that might be signature of new particles are desired. At the LHC, many of the particles considered to be heavy at previous accelerators will be frequently produced with a transverse momentum greatly exceeding their rest mass, like the electro-weak gauge bosons $W^{\pm}$ and Z, the top quark, the Higgs boson (or bosons) and possibly other new particles in the same mass range. These boosted objects, produced either by recoil against other energetic objects or from decays of even heavier BSM particles, upon decay can give rise to a highly collimated topology too close to be resolved by standard jet algorithms. A method for selecting these jets would allow for the study of their properties. This interest led to the development of a wide range of sophisticated tools in the last years [45, 46] that allow the analysis of the substructure of the ensuing jet and reveal its heavy-particle origin.

Jet substructure methods probe the internal structure of jets from a detailed study of its constituents. These techniques have been first implemented

for distinguishing boosted SM hadronic objects from the background of jets initiated by light quarks and gluons, see for example [47], but they have been also succesfully used in other applications, including separating quark jets from gluon jets [48] and identifying boosted decay producs in new physics searches [49].

Jet shapes, which are event shape-like observables applied to single jets, are an effective tool to measure the structure of individual jets [50].The shape of a jet not only depends on the type of parton (quark or gluon) but is also sensitive to non-perturbative fragmentation effects and underlying event contributions [51].

In chapter 2, several distinguishing characteristics between jets originating from single b-quarks and jets containing two close-by b-hadrons are determined using the techniques of jet substructure.

1.5 Heavy flavor jet production

Heavy flavor (HF) quarks enter in many collider searches, notably because they are produced in the decays of various SM particles (top quarks, the Z boson and the Higgs boson, if light), and of numerous particles appearing in proposed extensions of the SM. However, the most common process of HF production is QCD. Heavy Flavor QCD production can be classified into three processes depending on the number of heavy quarks participating in the hard scattering. The hard scatter is defined as the $2 \to 2$ subprocess with the largest virtuality (or shortest distance) in the hadron-hadron interaction [52].

- **Heavy flavor creation (FCR)**: two heavy quarks are produced in the hard scatter. Being b the heavy flavor quark, at leading order this process is described by $gg \to b\bar{b}$ and $q\bar{q} \to b\bar{b}$, with no b-quarks in the

initial state (IS).

- **Heavy flavor excitation (FEX)**: the heavy flavour quark excitation can be depicted as an initial state gluon splitting into a $b\bar{b}$ pair, where one the b-quarks subsequently enters the hard scatter, i.e., there is one b-quark in the IS and another one in the final state (FS).
- **Gluon splitting (GSP)**: no heavy quarks participate in the hard scatter in this case, but they are produced via subsequent $g \to Q\bar{Q}$ branchings.

Example of Feynman diagrams for QCD b-quark production up to NLO are shown in Fig. 1.3.

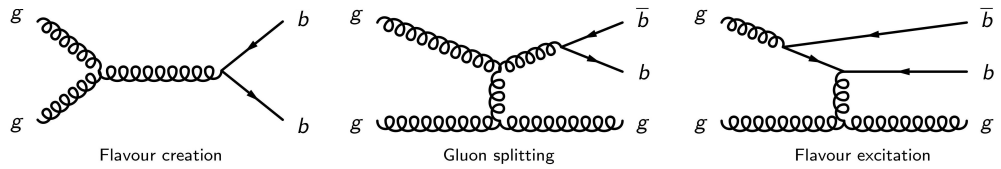

Figure 1.3: Representative diagrams of the three channels contributing to QCD b-quark production up to NLO. The flavour creation channel (left) is the only one present at LO. At NLO, two new channels open up, referred to as gluon splitting (center) and flavour excitation (right).

The definition above is not strict, but can be used as a basis for the understanding of the characteristics of heavy flavour quark production.

Final state b-quarks hadronize into b-hadrons. During the fragmentation process, other particles will also be produced along with the b-hadron, giving rise to b-jets. Directly produced b-jets are p_T balanced and back-to-back in the azimuthal angle ϕ. However they are not 3-D balanced because b-jets may be boosted in the z direction due to the different proton momentum fractions carried by the initial partons. In the flavor excitation process, the b-quark

which does not participate in the hard scatter belongs to the underlying event, resulting in a forward (large η) b-jet. The angular $\Delta\phi$ separation between the two b-jets is therefore expected to be flat. Gluon splitted b-jets are expected to be collinear since they originate from the splitting of a gluon and will tend to be identified as a same hadronic jet. The azimuthal separation between the two gluon splitted b-jets thus peaks at small angles.

The simplest and most fundamental measurement of heavy-quark jet production is the inclusive heavy-quark jet spectrum, which is dominated by pure QCD contributions. Studies of QCD bottom production are important in their own right because of the correspondence between parton level production and the observed hadron level: b-quarks give rise to observable b-hadrons, there is no such an association between light quarks (u, d and s) or gluons, and observed final state hadrons. In addition, the study of b-quark production have the potential to provide information on the b-quark parton distribution function, a component of the proton structure thought to be generated entirely perturbatively from the QCD evolution equations of the other flavours.

The theoretical calculation of the inclusive b-jet spectrum presents rather important uncertainties ($\sim 50\%$), considerably larger than those for the light jet inclusive spectrum ($\sim 10 - 20\%$) [53]. A review of the origin of these uncertainties is presented by Banfi, Salam and Zanderighi in reference [54]. They arise from the poor convergence of the perturbative series, as evidenced by a large value of the K-factor, the ratio of the next-to-leading order (NLO) to the leading order (LO) cross section. This is illustrated in Fig. 1.4 for the p_T range covered by the LHC. The observed K values (6 to 10) indicate that the NLO result cannot be an accurate approximation to the full result. It is for this reason that the scale dependence (middle panel in Fig. 1.4) is large.

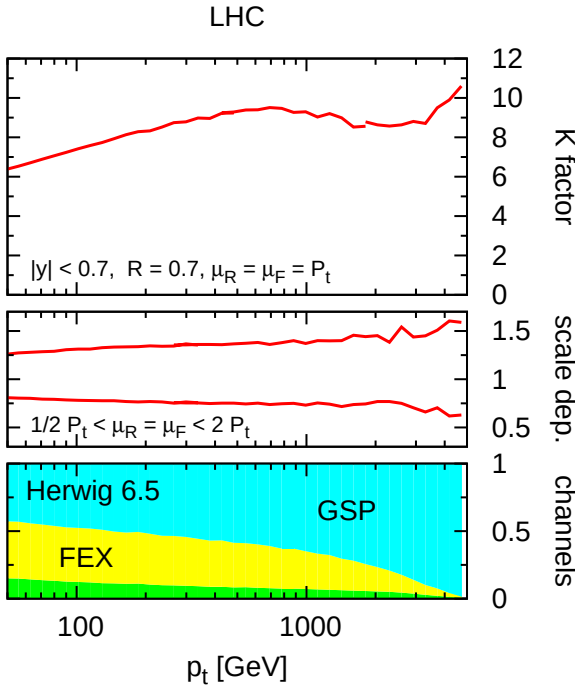

Figure 1.4: Top: K-factor for inclusive b-jet spectrum taken from [54], clustering particles into jets using the k_t jet-algorithm [38] with R=0.7, and selecting jets in the central rapidity region ($|y| < 0.7$). Middle: scale dependence obtained by simultaneously varying the renormalisation and factorisation scales by a factor two around p_T, the transverse momentum of the hardest jet in the event. Bottom: breakdown of the Herwig [55] inclusive b-jet spectrum into the three major underlying channels, flavor creation (FCR) flavor excitation (FEX) and gluon splitting (GSP).

The fact that the perturbative series is very poorly convergent is related to the different channels for heavy quark production. While at LO only the FCR channel is present, at NLO the FEX and GSP channels open up[9]. In the gluon splitting process, one of the final-state light partons (at NLO always a gluon) splits collinearly into a $b\bar{b}$ pair that a clustering algorithm can classify within the same jet. A jet containing both b and $\bar{b}$ is considered to be just a b-jet in standard definitions.

The various channels can be approximately separated in a parton shower Monte Carlo generator such as HERWIG or PYTHIA. These MC generators include NLO effects, and one can determine the underlying hard process from the event record. Their relative contributions to the total b-jet spectrum are shown in the bottom panel of Fig. 1.4. It is found that the LO channel has a much smaller contribution than the FEX and the GSP channels, which receive strong enhancement from collinear logarithms, going as $\alpha_s^2(\alpha_s \ln(p_T/m_b))^n$ for flavour excitation [23] and $\alpha_s^2 \cdot \alpha_s^n \ln^{2n-1}(p_T/m_b)$ for gluon splitting ($n \geq 1$) [56].

Ref. [54] proposes a new observable to free the heavy-flavour spectrum calculation from collinear logarithms, and improve the accuracy of the theoretical prediction, by not including in the production cross-section the contribution from double b jets. Final-state logarithms are removed by employing a recently developed jet reconstruction scheme, the flavour-k_t algorithm [57], which maintains the correspondence between partonic flavour and jet flavour. Specifically, jets containing a b-quark and a b-antiquark, which in a parton shower MC generator are produced $\sim 95\%$ of the time by the GSP chan-

[9]It is sometimes stated that it makes no sense, beyond LO, to separately discuss the different channels, for example because diagrams for separate channels interfere. However, each channel is associated with a different structure of logarithmic enhancements, $\ln^n(p_T/m_b)$, and so there is distinct physical meaning associated with each channel.

nel, are labeled in an IR-safe way as light jets and removed from the b-jet spectrum. The initial-state (FEX) collinear logarithms can be resummed by using a b-quark parton distribution functions. With this algorithm the K-factor for the differential heavy-jet spectrum cross-section is shown not to exceed a value of $K = 1.4$, with a factor of four reduction in the theoretical (scale variation) uncertainties.

1.6 Other application of $g \to b\bar{b}$

Succesfully identifying jets with two b-hadrons, the products of the b-quark or b-antiquark hadronization, can also provide an important handle to understand, estimate and/or reject b-tagged backgrounds to SM and new physics searches at the LHC.

SM physics analyses that rely on the presence of single b-jets in the final state, such as top quark physics, either in the $t\bar{t}$ or the single top channels, and associated Higgs production: $WH \to \ell\nu b\bar{b}$ and $ZH \to \nu\nu b\bar{b}$, suffer from the reducible background from QCD, which can produce double b-hadron jets as discussed above, and the irreducible background due to W bosons produced in association with b-quarks. Figure 1.5 shows the two diagramas for $W + b$ production.

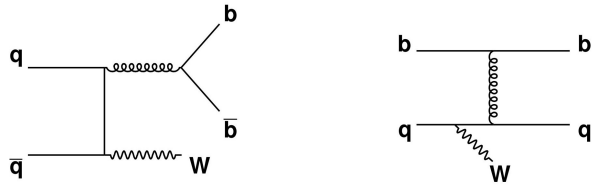

Figure 1.5: Feynman diagrams for W production in association with b quarks.

While at LO only single b-jets are present, at NLO jets containing two

b-hadrons are expected due to the contribution of a diagram containing a $gb\bar{b}$ vertex. The b-quark pair is produced at small angles and can be often reconstructed as one merged jet.

The relevance of double b-hadron jets is supported by NLO calculations of the production of W bosons and two jets with at least one b quark at the LHC for jet $p_T > 25$ GeV, and $|\eta| < 2.5$ [58] indicate that the cross section for $W(b\bar{b})j$ is almost a factor of two higher than $Wb\bar{b}$, and about a third of Wbj, where $W(b\bar{b})j$ denotes the case in which the two b quarks are merged into the same jet.

Jets containing a single b-quark or antiquark also enter in many BSM collider searches, notably because b-quarks are produced in the decays both of heavy SM particles (top quarks, the Z boson and the Higgs boson), and of particles appearing in proposed extensions of the SM. An example is the search for supersymmetry in the framework of generic R-parity conserving models [59]. The superpartners of quarks and gluons could be copiously produced via the strong interaction at the LHC. The partners of the right- and left-handed quarks, $\tilde{\mathrm{q}}_L$ and $\tilde{\mathrm{q}}_R$, can mix to form two mass eigenstates and, since mixing is proportional to the corresponding fermion masses, it becomes more important for the third generation producing sbottom and stop significantly lighter than the other squarks. In this model, thus, sbottom and stop production is expected to dominate. As they chain decay to b-quarks and the lightest supersymmetric particle, the signature for this channel is missing transverse energy plus (single) b-jets. The ability to distinguish single b-jets from jets containing two b-hadrons is thus here of wide application to reduce SM backgrounds giving rise to close-by $b\bar{b}$ pairs.

The study of $b\bar{b}$ jets from gluon splitting is an ideal testbed for exploring jet substructure in data, as it provides a large supply of boosted, merged

jets. Furthermore, understanding $g \to b\bar{b}$ jets is important as they are themselves the background to boosted object searches, like $Z \to b\bar{b}$ or $H \to b\bar{b}$. Understanding the much more common QCD events with double b-hadron jets will be essential before attempting to measure more rare final states.

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 1.3. Samples of dijet events from proton-proton collision processes were simulated with PYTHIA version 6.423 [24], 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 [29].

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 [60].

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

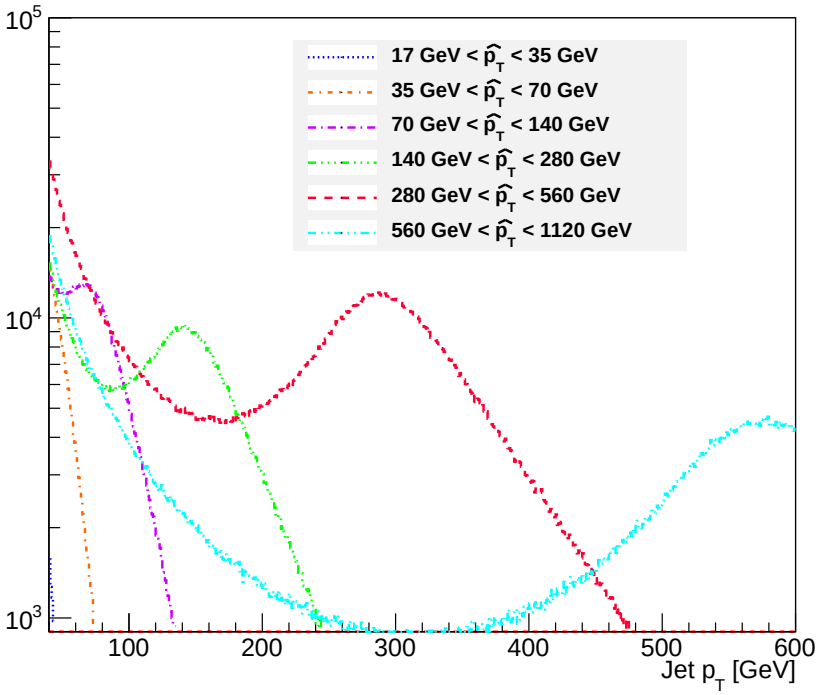

Figure 2.1: Calibrated jet p_T distribution for anti-k_t jets in a dijet Monte Carlo sample composed of different sub-samples generated with increasing thresholds of the hard-scattering partonic transverse momentum, $\hat{p}_T$.

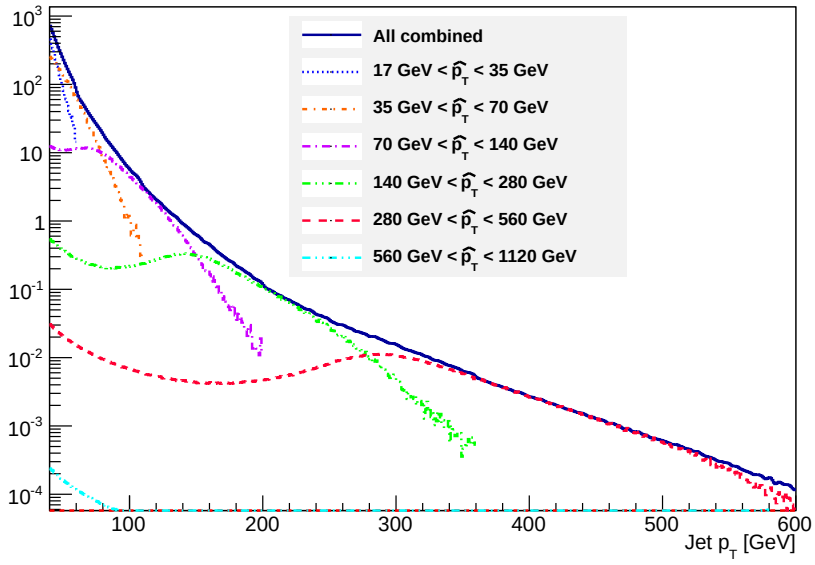

Figure 2.2: Calibrated jet p_T distribution for anti-k_t jets in a dijet Monte Carlo sample composed of different sub-samples generated with increasing thresholds of the hard-scattering partonic transverse momentum, $\hat{p}_T$. In order to obtain the falling p_T spectrum observed in data, the different 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 1.4.1), with a distance parameter $R = 0.4$, using calorimeter topological clusters as input (Section **??**). All jets were calibrated using the EM+JES scheme (Section **??**). 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 [61].

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 **??**). 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

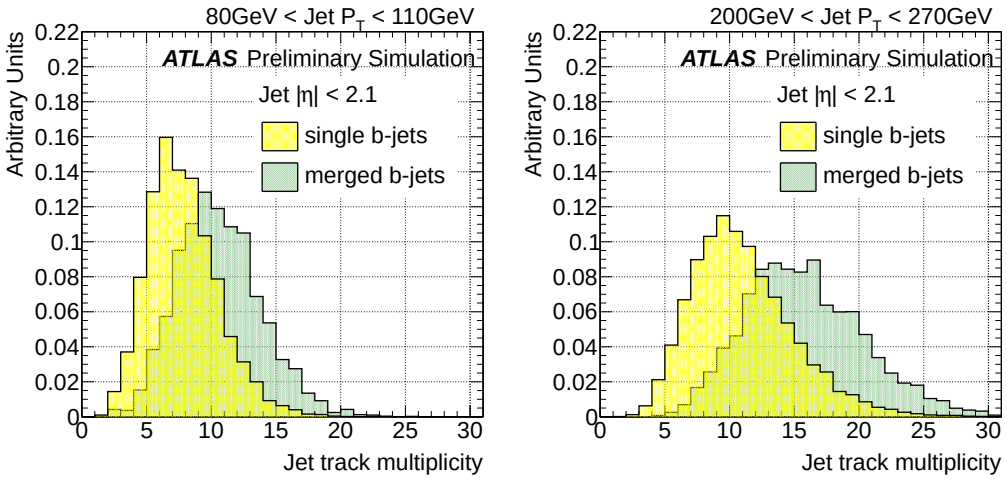

Figure 2.3: Distribution of the track multiplicity for single and merged b-jets from 80 GeV to 110 GeV (left) and 200 GeV to 270 GeV (right).

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.

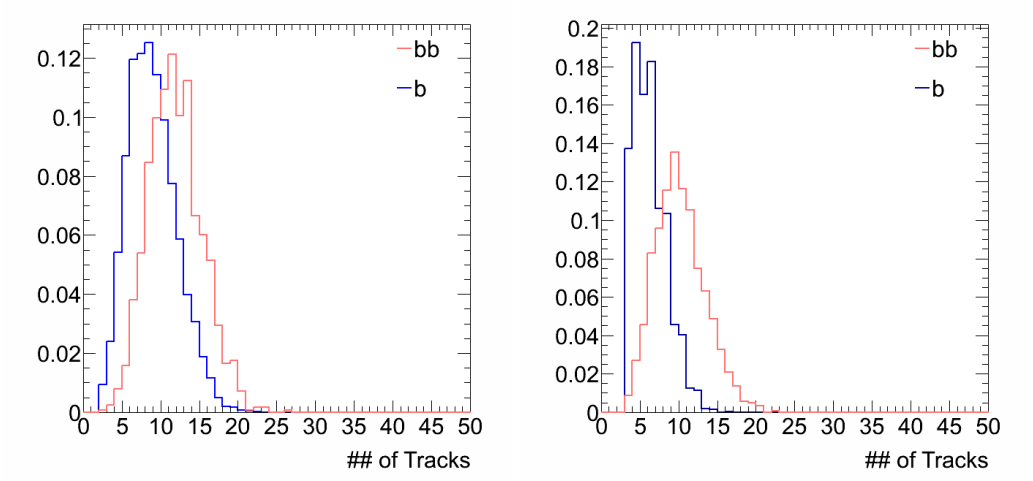

Figure 2.4: Distribution of the charged particle multiplicity for single (b) and merged (bb) jets from 80 GeV to 120 GeV in a sample of dijet events generated with PYTHIA and no detector simulation. Distributions are shown using all charged particles (left) and using only charged particles coming from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

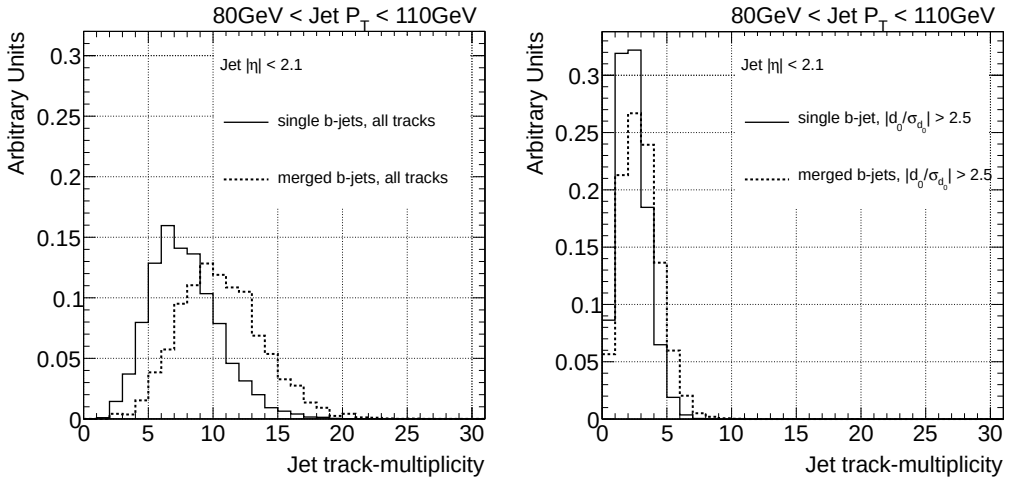

Figure 2.5: Distribution of the jet track multiplicity for single and merged b-jets from 80 GeV to 110 GeV, for all (left) and displaced tracks only (right). No improvement is obtained by using displaced tracks.

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 [48] and [62]. 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 [63]: 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

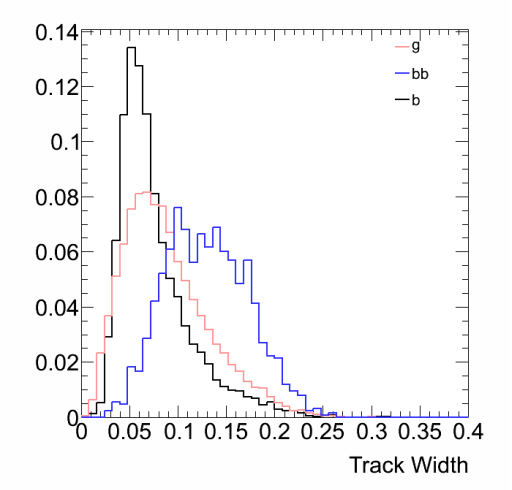

Figure 2.6: Distribution of track-jet width for gluon-initiated (g), single (b) and merged (bb) jets from 80 GeV to 120 GeV in a sample of dijet events generated with PYTHIA and no detector simulation.

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 [64].

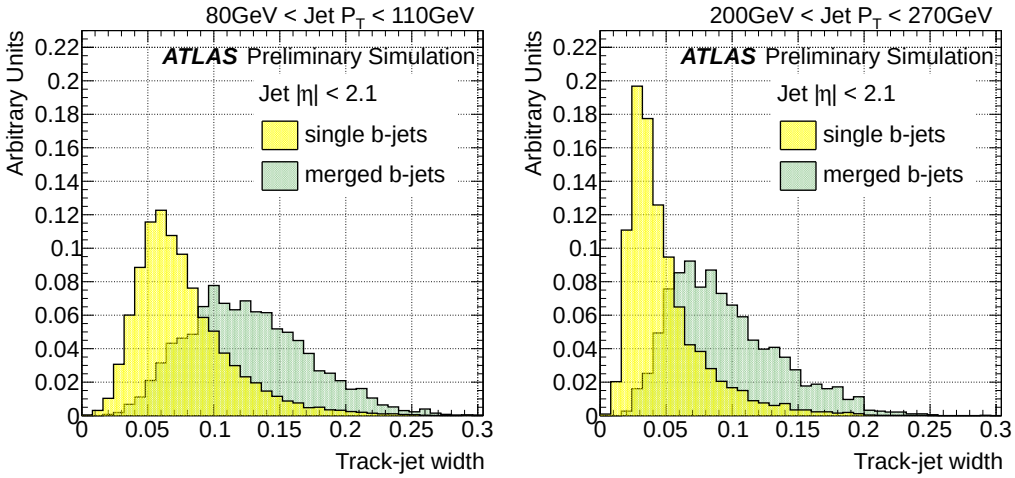

Figure 2.7: Distribution of track-jet width for single and merged b-jets from 80 GeV to 110 GeV (left) and 200 GeV to 270 GeV (right).

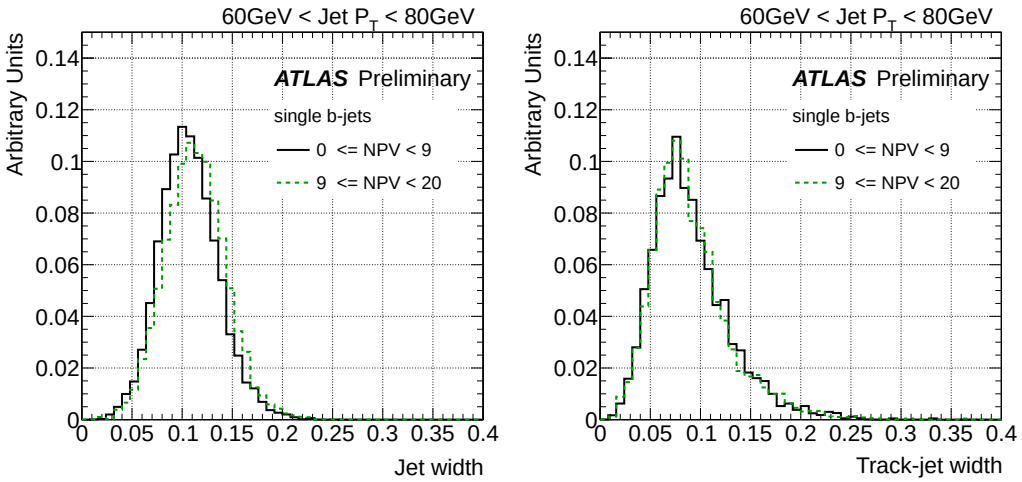

Figure 2.8: Distribution of jet width using topological clusters (left) and tracks (right) for single b-jets in two bins of number of primary vertices (NPV) for jets from 60 GeV to 80 GeV.

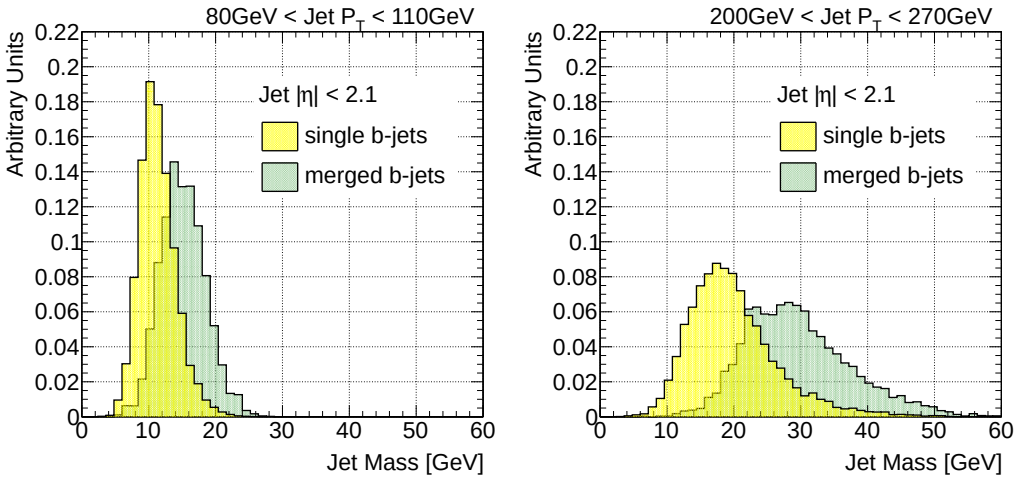

Figure 2.9: Distribution of jet mass in GeV for single and merged b-jets from 80 GeV to 110 GeV (left) and 200 GeV to 270 GeV (right).

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 [65].

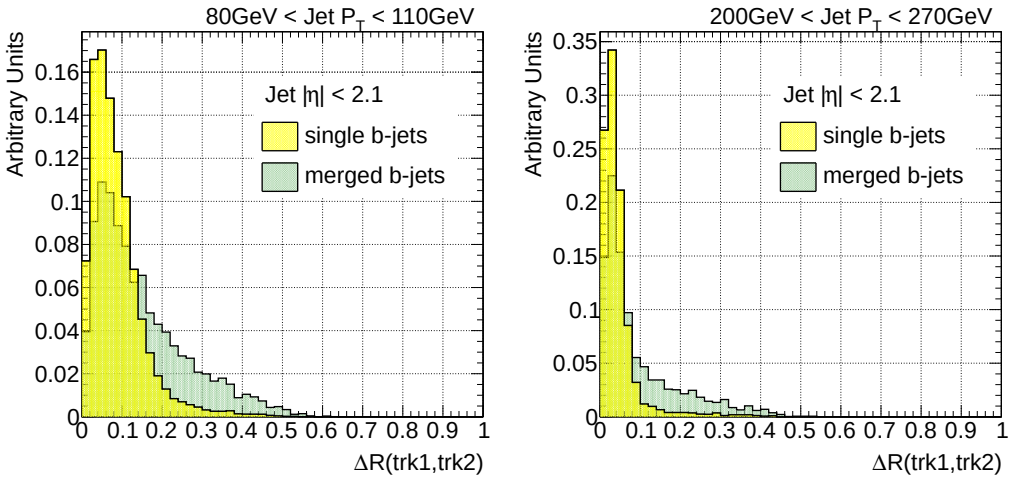

Figure 2.10: Distribution of ΔR between leading tracks for single and merged b-jets from 80 GeV to 110 GeV (left) and 200 GeV to 270 GeV (right).

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.

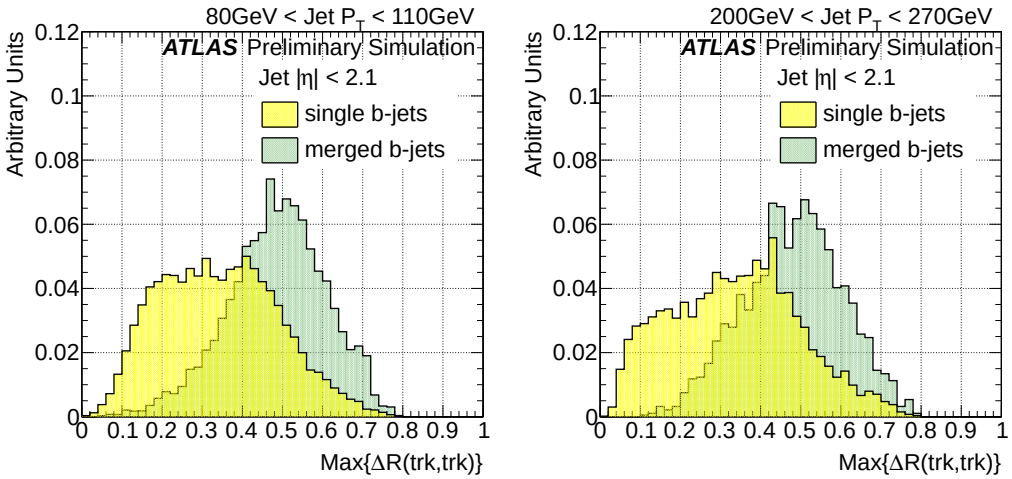

Figure 2.11: Distribution of the maximum ΔR between pairs of tracks for single and merged b-jets from 80 GeV to 110 GeV (left) and 200 GeV to 270 GeV (right).

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 [66]. 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 clustering 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 [68].

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 distance parameter $R = 0.2$. 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 1.4.1) 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 [67].

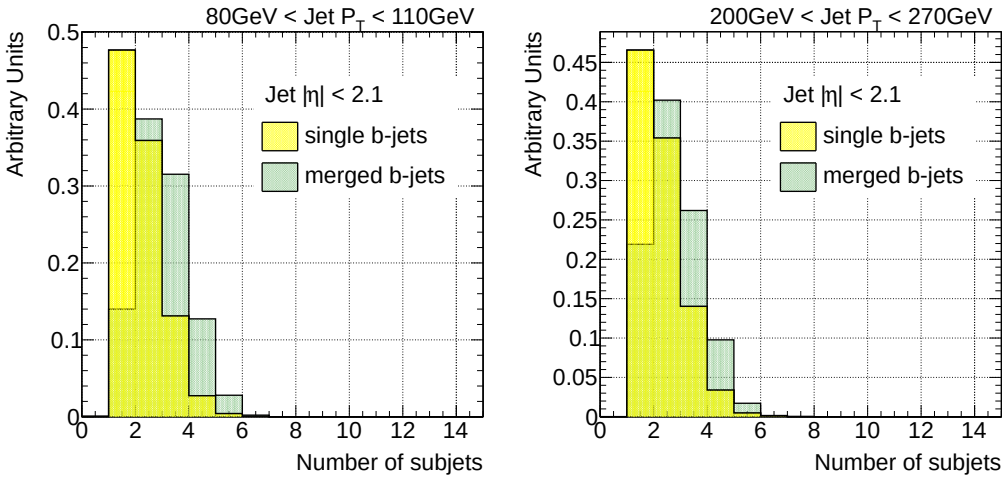

Figure 2.12: Distribution of the number of k_t sub-track-jets for single and merged b-jets from 80 GeV to 110 GeV (left) and 200 GeV to 270 GeV (right).

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. [41]. The event display indicates how the high energy cells in the jet with two b-hadorns are grouped around the b-hadrons directions, leading to the two-subjet substructure of merged jets.

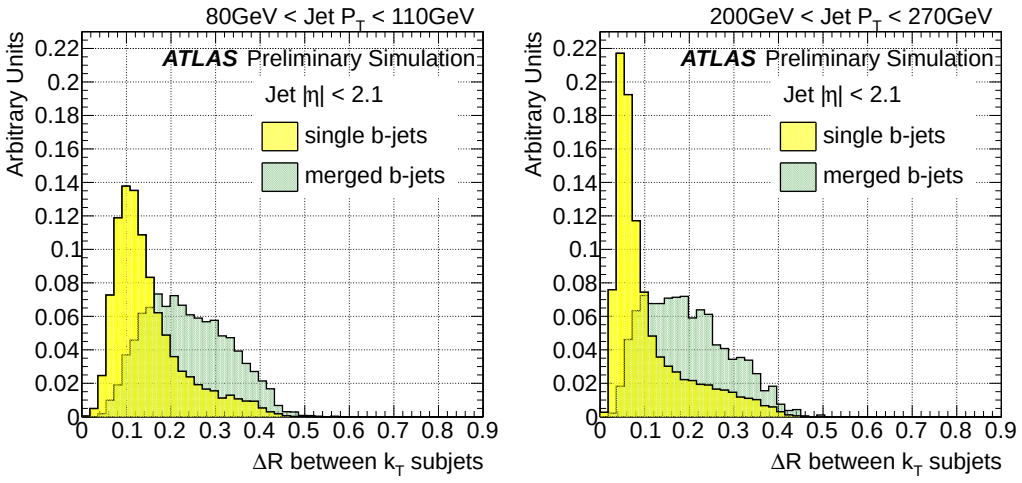

Figure 2.13: Distribution of the ΔR between the axes of the two k_t subjets in the jet for single and merged b-jets from 80 GeV to 110 GeV (left) and 200 GeV to 270 GeV (right).

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 [69]. 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 [70].

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

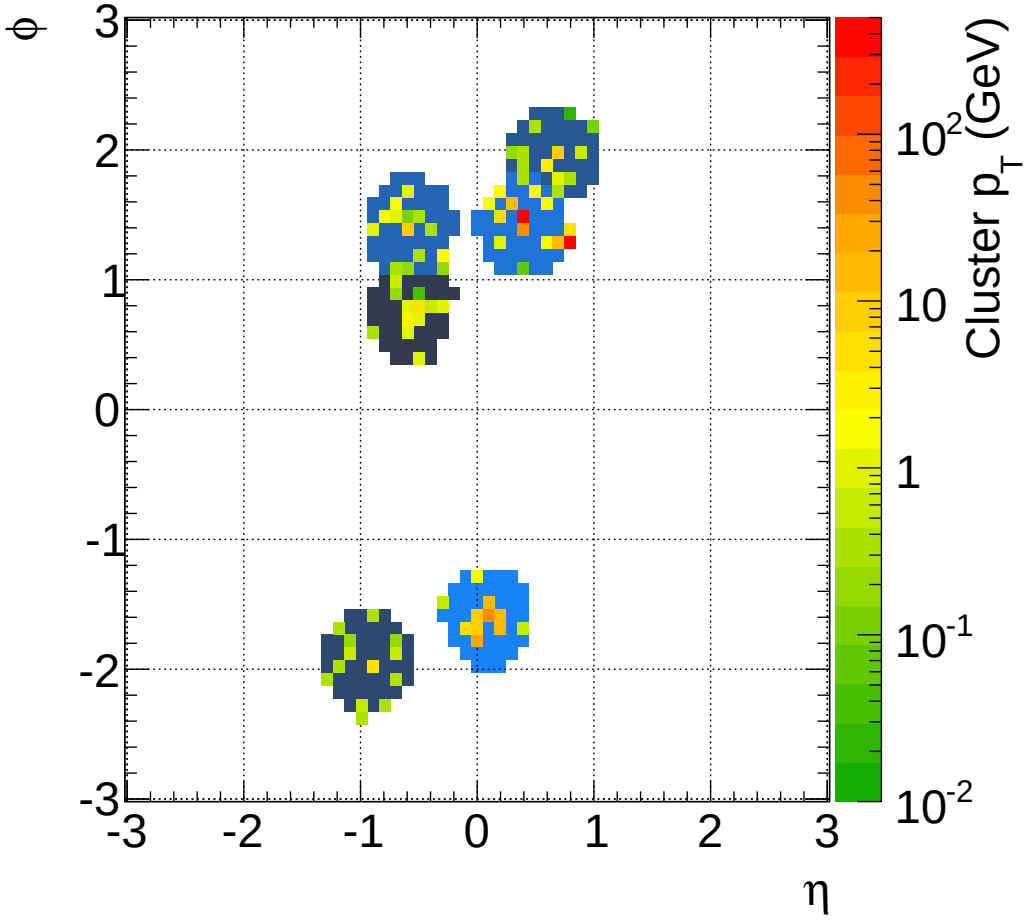

Figure 2.14: Event display of a merged b-jet in $(\eta, \phi) = (0.46, 1.41)$ and $p_T = 110$ GeV. The two b-hadrons are indicated as two red squares. The area of the jet is shown in blue and the topoclusters belonging to the jet are shown in different colors, from green to orange, depending on their 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 [69] . 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.

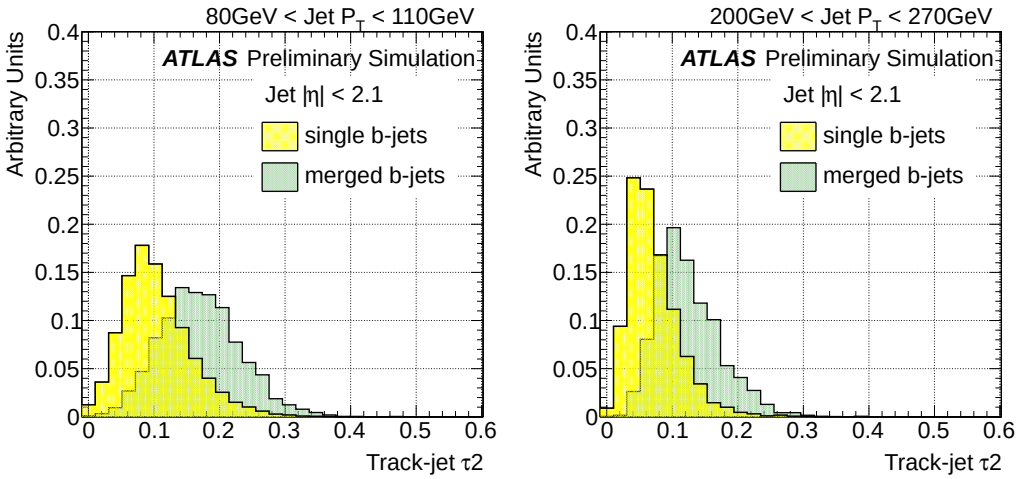

Figure 2.15: Distribution of τ_2 for single and merged b-jets from 80 GeV to 110 GeV (left) and 200 GeV to 270 GeV (right).

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,

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

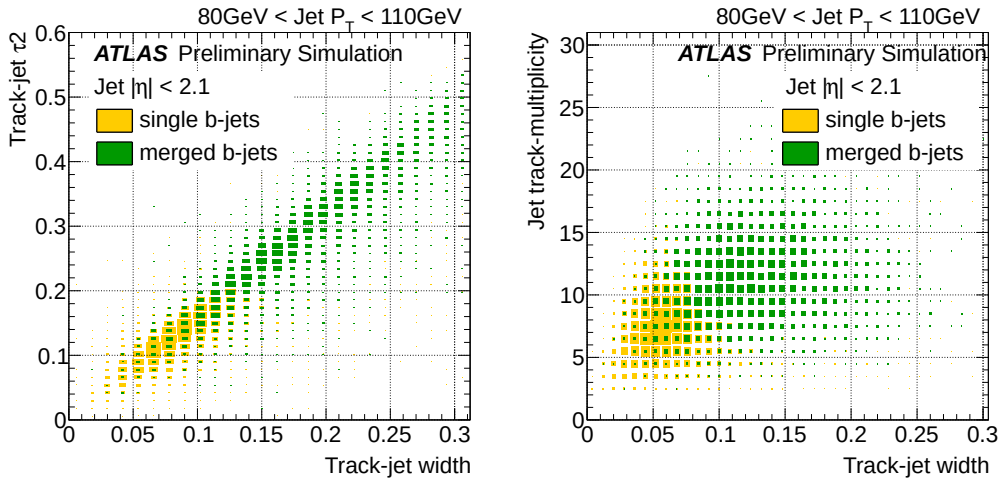

Figure 2.16: Correlation between τ_2 and track-jet width (left) and jet track multiplicity and track-jet width (right) for single and merged b-jets from 80 GeV to 110 GeV.

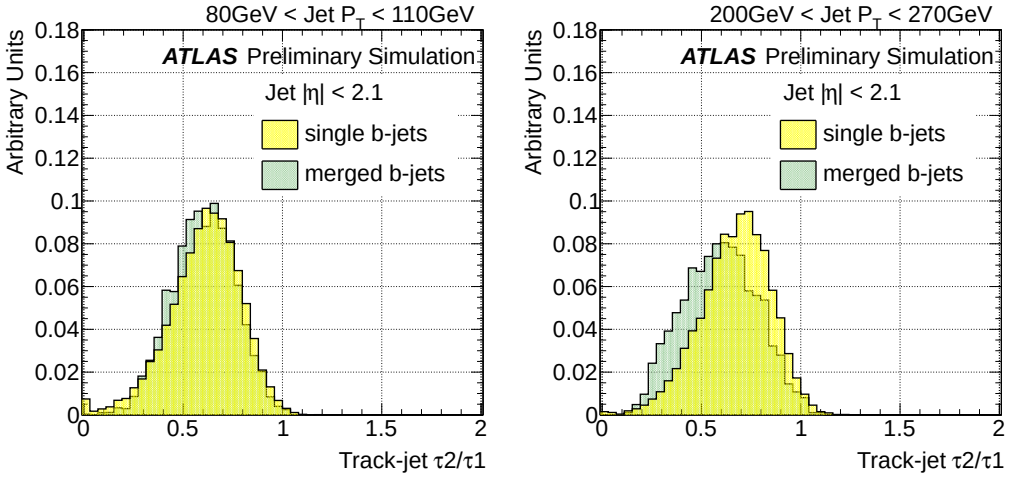

Figure 2.17: Distribution of τ_2/τ_1 for single and merged b-jets from 80 GeV to 110 GeV (left) and 200 GeV to 270 GeV (right).

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

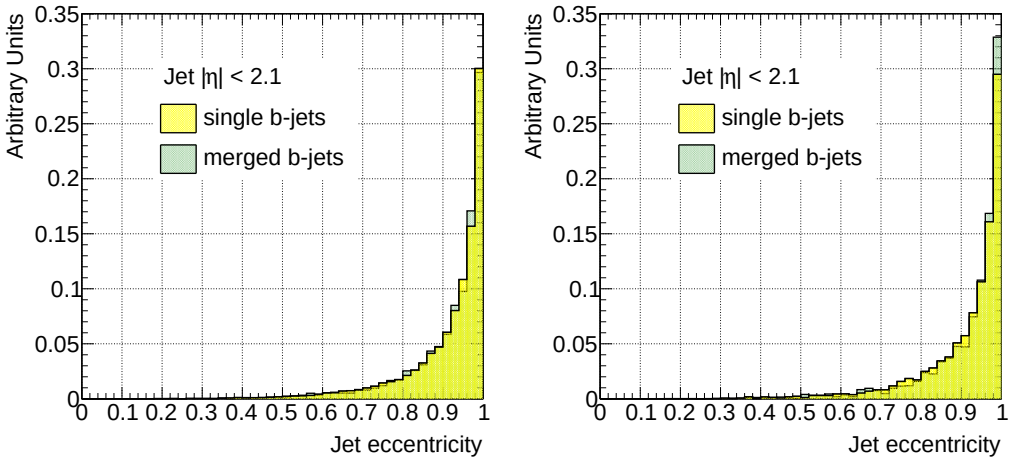

Figure 2.18: Distribution of the jet eccentricity for single and merged b-jets from 80 GeV to 110 GeV (left) and 200 GeV to 270 GeV (right).

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 2.25 and **??**.The agreement is still very good, suggesting that this compensation is not likely to occur in samples sufficiently enriched in b-jets.

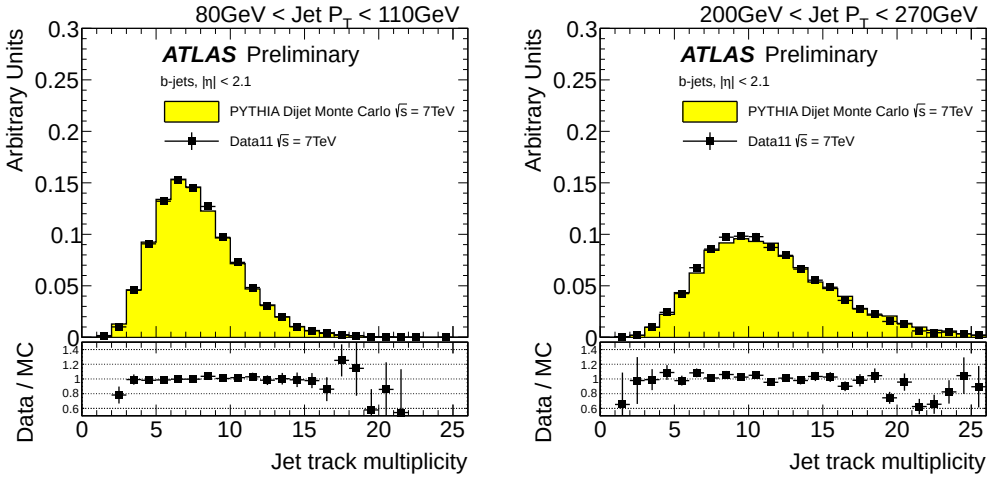

Figure 2.19: Distribution of the jet track multiplicity in 2 different jet p_T bins, for experimental data collected by ATLAS during 2011 (solid black points), and simulated data (filled histograms). The ratio data over simulation is shown at the bottom of each plot.

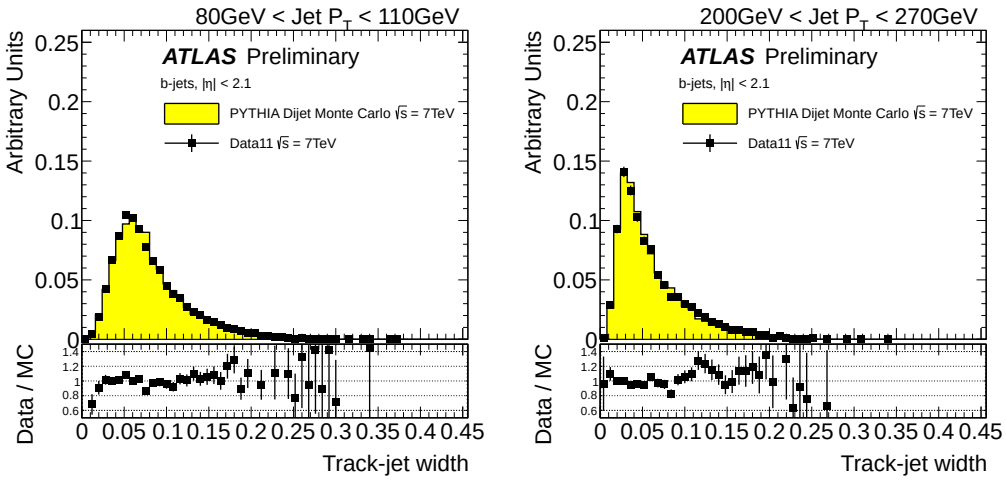

Figure 2.20: Distribution of the track-jet width in 2 different jet p_T bins, for experimental data collected by ATLAS during 2011 (solid black points), and simulated data (filled histograms). The ratio data over simulation is shown at the bottom of each plot.

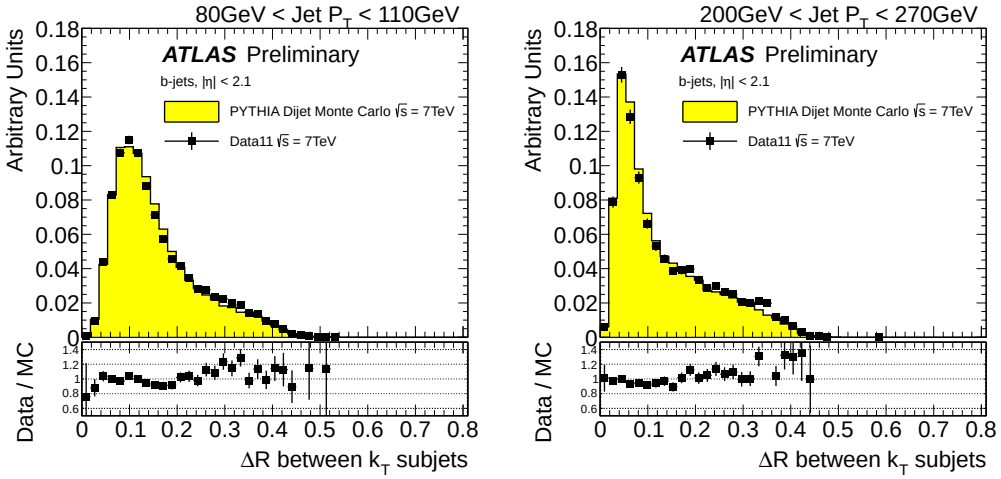

Figure 2.21: Distribution of the ΔR between the axes of the two k_t subjets in the jet in 2 different jet p_T bins, for experimental data collected by ATLAS during 2011 (solid black points), and simulated data (filled histograms). The ratio data over simulation is shown at the bottom of each plot.

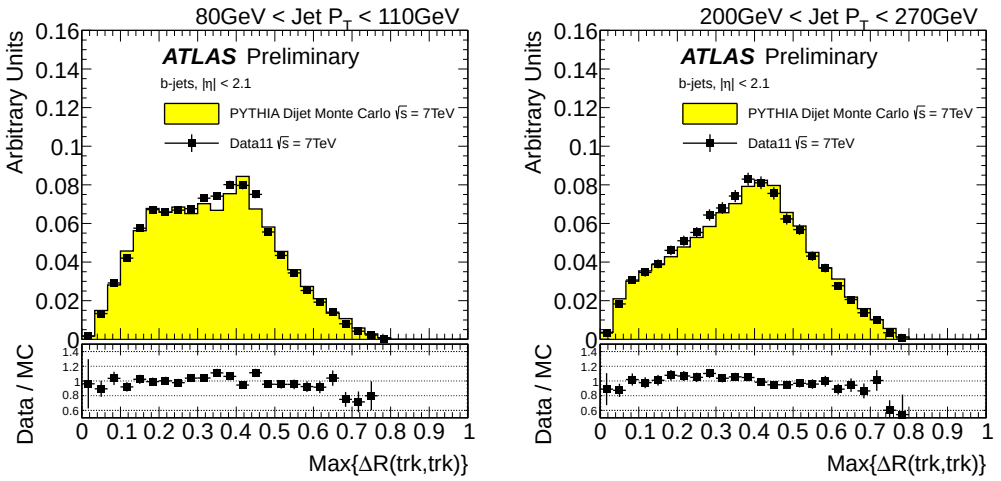

Figure 2.22: Distribution of maximum ΔR between pairs of tracks in two different jet p_T bins, for experimental data collected by ATLAS during 2011 (solid black points), and simulated data (filled histograms). The ratio data over simulation is shown at the bottom of each plot.

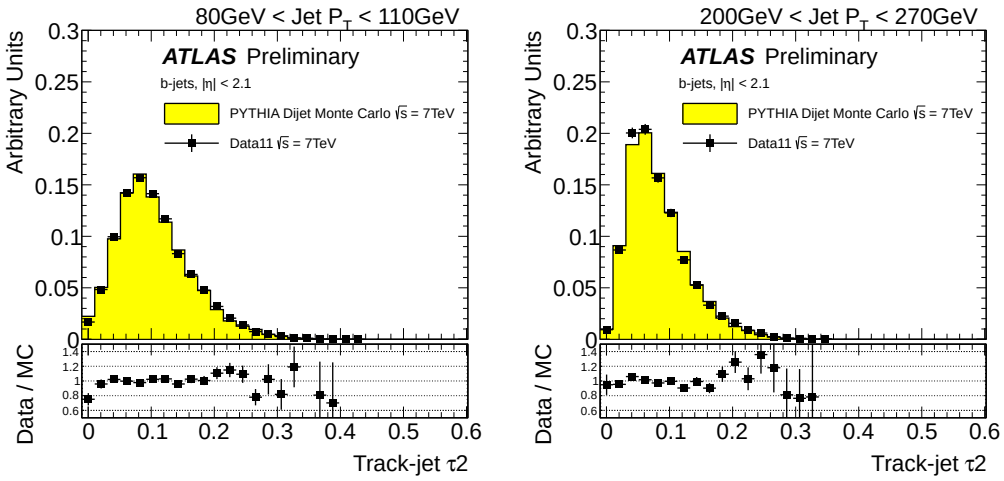

Figure 2.23: Distribution of τ_2 in two different jet p_T bins, for experimental data collected by ATLAS during 2011 (solid black points), and simulated data (filled histograms). The ratio data over simulation is shown at the bottom of each plot.

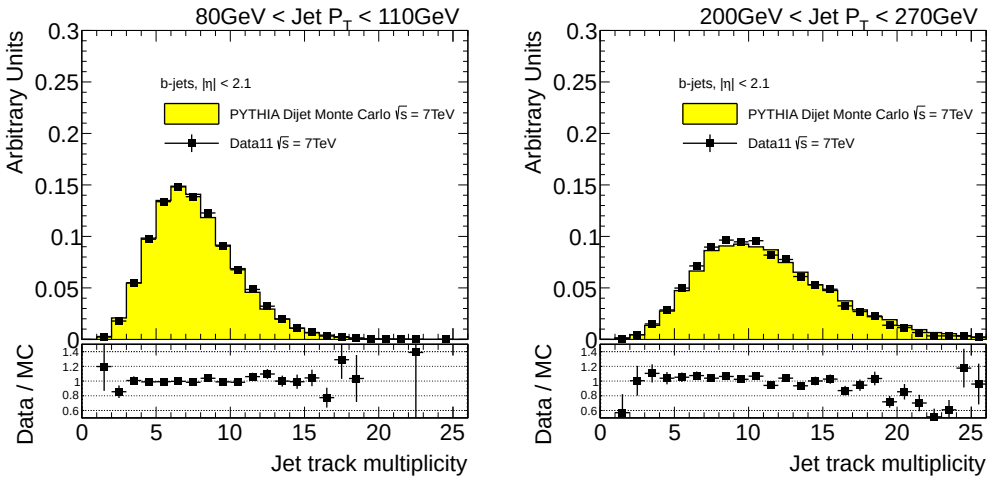

Figure 2.24: Distribution of the jet track multiplicity in 2 different jet p_T bins, for experimental data collected by ATLAS during 2011 (solid black points), and simulated data (filled histograms). Jets were selected using MV1 tagger at its 70% b-jet efficiency working point. The ratio data over simulation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

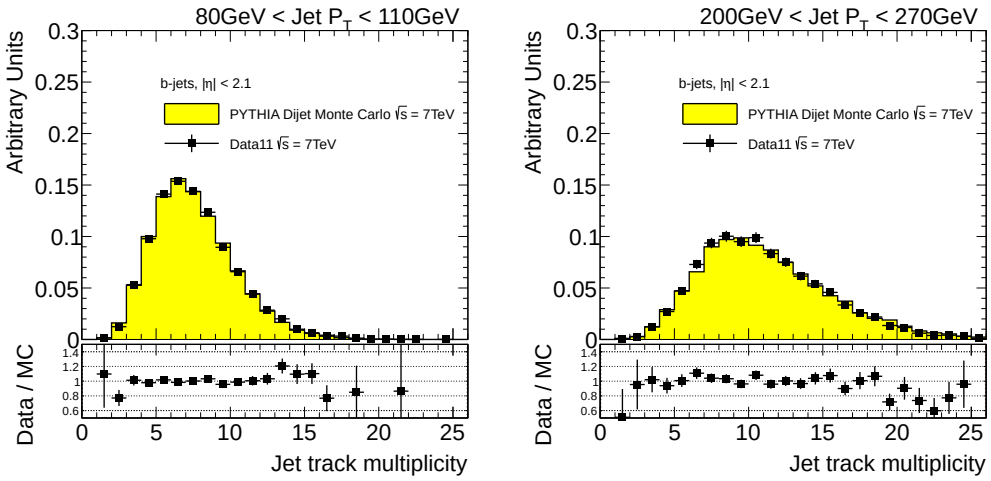

Figure 2.25: Distribution of the jet track multiplicity in 2 different jet p_T bins, for experimental data collected by ATLAS during 2011 (solid black points), and simulated data (filled histograms). Jets were selected using JetFitter tagger at its 60% b-jet efficiency working point. The ratio data over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

Bibliography

[1] F. Abe et al. Observation of top quark production in $\overline{p}p$ collisions with the collider detector at fermilab. *Phys. Rev. Lett.*, 74:2626–2631, Apr 1995.

[2] S. Abachi et al. Search for high mass top quark production in $p\overline{p}$ collisions at $\sqrt{s} =$ 1.8 tev. *Phys. Rev. Lett.*, 74:2422–2426, Mar 1995.

[3] P.W. Higgs. Broken symmetries, massless particles and gauge fields. *Physics Letters*, 12(2):132 – 133, 1964.

[4] Georges Aad et al. Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC. 2012.

[5] F. J. Dyson. The radiation theories of tomonaga, schwinger, and feynman. *Phys. Rev.*, 75:486–502, Feb 1949.

[6] Makoto Kobayashi and Toshihide Maskawa. CP-Violation in the Renormalizable Theory of Weak Interaction. *Progress of Theoretical Physics*, 49(2):652–657, 1973.

[7] Sheldon L. Glashow. Partial-symmetries of weak interactions. *Nuclear Physics*, 22(4):579 – 588, 1961.

[8] A. Salam and J.C. Ward. Electromagnetic and weak interactions. *Physics Letters*, 13(2):168 – 171, 1964.

[9] Steven Weinberg. A model of leptons. *Phys. Rev. Lett.*, 19:1264–1266, Nov 1967.

[10] M Banner et al. Observation of single isolated electrons of high transverse momentum in events with missing transverse energy at the cern pp collider. *Physics Letters B*, 122(5-6):476–485, 1983.

[11] G. Arnison et al. Experimental observation of lepton pairs of invariant mass around 95 gev/c2 at the cern sps collider. *Physics Letters B*, 126(5):398 – 410, 1983.

[12] M. Gell-Mann. A schematic model of baryons and mesons. *Physics Letters*, 8(3):214 – 215, 1964.

[13] Zweig, G. An SU(3) model for strong interaction symmetry and its breaking. 1964.

[14] G. Zweig. An SU(3) model for strong interaction symmetry and its breaking. 1964.

[15] Richard P. Feynman. Very high-energy collisions of hadrons. *Phys. Rev. Lett.*, 23:1415–1417, Dec 1969.

[16] Politzer, H. David. Reliable Perturbative Results for Strong Interactions? *Phys. Rev. Lett.*, 30:1346–1349, year =.

[17] Gross, David J. and Wilczek, Frank. Ultraviolet Behavior of Non-Abelian Gauge Theories. *Phys. Rev. Lett.*, 30:1343–1346, Jun 1973.

[18] John C. Collins, Davison E. Soper, and George F. Sterman. Factorization of Hard Processes in QCD. *Adv.Ser.Direct.High Energy Phys.*, 5:1–91, 1988.

[19] G. 't Hooft. Dimensional regularization and the renormalization group. *Nuclear Physics B*, 61(0):455– 468, 1973.

[20] Weinberg, Steven. New Approach to the Renormalization Group. *Phys. Rev. D*, 8.

[21] Ellis, R. K. and Stirling, W. J. and Webber, B. R. *QCD and Collider Physics*. Cambridge University Press, Cambridge, 1996.

[22] Mrinal Dasgupta and Gavin P Salam. Event shapes in e + e âĹŠ annihilation and deep inelastic scattering. *Journal of Physics G: Nuclear and Particle Physics*, 30(5):R143, 2004.

[23] G. Altarelli and G. Parisi. Asymptotic freedom in parton language. *Nuclear Physics B*, 126(2):298 – 318, 1977.

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

[25] M Bahr, S. Gieseke, M.A. Gigg, A. Grellscheid, K. Hamilton, O. Latunde-Dada, S Platzer, P Richardson, M.H Seymour, M Sherstnev, et al. Herwig++ physics and manual. *Eur.Phys.J.C*, 58:68, 2008.

[26] B. Andersson, G. Gustafson, G. Ingelman, and T. Sjostrand. Parton fragmentation and string dynamics. *Physics Reports*, 97(2-3):31–145, 1983.

[27] R. Corke and T. Sjöstrand. Improved parton showers at large transverse momenta. *European Physical Journal C*, 69:1, 2010.

[28] T. Sjöstrand and P. Z. Skands. Transverse-momentum-ordered showers and interleaved multiple interactions. *European Physical Journal C*, 39:129, 2005.

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

[30] Peter Z. Skands. The Perugia Tunes. 2009.

[31] Peter Zeiler Skands. Tuning Monte Carlo Generators: The Perugia Tunes. *Phys. Rev. D*, 82:074018, 2010.

[32] Manuel Bahr, Stefan Gieseke, and Michael H. Seymour. Simulation of multiple partonic interactions in herwig++. *Journal of High Energy Physics*, 2008(07):076, 2008.

[33] S. Agostinelli et al. Geant4 a simulation toolkit. *Nucl. Inst. Meth. Section A*, 506(3):250 – 303, 2003.

[34] G. Hanson, G. S. Abrams, A. M. Boyarski, M. Breidenbach, F. Bulos, W. Chinowsky, G. J. Feldman, C. E. Friedberg, D. Fryberger, G. Goldhaber, D. L. Hartill, B. Jean-Marie, J. A. Kadyk, R. R. Larsen, A. M. Litke, D. Lüke, B. A. Lulu, V. Lüth, H. L. Lynch, C. C. Morehouse, J. M. Paterson, M. L. Perl, F. M. Pierre, T. P. Pun, P. A. Rapidis, B. Richter, B. Sadoulet, R. F. Schwitters, W. Tanenbaum, G. H. Trilling, F. Vannucci, J. S. Whitaker, F. C. Winkelmann, and J. E. Wiss. Evidence for jet structure in hadron production by e^+e^- annihilation. *Phys. Rev. Lett.*, 35:1609–1612, Dec 1975.

[35] Salam, G.P. Ellements of QCD for hadron colliders. *CERN-2010-002*, Jan 2011.

[36] W. Bartel, L. Becker, R. Felst, D. Haidt, G. Knies, H. Krehbiel, P. Laurikainen, N. Magnussen, R. Meinke, B. Naroska, et al. Experimental studies on multijet production in e^+e annihilation at PETRA energies. *EPJ C Particles and Fields*, 33:8, 1986.

[37] Stephen D. Ellis and Davison E. Soper. Successive combination jet algorithm for hadron collisions. *Phys. Rev.*, D48:3160–3166, 1993.

[38] S. Catani, Y.L. Dokshitzer, H. Seymour, and B.R. Webber. Longitudinally invariant K(t) clustering algorithms for hadron hadron collisions. *Nucl. Phys.*, B406:187, 1993.

[39] Yu.L. Dokshitzer and G.D. Leder and S. Moretti and B.R. Webber. Better jet clustering algorithms. *Journal of High Energy Physics*, 1997(08):001, 1997.

[40] Matteo Cacciari, Gavin P. Salam, and Gregory Soyez. The anti-k_t jet clustering algorithm. *JHEP*, 04:063, 2008.

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

[42] M Cacciari and G.P. Salam. Dispelling the N^3 myth for the k_t jet-finder. *Phys. Lett. B*, 661:057, 2006.

[43] G. Abbiendi et al. Measurement of alpha(s) with Radiative Hadronic Events. 2007.

[44] Georges Aad et al. Measurement of event shapes at large momentum transfer with the ATLAS detector in pp collisions at $\sqrt{s} = 7$ TeV. 2012.

[45] A. Abdesselam et al. Boosted objects: a probe of beyond the standard model physics. *The European Physical Journal C - Particles and Fields*, 71:1–19, 2011.

[46] A. Altheimer et al. Jet Substructure at the Tevatron and LHC: New results, new tools, new benchmarks. 2012.

[47] ATLAS Collaboration. Atlas sensitivity to the standard model higgs in the hw and hz channels at high transverse momenta. *ATL-PHYS-PUB-2009-088*, Aug 2009.

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

[49] Graham D. Kribs, Adam Martin, Tuhin S. Roy, and Michael Spannowsky. Discovering higgs bosons of the mssm using jet substructure. *Phys. Rev. D*, 82:095012, Nov 2010.

[50] Stephen Ellis, Christopher Vermilion, Jonathan Walsh, Andrew Hornig, and Christopher Lee. Jet shapes and jet algorithms in scet. *Journal of High Energy Physics*, 2010:1–83, 2010. 10.1007/JHEP11(2010)101.

[51] G. Aad et al. Study of jet shapes in inclusive jet production in pp collisions at $\sqrt{s} = 7$ TeV using the atlas detector. *Phys. Rev. D*, 83:052003, Mar 2011.

[52] E. Norrbin and T. Sjostrand. Production and hadronization of heavy quarks. *Eur.Phys.J.*, C17:137–161, 2000.

[53] S. Frixione and M.L. Mangano. Heavy quark jets in hadronic collisions. *Nucl.Phys.*, B483:321–338, 1997.

[54] Andrea Banfi, Gavin Salam, and Giulia Zanderighi. Accurate qcd predictions for heavy-quark jets at the tevatron and lhc. *JHEP*, 0707:026, 2007.

[55] G. Corcella, I.G. Knowles, G. Marchesini, S. Moretti, K. Odagiri, et al. HERWIG 6: An Event generator for hadron emission reactions with interfering gluons (including supersymmetric processes). *JHEP*, 0101:010, 2001.

[56] M.H. Seymour. Heavy quark pair multiplicity in e^+e events. *Nuclear Physics B*, 436(1-2):163–183, 1995.

[57] Andrea Banfi, Gavin Salam, and Giulia Zanderighi. Infrared safe definition of jet flavour. *Eur.Phys.J.C*, 47:022, 2006.

[58] John M. Campbell, R.Keith Ellis, F. Maltoni, and S. Willenbrock. Production of a W boson and two jets with one b^- quark tag. *Phys.Rev.*, D75:054015, 2007.

[59] ATLAS Collaboration. Search for supersymmetry in pp collisions at $\sqrt{s} = 7$TeV in final states with missing transverse momentum, b-jets and no leptons with the ATLAS detector. *ATLAS-CONF-2011-098*, 2011.

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

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

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

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

[64] 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.

[65] 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.

[66] 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.

[67] 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.

[68] 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.

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

[70] 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.