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

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

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

Resumen

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

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

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

Abstract

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

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

Contents

Chapter 1

Introduction

The first years of proton-proton collisions at a centre of mass energy of 7 TeV delivered by the Large Hadron Collider and recorded by the ATLAS experiment have provided data to explore quantum chromodynamics (QCD) at scales never reached before. Precision measurements of strong interactions are interesting in their own right, but, in addition, QCD provides one of the main backgrounds to many New Physics measurements; furthermore, it is also through tests of QCD that New Physics may be discovered.

Due to QCD confinement the experimental signature of quarks and gluons are not the quarks and gluons themselves but a spray of "colorless" hadrons, that we call *jets*. Hadronic jets are a fundamental ingredient for precision tests of QCD: understanding and measuring their performance is crucial in the LHC environment. A wide range of physics signatures, within the Standard Model (SM) and Beyond the Standard Model (BSM) predictions, contain jets originating from bottom (b) quarks. The ability to identify jets containing b-hadrons, the product of the hadronization of b-quarks, is therefore important for the high-p_T physics program of the ATLAS experiment.

b-tagging algorithms rely on the relatively long decay length of b-hadrons

that gives rise to large impact parameter tracks and displaced decay secondary vertices; or on the presence of a soft lepton within the jet, the product of the semileptonic b-decay. These algorithms, however, do not provide information on the number of b-hadrons within the jet. In particular, they tag "merged" jets containing a $b\bar{b}$ pair, with no net heavy flavour, which do not correspond to the intuitive picture of a b-jet as a jet containing a single b-quark or antiquark.

Successfully tagging merged b-jets, which in QCD are produced mainly from gluon splitting $g \to b\bar{b}$, is important to reduce and to improve the estimation of the b-tag background to Standard Model analyses and to new physics searches involving b-jets in the final state. In particular, it has been shown that efficient tagging of gluon splitting jets can also help in reducing the theoretical uncertainties in the calculation of the inclusive b-jet spectrum [1].

There are two possible strategies to attempt to identify b-jets containing two b-hadrons in hadronic collisions. One of them, implemented at the CDF experiment at Fermilab [2], relies on the direct reconstruction of the two b-decay secondary vertices. This allows the measurement of the angular separation between the b-hadrons, but suffers from the low efficiency of a double b-tag requirement plus additional reconstruction inefficiencies at small angular separation between the two b-hadrons. In this thesis we develop for the first time an alternative method that does not rely on explicit vertex finding, but exploits the substructure differences between single and merged b-jets, combining them in a multivariate analysis. The method developed is then 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 in 2011.

The thesis is organized as follows: Chapter 2 describes the theoretical framework, with emphasis in the theory of the strong interactions and the aspects that are important for the understanding of the hadronic final state in hadronic collisions. The LHC and the ATLAS detector are described in Chapter 3, together with a summary of the experimental conditions during the 2011 data taking. Chapter 4 details how jet reconstruction and calibration are performed at ATLAS and describes the procedure for the identification of b-quark jets. Chapter 5 presents the analysis of jet shape and substructure variables for the discrimination between single and double b-hadron jets. The validation of the variables in 2011 data is also included. The construction of the multivariate discriminator and the discussion of its systematic uncertainties are presented in Chapter 6. Chapter 7 details the technique used for the measurement of the fraction of double b-hadron jets in QCD b-production and the associated systematic uncertainties. Finally, chapter 8 presents a summary and conclusion.

Chapter 2

Theoretical introduction

This chapter presents an introduction to the theoretical aspects involved in this thesis. After a brief overview of the Standard Model and perturbative Quantum Chromodynamics, Sections 2.1 and 2.2, the Monte Carlo tools to simulate QCD processes and the subjetc of jets and jet algorithms are discussed in some detail in Sections 2.3 and 2.4. Finally, the concepts specific to this thesis, the QCD production of heavy-flavor (HF) jets, and the motivation for studying HF jets originating from gluon splitting, are the subject of Sections 2.5 and 2.6, respectively.

2.1 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

[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.

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* 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

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

found [3, 4]. 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).

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 [5]. Very recently, a Higgs-like particle was discovered by ATLAS and CMS experiments at the LHC [6]. 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

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

density, which is given by

$$\mathcal{L}_{QED} = \bar{\psi}(i\gamma^\mu D_\mu - m)\psi - \frac{1}{4}F_{\mu\nu}F^{\mu\nu}. \tag{2.1}$$

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

$$D_\mu = \partial_\mu - ieA_\mu \tag{2.2}$$

$$F^{\mu\nu} = \partial^\mu A^\nu - \partial^\nu A^\mu \tag{2.3}$$

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 [7]. 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

$$P_L = \frac{1-\gamma^5}{2} \quad P_R = \frac{1+\gamma^5}{2} \tag{2.4}$$

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} [8], 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 [9, 10, 11]. The gauge fields corresponding to the generators of the gauge symmetry are W_μ^i 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

$$W_{\mu\nu}^i = \partial_\mu W_\nu^i - \partial_\nu W_\mu^i + g\epsilon_{ijk} W_\mu^j W_\nu^k \tag{2.5}$$

$$B_{\mu\nu} = \partial_\mu B_\nu - \partial_\nu B_\mu. \tag{2.6}$$

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

$$\mathcal{L}_{EW} = i \sum_f \bar{\psi}_f \gamma^\mu D_\mu \psi_f - \frac{1}{4} W_{\mu\nu}^i W^{i\mu\nu} - \frac{1}{4} B_{\mu\nu} B^{\mu\nu}, \tag{2.7}$$

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

$$D_\mu = \partial_\mu + \frac{1}{2} i g \tau^i W_\mu^i - \frac{1}{2} i g' Y B_\mu \tag{2.8}$$

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

$$Q = Y + I_3 \tag{2.9}$$

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 [12, 13]. 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 [14, 15], 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

$$\mathcal{L}_{QCD} = \sum_q \bar{\psi}_{q,a}(i\gamma^\mu(D_\mu)_{ab} - m_q\delta_{ab})\psi_{q,b} - \frac{1}{4}W^A_{\mu\nu}W^{A\mu\nu}. \tag{2.10}$$

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

$$D_\mu = \partial_\mu + ig_s t^A \mathcal{A}^A_\mu, \tag{2.11}$$

$$G^A_{\mu\nu} = \partial_\mu \mathcal{A}^A_\nu - \partial_\nu \mathcal{A}^A_\mu + g f^{ABC} \mathcal{A}^B_\mu \mathcal{A}^C_\nu, \tag{2.12}$$

where $\mathcal{A}^A_\mu$ 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 $gggg$.

2.2 Perturbative QCD

As described in section 2.1, the fundamental actors of the theory of the strong interactions are quarks and gluons or, collectively, partons [16]. Partons are confined in hadrons, but act quasi-free at sufficiently small scales. This behaviour is called asymptotic freedom. On the contrary, partons are coupled together more strongly as the distance between them increases. This effect, known as confinement, explains why quarks and gluons are only observed, at low energies, trapped together into color-neutral hadrons[4]. A quantitative

[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.

representation of the decreasing power of the strong force with increasing energy is given by the negative β-function of QCD [17, 18], which describes how $\alpha_s(\mu^2)$ decreases with energy, the so called "running" of the coupling constant.

The experimental consequence of asymptotic freedom is that the hard interactions of quarks and gluons at the energy scale probed by hadron colliders can be described by perturbative QCD, although their presence in the final state can only be inferred indirectly, as they appear confined in colorless hadrons. Each order of the perturbative expansion corresponds to an additional power in the coupling constant. This power is related to the number of vertices in the matrix element QCD Feynman diagrams, with $\sqrt{\alpha_s}$ per vertex, 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). Since the value of α_s varies with energy, it must be evaluated at the energy scale of the interaction. In particular, at the electroweak scale, $\alpha_s(M_Z) \sim 0.117$, and the perturbative expansion converges relatively fast, allowing all except the lower order terms to be ignored. The complexity of the process determines the precision of the calculation that has to be performed. For inclusive parton production, calculations are typically performed at next-to-leading order (NLO), only recently some processes have been extended up to third order, that is NNLO, like QCD $\gamma\gamma$ background to $H \to \gamma\gamma$ [19], or top QCD production [20].

Using this formalism, the two parton hard interaction cross section can be

computed at some fixed-order in perturbation theory. However, hadron 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 [21] 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 $f_i(x)$, the parton distribution functions (PDFs), where $f_i(x)dx$ gives the number of partons of type i that carry a fraction of the total hadron momentum between x and $x + dx$. Thus the total cross section, at some energy Q^2 that characterizes the interaction, can be written as:

$$\sigma(P_1, P_2) = \sum_{i,j} \int dx_1 dx_2 f_i(x_1, \mu_f^2) f_j(x_2, \mu_f^2) \hat{\sigma}_{ij}(p_1, p_2, \alpha_s(\mu_r^2), Q^2/\mu_r^2, Q^2/\mu_f^2). \tag{2.13}$$

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 interacting partons, and $p_1 = x_1 P_1$ and $p_2 = x_2 P_2$ are the interacting parton momenta. 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 [22, 23]. 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.

If the perturbative expansion were carried to all orders, the cross section

$\sigma(P_1, P_2)$ in Equation 2.13 would be independent of μ_F and μ_R. In actual finite order calculations this is not true. They are usually both taken to be equal, $\mu_F = \mu_R = \mu$, chosen at the typical scale Q^2 of the process, in order to minimize the contribution of (uncalculated) higher order terms which appear as logarithmic terms of the form $\log(Q^2/\mu_R^2)$ and $\log(Q^2/\mu_F^2)$. The dependence of the prediction on μ_F and μ_R is assigned as a theoretical uncertainty.

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 Giorgio Parisi [24]. 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.

When the process studied contains two or more natural scales, it is not possible to cancel the logarithms in the higher order terms with an adequate choice of the μ scale. This happens for instance when there both are natural mass and momentum scales, like (p_T, M_H) in the prediction of the Higgs transverse momentum distribution in $H + X$ production or (p_T, m_b) in the differential cross-section for QCD b production. These contain respectively $\log^n({p_T}^2/M_H^2)$ and $\log^n({p_T}^2/m_b^2)$ terms which cannot be cancelled by a scale choice and cannot be guaranteed to be small. Resummation techniques have been developed for this cases, which incorporate the leading (subleading) logarithmic terms at all orders, in the so called "leading logarithmic", LL, ("next-to-leading logarithmic", NLL) approximation.

2.3 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 [25] and HERWIG++ [26] 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 [27]. 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 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

[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.

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 $2\to1$ and $2\to2$, some $2\to3$, 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 [28] both for ISR and for FSR. Also MPIs are ordered in

p_T [29]. Hadronization is based solely on the Lund string fragmentation framework.

For the results presented in this thesis simulated samples of dijet (see Section 2.4) events from proton-proton collision processes were generated with PYTHIA 6.423 [25]. The ATLAS AMBT2 tune of the soft model parameters was used [30]. 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 [31] 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 [32] includes 7 TeV data from 2011 data taking.

Herwig++

HERWIG++ [26] 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 [33].

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 [34] 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.

2.4 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 [35].

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 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 de-

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

scribe the full process. The full parton shower and hadronization process then cannot be calculated from first principles, but has to be modelled.

2.4.1 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. 2.1.

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

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

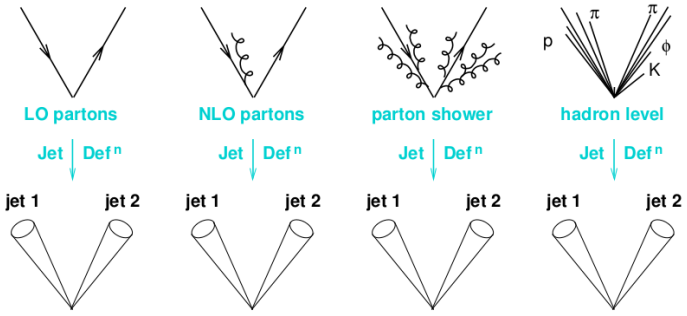

Figure 2.1: The application of a jet definition to a variety of events that differ just through soft/collinear branching and hadronization should give identical jets in all cases [36].

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 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

[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 3.2. 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.

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.

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.

[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.

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 [37] and subsequently extended for use at hadron colliders [38, 39]. 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:

$$d_{ij} = \frac{2E_iE_j(1-\cos\theta_{ij})}{Q^2}, \tag{2.14}$$

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,

$$d_{ij} = \min(p_{ti}^2, p_{tj}^2)\frac{\Delta R_{ij}^2}{R^2} \tag{2.15}$$

and the beam distance,

$$d_{iB} = p_{ti}^2. \tag{2.16}$$

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. 2.15.
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:

$$d_{ij} = \min(p_{ti}^{2n}, p_{tj}^{2n}) \frac{\Delta R_{ij}^2}{R^2} \tag{2.17}$$

$$d_{iB} = p_{ti}^{2p}. \tag{2.18}$$

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 [40], with $p = 0$, and the anti-k_t algorithm [41], with $p = -1$.

The Cambridge-Aachen algorithm. The C/A algorithm is obtained by choosing a value $p = 0$ in Equations 2.17 and 2.18. 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.

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 2.17 and 2.18, first clusters hard objects together which results in more regular jets with respect to the k_t and C/A

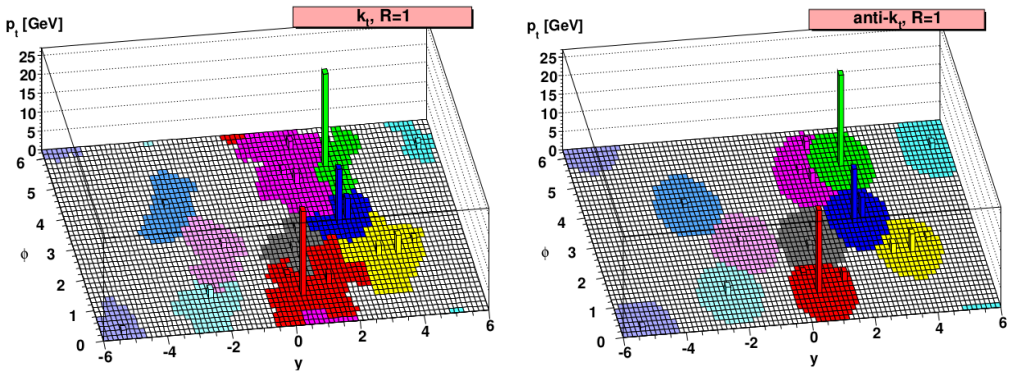

Figure 2.2: A sample parton-level event, generated with HERWIG, clustered with the k_t and anti-k_t algorithms, illustrating the active area of the resulting jets [42].

algorithms. This characteristic is illustrated for the k_t and anti-k_t algorithms in Fig. 2.2.

For this reason and the fact that this algorithm is less sensitive to soft emissions (see Chapter 4) 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 [43] software package for jet-finding.

2.4.2 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 [35]. 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 [44, 45].

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 [46, 47] 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 [48], but they have been also succesfully used in other applications, including separating quark

jets from gluon jets [49] and identifying boosted decay producs in new physics searches [50].

Jet shapes, which are event shape-like observables applied to single jets, are an effective tool to measure the structure of individual jets [51].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 [52].

In chapter 5, 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.

2.5 Production of b-jets

Jets produced by the fragmentation of b-quarks or b-jets, enter in many collider searches, notably because they are produced in the decays of various SM massive 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 mechanism of heavy-flavor production is Quantum Chromdynamics. Heavy-flavor QCD processes can be classified into three categories depending on the number of b-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 [53].

- **Heavy-flavor creation (FCR)**: two b-quarks in the hard scatter final state (FS), and no b-quarks in the initial state (IS). At leading order this process is described by $gg \to b\bar{b}$ and $q\bar{q} \to b\bar{b}$. It is called flavor creation because a $b\bar{b}$ pair is produced out of a light parton initial state. See Fig. 2.3a.

- **Heavy-flavor excitation (FEX)**: one b-quark in the IS and one b-quark in the FS. The process 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. Alternatively, if using a b-quark PDF, it can be described by a t-channel b – light-parton scattering. See Fig. 2.3b.
- **Gluon splitting (GSP)**: no heavy-quarks participate in the hard scatter, but a final state gluon produces a $b\bar{b}$ pair via a subsequent $g \to b\bar{b}$ branching. See Fig. 2.3c. The notation GSP can be extended to any QCD process (or any SM process) in which a gluon in the ensuing parton shower splits into a $b\bar{b}$ pair.

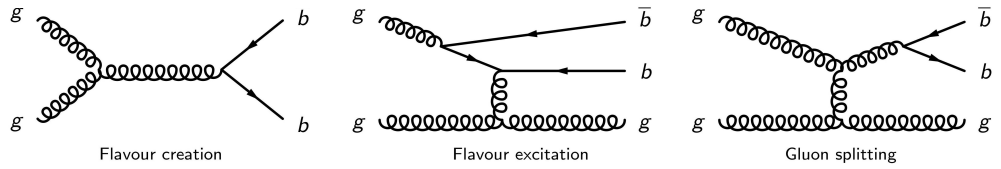

Figure 2.3: Representative diagrams of the three channels contributing to QCD b-quark production up to NLO. (a) *Left*: The flavour creation channel is the only one present at LO. At NLO, two new channels open up, referred to as (b) *Center*: flavour excitation (center) and (c) *Right*: gluon splitting.

Final state b-quarks hadronize into b-hadrons, either B mesons (B^+, B^0, B_s^0, B_c^+) or b-baryons. During the fragmentation process, other particles will also be produced along with the b-hadron, giving rise to b-jets. In the flavor creation case, 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 splitting is expected to give rise to close-by b-hadrons. Depending on how the separation between them compares to the jet size R parameter, they will clusterized within the same hadronic jet or identified as neighbor jets. 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, while there is no such an association between light partons (u, d, s, g) and observed final state hadrons. In addition, the study of b-quark production has 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 however rather important uncertainties ($\sim 50\%$), considerably larger than those for the light jet inclusive spectrum ($\sim 10 - 20\%$) [54]. These uncertainties are quantified by the renormalization and factorization scale dependence of the calculation. In varying μ_R and μ_F between $p_T/2$ and $2p_T$, p_T being the transverse momentum of the hardest jet in the event, the heavy flavor cross section varies by up to 50%, as shown in middle panel of Fig. 2.4. A review of the origin of these uncertainties is presented by Banfi, Salam and Zanderighi in reference [1]. They show that they arise from the poor convergence of the perturbative series, as evidenced by a rather large value of the K-factor,

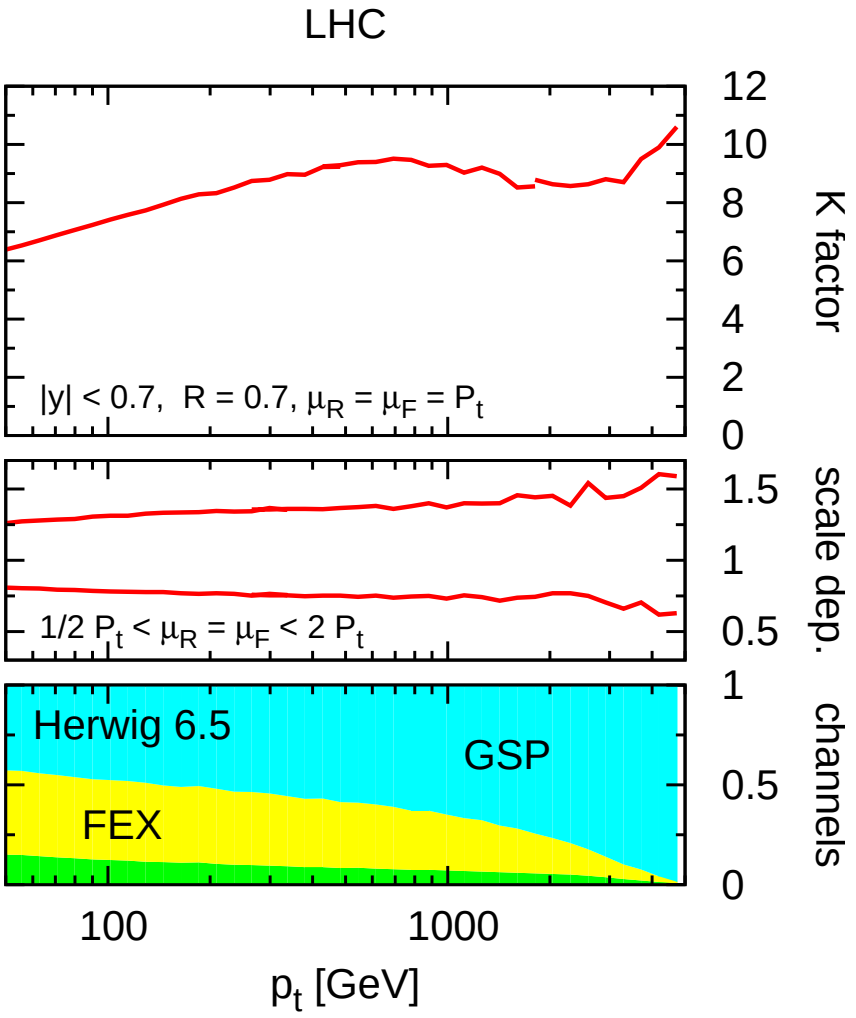

Figure 2.4: *Top*: K-factor for the inclusive b-jet spectrum taken from [1], clustering particles into jets using the k_t jet-algorithm [39] with R=0.7, and selecting jets in the central rapidity region ($|y| < 0.7$). *Middle*: scale dependence obtained by simultaneously varying the renormalisation and factorisation scales by a factor two around p_T, the transverse momentum of the hardest jet in the event. *Bottom*: breakdown of spectrum into the three major underlying channels, flavor creation (FCR) flavor excitation (FEX) and gluon splitting (GSP) as predicted by a parton shower MC, Herwig [55].

the ratio of the next-to-leading order (NLO) to the leading order (LO) cross section. This is illustrated in the upper panel of Fig. 2.4 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.

The fact that the perturbative series is very poorly convergent can be explained in terms of 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]. The various channels can be approximately separated with a parton shower Monte Carlo generator such as HERWIG or PYTHIA, where one can determine the underlying hard process from the event record. These MC generators effectively include NLO effects via parton showers. The relative importance of each channel in the b-jet spectrum is shown in the bottom panel of Fig. 2.4. It is found that the supposedly LO channel (FCR) channel has a much smaller contribution than the two channels that at fixed order enter only at NLO (FEX and GSP). This is because both NLO channels receive strong enhancement from collinear logarithms, going as $\alpha_s^2(\alpha_s \ln(p_T/m_b))^n$ for flavour excitation [24] and $\alpha_s^2 \cdot \alpha_s^n \ln^{2n-1}(p_T/m_b)$ for gluon splitting ($n \geq 1$) [56].

As such, this problem is not solved yet. The obvious approach of carrying out the full massive next-to-next-to-leading order calculation is beyond the limit of today's technology. A second approach would be to carry out the explicit resummation of both the incoming and outgoing collinear logarithms. Although the technology for each resummation on its own is well-known at NLL accuracy, significant effort would be necessary to assemble them

[9]It is sometimes stated that it makes no sense, beyond LO, to separately discuss the different channels, for example because diagrams from 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.

together effectively.

In both approaches, the largest residual uncertainties are likely to be associated with the channel with the most logarithms, gluon splitting, and the presence of $g \to b\bar{b}$ jets.

2.6 Identification of b-jets from gluon splitting

As discussed in the previous Section, jets stemming from the hadronization of b-quarks, *i.e.* b-jets, can have two possible origins: the fragmentation of a single b-quark or the fragmentation of a gluon via a $b\bar{b}$ pair, $g \to b\bar{b}$, producing respectively jets containing a single b-hadron or a pair of b-hadrons. The main subject of this thesis is precisely the design, development ad tuning of an algorithm to distinguish between these to cases. In this section we present the theoretical motivation and applications of a tool with the ability to separate genuine b-quark b-jets from those produced via gluon splitting.

2.6.1 The measurement of the inclusive b-jet spectrum

The large theoretical uncertainties in the predicion of the QCD inclusive b-jet spectrum, Section 2.5, arise from the strong enhancement from collinear logarithms in the flavor excitation ($\sim \ln^n(p_T/m_b)$) and in particular in the gluon splitting ($\sim \ln^{2n-1}(p_T/m_b)$) processes. This last channel however does not even correspond to one's physical idea of a b-jet, i.e. one induced by a hard b-quark, and it seems somehow unnatural to include it at all as part of one's b-jet spectrum. Ref. [1] proposes a new observable to free the heavy-flavor spectrum calculation from collinear logarithms, and improve the accuracy of the theoretical prediction. At a theoretical level this is accomplished by introducing a new jet clustering procedure, the "flavour-k_t" jet algorithm [57],

which maintains in an infrared-safe way the correspondence between partonic flavour and jet flavour: a jet containing equal number of b quarks and b antiquarks is considered to be a light jet. In this way, jets that contain a b and $\bar{b}$, which in a parton shower MC generator are produced $\sim 95\%$ of the time from the gluon splitting channel, do not contribute to the b-jet spectrum. From an experimental side, this requires the separation of single and merged b-jets.

Further improvement can be obtained by exploiting the fact that the logarithms of p_T/m_b that remain are those associated with flavour excitation, which coincide with those resummed in the b-quark parton distribution function (PDF) at scale p_T. If one uses a b-quark PDF to resum these logarithms, no other logarithms $\ln(p_T/m_b)$ appear in the rest of the calculation

With this procedure, the K-factor for the differential heavy-jet spectrum cross-section can be shown not to exceed a value of $K = 1.4$, with a factor of four reduction in the theoretical (scale variation) uncertainties, allowing a much stricter comparison between theory and experiment.

2.6.2 Rejection of background in Standard Model analyses and beyond-SM searches

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 2.5 shows the two diagramas for $W + b$ production.

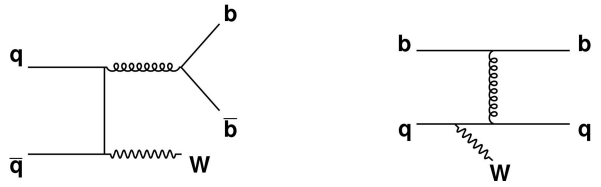

Figure 2.5: Feynman diagrams for W production in association with b quarks.

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], which 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.

2.6.3 Jet substructure and boosted objects

At the LHC, many of the particles considered to be heavy at previous accelerators are frequently produced with a transverse momentum greatly exceeding their rest mass. Good examples are the electro-weak gauge bosons $W^{\pm}$ and Z^0, the top quark, the Higgs boson or bosons and possibly other new particles in the same mass range. These boosted objects, produced either because they recoil against other energetic objects or because they arise from decays of even heavier BSM particles, can form upon decay a highly collimated topology too close to be resolved by a jet algorithm. For theses cases, sophisticated tools have been developed in the last years [60] to analyse the substructure of the ensuing jet and reveal its heavy-particle origin.

The study of $b\bar{b}$ jets from gluon splitting is an ideal testbed for studying 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 $H \to b\bar{b}$. Boosted object techniques have been already applied in ATLAS analyses like the ZZ resonance search in the $\ell\ell bb$ channel or gluino pair production in the fully hadronic channel [61], and it is investigated as a potential analysis

channel for WH and ZH production restricting to events in which the vector and Higgs bosons have large transverse momentum, ${p_T}^H \gtrsim 200$ GeV [62]. Understanding the much more common QCD events with merged $b\bar{b}$ jets will be essential before attempting to measure these rare final states.

Chapter 3

The ATLAS Detector at the LHC

3.1 The Large Hadron Collider

The Large Hadron Collider (LHC) [63] is a proton-proton (pp) synchrotron located in the previous Large Electron Positron (LEP) collider tunnel at CERN Laboratory, just outside the city of Geneva (Switzerland), approximately 100 m underground. It is designed to collide bunches of up to $\sim 10^{11}$ protons every 25 ns at a center-of-mass energy of 14 TeV (seven times the 2 TeV reached by the Tevatron accelerator at Fermilab Laboratory, in Chicago).

The experiments analyzing the collisions produced by the LHC are distributed around the 27 km ring at the various interaction points. The ATLAS experiment is located at Point 1, which is closest to the main CERN site. Point 5 houses the other general purpose detector, CMS. ALICE and LHCb experiments are located at Point 2 and Point 8, respectively. The former is designed to investigate heavy ion collisions; the latter, to investigate rare decays of b-mesons. The layout of these four experiments along the LHC ring is shown in Fig. 3.1.

Proton beams are formed, before insertion into the main LHC ring, using

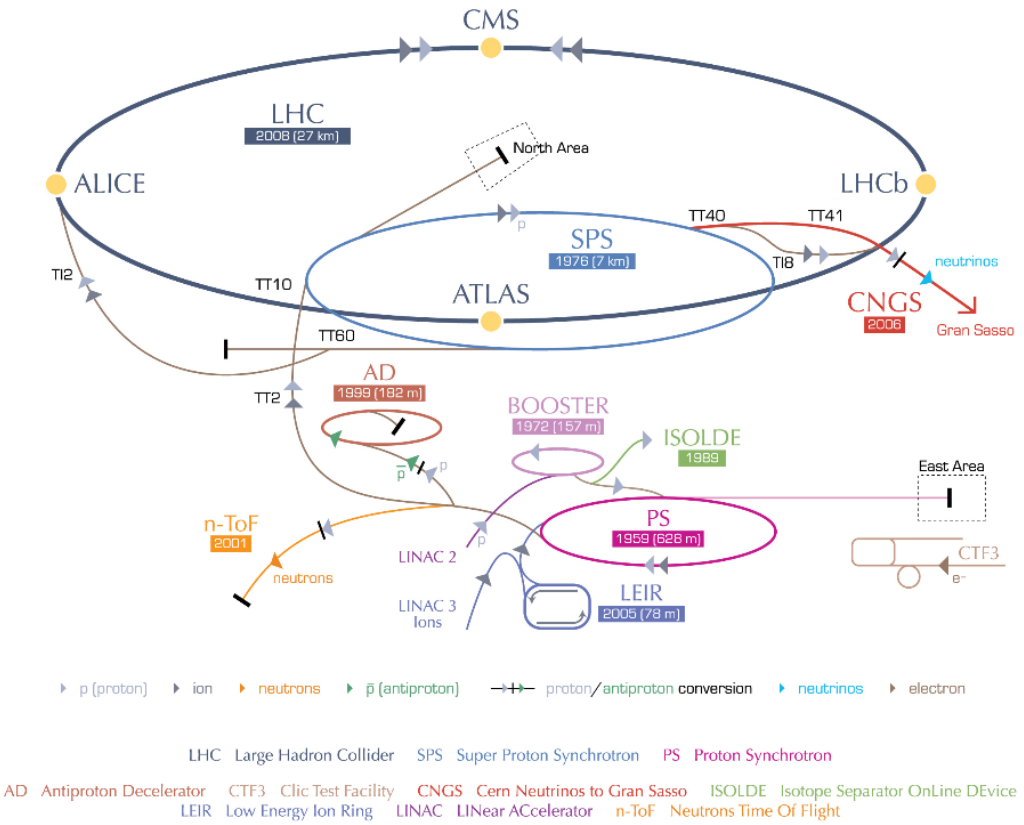

Figure 3.1: The CERN accelerator complex, showing the injection system, along with each component's date of construction, and the placement of the four main experiments.

a succession of smaller machines with increasingly higher energies, as shown in Fig. 3.1. The chain begins as protons are injected into the PS Booster (PSB) at an energy of 50 MeV from Linac2. The booster accelerates them to 1.4 GeV. The beam is then fed to the Proton Synchroton (PS) where it is accelerated to 25 GeV. At desin strength, the bunch structure, known as a bunch train, contains 72 bunches of protons upon entry to the Super Proton Synchrotron (SPS). The SPS accumulates up to four fills of 72 bunches from the PS and accelerates them to 450 GeV, with a bunch spacing of $\sim$ 25 ns. They are finally transferred to the LHC (both in a clockwise and an anticlockwise direction) where they are accelerated for 20 minutes to their nominal energy of 7 TeV. Beams will circulate for many hours inside the LHC beam pipes under normal operating conditions.

The bunch structure is a direct consequence of a radio frequency (RF) acceleration scheme used to attain the desired high proton beam energy. In RF acceleration, particles travel through a series of time-varying electrical fields and they can only be accelerated when the RF field has the correct orientation when particles pass through an accelerating cavity, which happens at well specified moments during an RF cycle. The result of a sequence of RF accelerations is several bunches of protons. It is important to note that when we speak about "beams" we refer to many bunches of protons separated by some uniform distance. Increasing the number of bunches is one of the ways to increase luminosity in a machine (see Section 3.1.1). At desinged beam intensity, when the bunches cross, there will be a maximum of about 20 collisions.

A large magnetic field is needed to guide and maintain the beam particles in their circular orbit. The needed field is achieved by using superconducting electromagnets built from NbTi coils that operate in a superconducting

state, efficiently conducting electricity without resistance or loss of energy. The currents through the coils produce magnetic fields perpendicular to the direction of motion of the protons that deflect the protons into their orbits. The whole magnetic system comprises 1232 dipole magnets of 15 m length which are used to bend the beams, and 392 quadrupole magnets, each 5-7 m long, to focus the beams. At a peak beam energy of 7 TeV, the dipoles need to produce an 8.33 T magnetic field, requiring a current of $\sim$ 12 kA. In order to deliver the current densities and magnetic field required for 7 TeV proton beams, the magnets are kept at 1.9 K by circulating superfluid helium.

The first pp collisions produced by the LHC ocurred on November 23 2009, at the SPS extraction energy of 450 GeV per beam. Very quick after, on December 8, ATLAS and CMS detectors started recording data at energy of 2.36 TeV. By this time the LHC became the highest energy accelerator in the world. During this period, bunch intensities were limited by machine-protection considerations to 1.5×10^{10} protons.

In February 2010, the LHC was commissioned once more with 450 GeV beams, and a series of tests were performed to ensure that the magnet systems could operate safely at the currents necessary to control 3.5 TeV beams. This was followed by the very first collisions at 7 TeV center-of-mass energy on March 30. During the 2010 run the beam parameters were tuned (the beam widths squeezed and the number of protons per bunch and the number of bunches in each beam increased) in order to increase the beam intensity. In particular, as the intensity of the beams increased, the mean number of interactions per bunch crossing augmented.

The data samples analyzed in this thesis 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.

Table 3.1 summarizes the basic beam parameters expected for design energy and luminosity and the beam parameters as of May 2011. The LHC performance steadily improved during 2011. The average number of interactions per bunch crossing throughout the data-taking period considered rapidly increased from ∼3 to 8 until July 2011, with a global average for this period of ≈ 6. Starting in August 2011 and lasting through the end of the proton run, this number ranged from approximately 5 to 17, with an average of about 12. This evolution is illustrated in Fig. 3.2, which shows the maximum mean number of collisions per beam crossing versus day in 2011.

Parameter	2011 runs	Design
Center-of-mass energy [TeV]	7	14
Instantaneus luminosity [$\mathrm{cm^{-2}s^{-1}}$]	3.65 10^{33} (year peak)	10^{34}
Bunches per beam	38 (May)	2808
Protons per bunch	0.8×10^{11} (May)	1.5×10^{11}
Mean interactions per crossing	6 to 12 (year average)	23

Table 3.1: Summary of beam conditions during the 2011 7 TeV runs and those foreseen at design energy and luminosity.

3.1.1 Luminosity and pile-up

The rate of events produced by the colliding beams depends on the luminosity of the collisions, which is a measure of the number of events per second per unit cross section, typically measured in units of $\mathrm{cm^2s^{-1}}$. The number of events of a particular process, then, is given by the product of the integrated luminosity, $\int dtL$, and the cross section of the process, σ_{event}. The integrated luminosities are typically quoted in units of inverse picobarns,

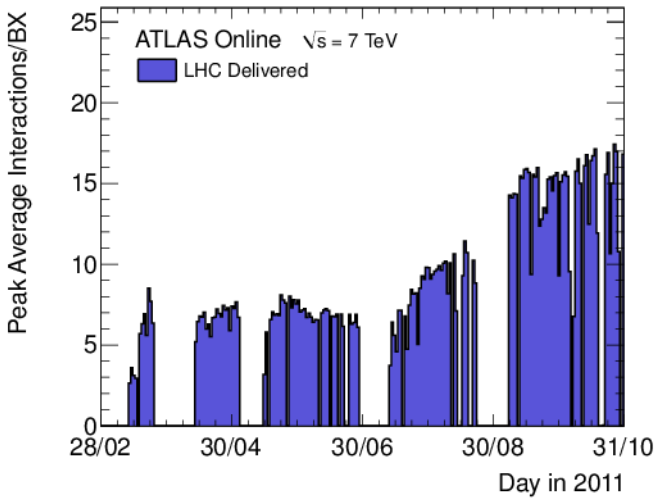

Figure 3.2: The maximum mean number of events per beam crossing versus day in 2011.

$\text{pb}^{-1} = 10^{-36}\text{cm}^2$. In order to measure processes with very little cross sections a very high luminosity is required.

The delivered luminosity can be written as [64]:

$$L = \frac{n_b f_r n_1 n_2}{2\pi \Sigma_x \Sigma_y} \tag{3.1}$$

where n_b is the number of colliding bunch pairs, n_1 and n_2 are the bunch populations (protons per bunch) in beam 1 and beam 2 respectively (together forming the bunch charged product), f_r is the machine revolution frequency, and Σ_x and Σ_y are the width and the height of the proton beams.

The number of protons per bunch, the number of bunches per beam, and the revolution frequency are all set by the beam operators. The widths of the proton beams are measured in a process known as a Van der Meer (vdM) scan [65]. In a vdM scan, the beams are separated by steps of a known distance. The collision rate is measured as a function of this separation, and

the width of a gaussian fit to the distributions yields the width of the beams in the direction of the separation.

The total integrated luminosities provided by the LHC and recorded by ATLAS in 2011 are shown in Figure 3.3. These events form the dataset analyzed in this thesis. By means of the beam-separation or *vdM* scans, as well as other techniques to measure the bunch charged product, the ATLAS Collaboration has determined that the uncertainty on its luminosity measurement is $\delta L = \pm 3.7\%$. For a complete description of the methods used and the systematic erros evaluated see reference [64].

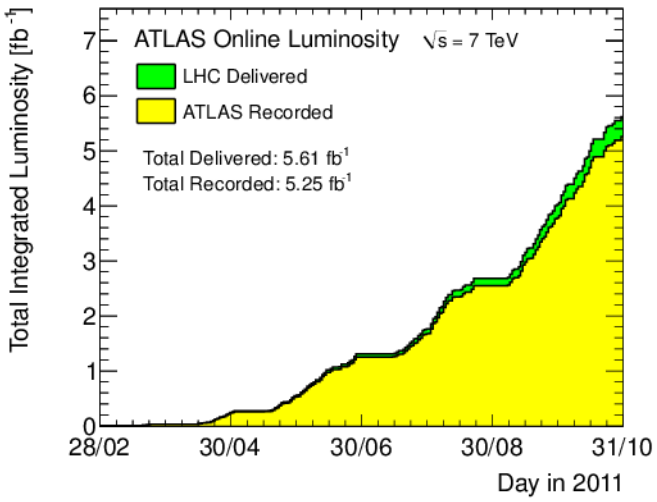

Figure 3.3: Total luminosity delivered by the LHC and recorded by ATLAS during the 2011 $\sqrt{s} = 7$ TeV proton-proton run.

Due to the cross-section for interaction and the large number of protons per bunch, the possibility to observe multiple *pp* interactions per bunch crossing increases proportionally. This phenomenon, referred to as "pile-up", can really occur in two distinct forms. The first form is the presence of multiple *pp* collisions (different from the interaction of interest) in the same bunch

crossing, referred to as "in-time" pile-up. The second form of pile-up takes place due to electronic integration times within the detector. Certain detector components are actually sensitive to multiple bunch crossings due to the long electronic signals generated in the response to energy depositions or charge collection. One or more *pp* collisions in a bunch-crossing different from that which produced the collision of interest can then affect the measurement. This form of pile-up is referred to as "out-of-time" pile-up and will become more important as the LHC bunch spacing gets closer to the nominal value, 25 ns.

The fraction of events with pile-up increased significatively since the data taking started. The experimental signature of this fact is obtain via the number of reconstructed primary vertices, or NPV. The effect of the event NPV is an important concern for the measurement of jet properties and will be discussed in the next chapters.

3.2 The ATLAS Detector

The ATLAS detector [66] is one of the two general purpose particle detectors built for probing *pp* collisions at the LHC. Inside the LHC, bunches of up to 10^{11} protons will collide 40 million times per second to provide 14 TeV proton-proton collisions at a nominal luminosity of $10^{34}\text{cm}^{-2}\text{s}^{-1}$; these high interaction rates and energies, as well as the requirements for high precision physics measurements set the standars for the design of the detector. At even 7 TeV center-of-mass energy, the LHC interactions result in high particle multiplicity, requiring fine detector granularity; and, particle production at forward rapidity, requiring large detector angular coverage.

To achieve these performance goals, a design consisting of multiple detector sub-systems with cylindrical symmetry around the incoming beams is used as shown in Fig. 3.4. Closest to the interaction point the inner tracking detector is placed, providing charged particle reconstruction. The magnet configuration comprises a thin superconducting solenoid surrounding the inner detector cavity, and three large superconducting toroids (one barrel and two end-caps) arranged with an eight-fold azimuthal symmetry around the calorimeters. This fundamental choice has driven the design and size (44 m in length and 25 m in height) of the rest of the detector. Outside the solenoid, a calorimeter system performs electron, photon, tau, and jet energy measurements. Finally, the calorimeter is surrounded by the muon spectrometer where an array of muon drift chambers perform muon identification and momentum measurements.

The ATLAS detector coordinate system is used to describe the position of particles as they traverse these subdetectors. It is a right-handed coordinate system, with z pointing along the beam direction, positive x pointing toward the center of the LHC ring, and positive y pointing up. The $x-y$ plane is referred to as the transverse plane, and the z direction as the longitudinal direction. The azimuthal angle ϕ is measured as usual around the beam axis, and the polar angle θ is the angle from the beam axis. The pseudorapidity is defined as $\eta = -\ln(\tan(\frac{\theta}{2}))$, regions of low η are referred to as "central", and regions of high η are referred to as "forward". The transverse momentum p_T is defined in the $x-y$ plane unless stated otherwise. The distance ΔR in the pseudorapidity-azimuthal angle space is defined as $\Delta R = \sqrt{\Delta\eta^2 + \Delta\phi^2}$.

To meet the extremely high demands that the LHC luminosity places on the speed with which ATLAS must record data, a dedicated trigger and data acquisition (TDAQ) system is used. The interaction rate at the design

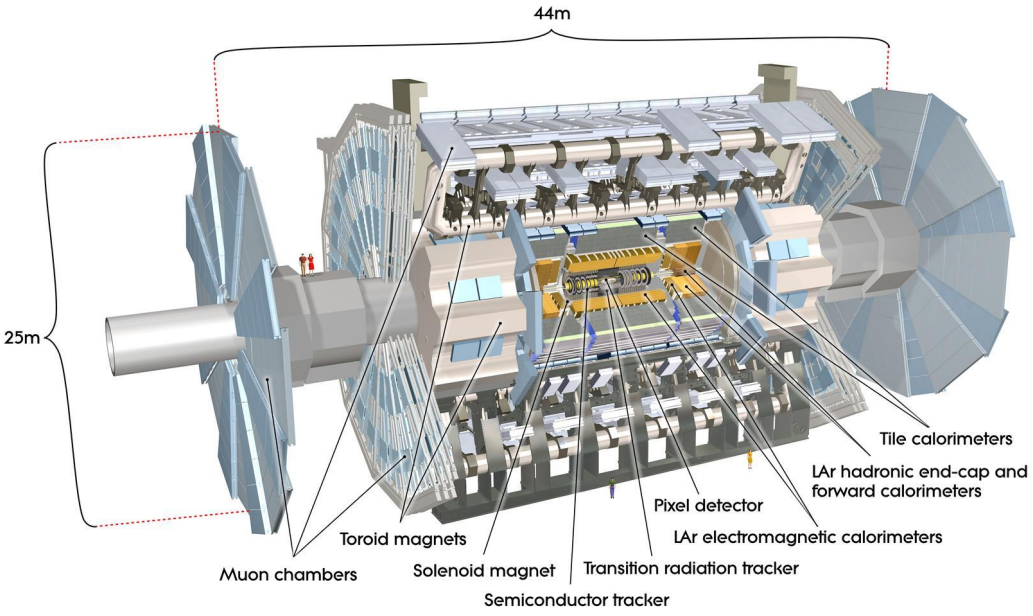

Figure 3.4: The ATLAS Detector.

luminosity is approximately 1 GHz, while the event data recording, based on technology and resource restrictions, is limited to about ∼200 Hz. This requires a high rejection of minimum-bias processes while maintaining maximum efficiency for the new physics. The Level-1 (L1) trigger system uses a subset of the total detector information to make a decision on whether or not to continue processing an event, reducing the data rate to approximately 75 kHz (limited by the bandwidth of the readout system, which is upgradeable to 100 kHz). The subsequent two levels, collectively known as the high-level trigger (HLT), are the Level-2 (L2) trigger and the Event Filter (EF). They provide the reduction to a final data-taking rate of approximately 200 Hz.

3.2.1 Inner Tracking System

The inner tracking system or Inner Detector (ID) is composed of three subdetectors: the pixel detector, the semiconductor tracker (SCT) and the transition radiation tracker (TRT). The goal of these three is to provide charged particle trajectory reconstruction and momentum measurements with an overall acceptance in pseudorapidity of $|\eta| < 2.5$ and full ϕ coverage.

The sensors which built this system register signals, referred to as "hits", in response to the passage of charged particles. The ID is inmersed in a 2 T magnetic field, generated by the central solenoid. The positions of the registered hits are combined to form tracks, with the radius of curvature of the tracks (caused by the presence of the magnetic field) providing a measurement of the particle's transverse momentum. The track reconstruction efficiency ranges from 78% at $p_T^{track} = 500$ MeV to more than 85% above 10 GeV, averaged across the full η coverage [67]. A transverse momentum resolution of $\sigma_{p_T}/p_T \lesssim 0.05$ [68] and a transverse impact parameter resolution of ∼20 μm

for tracks in the central η region [69] are primarily achieved through the use of high precision subsystems within the ID.

The pixel detector, SCT, and TRT sensors are arranged on concentric cylinders around the beam axis, known as barrel layers, and on disks perpindicular to the beam at either end of the barrel, known as end-caps. A more complete description of these systems is given below. The overall layout of the inner detector is shown in Fig. 3.5.

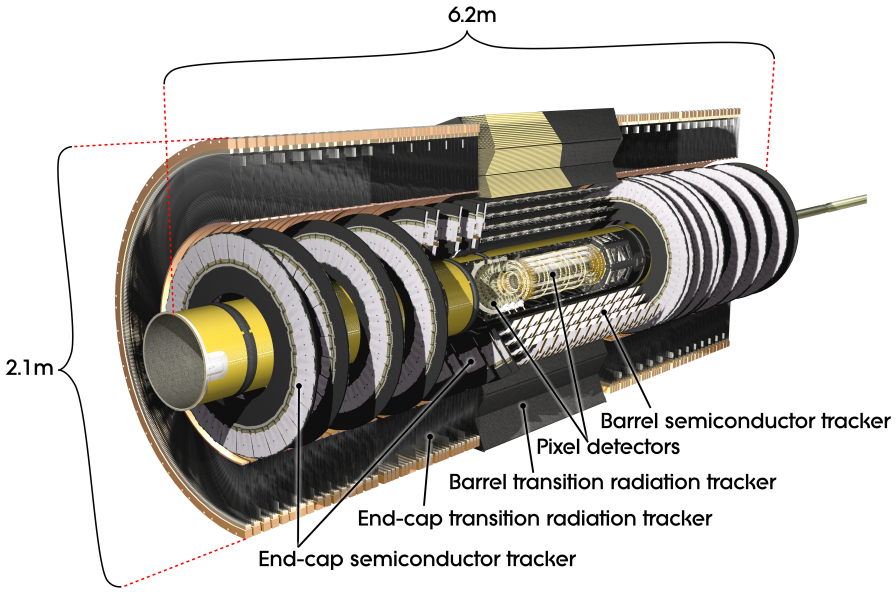

Figure 3.5: Layout of the ATLAS Inner Detector.

The Pixel detector

The pixel detector consists of three concentric barrel layers. The innermost one, the so called "b-layer" due to its role in identifying b-quarks initiated jets, is located at 5 cm from the interaction region. Three additional disks are located at each end-cap, producing typically three pixel position measure-

ments per charged particle track. Each layer or disk is instrumented with modules that form the basic unit of data acquisition, each with 47,232 pixels. All pixel sensors are identical and have a minimum pixel size in $r-\phi \times z$ of 50×400 μm^2. The intrinsic accuracies in the barrel are 10 μm in $r-\phi$ and 115 μm along z, or along r in the end-caps. The pixel detector has approximately 80.4 million readout channels, an order of magnitude more readout channels than the rest of ATLAS combined, and it extends to a total length of $z \sim \pm 650$ mm and radius of $r \sim 150$ mm, providing good reconstruction efficiency for tracks up to $|\eta| < 2.5$.

The SCT

The SCT consists of four barrel layers and nine end-cap layers surrounding the pixel detector, resulting in at least four hits along every charged particle track. The SCT barrel reaches to $z \sim \pm 750$ mm and $r \sim 515$ mm, while the end-cap covers out to $z \sim \pm 2720$ mm and $r = 560$ mm. There are 15,912 SCT module sensors, each 12.8 cm long and approximately 285 μm thick.

In the barrel region, these modules use small-angle (40 mrad) stereo strips to measure both coordinates, with one set of strips in each layer parallel to the beam direction, measuring the ϕ coordinate directly . In the end-cap region, the detectors have a set of strips running radially and a set of stereo strips at an angle of 40 mrad. The mean pitch of the strips is 80 μm. The intrinsic accuracies per module in the barrel are 17 μm in $r-\phi$ and 580 μm in z (or r in the end-caps). The total number of readout channels in the SCT is approximately 6.3 million. A hit is registered only if the pulse height in a channel exceeds a preset threshold ($\sim$ 1 fC). The charged measured in the strip is then recorded into a memory buffer that is only read out and used for tracking if a trigger is received signaling that the event should be considered

in more detail.

The TRT

The TRT surrounds the silicom detectors and is comprised of up to 76 layers of longitudinal straw tubes in the barrel, extending to $z \sim \pm 710$ mm and $r \sim$1060 mm, and 160 radial straw planes in each end-cap cylinders, reaching $z \sim \pm 2710$ mm and $r \sim$1000 mm.

The TRT sensors are thin drift tubes consisting of cathode metal straws filled with an ionizing gas mixture of xenon, oxygen, and CO_2, with an anode wire running down the center of the straw. The passage of a charged particle through the gas produces positive ions and free electrons, which travel to the cathode and anode, respectively, under the influence of an applied voltage of 1600 V. Comparing the time that the signals are received at the cathode and the anode gives a drift time measurement that can be used to calculate the impact parameter of the particle. This method gives no information on the position along the length of the straw.

To give the best resolution of particle trajectories as they bend in the solenoidal field, the straws lie along the beam direction in the barrel and radially in the end-caps. The straw diameter of 4 mm causes a maximum drift time of approximately 48 ns and an intrinsic accuracy of 130 μm along the radius of the straw.

In addition to directly detecting charged particles produced by the collision, the TRT also measures the transition radiation induced by the passage of these particles through polypropylene sheets placed between the drift tube straws. Transition radiation refers to the photons emitted by charged particles as they pass from one material into another with a different dielectric constant. These photons yield a much larger signal amplitude than the

charged particles, so separate thresholds in the electronics can be used to distinguish the two.

One of the most important tasks of the inner detector is to provide accurate collision vertex identification, exploiting the excellent position resolution and tracking efficiency. Vertices are reconstructed by matching inner detector tracks with $p_T > 150$ MeV back to a common origin.

3.2.2 The Calorimeter System

The purpose of the ATLAS calorimeter system is to measure the energy of electrons, photons, taus and jets, within the pseudorapidity region of $|\eta| < 4.9$ and with full ϕ symmetry and coverage around the beam axis. It also provides fast position and energy measurements to serve as trigger signals for these objects as well as the missing transverse energy.

The calorimeter detector consist of electromagnetic (EM) calorimeter and hadronic calorimeter components. The EM calorimeter provides fine granularity measurements of electrons and photons. Each calorimeter is segmented both transverse to the particle direction, to give position information, and along the particle direction, to chart the development of the particle shower. This permits detailed mapping of EM and hadronic showers in the calorimeter, allowing for studies of the internal structure of hadronic jets and partially giving rise to the high resolution measurements of their energy.

The EM and hadronic calorimeters are sampling calorimeters meaning that they utilize alternating layers of absorber material, composed of heavy atoms that interact with energetic particles and cause them to loose energy, and an active material, that produces a signal in response to the deposited energy.

The calorimeters closest to the beam-line are housed in three cryostats,

one barrel and two end-caps. The barrel cryostat contains the electromagnetic barrel calorimeter, and the two end-cap cryostats each contain an electromagnetic end-cap calorimeter (EMEC), a hadronic end-cap calorimeter (HEC), located behind the EMEC, and a forward calorimeter (FCal) to cover the region closest to the beam. These calorimeters use liquid argon as the active detector medium and need to be mantained at a constant temperature of ∼88K. Liquid argon (LAr) has been chosen for its intrinsic linear behaviour (production of ionization charge as a function of incident charge), its stability of response over time and its intrinsic radiation-hardness.

An illustration of all these components can be found in Fig. 3.6. Further specifications are given in the next sections.

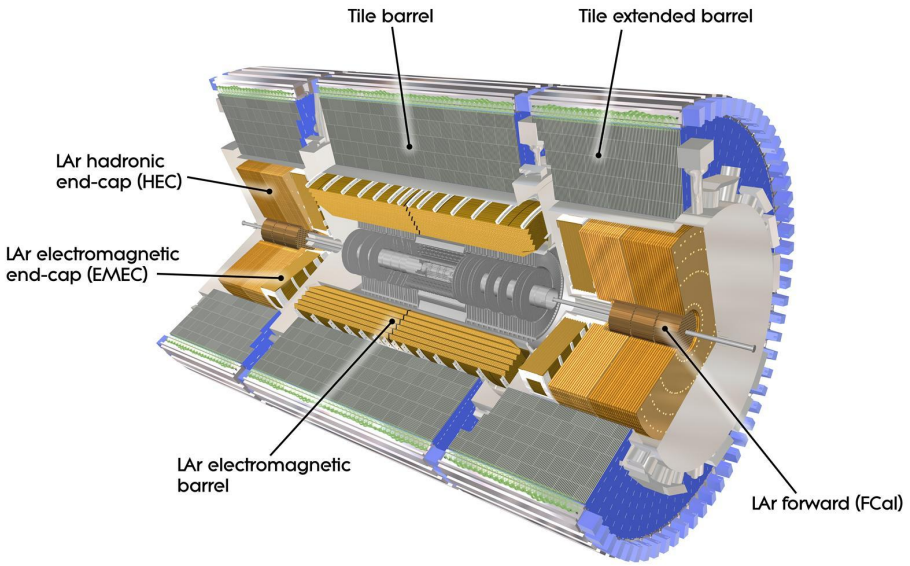

Figure 3.6: Layout of the ATLAS electromagnetic and hadronic calorimeter systems. The total length is ∼ 12 m, extending to a maximum radius of 4.25 m.

Liquid argon EM calorimeter

The EM calorimeter uses lead as the absorber and liquid Argon as the active material. A photon traversing the absorber will interact with the heavy nucleus via Compton scattering or the photo-electric effect, producing low-energy electrons or; pair production, producing electron/positron pairs. An electron or positron, in turn, can produce bremsstrahlung photons as it is deflected by the nuclei or produce more charged particles via ionization. Thus each incident photon, electron, or positron produces a shower of photons, electrons, and positrons that lose their energy through successive interactions in the absorber. The produced particles ionize the liquid argon, and the charge is collected by electrodes located in the liquid argon gap. These electrodes consist of three layers of copper sheets, the outer two kept at high-voltage potential and the inner one used to readout the signal.

To provide full coverage in ϕ without any cracks, an accordion-shaped absorber and electrode geometry is used, shown in Fig. 3.7. This design was chosen to ensure high azimuthal uniformity, a regular liquid argon ionization gap, and a constant sampling fraction within a given detector region. The figure highlights how this geometry is divided among rectangular cells in $\eta \times \phi$ space, the individual readout elements of varying size, finely segmented both laterally and longitudinally. Such fine segmentation $\Delta\eta \times \Delta\phi = 0.025 \times 0.025$ in the second layer of the EM barrel, for example permits a detailed mapping of the electromagnetic and hadronic showers.

The position resolution of the EM is driven by the readout geometry (rectangular cells). There are three layers of cells, segmented along the particle's direction of motion. The ϕ segmentation comes from grouping the accordion-shaped electrodes together into a common read out channel.

In the region $0 < |\eta| < 1.8$ the electromagnetic calorimeters are com-

plemented by a "presampler" detector, an instrumented argon layer, which provides a measurement of the energy lost in the solenoid and the outer wall of the barrel cryostat.

The EMEC uses the same accordion geometry as the EMB, whereas the granularity is typically slightly larger than in the barrel.

The signal readout chain for the LAr calorimeter (actually for all calorimeter systems) is divided into a fast analog readout for the trigger system and a slower digital readout used for more redefined trigger decisions and the offline reconstruction. However, regardless of the readout path, the signal is initiated within the active LAr medium. To minimize noise and increase speed the first level of readout is located on the detector (both for LAr and Tile calorimeter, see 3.2.2). The front-end electronics amplify and shape the signal. Shaping electronics induce a bipolar puse shape in the ionization signal. This shape is characterized by having both a positive and a negative component, which renders the integral of the signal exactly equal to zero.

The performance of the shaping electronics is critical for a correct energy calibration of the detector since the energy is primarily determined from the peak height of the pulse. In each calorimeter region, the overall pulse shape and duration are optimized to approximately cancel a constant injection of energy into the detector. The motivation for this approach is to effectively redefine the baseline of the energy measurement. In the high luminosity environment of the LHC, this reduces the sensitivity to the background from multiple pp interactions on average.

To translate these analog signals to digital signals that can be transmitted long distances to the next stage of the readout system, the pulse shape is measured over several 25 ns (nominal) time intervals, known as samples. The challenge of calorimeter calibration is to map these measured signals to the

energy deposited in the active detector medium, known as the visible energy. This calibration is established using test-beam measurements of electrons in the EMB [70, 71, 72, 73] and EMEC calorimeters [74, 75].

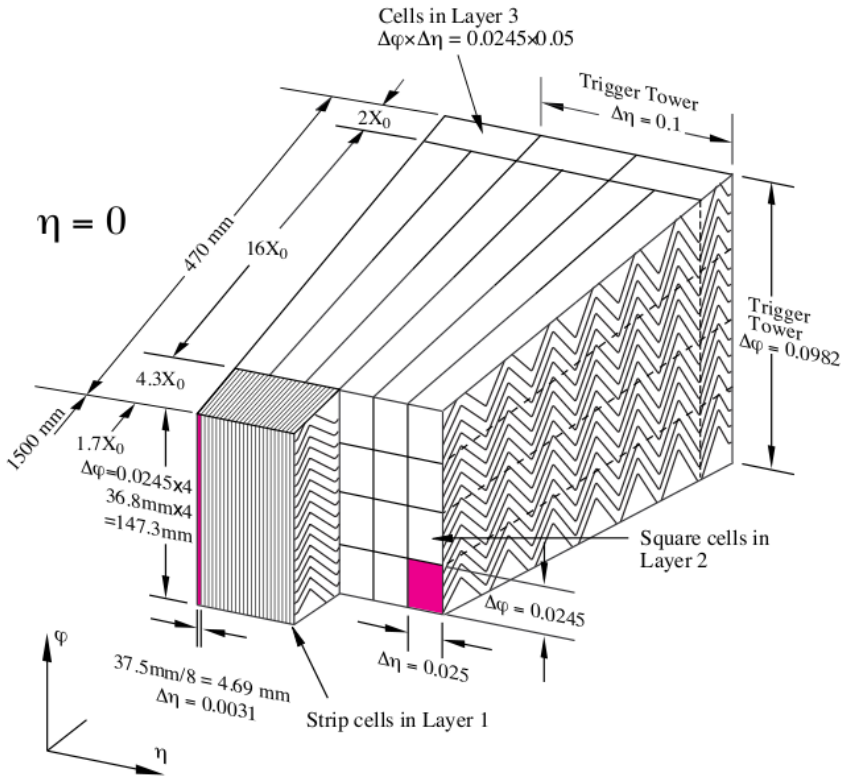

Figure 3.7: Cross section of the LAr barrel calorimeter where the different layers are visible. The granurality in η and ϕ of the cells of each of the three layers is also shown.

The hadronic calorimeter

Outside the EM calorimeter lies the system of hadronic calorimeters. The barrel portion, known as the Tile calorimeter, uses iron absorber slabs interspersed with scintillating tiles. The Tile calorimeter is most notable for its

depth of 7.4 radiation lengths (λ[1]). The hadronic end-cap and the forward calorimeter, which need to absorb the more energetic particles that are produced at large $|\eta|$, are made of copper and tungsten abosrbers, respectively, with liquid argon as the active material.

The tile calorimeter is composed of 3 mm thick scintillating tiles, arranged to lie parallel to the incoming particle direction, interleaved with 14 mm thick iron plates. It is divided into the barrel calorimeter, covering $|\eta| < 1.0$, and two extended barrel calorimeters, covering $0.8 < |\eta| < 1.7$. Each tile is read out by two wavelength-shifting fibers, which convert the scintillator signal to visible light. The readout fibers of several tiles are grouped to a single photomultiplier tube forming cells in $\eta \times \phi$ space. As in the EM calorimeter, these cells are segmented into three layers, the first two of size $\Delta\eta = 0.1$ and $\Delta\phi = 0.1$ and the last of size $\Delta\eta = 0.2$ and $\Delta\phi = 0.1$. Towers to provide information to the trigger system are formed from 0.1×0.1 grouping of all three layers.

The HEC uses the LAr active readout design due to the higher radiation tolerance required for the forward regions. Although housed in the same cryostat as the accordion geometry EMEC, the HEC implements a flat-plate design.

The forward calorimeter extends to cover the region $3.1 < |\eta| < 4.9$. Since it is the only calorimeter that covers this very forward region, it must provide both electromagnetic and hadronic measurements. In addition, the high particle fluxes in this region necessitate a finely granulated design. The FCal is approximately 10 interaction lenghts deep, and consists of three modules in

[1]To quantify the amount of material needed to capture a particles's energy, the unit of an interaction length, which is the distance over which a high energy charged particle loses $1 - \frac{1}{e} \sim 63\%$ of its energy , is commonly used.

each end-cap: the first, made of copper, is optimised for electromagnetic measurements, while the other two, made of tungsten, measure predominantly the energy of hadronic interactions.

The hadronic calorimeters are calibrated using muons in test-beam experiments and those muons produced by cosmic-rays in situ. The invariant mass of the Z boson in $Z \to ee$ events measured in-situ in the 2010 pp collisions is used to adjust the calibration derived from test-beams and cosmic-muons.

3.2.3 The Muon System

The muon system gives the ATLAS detector its overall shape and imposing nature, as depicted in Fig. 3.8. Muons have much smaller cross section to interact in material than electrons and hadrons, for this, they do not deposit all their energy in the calorimeters. The muon spectrometer is designed to detect muons within $|\eta| < 2.7$. Because many new physics signatures involve high-momentum muons, the system is also required to provide trigger signals based on the particle p_T for $|\eta| < 2.4$.

To provide a momentum measurement, the muons trajectorires are bent in a toroidal magnetic field. This field is provided by one large barrel toroid and two large end-cap toroids, each toroid consisting of eight coils arranged symmetrically around the beam axis. The toroid system produces a magnetic field that is typically oriented in the ϕ direction and that is measured with over 1800 Hall sensors placed throught the magnets. Under the influence of this field, muons are deflected in the $r-z$ plane and the transverse momentum of the muons is given then by the radius of curvature of the tracks. Since the highly-energetic muons bend very little even in this high magnetic field, the muon system is the largest of all the ATLAS sub-detectors, covering a radius from $\sim$4.5 m to $\sim$12.5 m.

Four primary subsystems comprise the integrated muon spectrometer: monitored drift tubes (MDT), cathode strip chambers (CSC, which are multiwired proportional chambers with cathodes segmented into strips), resistive plate chambers (RPC) and thing gap chambers (TGC). The MDT and CSC subsytems are primarily designed for precision measurements of muon tracks, width the MDT system providing coverage for the more central region ($|\eta| < 2.7$, with full coverage only in $|\eta| < 2.0$), whereas the CSC is located in the more forward region ($2.0 < |\eta| < 2.7$) due to its ability to cope with higher background rates. The RPC and TGC muon subsystems are desinged to provide fast, robust readout for use in the trigger and data acquisition system. A detailed description of the subsystems can be found elsewhere [66].

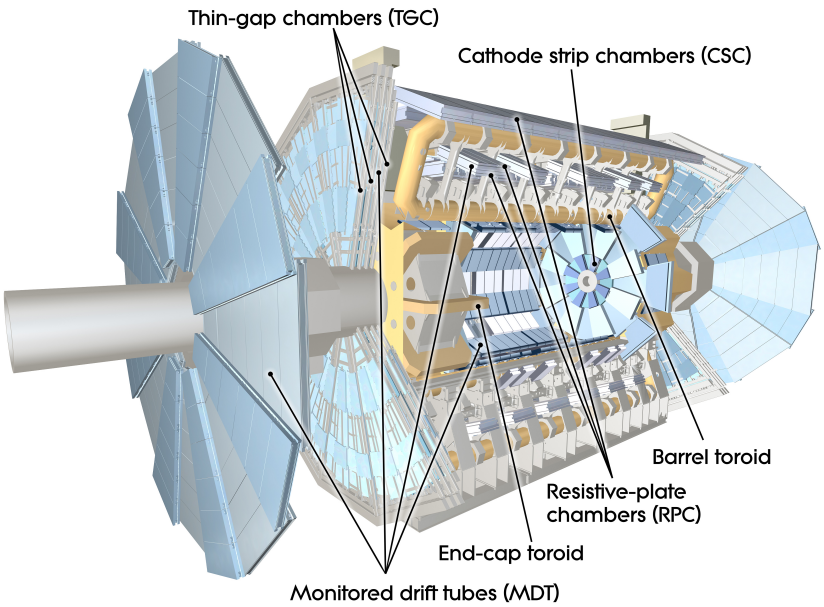

Figure 3.8: The Muon Chamber.

3.2.4 Forward Detectors

Three smaller detector systems cover the ATLAS forward region. The main function of the first two systems is to determine the luminosity delivered to ATLAS. At ± 17 m from the interaction point lies LUCID (LUminosity measurement using Cerenkov Integrating Detector). The principle of LUCID is to detect inelastic $p - p$ scattering in the forward region, exploiting the fact that the number of particles detectted is proportional to the total, both primary and pile-up, interactions in a bunch-crossing. LUCID thus provides a relative luminosity measurement, in which the detected number of particles must be translated to the total number of proton-proton interactions via calibration runs. The second detector is ALFA (Absolute Luminosity For ATLAS). Located at ± 240 m, it consists of scintillating fibre trackers located inside Roman pots which are designed to approach as close as 1 mm to the beam. The third system is Zero-Degree Calorimeter (ZDC), which place a role in determining the centrality of heavy-collisions.

3.2.5 Trigger and Data Adquisition System

At design luminoisty, the LHC will deliver approximately 40 million collision events every second. With an average ATLAS event size of $\sim$1.5 MB, this is far more information than can be saved into the finite data storage resources available. The goal of the trigger system is to move interesting physics events to permanent storage, while rejecting the vast majority of other events.

The online selection is done in three stages: the Level 1 (L1), Level 2 (L2) and Event Filter (EF) stages. Each trigger level refines the decisions made at the previous level and, where necessary, applies additional selection criteria. The data acquisition (DAQ) system receives and buffers the event data from

the detector-specific readout electronics, at the L1 trigger accept rate, over 1600 point-to-point readout links. The L1 trigger uses a limited amount of the total detector information (only from the calorimeter and the muon systems) using only simple hardware based algorithms to make a decision in less than 2.5 μs, reducing the rate to about 75 kHz. The L2 and EF, collectively referred to as the High Level Trigger (HLT), are based on fast software algorithms running on large farms of commercial processors. The L2 is the first stage of the ATLAS DAQ system that has access to data from the ID and is capable of doing partial reconstruction of events up to the L1 accept rate. L2 trigger is designed to reduce the rate to approximately 3.5 kHz, with an event processing time of about 40 ms, averaged over all events. The EF reduces the rate to roughtly 200 Hz. Its selections are implemented using offline analysis procedures within an average event processing time of the order of 4 s.

The L1 trigger is desinged to accept high-p_T muons, electrons, photons, jets, and taus, as well as events with large missing transverse energy or sum energy. It uses signals from the TGCs and RPCs from muon triggers and reduced granularity calorimeter information for electron, photon, jet, tau and total energy triggers. The calorimeter trigger system, which maintains a fast readout independent from the remainder of the calorimeter is known as the Level-1 Calorimeter. At this level coarse calorimeter information is available in the form of jet elements with $\Delta\eta \times \Delta\phi = 0.2 \times 0.2$ for $|\eta| < 3.2$. Jets are reconstructed using a square sliding window algorithm. In addition to coarse jets, the total transverse energy is also measured at the L1. The region of the detector corresponding to the location where the L1 thresholds were passed − so called "region of interest" (RoI) − are then delivered to the L2 sofware algorithms.

The L2 trigger applies additional energy thresholds and multiplicity requirements using the RoI around triggered L1 objects. For example, the L2 jet trigger retrieves the data from cells surrounding the L1 RoI and constructs jets using a simplified cone jet algorithm.

The next step and last stage in the trigger chain is the EF, which receives events that have been selected by the L2 triggers and processes the entire event with the full detector granularity instead of only a restricted region.

The monitoring infrastructure of the HLT supports the real-time accumulation of histograms, and their aggregation across the farm, so that parameters can be extracted from cumulative distributions that contain events from all processor nodes. Beam parameters determined from those live histograms are transmitted online to the LHC and are also available to fed back into the HLT itself for use by its own trigger algorithms that depend on the precise knowledge of the luminous region.

3.2.6 ATLAS performance and data quality

The ATLAS detector has been operational for a number of years collecting large amounts of data. Before the start-up of the LHC, the detector measured muons from cosmic rays; which were used to test, understand, and align the detector. In 2010 and 2011 ATLAS recorded over 5.2fb^{-1} of collision data. Fig. 3.3 presents the luminosity delivered by the LHC in 2011 as well as the recorded luminosity by the detector, showing a good performance of the ATLAS Experiment. The fraction of time that each subdetector system was operational during data-taking is shown in Table 3.2.

The Data Quality (DQ) selection within ATLAS is based on the inspection of a standard set of distributions that lead to a data quality assesment which is encoded in so-called DQ flags. DQ flags are issued for each de-

tector, usually segmented in subdetectors like barrel, end-caps and forward. DQ flags are also issued for trigger slices and for each physics object reconstruction. In this way, the state of the ATLAS detector from hardware to physics object reconstruction is expressed through DQ flags, which are saved per luminosity block. A luminosity block is a time interval of typically two minutes.

The DQ information is used in analyses through dedicated lists of good runs/luminosity blocks. Good run lists are formed by DQ selection criteria in addition to other criteria, such as run range, magnetic field configuration and beam energy. A complete list of valid physics runs and luminosity blocks is used in each analysis.

Detector component	operational
Inner Detector	
Pixel	≈96.4%
SCT	≈99.2%
TRT	≈97.5%
Calorimeter	
EM	≈99.8%
Tile	≈96.2%
Hadronic, end-cap	≈99.6%
Forward calorimeter	≈99.8%
Muon Spectrometer	
MDT	≈99.7%
CSC	≈97.7%
RPC	≈97.0%
TGC	≈97.9%

Table 3.2: The approximate fraction of time that each individual subdetector system was operational during data-taking.

Chapter 4

Event reconstruction and b-Tagging

The event reconstruction packages, which in ATLAS are implemented in the software framework ATHENA [76], process the events, starting from the raw data obtained from the various sub-detectors (energy deposits and hits), through different stages to finally interpreting them as a set of charged tracks, electrons, photons, jets, muons and, in general, of possible kinds of final state objects with related four momenta. In this chapter the reconstruction of these objects is briefly described together with the algorithms for the identification of b-quark jets. These algorithms are mainly based on the reconstruction of the primary interaction vertex, on the reconstruction of charged particles in the Inner Detector and on the reconstruction of jets in the calorimeter.

4.1 Jet reconstruction and calibration

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

in the calorimeters. The ATLAS performance group, addressing the calibration of jets and the missing transverse energy (Jet/Etmiss), has made the decision to adopt the anti-k_t algorithm (Chapter 2) as its default jet algorithm. This choice was driven by multiple requirements ranging from physics performance to those intimately involved with the computing, trigger and detector: the anti-k_t algorithm is fast and its memory consumption is low, it is well adapted to algorithms used in the trigger, and it has the best jet reconstruction efficiency at low p_T. Moreover, this algorithm exhibits the smallest fluctuations of the jet area showing good stability under pile-up [77].

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

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

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

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

The results presented in this thesis use jets built from noise-suppressed topological clusters, also known as "topo-clusters" [78]. Topological clusters are groups of calorimeter cells that are designed to follow the shower development taking advantage of the fine segmentation of the ATLAS calorimeters. The topological cluster formation starts from a seed cell with $|E_{cell}| > 4\sigma$ above the noise. In a second step, neighbor cells that have an energy at least 2σ above their mean noise are added to the cluster. Finally, all nearest-neighbor cells surrounding the clustered cells are added to the cluster, regardless of the signal-to-noise ratio[1]. The position of the cluster is assigned as the energy-weighted centroid of all constituent cells (the weight used is the absolute cell energy).

Jet calibration

The baseline EM energy scale of the calorimeters is the result of the calibration of the electronics signal to the energy deposited in the calorimeter by electromagnetic showers (see Chapter 3). The purpose of the jet energy calibration, or jet energy scale (JES), is to correct the measured EM scale energy

[1]Noise-supressed towers also make use of the topological clusters algorithm [78] to select cells, i.e. only calorimeter cells that are included in topo-clusters are used.

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

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

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

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

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

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

$$\mathcal{R} = E_{reco}/E_{truth} \tag{4.1}$$

may be defined at any energy scale. In Equation 4.1, E_{truth} is the energy of the closest isolated truth jet, within $\Delta R < 0.3$[2]. The isolation requirement is applied in order to factorize the effects due to close-by jets from those due to purely detector effects such as dead material and non-compensation. The isolation criterion requires that no other jet with a $p_T > 7$ GeV be within $\Delta R < 2.5R$, where R is the distance parameter of the jet algorithm.

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

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

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

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

$$E^{EM+JES} = \mathcal{C} \cdot E^{EM}. \tag{4.2}$$

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

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

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

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

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

Jet energy scale uncertainties for the EM+JES scheme

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

justing the JES calibration methods itself. The final JES uncertainty in the central region, $|\eta| < 0.8$, is determined from the maximum deviation in response observed with respect to the response in the nominal sample. For the more forward region, the so called "η-intercalibration" contribution is estimated. This is a procedure that uses direct di-jet balance measurements in two-jet events to measure the relative energy scale of jets in the more forward regions compared to jets in a reference region. The technique exploits the fact that these jets are expected to have equal p_T due to transverse momentum conservation. Figure 4.1 shows the final fractional jet energy scale uncertainty and its individual contributions as a function of p_T for a central η region. The JES uncertainty for anti-k_t jets with $R = 0.4$ is between $\approx$4% (8%, 14%) at low jet p_T and $\approx$2.5%-3% (2.5%-3.5%, 5%) for jets with $p_T > 60$ GeV in the central (endcap, forward) region.

In addition to the tests above, *in situ* tests of the JES using direct γ-jet balance, multi-jet balance, and track-jets indicate that the uncertainties in Fig. 4.1 reflect accurately the true uncertainties in the JES.

In the case of jets induced by bottom quarks (b-jets), the calorimeter response uncertainties are also evaluated using single hadron response measurements *in situ* and in test beams [79]. For jets within $|\eta| < 0.8$ and $20 \leq p_T < 250$ GeV the expected difference in the calorimeter response uncertainty of identified b-jets with respect to the one of inclusive jets is less than 0.5%. It is assumed that this uncertainty extends up to $|\eta| < 2.5$.

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

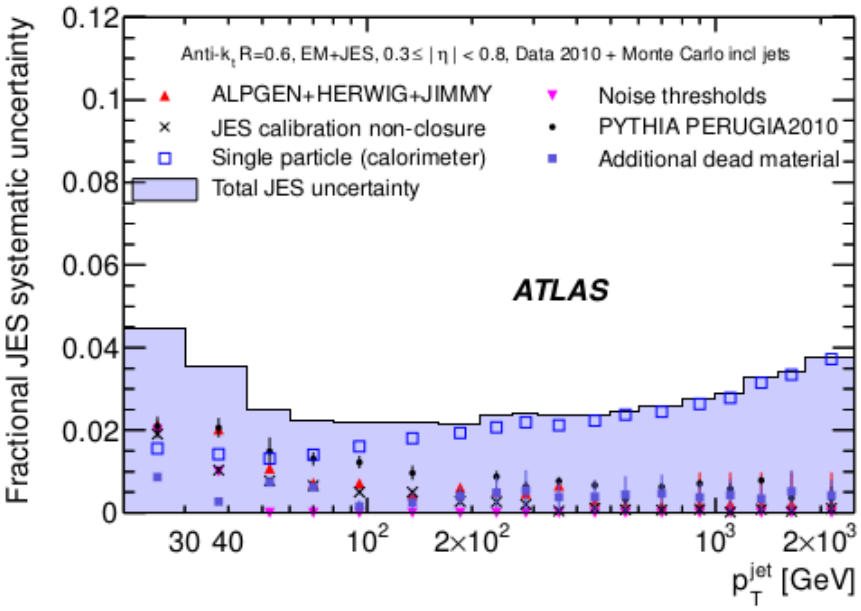

Figure 4.1: Fractional jet energy scale uncertainty as a function of jet p_T for jets in the pseudorapidity region 0.3< $|\eta|$ <0.8 in the calorimeter barrel. The total uncertainty is shown as the solid tight blue area. The individual sources are also shown.

a detailed study of the b-quark fragmentation function in comparison with OPAL [80] and SLD [81] data. To assess the impact of the b-fragmentation, the nominal parameters of the PYTHIA fragmentation function are replaced by the values from a tune using the Professor framework [82]. In addition, the nominal fragmentation function is replaced by the modified Bowler-Lund fragmentation function [83]. The b-jet response uncertainty is evaluated from the ratio between the response of b-jets in the varied Monte Carlo samples to the nominal PYTHIA. The response variations are well within 2%.

The b-jet JES uncertainty is obtained adding the calorimeter response uncertainty and the uncertainties from the systematic Monte Carlo variations in quadrature. The resulting additional JES uncertainty for b-jets is shown in Fig. 4.2. It is about 2% up to $p_T \approx 100$ GeV and below 1% for higher p_T. To obtain the overall b-jet uncertainty this uncertainty is added in quadrature to the JES uncertainty for inclusive jets.

4.2 Reconstruction of charged particle tracks

The Inner Detector layout and the characteristics of its main sub-detectors were presented in Section 3.2.1 of Chapter 3. The algorithm used for track reconstruction is based on a modular software framework, which is described in more detail in Ref. [84]. The main steps are the following:

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

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

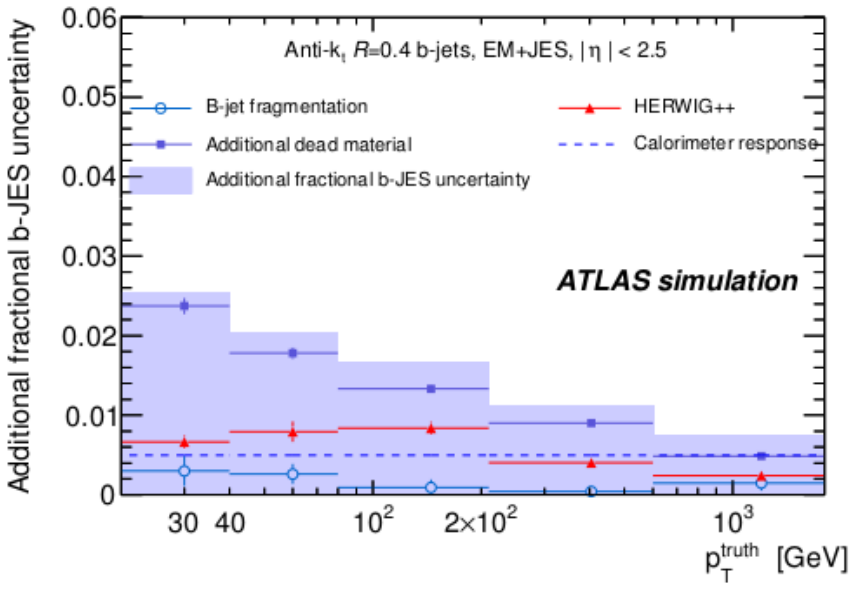

Figure 4.2: Additional fractional b-jet JES uncertainty as a function of the truth jet transverse momentum for anti-k_t jets with $R = 0.4$ calibrated with the EM+JES scheme for $|\eta| < 2.5$. Shown are systematic Monte Carlo variations using different modelling of the b-quark fragmentation and physics effects as well as variations in the detector geometry and the uncertainty in the calorimeter response to b-jets as evaluated from single hadron response measurements. Uncertainties in the individual points are statistical only.

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

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

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

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

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

The track reconstruction efficiency is defined as the fraction of primary particles with $p_T > 400$ MeV and $|\eta| < 2.5$ matched to a reconstructed track. The reconstruction efficiency for primary tracks with transverse momentum above 1 GeV and central η is above 80%, going down to values below 70% for tracks at the edge of the Inner Detector acceptance [67]. The dense environment of a jet decreases the track reconstruction efficiency and increases the fake rate. This is caused by the ocurrence of shared hits between different tracks, which makes the pattern recognition and track fitting tasks more difficult.

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

$$ {p_T}^{MC} \times (1/{p_T}^{MC} - 1/{p_T}^{reco}) = 1 - \frac{{p_T}^{MC}}{{p_T}^{reco}} \tag{4.3} $$

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

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

4.3 Vertex reconstruction

Primary vertices are reconstructed using an iterative vertex finding algorithm [85]. In a first step, a dedicated vertex finding algorithm associates tracks to vertex candidates. Vertex seeds are obtained by looking for the global maximum in the distribution of the z coordinates of the tracks. In a second stage, an iterative χ^2 fit is made using the seed and nearby tracks. Each track carries a weight which is a measure of its compatibility with the fitted vertex depending on the χ^2 of the fit. Tracks displaced by more than 7σ from the vertex are used to seed a new vertex and the procedure is repeated until no additional verteces can be found. The parameters of the beam spot are used both during the finding to preselect compatible tracks and during the fitting step to constrain the vertex fit.

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

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

4.4 b-jet Tagging

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

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

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

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

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

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

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

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

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

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

Association of tracks to jets

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

$$\Delta R(jet, track) < \Delta R_{cut} \tag{4.4}$$

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

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

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

4.4.1 b-tagging algorithms

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

Impact parameter based b-tagging algorithms

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

both in the transverse and longitudinal plane. With

$$IP_{r\phi} = d_0 \text{ and } IP_z = z_0 \sin\theta, \tag{4.5}$$

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

$$IP_{r\phi}/\sigma(IP_{r\phi}) \text{ and } IP_z/\sigma(IP_z). \tag{4.6}$$

On the basis that the decay point of the b-hadron must lie along its flight path, and in order to increase the discriminating power of the impact parameter significance, a lifetime sign is assigned to these variables (replacing the sign of the geometrical definition of the impact parameter). The sign is positive if the track extrapolation crosses the jet direction in front of the primary vertex (i.e. is more compatible with having its origin in a secondary decay vertex in the direction of flight expected for the b-hadron) or negative if the track is more likely to intersect the flight axis behind the primary vertex, opposite to the jet direction. Both cases are illustrated in Fig. 4.3.

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

$$\text{sign}_{3D} = \text{sign}([\vec{p}_{trk} \times \vec{p}_{jet}] \cdot [\vec{p}_{trk} \times \vec{\Delta r}_{IP}]). \tag{4.7}$$

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

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

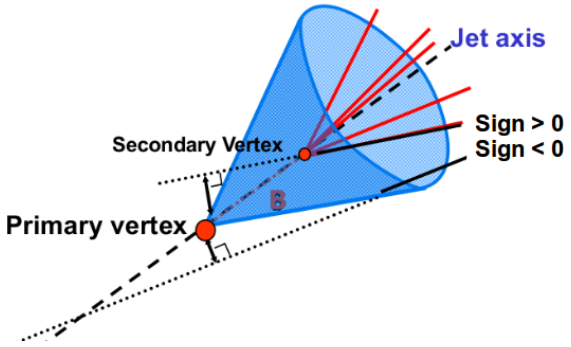

Figure 4.3: Lifetime sign of tracks. A positive and a negative lifetime signed track is shown.

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

$$\text{sign}_{r\phi} = \text{sign}(\sin(\phi_{jet} - \phi_{trk}) \cdot d_{0,trk}); \text{ and } \text{sign}_z = \text{sign}((\eta_{jet} - \eta_{trk}) \cdot z_{0,trk}). \tag{4.8}$$

Distributions of the signed transverse impact parameter and signed transverse impact parameter significance for light, c-, and b-jets, are shown in Fig. 4.4 for experimental data and for simulation; the sign is defined by "$\text{sign}_{r\phi}$". Tracks from the fragmentation in light-jets tend to have a signed impact parameter distribution which is symmetric around 0, since they have no correlation with the jet direction. Tracks from b- and c-hadron decays, as expected, have an asymmetric distribution, with the most significant contribution at positive significances; however a negative tail extending beyond the pure fragmentation contribution is also seen, corresponding to resolution effects and to an eventual mismatch between the b-jet and the b-hadron directions.

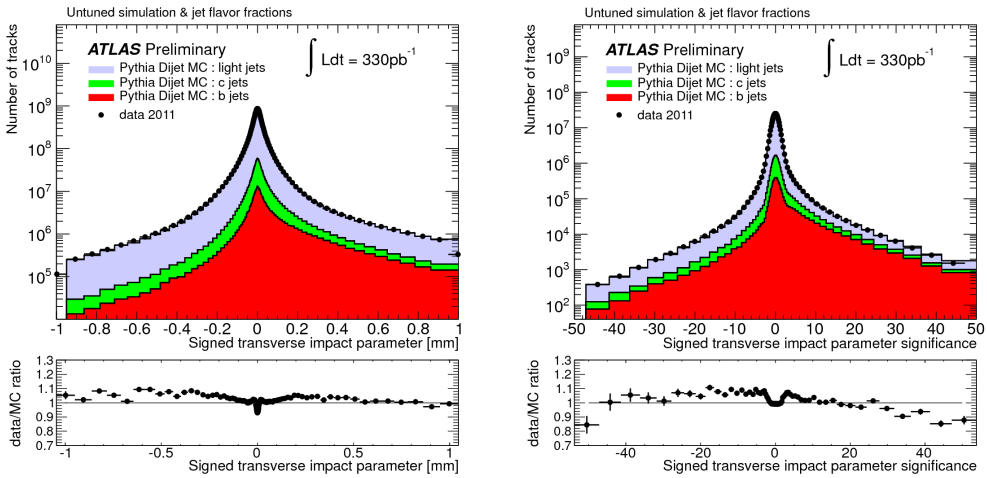

Figure 4.4: Distribution of the signed transverse impact parameter (left) and signed transverse impact parameter significance with respect to primary vertex for tracks associated to jets, for experimental data (solid black points) and for simulated data (filled histograms for the various flavors). The ratio data over simulation is shown at the bottom of the plot.

The significance, which gives more weight to tracks measured precisely, is the main ingredient of the tagging algorithms based on impact parameters. Now, the impact parameter significance of all N tracks associated to the jet to tag need to be combined into a single discriminating variable. It is assumed that tracks are uncorrelated, so their probability density functions (PDF), defined based on the transverse and/or longitudinal impact parameter significance distributions for the different hipothesis, are uniquely defined as a function of the jet flavour. Using a likelihood function defined according to the product of these PDFs, under the hypothesis of uncorrelated tracks, the following likelihood ratio provides the optimal separation, according to Neyman-Person lemma [86]:

$$\mathrm{LR}(IP_1, IP_2, ..., IP_N) = \frac{\prod_{i=1}^{N} \mathrm{PDF}_b(IP_i)}{\prod_{i=1}^{N} \mathrm{PDF}_l(IP_i)} \tag{4.9}$$

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

$$\mathrm{weight}(IP_1, IP_2, ..., IP_N) = \log(\mathrm{LR}(IP_1, IP_2, ..., IP_N)) \tag{4.10}$$

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

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

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

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

smooth Monte Carlo PDFs for both b-jet and light jet hypotheses [87]. Prior to the use of these advanced tagger, a simpler tagging algorithm, the **JetProb**, combining the impact parameter significances of all tracks associated to the jet was devised to be used for early data, being extensively used during 2010 [88].

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

Secondary vertex based b-tagging algorithms

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

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

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

The inclusive fit of a single displaced vertex in b-jets is based on the VKalVrt [89] reconstruction package. The main idea of the algorithm is to maximise the b/c-hadron vertex detection efficiency, keeping at the same time the probability to find a vertex inside a light jet low.

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

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

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

The three-dimensional decay length significance alone, signed with respect to the jet direction can be used as a discriminating variable between b-jets and light jets: this is the principle of the early data **SV0** tagger, extensively used as well with the 2010 and 2011 data [90].

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

Algorithm combinations and performance

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

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

Figure 4.5 compares the performance for the various ATLAS b-tagging algorithms described in a simulated sample of $t\bar{t}$ events. It can be seen that by combining the vertexing techniques and the impact parameter information, the IP3D+SV1 and IP3D+JetFitter algorithms can reach very high tagging efficiencies.

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

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

quarks at this stage.

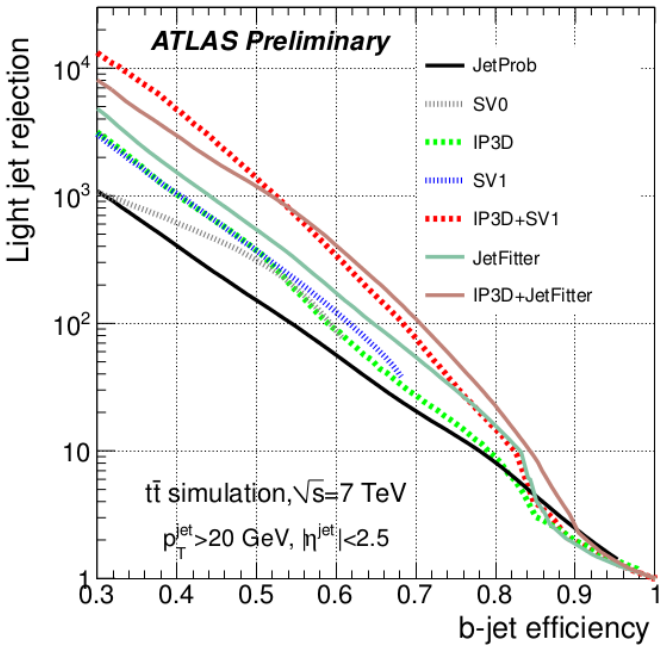

Figure 4.5: Light-jet rejection as a function of the b-jet tagging efficiency for the early tagging algorithms (JetProb and SV0) and for the high-performance algorithms, based on simulated $t\bar{t}$ events.

The MV1 tagging algorithm

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

4.4.2 b-tagging calibration

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

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

$$\kappa_{\epsilon_b}^{data/sim} = \frac{\epsilon_b^{data}}{\epsilon_b^{sim}}, \quad \kappa_{\epsilon_l}^{data/sim} = \frac{\epsilon_l^{data}}{\epsilon_l^{sim}}, \tag{4.11}$$

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

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

The main b-tagging efficiency calibration methods, the so called *system8* and $p_{T_{rel}}$ methods, are described in detail in ref [91].These measurements are based on a sample of jets with muons inside, where the muons are serving as a reference b-tagging algorithm to obtain a b-jet sample on which the calibrations can be performed. At the LHC, the large $t\bar{t}$ production cross section of

$\sigma_{t\bar{t}} = 177 \pm 3(\text{stat.}) \, ^{+8}_{-7}(\text{syst.}) \pm 7(\text{lum.})$ pb [92] offers an alternative source of events enriched in b-jets. Calibrations using samples of $t\bar{t}$ events have been obtained for SV0, IP3D+SV1, JetFitter and MV1 b-tagging algorithms [93]. All these algorithms provide an output weight w, discriminating between b-jets and non-b-jets. Lower values of w are assigned to c- and light-flavour jets, whereas the purity of b-jets increases with w. For each b-tagging algorithm a set of operating points, corresponding to a certain w cut value, are defined and calibrated:

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

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

The mistag rate is measured in data using two methods, both based on an inclusive sample of jets, referred to as the *negativetag* and *sv0mass* methods [94]. The first method uses the invariant mass spectrum of tracks associated with reconstructed secondary vertices to separate light and heavy-flavour jets, and the other is based on the rate at which secondary vertices with negative decay length, or tracks with negative impact parameter, are present in the data.

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

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

Chapter 5

Single and double b-hadron jet properties

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 6.

5.1 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 3 to 20 (see Fig.3.2). 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 3.2.5), 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.

5.2 Monte Carlo sample

The Monte Carlo samples employed were produced with the event generators discussed in Section 2.3. Samples of dijet events from proton-proton collision processes were simulated with PYTHIA version 6.423 [25], 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 [30].

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 (raw) generated distributions are shown in Fig. 5.1. The p_T spectrum obtained after weighting is displayed in Fig. 5.2, showing that a smooth distribution is obtained, with an even population of events over the whole p_T range.

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.

5.3 Event and jet selection

The data sample in the analysis is selected online using a set of single jet triggers as described in Section 5.1. 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 [95].

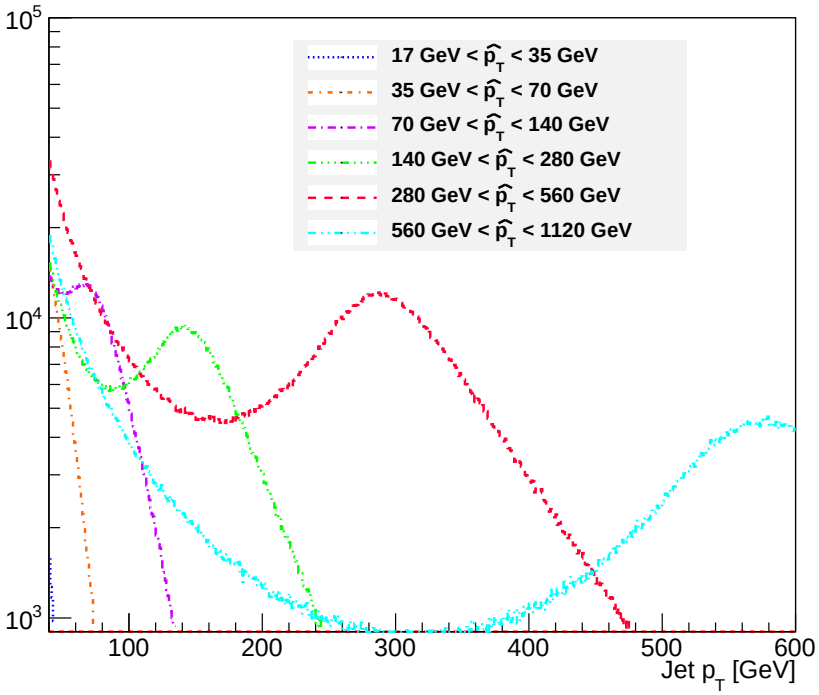

Figure 5.1: Calibrated jet p_T distribution for anti-k_t jets in a dijet Monte Carlo sample composed of different sub-samples generated with increasing thresholds of the hard-scattering partonic transverse momentum, $\hat{p}_T$.

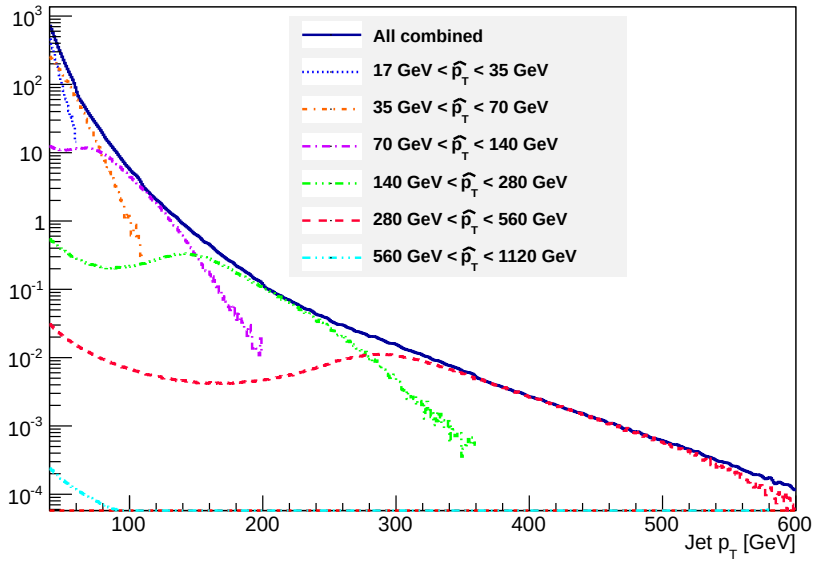

Figure 5.2: Calibrated jet p_T distribution for anti-k_t jets in a dijet Monte Carlo sample composed of different sub-samples generated with increasing thresholds of the hard-scattering partonic transverse momentum, $\hat{p}_T$. In order to obtain the falling p_T spectrum observed in data, the different samples were weighted by their respective production cross sections.

The offline event selection comprises an additional set of cuts on the 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 2.4.1), with a distance parameter $R = 0.4$, using calorimeter topological clusters as input (Section 4.1). All jets were calibrated using the EM+JES scheme (Section 4.1). A high cut on the minimum jet p_T is implemented to select jets in the region where the triggers used in the analysis are most efficient. Jets are required to have a minimum p_T of 40 GeV. Jets with transverse momentum above this threshold were also required to be in a region with full tracking coverage, $|\eta_{jet}| < 2.1$. Although the Pixel and SCT detectors cover up to $|\eta| < 2.5$, a lower pseudorapidity cut is used in order to account for the size of the calorimeter jets, $R = 0.4$. Jets passing this selection were classified in eight p_T bins chosen such as to match the jet trigger 99% efficiency thresholds (in GeV): 40, 60, 80, 110, 150, 200, 270, 360. An event is used if it satisfies the highest threshold trigger that is 99% efficient for the p_T bin that corresponds to the p_T of its leading jet. The upper limit of our highest p_T bin was set to 480 GeV; beyond this energy the b-tagging efficiency becomes very poor.

Several quality criteria are applied to jets to eliminate "fakes" that are caused by noise bursts in the calorimeters and energy depositions belonging

to a previous bunch crossing. A detailed description of these quality cuts can be found in reference [96].

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:

$$\text{single } b\text{-jets: } \Delta R(j, B_i) < 0.4 \;\&\; \Delta R(j, B_j) > 0.4 \;\; for\; i \neq j \tag{5.1}$$

$$\text{merged } b\text{-jets: } \Delta R(j, B_i) < 0.4 \;\&\; \Delta R(j, B_j) < 0.4 \;\; for\; i \neq j \tag{5.2}$$

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 5.1 and 5.2 change accordingly.

5.3.1 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 4.4). Cuts on ${p_T}^{trk} > 1.0$ GeV and the χ^2 of the track fit, $\chi^2/ndf < 3$, are applied. The effect of a lower cut on the track transverse momentum, ${p_T}^{trk} > 0.5$ GeV, is discussed in the next section. In addition, tracks are required to have a total of at least seven precision hits (pixel or micro-strip) in order to guarantee at least 3 z-measurements. As cutting on impact parameter (IP) might be detrimental for b-jets, where large IP values are expected, relaxed cuts were used, $|d_0| <$ 2 mm, and $|z_0 \sin\theta| < 2$ mm, with θ being the polar angle measured with respect to the beam axis. The track quality cuts are summarized in table 5.1.

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 5.1: 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.

5.4 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 5.3.1, and contained within a cone of radius $R = 0.4$ around the jet axis. Figure 5.3 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

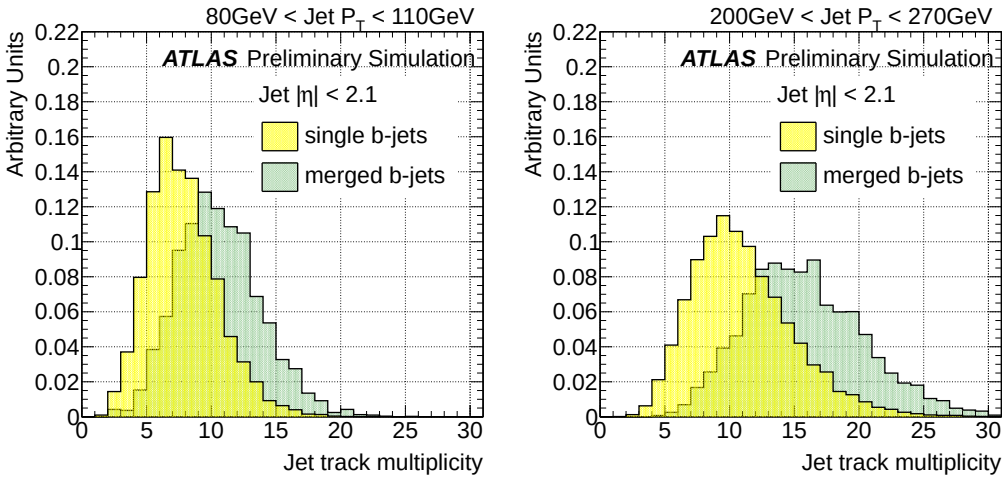

Figure 5.3: Distribution of the track multiplicity for single and merged b-jets from 80 GeV to 110 GeV (left) and 200 GeV to 270 GeV (right).

particles and using only charged particles coming from the b-hadron decay ("b-tracks") are illustrated in Fig. 5.4. 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. 5.5. 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.

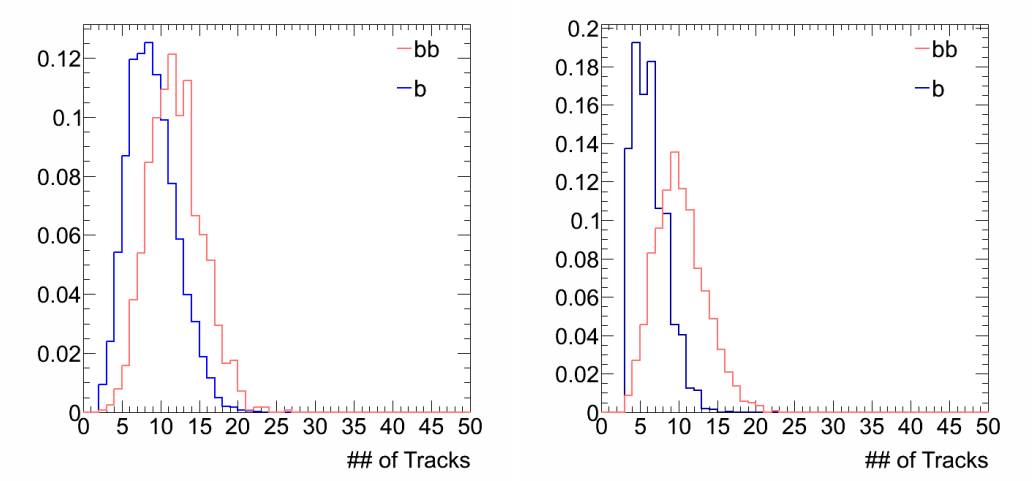

Figure 5.4: Distribution of the charged particle multiplicity for single (b) and merged (bb) jets from 80 GeV to 120 GeV in a sample of dijet events generated with PYTHIA and no detector simulation. Distributions are shown using all charged particles (left) and using only charged particles coming from b-hadron decay (right). A better discrimination between single and merged b-jets is obtained when using tracks from b-decay only.

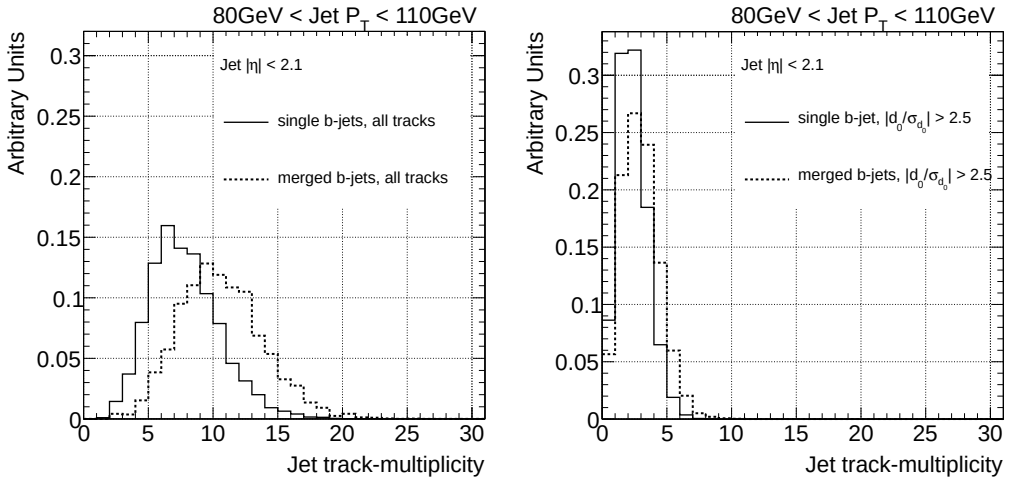

Figure 5.5: Distribution of the jet track multiplicity for single and merged b-jets from 80 GeV to 110 GeV, for all (left) and displaced tracks only (right). No improvement is obtained by using displaced tracks.

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,

$$Jet\ width = \frac{\sum_{i=1}^{N} {p_T}^{const_i}\, \Delta R(const_i, jet)}{\sum_{i=1}^{N} {p_T}^{const_i}} \tag{5.3}$$

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 [49] and [97]. 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 [98]: 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 5.6 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. 5.7. In this case the sum in equation 5.3 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. 5.8, 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

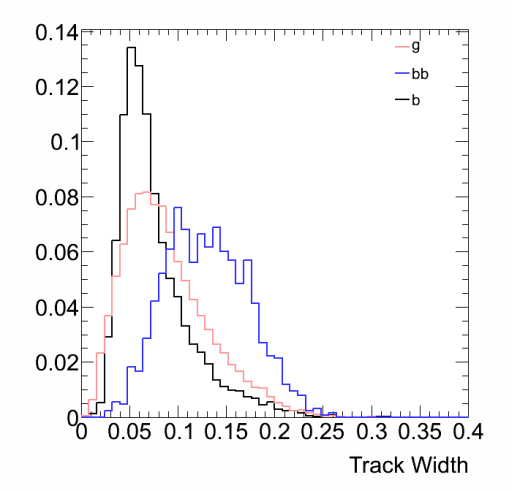

Figure 5.6: Distribution of track-jet width for gluon-initiated (g), single (b) and merged (bb) jets from 80 GeV to 120 GeV in a sample of dijet events generated with PYTHIA and no detector simulation.

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

$$Jet\ mass = E_{jet}^2 - \ \mathbf{p}^2 = (\sum_i E_i)^2 - (\sum_i \mathbf{p}_i)^2 \tag{5.4}$$

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 [99].

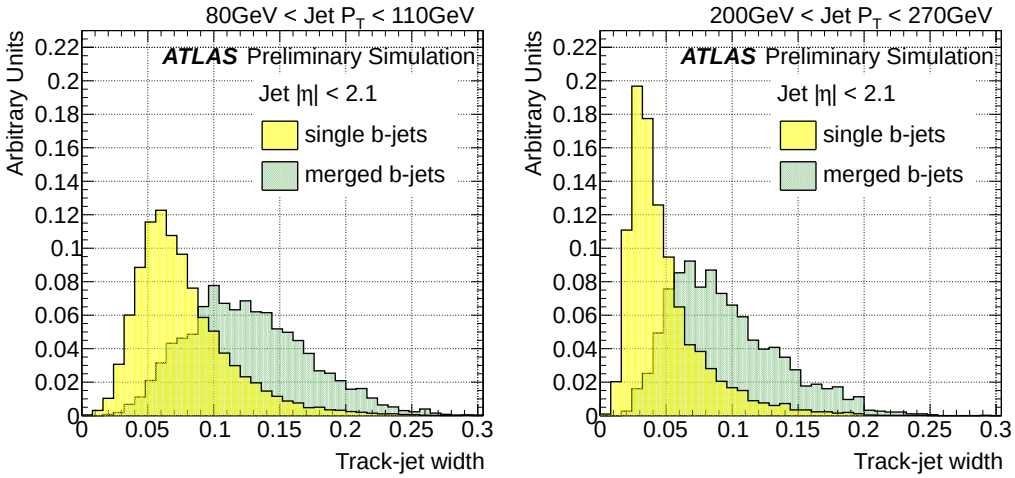

Figure 5.7: Distribution of track-jet width for single and merged b-jets from 80 GeV to 110 GeV (left) and 200 GeV to 270 GeV (right).

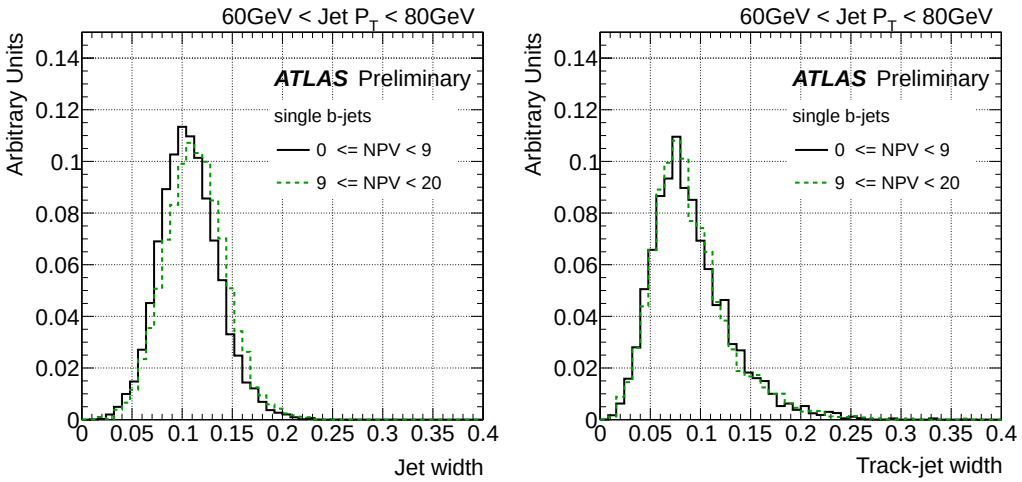

Figure 5.8: Distribution of jet width using topological clusters (left) and tracks (right) for single b-jets in two bins of number of primary vertices (NPV) for jets from 60 GeV to 80 GeV.

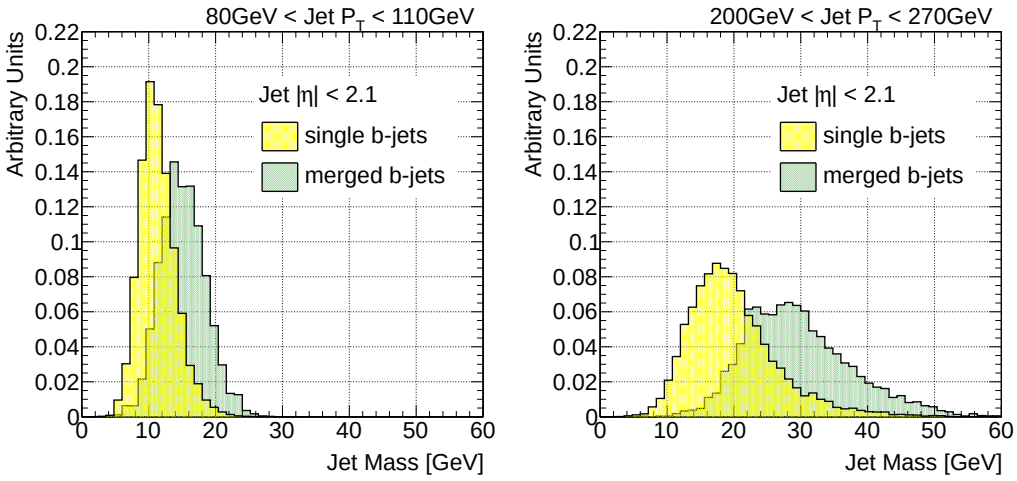

Figure 5.9: Distribution of jet mass in GeV for single and merged b-jets from 80 GeV to 110 GeV (left) and 200 GeV to 270 GeV (right).

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. 5.9, 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 [100].

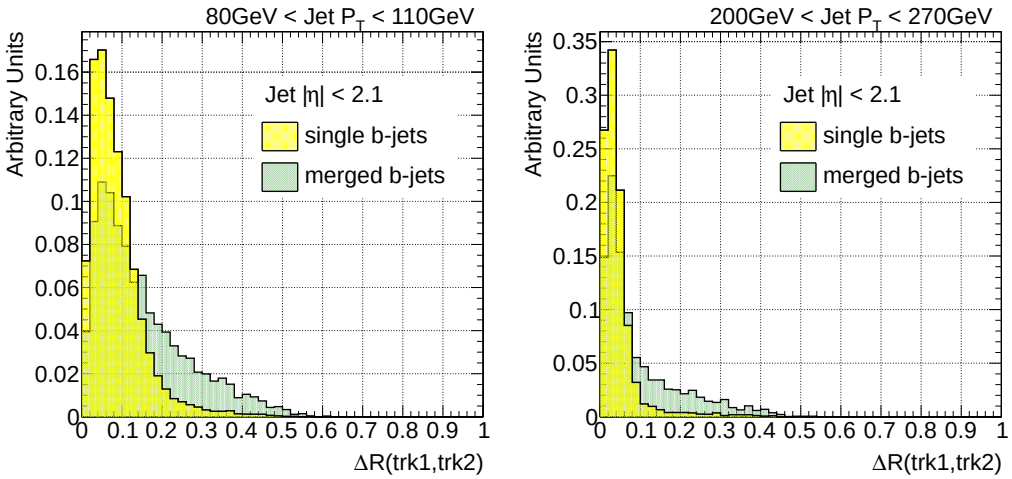

Figure 5.10: Distribution of ΔR between leading tracks for single and merged b-jets from 80 GeV to 110 GeV (left) and 200 GeV to 270 GeV (right).

and focuses on the hard components of the jet.

Figure 5.10 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. 5.11, 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.

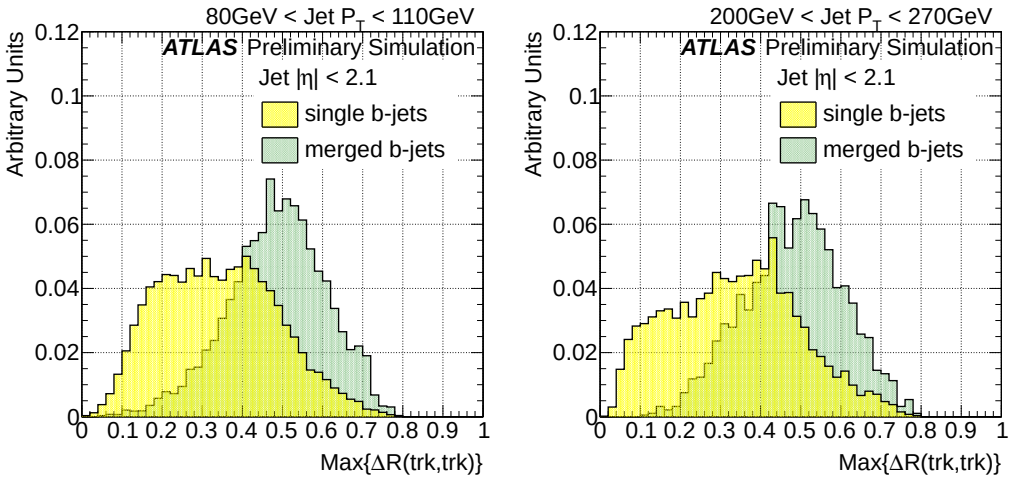

Figure 5.11: Distribution of the maximum ΔR between pairs of tracks for single and merged b-jets from 80 GeV to 110 GeV (left) and 200 GeV to 270 GeV (right).

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 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 [101]. This result is consistent with the QCD prediction that gluons radiate more than quarks.

The subjets were resolved by use of the inclusive k_t jet algorithm on the jet constituents with a fixed distance parameter. The k_t algorithm is the only jet algorithm that correctly identifies the resulting substructure as physical objects and therefore is the algorithm used for substructure analysis. As an alternative to fixed distance parameter subjets, it is also possible to undo the last step in the recombination sequence in order to identify the decay products of an object. This corresponds conceptually to undoing the first step in the fragmentation process that leads from interacting partons to jets. This approach is used in several jet grooming procedures[2], see for instance [103].

Figure 5.12 shows the distribution of the number of subjets for single and merged b-jets. The subjets in this case were built using the associated tracks as constituents, clustered by the inclusive k_t algorithm with a fixed distance parameter of $R = 0.2$. Merged jets tend to have on average one more subjet than single b-jets. The discrimination power of this variable is very poor and has the problem of being discrete with small numbers.

ΔR between the axes of two k_t subjets

The ΔR between k_t subjets is obtained by applying the k_t algorithm to the tracks associated to the jet using a large k_t distance parameter in order to ensure that all tracks get combined. The clustering is stopped once it reaches exactly two jets. This is done in Fastjet (see Section 2.4.1) 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, upon merging, give rise to the final jet. This cannot be done with anti-k_t.

[2]Jet grooming comprises dedicated techniques to remove uncorrelated radiation within a jet. A review of these procedures can be found in [102].

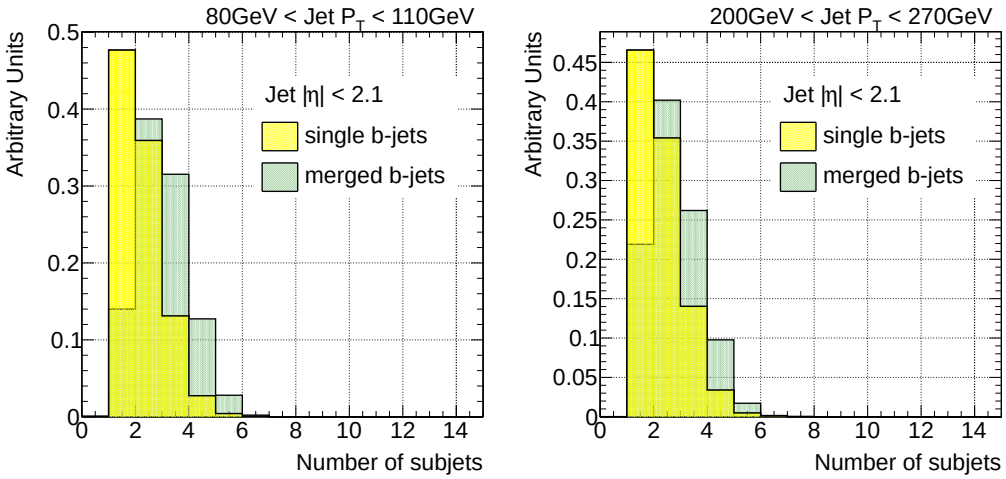

Figure 5.12: Distribution of the number of k_t sub-track-jets for single and merged b-jets from 80 GeV to 110 GeV (left) and 200 GeV to 270 GeV (right).

The ΔR between the axes of the two exclusive subjets is shown in Fig. 5.13. 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. 5.14. 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. [42]. The event display indicates how the high energy cells in the jet with two b-hadrons (in red) are grouped around the b-hadrons directions, leading to the two-subjet substructure of merged jets.

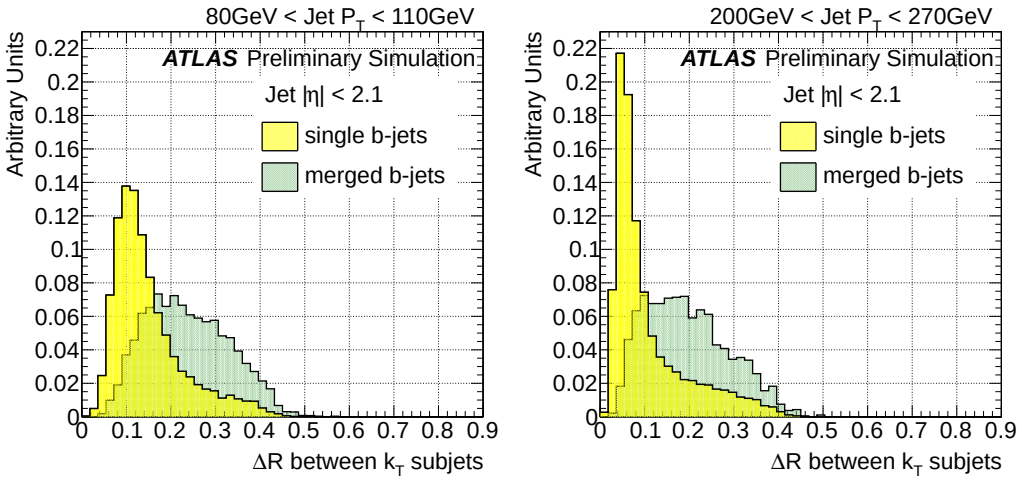

Figure 5.13: Distribution of the ΔR between the axes of the two k_t subjets in the jet for single and merged b-jets from 80 GeV to 110 GeV (left) and 200 GeV to 270 GeV (right).

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 [104]. 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 [105].

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

$$\tau_N^{(\beta)} = \frac{1}{\sum_k p_{Tk} (R_0)^\beta} \sum_k p_{Tk} (\min\{\Delta R_{j1,k}, \Delta R_{j2,k}, ..., \Delta R_{jN,k}\})^\beta. \quad (5.5)$$

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

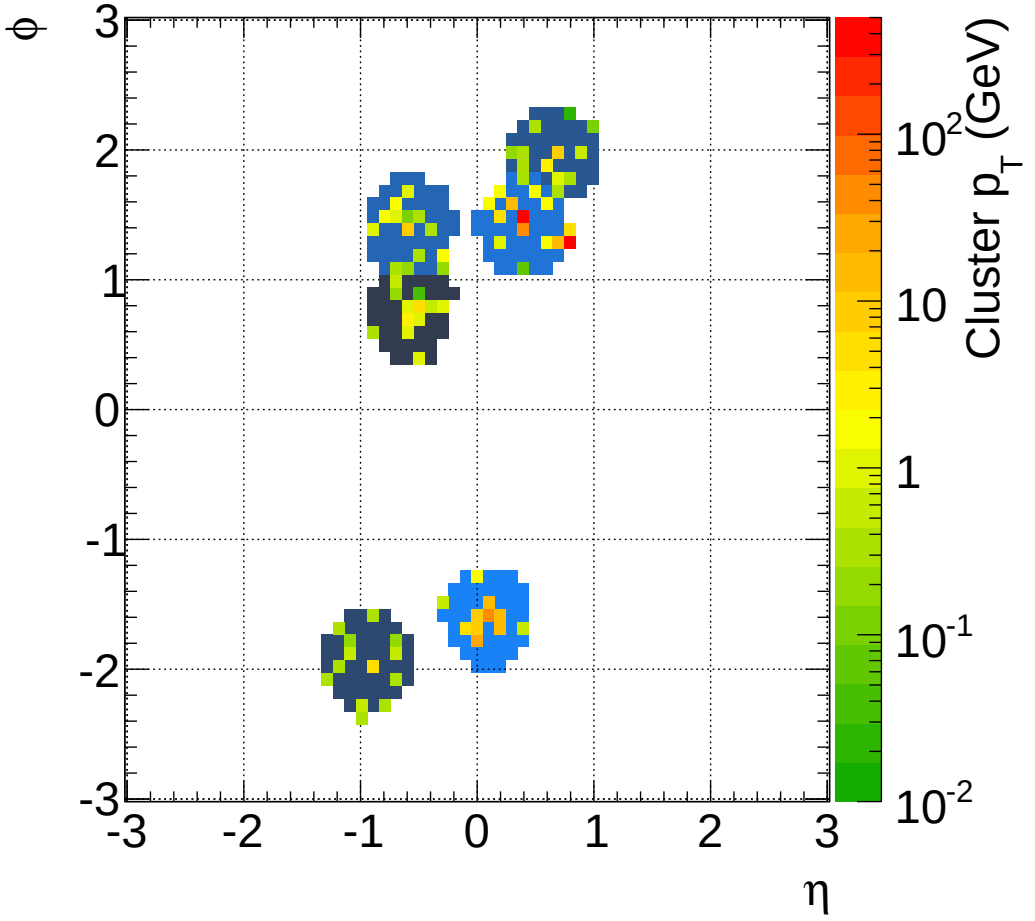

Figure 5.14: Event display of a merged b-jet in $(\eta, \phi) = (0.46, 1.41)$ and $p_T = 110$ GeV. The two b-hadrons are indicated as two red squares. The area of the jet is shown in blue and the topoclusters belonging to the jet are shown in different colors, from green to orange, depending on their 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. 5.5 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 [104] . 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 5.5 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. 5.15, 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. 5.16a, to be compared to the much lower correlation presented, for instance, between track-jet width and jet track multiplicity, shown in Fig. 5.16b.

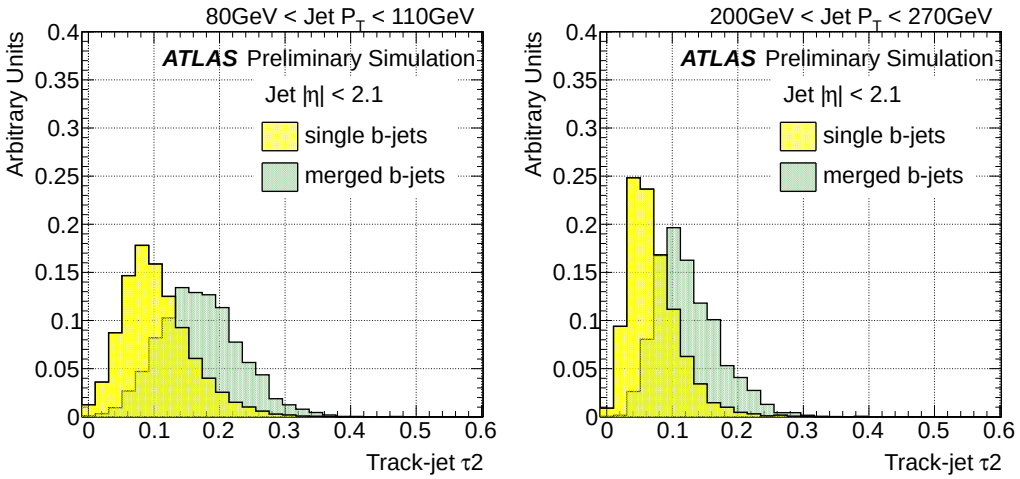

Figure 5.15: Distribution of τ_2 for single and merged b-jets from 80 GeV to 110 GeV (left) and 200 GeV to 270 GeV (right).

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. 5.17. 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,

$$\begin{bmatrix} \sum p_{Ti}\delta\eta_i^2 & -\sum p_{Ti}\delta\eta_i\delta\phi_i \\ -\sum p_{Ti}\delta\eta_i\delta\phi_i & \sum p_{Ti}\delta\phi_i^2 \end{bmatrix} \tag{5.6}$$

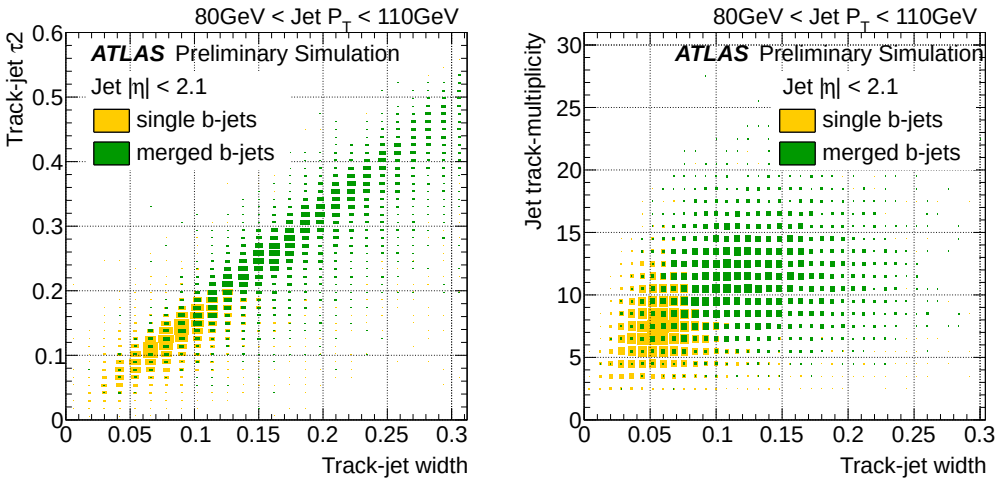

Figure 5.16: Correlation between τ_2 and track-jet width (left) and jet track multiplicity and track-jet width (right) for single and merged b-jets from 80 GeV to 110 GeV.

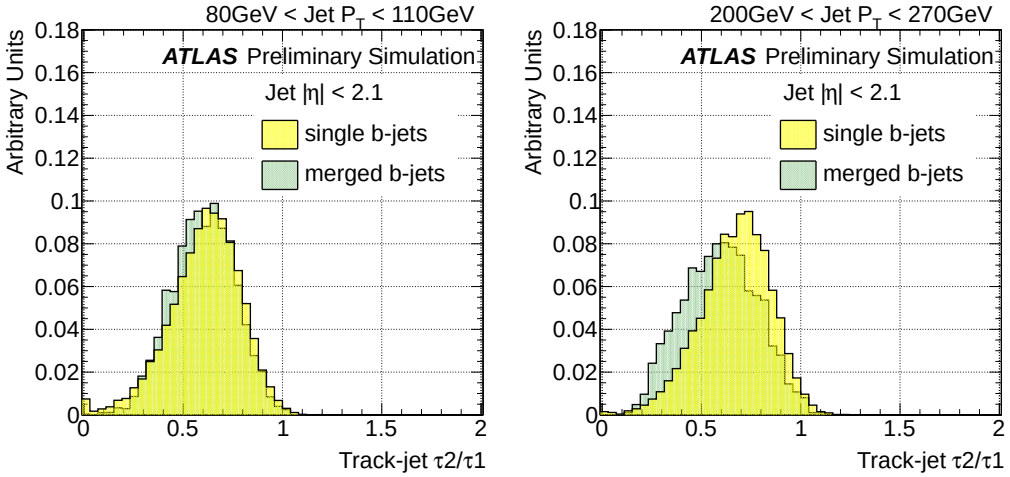

Figure 5.17: Distribution of τ_2/τ_1 for single and merged b-jets from 80 GeV to 110 GeV (left) and 200 GeV to 270 GeV (right).

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,

$$e = \sqrt{1 - r^2} \tag{5.7}$$

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

$$r = \frac{\lambda_m}{\lambda_p} = \frac{\sum p_{Ti}\delta\eta_i^2 + \sum p_{Ti}\delta\phi_i^2 - \sqrt{(\sum p_{Ti}\delta\eta^2 - \sum p_{Ti}\delta\phi_i^2)^2 + 4(\sum p_{Ti}\delta\eta_i\delta\phi_i)^2}}{\sum p_{Ti}\delta\eta_i^2 + \sum p_{Ti}\delta\phi_i^2 + \sqrt{(\sum p_{Ti}\delta\eta^2 - \sum p_{Ti}\delta\phi_i^2)^2 + 4(\sum p_{Ti}\delta\eta_i\delta\phi_i)^2}}. \tag{5.8}$$

Figure 5.18 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 5.7, 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.

5.5 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 5.19 to 5.23 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

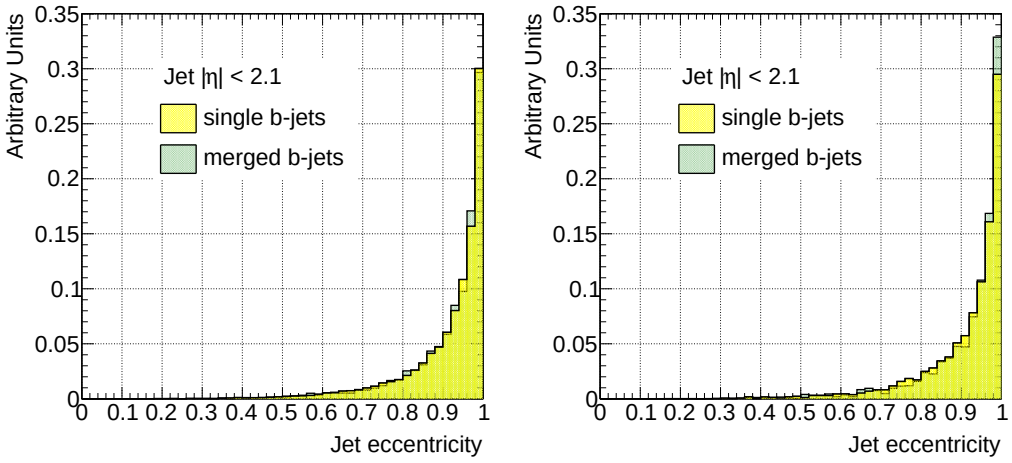

Figure 5.18: Distribution of the jet eccentricity for single and merged b-jets from 80 GeV to 110 GeV (left) and 200 GeV to 270 GeV (right).

Monte Carlo and data events passing the selection described in Section 5.3. 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 5.24 and 5.25.The agreement is still very good, suggesting that it is not the result of a compensation.

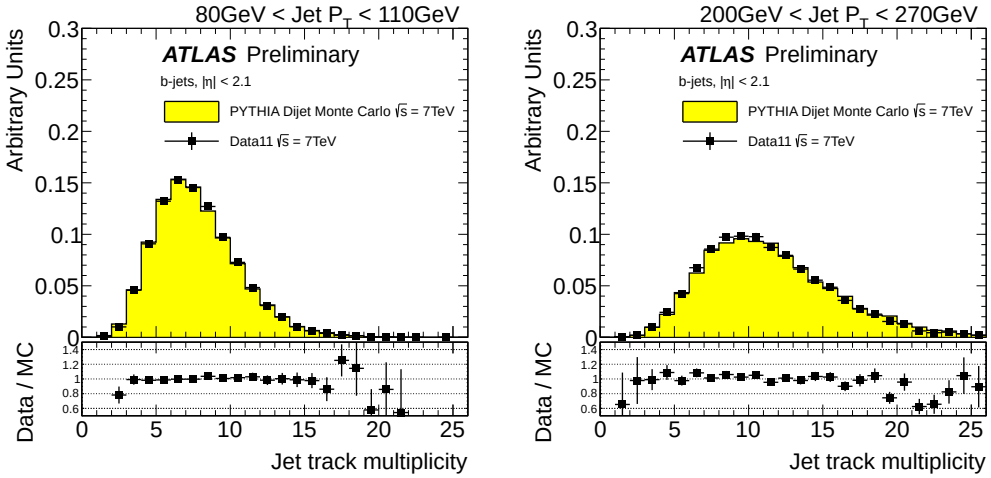

Figure 5.19: Distribution of the jet track multiplicity in 2 different jet p_T bins, for experimental data collected by ATLAS during 2011 (solid black points), and simulated data (filled histograms). The ratio data over simulation is shown at the bottom of each plot.

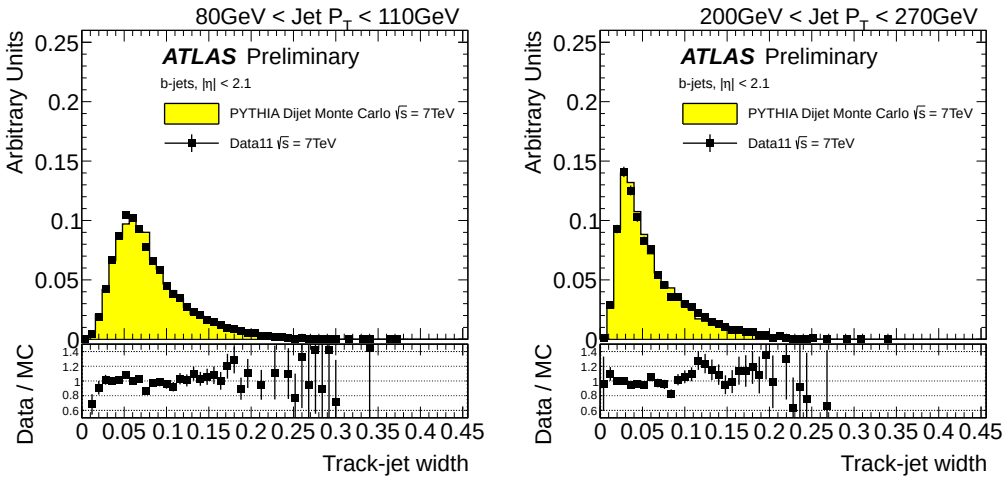

Figure 5.20: Distribution of the track-jet width in 2 different jet p_T bins, for experimental data collected by ATLAS during 2011 (solid black points), and simulated data (filled histograms). The ratio data over simulation is shown at the bottom of each plot.

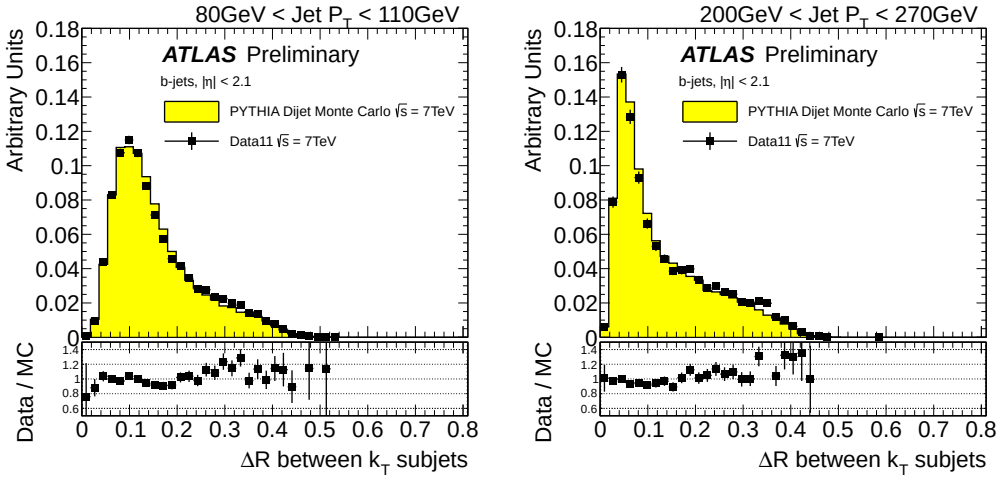

Figure 5.21: Distribution of the ΔR between the axes of the two k_t subjets in the jet in 2 different jet p_T bins, for experimental data collected by ATLAS during 2011 (solid black points), and simulated data (filled histograms). The ratio data over simulation is shown at the bottom of each plot.

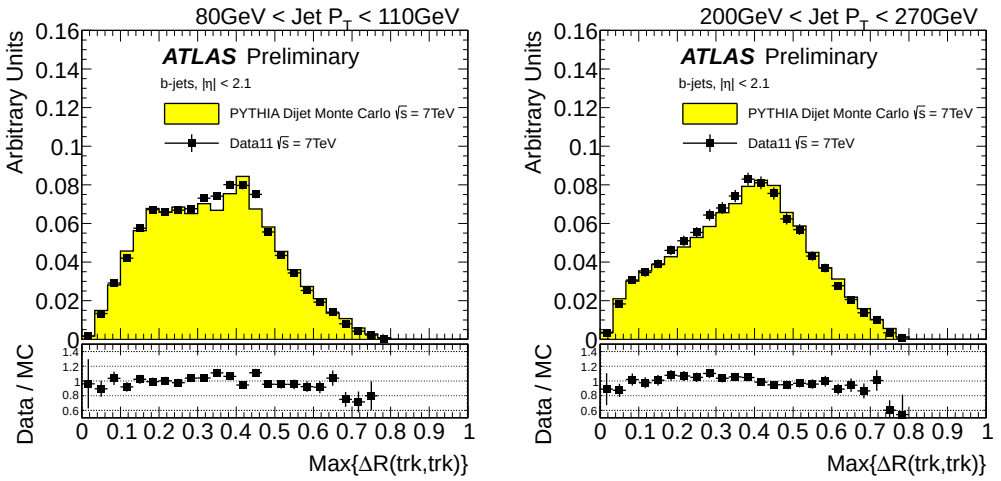

Figure 5.22: Distribution of maximum ΔR between pairs of tracks in two different jet p_T bins, for experimental data collected by ATLAS during 2011 (solid black points), and simulated data (filled histograms). The ratio data over simulation is shown at the bottom of each plot.

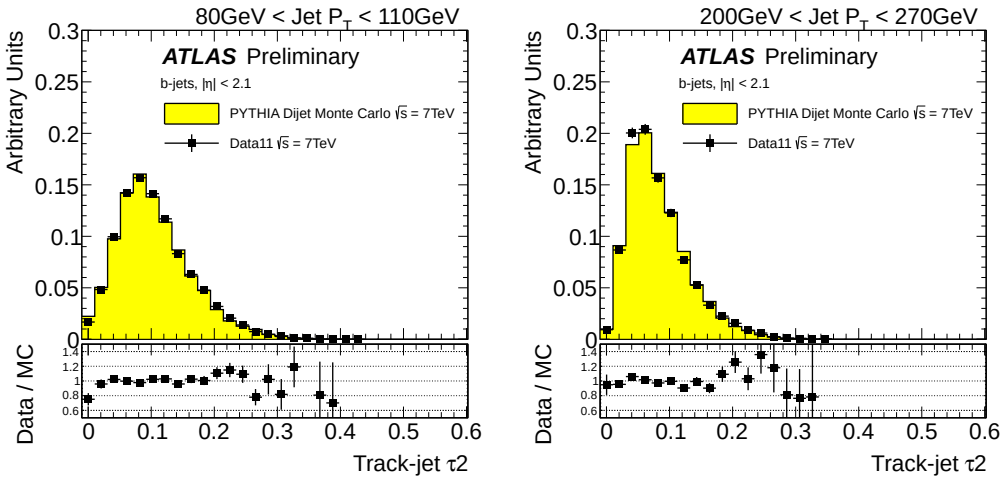

Figure 5.23: Distribution of τ_2 in two different jet p_T bins, for experimental data collected by ATLAS during 2011 (solid black points), and simulated data (filled histograms). The ratio data over simulation is shown at the bottom of each plot.

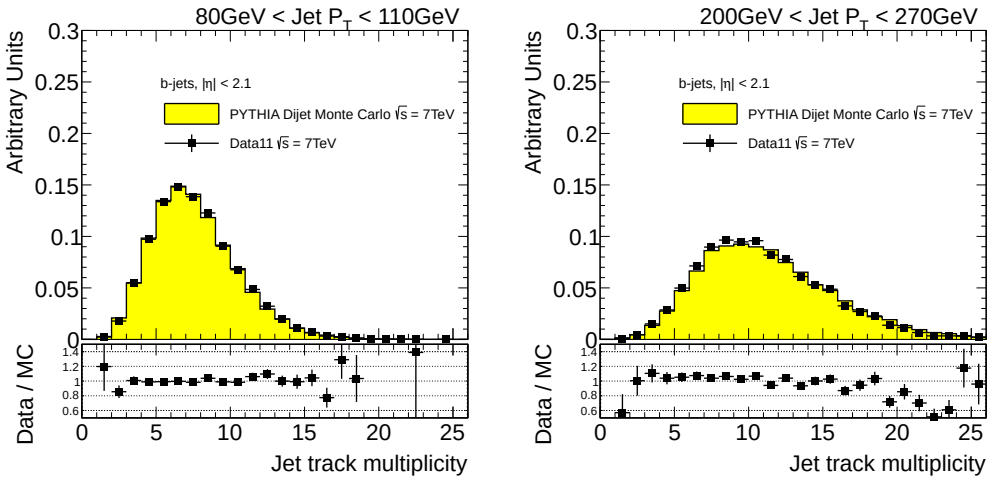

Figure 5.24: Distribution of the jet track multiplicity in 2 different jet p_T bins, for experimental data collected by ATLAS during 2011 (solid black points), and simulated data (filled histograms). Jets were selected using MV1 tagger at its 70% b-jet efficiency working point. The ratio data over simulation is shown at the bottom of each plot.The agreement is very good.

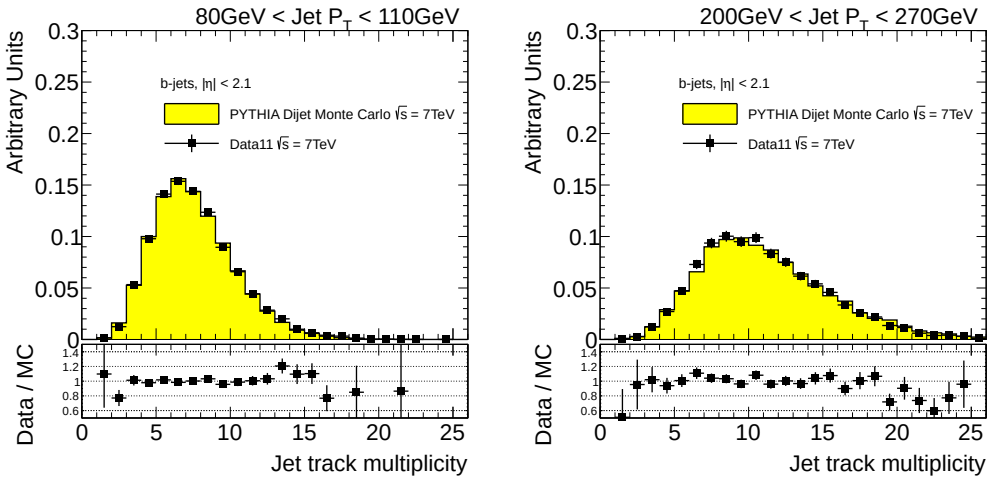

Figure 5.25: Distribution of the jet track multiplicity in 2 different jet p_T bins, for experimental data collected by ATLAS during 2011 (solid black points), and simulated data (filled histograms). Jets were selected using JetFitter tagger at its 60% b-jet efficiency working point. The ratio data over simulation is shown at the bottom of each plot.The agreement is very good.

Chapter 6

Identification of double b-hadron jets

After the evaluation of the best discriminating variables, this chapter presents their combination into a tagging algorithm, capable of efficiently identifying single b-jets while rejecting merged b-jets. First, the different multivariate techniques explored for this analysis are presented, followed by a study and comparison of their application to the case of the double b-hadron tagger. The assessment of the performance for the chosen technique, the likelihood ratio, is studied next. Finally, the systematic uncertainties and their influence on the tagger are discussed in detail.

6.1 Multivariate methods

Multivariate data analysis refers to a statistical technique used to analyze data that is composed of more than one variable. Classification is done through learning algorithms that make use of training events, for which the desired output is known, to determine the mapping function that describes

a decision boundary. The following multivariate methods were explored:

- Likelihood ratio estimators (LLR)
- Neural networks (NN)
- Boosted decision trees (BDTs)

A description of each of these methods is presented in the next sections.

Likelihood ratio estimators

The method of LLR consists of building a model out of probability density functions (PDF) that reproduce the distributions of the input variables for signal and background. The likelihood ratio $y_L(i)$ for event i is defined by:

$$y_L(i) = \frac{L_S(i)}{L_S(i) + L_B(i)}, \tag{6.1}$$

where L_s and L_B are the likelihoods of event i under the signal and background hypothesis respectively. In the case of poorly correlated variables, the likelihoods are obtained by multiplying the probability densities of all input variables and normalising this by the sum of the signal and background likelihoods,

$$L_{S(B)}(i) = \prod_{k=1}^{n_{var}} p_{S(B),k}(x_k(i)), \tag{6.2}$$

and where $p_{S(B),k}(x_k(i))$ is the signal (background) PDF for the kth input variable x_k. All the PDFs are normalized to one.

The parametric form of the PDFs is generally unknown, however it is possible to empirically approximate its shape by nonparametric functions. Nonparametric models differ from parametric models in that the model structure is not specified a priori but is instead determined from the data sample used for training. A histogram is a simple example of a nonparametric estimate of a probability distribution. The nonparametric functions can be chosen

individually for each variable and can be either polynomial splines of various degrees fitted to binned histograms[1] or unbinned kernel density estimators (KDE). The basic idea in the KDE approach is to extract the PDF from the trainig events themselves, x_i with $i = 1, ..., N$, via the sum of individual equal-area Gaussian kernels, G, for each event. For a PDF $p(x)$ of a variable x, one finds [106],

$$p(x) = \frac{1}{Nh} \sum_{i=i}^{N} G(\frac{x - x_i}{h}) = \frac{1}{N} \sum_{i=i}^{N} G_h(x - x_i), \tag{6.3}$$

where N is the number of training events, $G_h(t) = G(t/h)h$ is the kernel function, and the extent of each event's contribution, h, can be kept constant for the entire training sample or be calculated adaptively as a function of the local densitiy of events. h is termed the "bandwidth" of the kernel or, also, the "smoothing parameter". The optimum non-adaptive (NA) bandwidth, h_{NA}, for a Gaussian kernel function, can be shown to be

$$h_{\mathrm{NA}} = (\frac{4}{3})^{1/5} \sigma_x N^{-1/5} \tag{6.4}$$

where σ_x is the RMS of the variable x. The adaptive (A) approach uses as input the result of the non-adaptive KDE, but also takes into account the local event density. The adaptive bandwidth h_{A} then becomes a function of $p(x)$

$$h_{\mathrm{A}} = \frac{h_{\mathrm{NA}}}{\sqrt{p(x)}}. \tag{6.5}$$

The adaptive approach improves the shape estimation in regions with low event density.

Instead of unbinned training data, the KDE approach we implemented uses a finely-binned histogram as input, which allows to significantly speed

[1] A spline is a sufficiently smooth polynomial function that is piecewise-defined, and possesses a high degree of smoothness at the places where the polynomial pieces connect. It is often referred to as polynomial interpolation.

up the algorithm. The calculation of the optimal bandwidth is the first step. Subsequently, the smothed high-binned histogram estimating the PDF shape is created by looping over the bins of the input histogram and summing up the corresponding kernel functions using h_{A} (h_{NA}) in de case of the non-adaptive (adaptive) mode. This output is the final estimate for the PDF.

The smoothness of the kernel density estimate is evident compared to the discreteness of a histogram; kernel density estimates converge faster to the true underlying density for continuous random variables.

A generalization of the LLR estimator to n_{var} dimensions, where n_{var} is the number of correlated input variables, exist in the theory today. If the multidimensional PDF for a signal and background were known, this classifier would exploit the full information, and would hence be optimal. In practice however, huge training samples are necessary to sufficiently populate the phase space[2]. Kernel estimation methods may be used to approximate the shape of the PDF for finite training samples.

A simple probability density estimator denoted as Projective likelihood estimator *range search*, or PDE-RS, has been suggested in Ref. [107]. The PDE for a given test event is obtained by counting the number of training events that occur in the "vecinity" of the test event. The classification of the test event as being either signal or background type can be conducted by a local estimate of the probability density of it belonging to either class. The n_{var}-dimensional volume that encloses the "vecinity" is user-defined and adaptive: the volume is defined in each dimension with respect to the RMS of that dimension, estimated from the training sample. Although the adaptive volume adjustment is flexible and should perform better, it significantly

[2]Due to correlations between the input variables, only a sub-space of the full phase space may be populated.

increases the computing time of the PDE-RS discriminant.

One of the shortcomings of the original PDE-RS implementation is its sensitivity to the exact location of the sampling volume boundaries: an infinitesimal change in the boundary placement can include or exclude a training event. Kernel functions mitigate these problems by weighting each event within the volume as a function of its distance to the test event.

PDE-RS can yield competitive performance if the number of input variables is not too large and the statistics of the training sample is ample; on the other hand, it is a slowly responding classifier.

Neural networks

An artificial Neural Network (NN) is a nonlinear discriminant. It is, most generally speaking, a simulated collection of interconnected neurons, with each neuron producing a certain response at a given set of input signals. It can be viewed as a mapping from a space of input variables $x_1, ..., x_{n_{var}}$ onto, in the case of a signal-versus-background discimination problem, a one-dimensional output variable. The behaviour of an artificial neural network is determined by the layout of the neurons, the weights of the inter-neuron connections, and by the response of the neurons to the input, described by the neuron response function. The neuron response function maps the neuron input (in R^n) onto the neuron output (R); often it can be separated into a synapse function ($R^n \rightarrow R$) and a neuron activation function ($R \rightarrow R$). The neuron activation function can be a either a *linear*, *sigmoid*, *tanh*, or a *radial* function.

While in principle a neural network with n neurons can have n^2 directional connections, the complexity can be reduced by organising the neurons in layers and only allowing direct connections from a given layer to the following

one. This kind of neural network is termed multi-layer perceptron. The first layer of a multilayer perceptron is the input layer, the last one the output layer, and all others are hidden layers. For a classification problem with n_{var} input variables the input layer consists of n_{var} neurons that hold the input values, $x_1, ..., x_{n_{var}}$, and one neuron in the output layer that holds the output variable, the neural net estimator y_{NN} .

Decision trees

A decision tree is a binary tree structured classifier similar to the one sketched in Fig. 6.1. The training, building or growing of a decision tree is the process that defines the splitting criteria for each node. The training starts with the root node, where an initial splitting criterion for the full training sample is determined. The split results in two subsets of training events that each go through the same algorithm of determining the next splitting iteration. This procedure is repeated until the whole tree is built. At each node, the split is determined by finding the variable and corresponding cut value that provides the best separation between signal and background. The phase space is split this way into many regions that are eventually identified as "signal-like" or "background-like", depending on the majority of training events that end up in the final *leaf* node. The boosting of a decision tree extends this concept from one tree to several trees which form a *forest*. Boosting increases the statistical stability of the classifier and typically also improves the separation performance compared to a single decision tree [108].

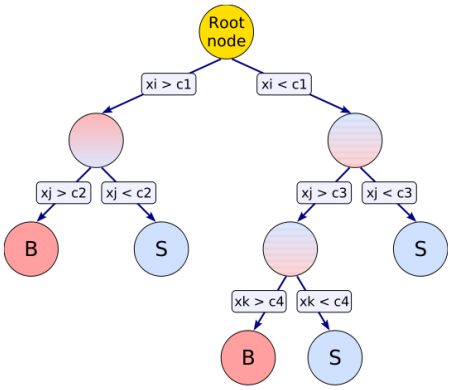

Figure 6.1: Schematic view of a decision tree. Starting from the *root node*, a sequence of binary splits using the discriminating variable x_i is applied to the data. The leaf nodes at the bottom end of the tree are labeled "S" for signal and "B" for background depending on the majority of events that end up in the respective nodes [109].

6.2 The double b-hadron jet tagger

Three variables were selected for training the multivariate methods described in the previous section, based on discrimination power, correlation and pile-up dependence: the jet track multiplicity, the track-jet width and the ΔR between the axes of two k_t subjets in the jet. Signal and background datasets were given to the multivariate methods as input, containing a list of the b-tagged jets with the information of their p_T and values for the chosen tracking variables. Other variables such as τ_2 or $\max\{\Delta R(trk, trk)\}$ were also tested leading to no gain in performance.

Selection of an MVA method

A sub-set of the dijet Monte Carlo sample was used for training the methods in the context of the Toolkit for Multivariate Data Analysis, TMVA [109],

written in C++ language. After the event and jet selections, described in Section 5.3, were performed, the b-tagged jets were classified as signal (single b-jets) or background (merged b).

Two different likelihood configurations were evaluated: the simple likelihood ratio estimator, using an adaptive Gaussian KDE strategy for the estimation of the PDFs of the input variables; and the more sophisticated PDE-RS approach, adopting an adaptive mode for the volume search. The multidimentional PDE-RS classifier offered no gain in discrimination with respect to the more simpler likelihood method, with the further disadvange of being more time consuming. The likelihood method shows good performance, and, given the low correlation of the input variables accross p_T, constitutes an adequate method for single-merged discrimination with a fast training step.

An Multi-layer perceptron (MLP) neural net, with two hidden layers of n_{var} and $n_{var} - 1$ neurons respectively, was also trained ($n_{var} = 3$). Two different neuron activation functions were tested, *tanh* and *sigmoid*, with the latter showing better performance. The initial NN training was carried out in 600 cycles, also termed "epochs", for a fast implementation of the method. Although, faster, this configuration led to an irregularly shaped output. This was only fixed with a 3000-epoch training.

The Boosted Decision Tree (BDT) approach was implemented with 400 trees in the BDT forest. Due to the simplicity of the method where each training step (node splitting) involves only a one-dimensional cut optimisation, little tuning was required in order to obtain reasonably good results.

Examples of the distributions of the final output for these methods, evaluated in an orthogonal sample of simulated dijet events, are displayed in Fig. 6.3 for a medium p_T bin.

The outputs of the explored MVA discriminants are different in terms of shape and range, although the latter could be rearranged with a suitable variable transformation. In spite of these distinct features the performances of the different methods agree within statistics, see Fig. 6.3. The performance of a classifier algorithm can be assessed by a curve of rejection of merged b-jets, $(1/\epsilon_{bkg})$, as a function of single b-jet efficiency, ϵ_{sig}; where ϵ_{bkg} (ϵ_{sig}) is the probability that a double (single) b-hadron jet passes the single b-jet tagger. The different points in the curve are obtained by varying the likelihood value above which a jet was classified as single (see Section 4.4.1).

As opposed to NN discriminants with large number of training cycles, the training and the application of the likelihood are very fast operations that are suitable for very large data sets and tuning of the training parameters. Although also very fast, a shortcoming of decision trees is their instability with respect to statistical fluctuations in the training sample from which the tree structure is derived. If two input variables exhibit similar separation power, a fluctuation in the training sample may cause the tree growing algorithm to decide to split on one variable, while the other variable could have been selected without that fluctuation. In such a case the whole tree structure is altered below this node, possibly resulting also in a substantially different classifier response [109]. It is for these reasons that the likelihood classifier is the selected method for our tagger.

6.3 Results obtained

A discriminant between single and merged b-jets was built by training a likelihood ratio estimator, with the following three variables as input,

1. Jet track multiplicity

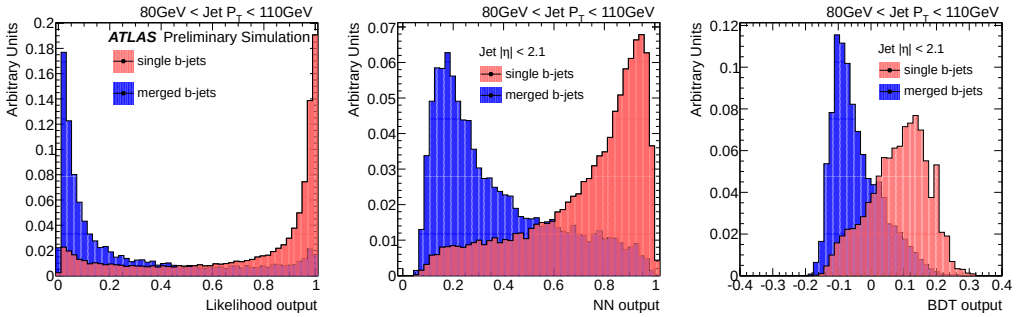

Figure 6.2: Distribution of the MVA discriminant outputs for the Likelihood (a), Neural Network (b) and Boosted Decision Trees (c) classifiers, for single and merged b-jets between 80 GeV and 110 GeV.

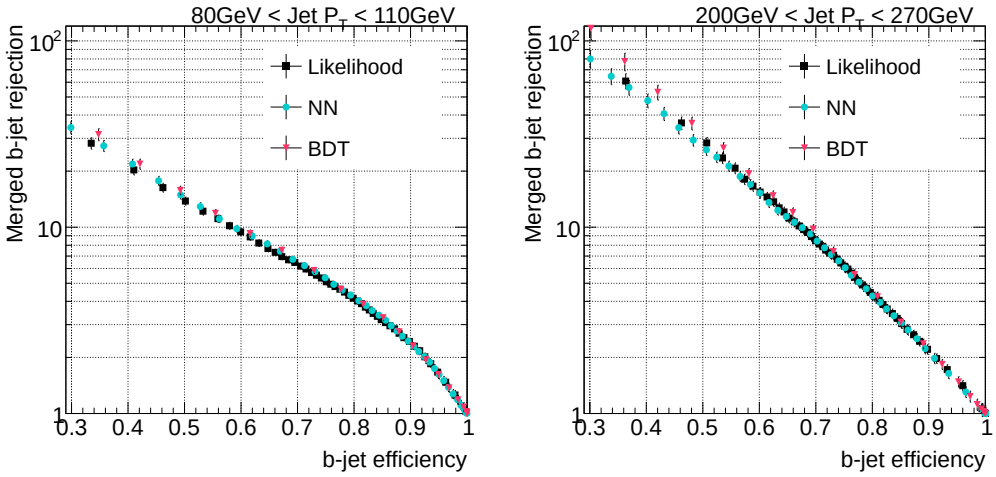

Figure 6.3: Rejection of merged b-jets as a function of single b-jet efficiency for the the different MVA methods evaluated for medium and high jet p_T.

2. Track-jet width
3. ΔR between the axes of 2 k_t subjets within the jet

Given the correlation of the variables with the jet transverse momentum, the training sample was categorized in bins of calorimeter jet p_T, and independent likelihood classifiers were built for each category.

Due to the lack of statistics of merged jets in the low p_T bins, signal and background jets were not weighted by the dijet samples cross-sections to allow the contribution of subleading lower p_T jets from high p_T events. The gain in statistics in merged b-jets for the first p_T bin was of more than 500%. It is important to stress that, although essential for data to MC comparisons (see Section 5.2), the weighting of the dijet Monte Carlo samples by their respective cross-sections is not necessary for studies performed at simulation level only. For the evaluation of the method the same procedure was followed.

Example distributions of the likelihood output for single and merged b-jets are displayed in Fig. 6.4 for low and high transverse momentum jets. The performance plot including the rejection vs. efficiency curves for each of the eight p_T bins studied (see Section 5.3) is shown in Fig. 6.5. The performance of the tagger improves with p_T:

- $p_T > 40$ GeV: rejection above 8 at 50% eff.
- $p_T > 60$ GeV: rejection above 10 at 50% eff.
- $p_T > 200$ GeV: rejection above 30 at 50% eff.

The rejection of merged jets attained as a function of p_T for the 50% and 60% single b-jet efficiency working points are summarized in Table 6.1, together with their relative statistical error. These are propagated from the Poisson fluctuations of the number of events in the merged and single b distributions. The error is slightly lower for the 60% efficiency working point

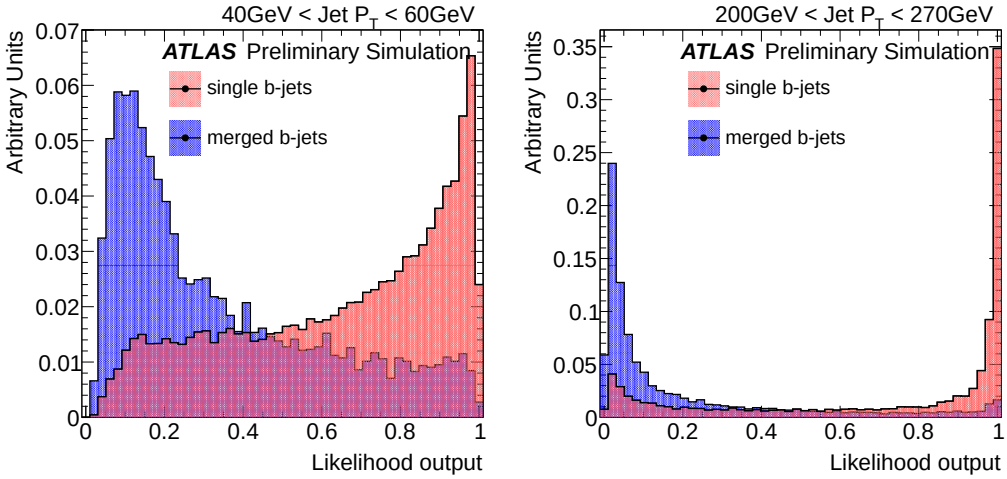

Figure 6.4: Distribution of the likelihood output for single and merged b-jets for low and high p_T jets.

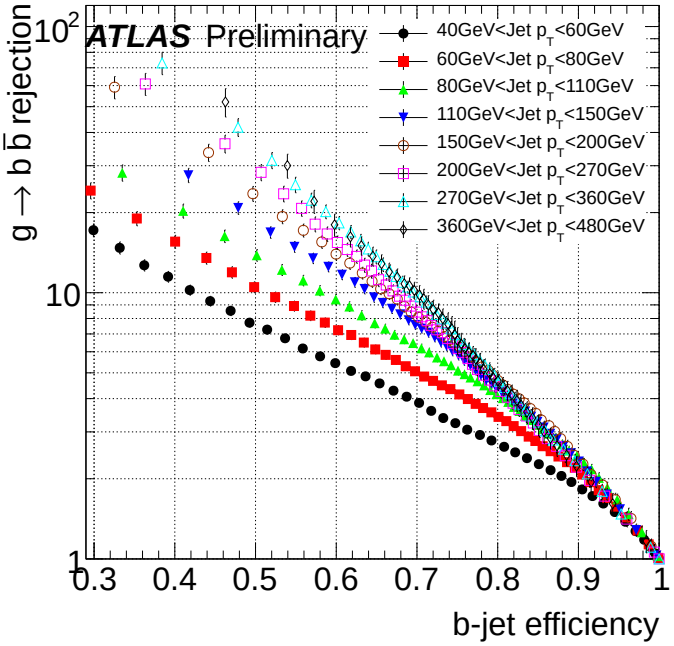

Figure 6.5: Rejection of merged b-jets as a function of single b-jet efficiency for dijet events in 8 jet p_T bins.

because a higher efficiency allows for a greater number of Monte Carlo events to measure the performance.

Jet p_T	single b-jet efficiency 50%		single b-jet efficiency 60%	
(GeV)	Rejection	stat.err.	Rejection	stat.err.
40 - 60	8	4%	5	3%
60 - 80	10	4%	7	4%
80 - 110	14	5%	9	4%
110 - 150	19	5%	12	4%
150 - 200	23	5%	14	5%
200 - 270	30	7%	16	6%
270 - 360	36	7%	19	6%
360 - 480	41	8%	18	8%

Table 6.1: The merged b-jet rejection for the 50% and 60% efficiency working points in bins of p_T.

6.4 Systematic uncertainties

The development, training and performance determination of the tagger is based on simulated events. Although the agreement between simulation and data explored in section 5.5 is a necessary validation condition, it is also important to investigate how the tagger performance depends on the systematic precision with which the MC simulates the data. In particular we have considered:

- presence of additional interactions (pile-up);

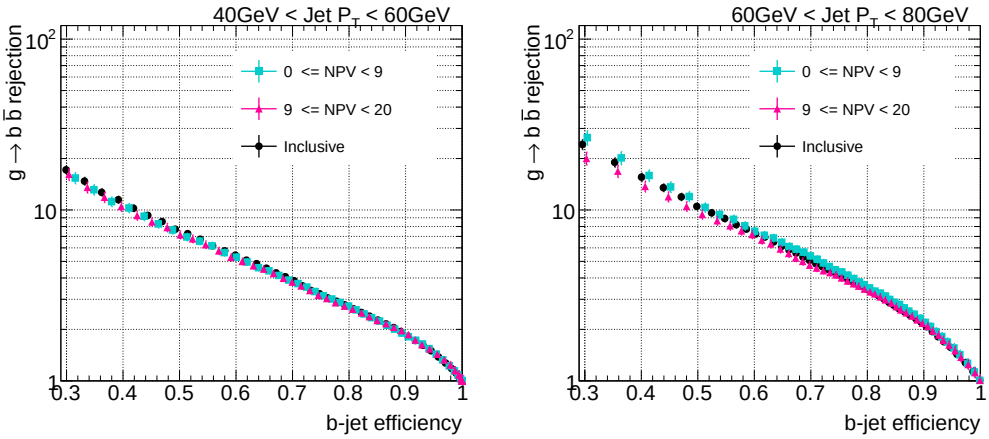

Figure 6.6: Rejection of merged b-jets as a function of single b-jet efficiency in bins of NPV for two low jet p_T bins.

- uncertainty in the b-jet tagging efficiency;
- uncertainty in the track reconstruction efficiency;
- uncertainty in the track transverse momentum resolution;
- uncertainty in the jet transverse momentum resolution;
- uncertainty in the jet energy scale.

I. Pile-up

The size of this effect was studied by comparing the performance of the likelihood discriminant with b-jets in events with small (1-9) and large (9-20) number of primary vertices. A comparison of the performance in these two sub-samples relative to the inclusive sample is shown in Fig. 6.6 for the two lowest p_T bins, where the effect of pile-up is more important. As expected from the use of tracking (as opposed to calorimeter) variables, no significant dependence with pile-up is observed. Performance differences between high and low number of primary vertices events are $\leq 2\%$. The impact of pile-up might be larger in 2012 data.

II. b-tagging efficiency

The performance of heavy-flavor tagging in Monte Carlo events is calibrated to experimental data by means of scale factors (SFs). The SFs are defined as the ratio of the heavy-flavor tagging efficiency in data over that in Monte Carlo for the different jet flavors. They are measured by the ATLAS Flavour Tagging Working group, and their measurement carries a systematic uncertainty, see Section 4.4.2.

To estimate the impact of this uncertainty a conservative approach is followed: the SFs are varied in all the p_T bins simultaneously by one standard deviation both in the up and down directions. The MC distributions weighted by the varied SFs show no major deviations from the nominal, see Fig. 6.7. In the same manner, the effect of the b-tagging calibration uncertainty on the likelihood peformance, shown in Fig. 6.8, is $< 1\%$, negligible with respect to the statistical uncertainty. This was indeed expected. The scale factors depend on the true flavor of the jet and on its p_T, but these are basically constant in the performance determination, which is based on single flavor (true b-) jets classified in p_T-bins.

III. Track reconstruction efficiency

The track reconstruction efficiency, ϵ_{trk}, parametrised in bins of p_T and η, is defined as:

$$\epsilon_{trk} = \frac{N_{rec}^{matched}(p_T, \eta)}{N_{gen}(p_T, \eta)} \tag{6.6}$$

where $N_{rec}^{matched}$ is the number of reconstructed tracks matched to a generated charged particle, and $N_{gen}(p_T, \eta)$ is the number of generated charged particles

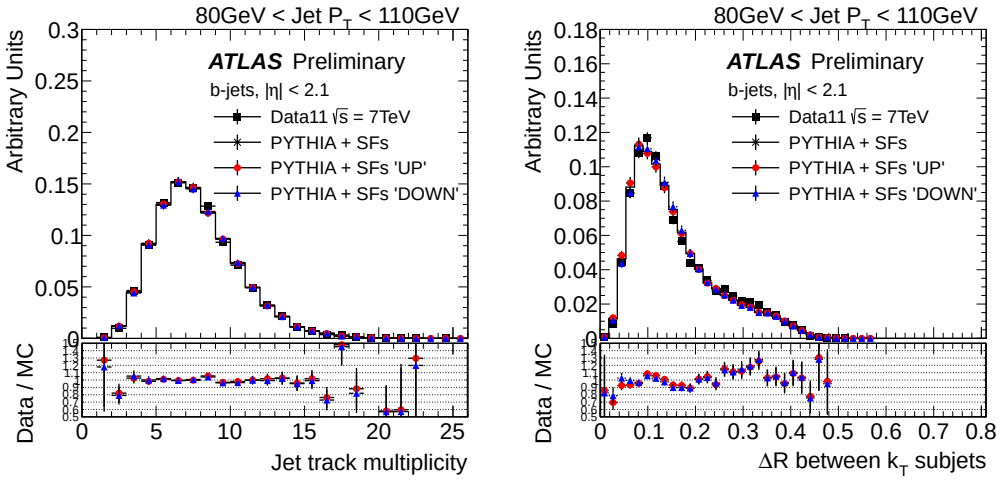

Figure 6.7: The effect of a variation in the b-tagging Scale Factors on the tracking variables distributions. Scale Factors were varied up (down) by 1-sigma to evaluate the systematic uncertainty from this source. The ratio data over MC is shown for MC PYTHIA with SFs varied up (circles) and down (triangles).

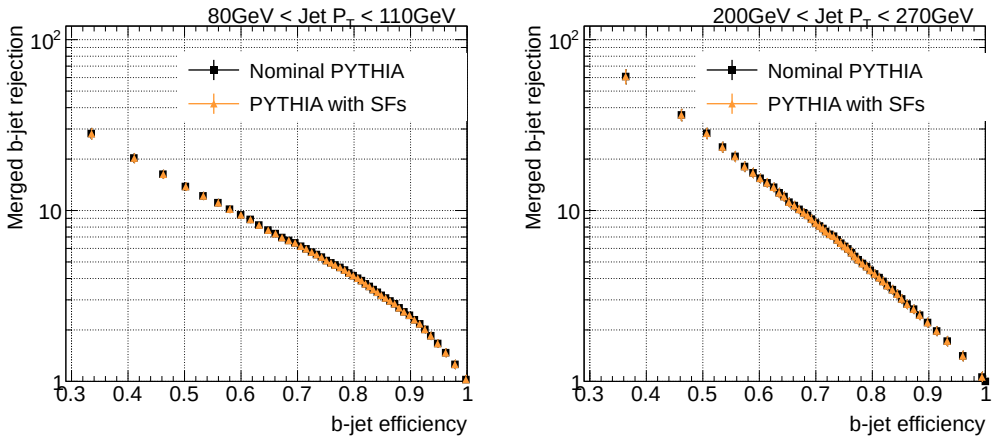

Figure 6.8: Rejection of merged b-jets as a function of single b-jet efficiency with and without scale factors as weights.

in that bin[3]. As the track reconstruction efficiency is determined from MC, the main systematic uncertainty results from the level of agreement between data and MC. Since charged hadrons are known to suffer from hadronic interactions with the material in the detector, a good description of the material in MC is needed to get a good description of the track reconstruction efficiency. An increase (decrease) in material leads to an increase (decrease) in the number of hadronic interactions, hence to a decrease (increase) in the reconstruction efficiency.

The contribution to the tracking reconstruction efficiency systematics of the imperfect description of the detector, in particular the knowledge of the material in the inner tracker, was measured with a data-driven method [67]. The results are given in bins of track η. For tracks with $p_{\mathrm{T}}^{\mathrm{track}} > 500$ MeV the uncertainties are independent of p_T: 2% for $|\eta^{\mathrm{track}}| < 1.3$, 3% for $1.3 < |\eta^{\mathrm{track}}| < 1.9$, 4% for $1.9 < |\eta^{\mathrm{track}}| < 2.1$, 4% for $2.1 < |\eta^{\mathrm{track}}| < 2.3$ and 7% for $2.3 < |\eta^{\mathrm{track}}| < 2.5$. All numbers are relative to the corresponding tracking efficiencies.

To test the impact of these uncertainties, a fraction of tracks determined from the track efficiency uncertainty was randomly removed. The tracking variables were re-calculated and the performance of the nominal likelihood was evaluated in the new sample with worse tracking efficiency. The rejection-efficiency curves show a small degradation of the performance which is comparable to the statistical uncertainty. The effect is however systematically present over all 16 p_T bin/working points, without a clear p_T dependence. We have thus taken the average over p_T, and obtained a global

[3]The matching between a generated particle and a reconstructed track uses a cone-matching algorithm, associating the particle to the track with the smallest ΔR within a cone of radius 0.15.

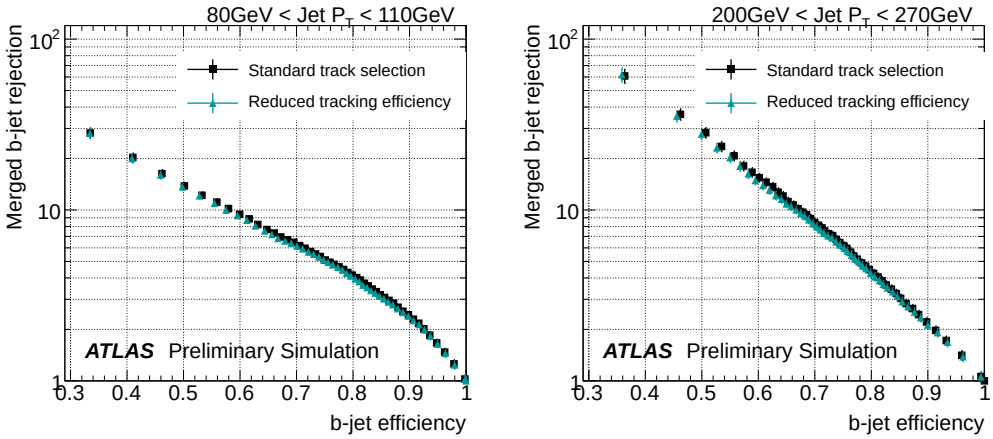

Figure 6.9: Rejection of merged b-jets as a function of the single b-jet efficiency showing shift in likelihood performance caused by a reduction in the tracking efficiency.

systematic uncertainty of 4% both for the 50% and 60% efficiency working points. The performance comparison is shown in Fig. 6.9 for two p_T bins.

IV. Track momentum resolution

The knowledge of the track momentum resolution is limited by the precision both in the material description of the Inner Detector and in the mapping of the magnetic field. Its uncertainty propagates to the kinematic variables used in the double b-hadron jet tagger. In order to study this effect, track momenta are over-smeared according to the measured resolution uncertainties, before the track selection cuts are applied. The actual smearing is done in $1/p_T$, which has a gaussian distribution, with an upper bound to the resolution uncertainty given by $\sigma(1/p_T) = 0.02/p_T$ [68]. The effect is found to be negligible.

V. Jet energy scale and momentum resolution

The jet energy scale (JES) uncertainty for light jets reconstructed with the anti-k_t algorithm with distance parameter $R = 0.4$ and calibrated to the EM+JES scale is between ∼4% at low p_T and ∼2.5% for jets with p_T >60 GeV in the central region [110]. In the case of b-jets, an additional uncertainty arising from the modelling of the b-quark production mechanism and the b-quark fragmentation was determined from systematic variations of the Monte Carlo simulation. The resulting fractional additional JES uncertainty for b-jets has an upper bound of 2% for jets with p_T ≤100 GeV and it is below 1% for higher p_T jets. To obtain the overall b-jet uncertainty this needs to be added in quadrature to the light JES uncertainty.

The systematic uncertainty originating from the jet energy scale is obtained by scaling the p_T of each jet in the simulation up and down by one standard deviation according to the uncertainty of the JES. The result is shown in Fig. 6.10a for a medium p_T bin. The effect on the likelihood performance is an average variation of 5% for the 50% and 60% efficiency working points.

The jet momentum resolution was measured for 2011 data and found to be in agreement with the predictions from the PYTHIA-based simulation [111]. The precision of this measurement, determined in p_T and η bins, is typically 10%. The systematic uncertainty due to the calorimeter jet p_T resolution was estimated by over-smearing the jet 4-momentum in the simulated data, without changing jet η or ϕ angles. The performance, shown in Fig. 6.10b, is found to globally decrease by 5%, without a particular p_T dependence.

The different contributions to the systematic uncertainty on the merged b-jet rejection are summarized in Table 6.2.

Although the likelihood training was peformed in EM+JES calibrated jets, the performance of the tagger was also evaluated in jets calibrated with

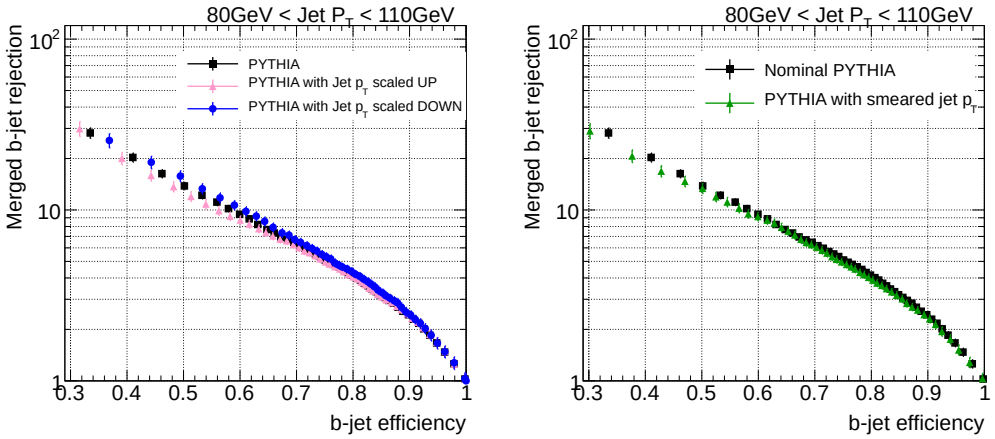

Figure 6.10: Rejection of merged b-jets as a function of single b-jet efficiency for (a) jets with smeared p_T and (b) for jets with varied energy scale compared to nominal.

Systematic source	Uncertainty
pile-up	2%
b-tagging efficiency	neglible
track reconstruction efficiency	4%
track p_T resolution	neglible
jet p_T resolution	5%
jet energy scale	5%

Table 6.2: Systematic uncertainties in the merged b-jet rejection (common to both the 50% and the 60% efficiency working points).

the LC+JES scheme, described in Section 4.1. A small degradation of the performance is observed, but comparable with the statistical uncertanties. A comparison of the performances is shown in Fig. 6.11 for two p_T bins, representative of the jet momentum range covered.

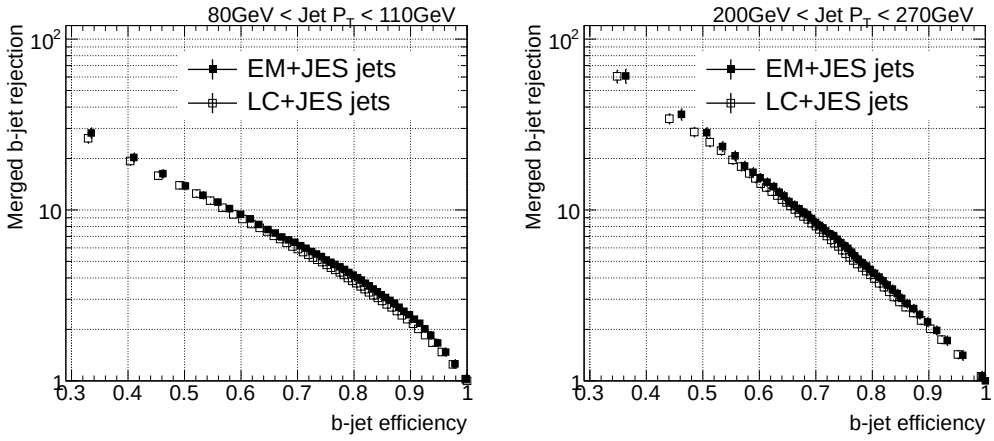

Figure 6.11: Rejection of merged b-jets as a function of single b-jet efficiency for jets calibrated to the EM+JES (LC+JES) scale, between 80 GeV and 110 GeV and 200 GeV and 270 GeV.

6.5 Other Monte Carlo generators

The development, training and performance determination of the tagger has been done using Monte Carlo events generated with the PYTHIA event simulator, interfaced to the GEANT4 based simulation of the ATLAS detector. An immediate question is what the performance would be if studied with a different simulation. In this section we investigate this question for the PYTHIA Perugia tune and the HERWIG++ event generators (Section 2.3).

Fig. 6.12 shows a comparison of the likelihood rejection, at the 50% efficiency working point, between nominal PYTHIA and the alternative simulations as a function of the jet p_T . The larger errors are due to the reduced statistics available, which are even lower for the Perugia case than for HERWIG.

The performance in HERWIG shows a systematic trend, with agreement at low p_T and increasingly poorer performances compared to PYTHIA as p_T

grows. For the Perugia tune, on the other hand, there is no definite behavior, with the performance fluctuating above or below the nominal simulation for different p_T bins consistently with the statistical uncertainties.

The reason for the systematic difference observed between the performances of PYTHIA and HERWIG can be traced to the extent with which jets are accurately modelled. Fig. 6.13 compares the measured jet track multiplicity distributions in b-tagged jets and the prediction from both simulations, for low and high p_T jets. It is observed that indeed HERWIG++ does not correctly reproduce the data, particularly at high p_T. The level of agreement is found to be better for track-jet width and the ΔR between the axes of the two k_t subjets in the jet, the two other variables used for discrimination.

The more accurate description of b-jet substructure in PYTHIA than in HERWIG++ had been previously observed in ATLAS. It is due to the inclusion in PYTHIA of a detailed study of the b-quark fragmentation function based on LEP and SLD data [110].

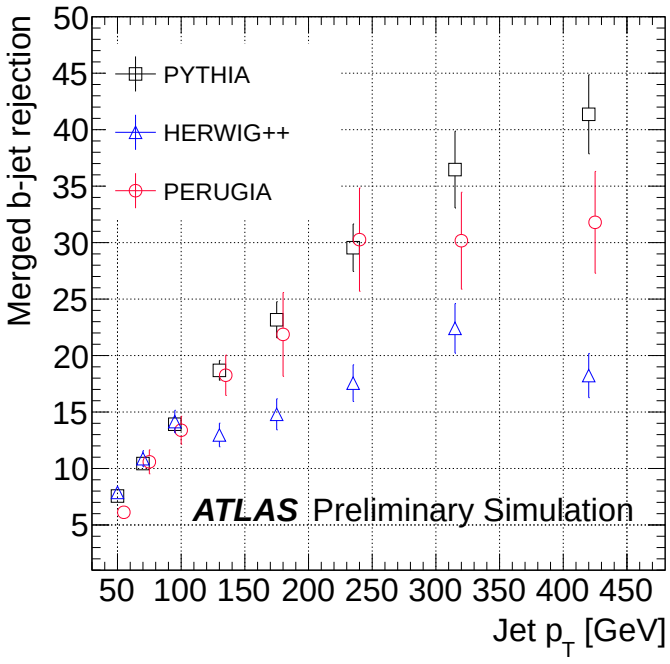

Figure 6.12: Rejection of merged b-jets as a function of jet p_T for different Monte Carlo generators, at the 50% efficiency working point. The Perugia points are shifted by +5 GeV for display purposes.

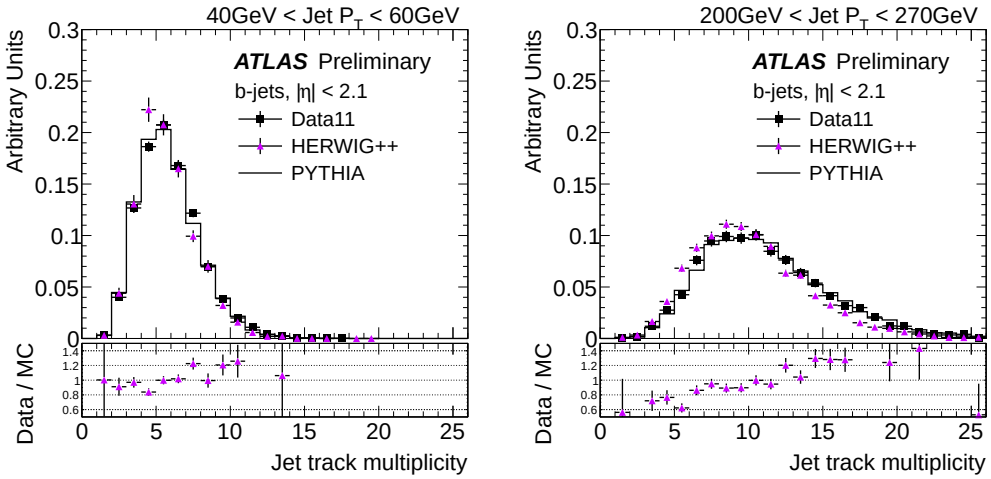

Figure 6.13: Distribution of the jet track multiplicity in 2 different jet p_T bins, for experimental data collected during 2011 (solid black points) and HERWIG++ events (solid violet triangules). The ratio data over HERWIG++ simulation is shown at the bottom of the plot. PYTHIA distribution is also shown for reference.

Chapter 7

Fraction of double b-hadron jets in QCD b-production

In this chapter we apply the newly developed $g \to b\bar{b}$ tagging tool to measure the fraction of merged b-jets in QCD b-jet production. The fractions are determined both for an inclusive b-jets sample with $|\eta| < 2.1$, and for exclusive samples enriched in single or in merged b-jets. The measured fractions are in excellent agreement, within the experimental uncertainties, with the theoretical predictions from a simulation of hadronic collisions.

7.1 Introduction

The $g \to b\bar{b}$ tagger developed and described in the previous chapters produces for every b-tagged jet a number between 0 and 1, the double b-hadron likelihood (LL). The closer this number is to 1 (0), the more likely the b-tagged jet is single (merged). When used as a tagger, a working point (Wp) is chosen so that if LL$\geq$Wp the jet is flagged as single. The value of Wp is chosen as a compromise between good efficiency (the lower the Wp, the

lower the probability that an actual single b-jet will be missed by the tagger), and rejection power (the higher the Wp the lower the probability that a non-single b-jet will be incorrectly flagged as single). Depending on the necessities of the particular analysis, an appropriate Wp is to be chosen from the plot in Figure 6.5. In particular, the performance results presented in Chapter 6 correspond to the 50% and 60% efficiency working points, two reasonable choices.

However, the values of LL in a given sample offer more information than just a jet-by-jet tagger: the distribution of LL allows to measure the composition of the particular sample. In effect, a b-tagged jet has a certain probability to actually originate from the hadronization of:

- b: a b-quark
- $b\bar{b}$: gluon splitting into a $b\bar{b}$ pair
- c: a c-quark
- $c\bar{c}$: gluon splitting into a $c\bar{c}$ pair
- ℓ: a light parton (u, d, s quarks, or a g not splitting into heavy flavor)

The expected distribution of LL is different for each of the five cases. This is illustrated in Figure 7.1, which plots LL for each hypothesis for jets in a p_T range representative of the energy covered in the analysis ($80 < p_T \leq 110$ GeV). These distributions are henceforth called "templates". The shape of the distributions in Figure 7.1 can be intuitively understood. The b and $b\bar{b}$ templates behave as expected, respectively peaking at high and low values of LL. The c template resembles its b counterpart. The $c\bar{c}$ although similar to the $b\bar{b}$ template, also exhibits a spike for large values of LL. This was to be expected, given that c-jets have less tracks than b-jets: they decay to D

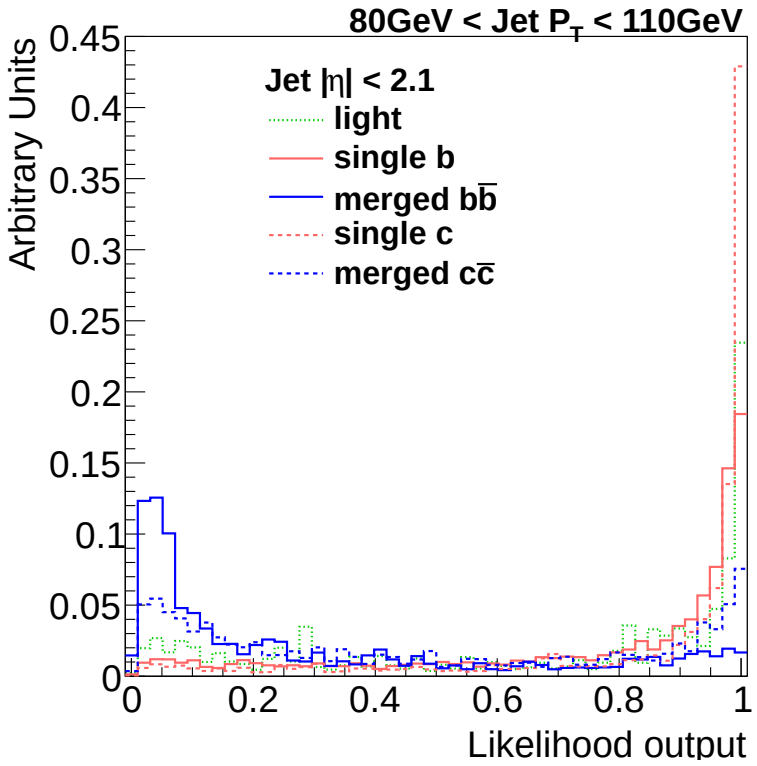

Figure 7.1: Likelihood distribution for the five flavours, b, c, $b\bar{b}$, $c\bar{c}$ and ℓ are shown for b-tagged jets in the simulated QCD sample. The templates are all normalized to unit area to allow the comparison.

mesons, $c \to D \to$ light hadrons, as opposed to b-jets which present a longer decay chain, $b \to B \to D \to$ light hadrons. They also show smaller angular separation (width), since $m_D < m_B$ ($m_D \sim 1.9$ GeV and $m_B \sim 5.3$ GeV). The template of light jets has a larger peak at LL$\to 1$ than for single b-jets. This can be traced to the observation that gluons and light-quarks jets tend to be narrower and have lower track multiplicity than heavy-flavour jets, leading to a more single-like likelihood distributions (see Section 5.4).

One can determine the composition of a given sample by measuring the values of LL of the b-tagged jets and estimating the fractions needed from

each of the templates to accurately describe the experimental LL distribution. This process is known as "template fitting", see Section 7.3. The measured composition can then be compared to the theoretical prediction from QCD. Figure 7.2 shows the expected fractions as a function of p_T from a PYTHIA

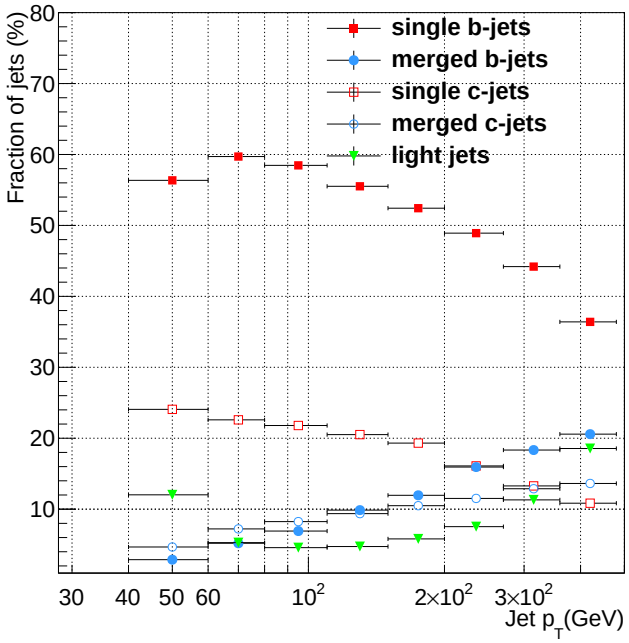

Figure 7.2: Theoretical predictions of the fractions of b-, $b\bar{b}$-, c-, $c\bar{c}$-, and ℓ-jets in b-tagged jets from a PYTHIA simulation of QCD jet production.

simulation of QCD jet production in pp collisions at 7 TeV. We observe that the fraction of single (merged) b-jets tends to decrease (increase) with p_T. This is expected as, with increasing p_T, the larger the boost of the $b\bar{b}$ pair produced by gluon splitting and the higher the probability that they will be produced at a small angle and reconstructed within the same jet. The fraction of merged jets is slightly smaller for the first p_T bin because the b-tagging efficiency drops at low p_T, bringing about a larger fraction of light jets.

7.2 Unbinned maximum likelihood fits

The analysis of experimental data often involves the estimation of the composition of a sample, based on Monte Carlo description of the various sources. We measure a number of observables x_i and we want to determine one or more parameters p_i from the data, such as the number of signal and background events. The distribution of the observables is described by a probability density function (PDF), which is a function of the parameters, $F(\vec{x}, \vec{p})$. We choose the PDF based on some hypothesis about what function would match the data, and vary the parameters in order to make the PDF match the distribution of the observables as well as possible.

In the case of data binned into a histogram, one approach is to use a least-squares fitting technique to estimate the parameters. They are adjusted to minimize

$$\chi^2 = \sum_i^n \frac{(d_i - f_i)^2}{d_i} \tag{7.1}$$

where d_i is the number of data events that fall into bin i; n is the number of bins, and f_i is the predicted number of events in bin i, defined by

$$f_i = N_D \sum_{j=1}^{m} p_j \cdot a_{ij}/N_j \tag{7.2}$$

with p_j, the proportions of the different m sources; a_{ij}, the number of Monte Carlo events from source j in bin i, with $i = 1, 2, ..., n$; N_D, the total number of events in the data sample; and N_j, the total number in the MC sample for source j.

This χ^2 assumes that the distribution for d_i is Gaussian and that a_{ij} has no uncertainty. The distribution of d_i is of course Poisson, but the Gaussian $N(\mu{=}d_i, \sigma{=}\sqrt{d_i})$ is a good approximation for large d_i. Unfortunately it often happens that many of the d_i are small, making the χ^2 value given

in Equation 7.1 inappropiate to describe the problem. Instead one can go back to the original Poisson distribution, and write down the probability for observing a particular d_i as

$$e^{-f_i}\frac{f_i^{d_i}}{d_i!} \tag{7.3}$$

and the estimates of the proportions p_j are found by maximizing the total likelihood,

$$\mathcal{L} = \prod_{i=1}^{n} e^{-f_i}\frac{f_i^{d_i}}{d_i!}. \tag{7.4}$$

This accounts correctly for bins with small numbers of events. It is often referred to as a "binned maximum likelihood" fit. Actually this formalism does not account for fluctuations in the a_{ij} due to finite Monte Carlo samples. A similar methodology that correctly describes this scenario exists, see Ref. [112]. The effects of finite MC data size can be considered small for MC samples ten times larger than the data sample.

The technique of binned maximum likelihood fit is fast and analytical; however, we observed that the obtained uncertainties were unnaturally large. This was traced to the use of events with rather different weights. In effect, Ref. [112] reports that this method only works satisfactory with weighted events if the weights do not differ very much [112]. We had thus to move to a different, more general technique, an "unbinned maximum likelihood fit", which allows arbitrary weights and has the further advantage of using all the information contained in the data sample; although it is not analytical but numerical and iterative.

The likelihood to be maximized in an unbinned dataset of events $\{\vec{x}_k\}_{k=1}^{N}$ is the product of the $F(\vec{x},\vec{p})$ PDF over all events

$$\mathcal{L}(\vec{x};\vec{p}) = \prod_{k=1}^{N} F(\vec{x}_k;\vec{p}) \tag{7.5}$$

which can be rewritten in terms of the probability distributions of observing an event from source j in the sample,

$$\mathcal{L} = \prod_{k=1}^{N} \sum_{j=1}^{m} p_j \mathcal{T}_j(\vec{x}_k) \tag{7.6}$$

where $\mathcal{T}_j$ are the PDFs that represent the distribution of $\vec{x}_k$ for each of the m hypothesis (the "templates"), p_j are the parameters representing the proportions for the j^{th} hypothesis, and N is the total number of input data points.

The PDF in Equation 7.6 is the sum of multiple probability density functions. Mathematically, the sum of two probability density functions is also a normalized probability density function as long as the coeffients add up to 1

$$F(\vec{x}_k) = p_0 \cdot \mathcal{T}_0(\vec{x}_k) + p_1 \cdot \mathcal{T}_1(\vec{x}_k) + ... + p_{m-1}\mathcal{T}_{m-1}(\vec{x}_k) + (1 - \sum_{i=1}^{m-1} p_i)\mathcal{T}_m(\vec{x}_k). \tag{7.7}$$

If the sum of these coefficients becomes larger than one, the remainding coefficient will be assigned a negative fraction. As long as the summed p.d.f is greater than zero everywhere, this is not ill-defined. In the case of the present analysis $\vec{x}_k$ represents a one-dimensional variable, the double b-hadron likelihood LL.

The fits were performed in this thesis by means of the RooFit Toolkit for data modelling [113]. Performing a fit consists of minimizing the negative log-likelihood of a PDF calculated over the data set

$$-\log \mathcal{L}(\vec{p}) = \sum_{k} F(\vec{x}_k; \vec{p}) \tag{7.8}$$

with respect to the model's parameters. The RooFitTools package uses the MINUIT[114] algorithms to find the minimum of this function and estimate the errors in each parameter.

7.3 Measurement for the inclusive QCD sample

Likelihood Monte Carlo templates were derived from the simulated QCD samples described in Section 5.2, using all jets passing the selection criteria defined in Section 5.3. Likelihood templates were constructed for b, c, $b\bar{b}$, $c\bar{c}$ and ℓ jets separately, and these were fitted to the likelihood distribution in data in order to respectively obtain the fractions of single b, merged b, single c, merged c and light jets in the QCD sample. Merged c-jets (single c-jets) are defined as those matching exactly two (only one) D hadrons, the products of the fragmentation of c-quarks. A jet is classified as light when it has no B nor D hadrons within a cone of 0.4 around its axis.

The likelihood template fits are performed using the unbinned maximum likelihood technique (see Section 7.2) separately for each p_T bin. The functional form used is

$$F(x) = p_s \cdot S(x) + p_m \cdot M(x) + p_\ell \cdot L(x) + p_{sc} \cdot S_c(x) + p_{mc} \cdot M_c(x) \quad (7.9)$$

with $S(x)$, $M(x)$, $L(x)$, $S_c(x)$ and $M_c(x)$ corresponding to the template distributions for the different hypothesis; and p_s, p_m, p_ℓ, p_{sc} and p_{mc} the parameters representing the respective fractions of expected events.

The fit to the full sample of 4.7 fb^{-1} of pp collision data at $\sqrt{s} = 7$ TeV collected by the ATLAS experiment during 2011 is displayed in Figures 7.3 and 7.4 for two representative p_T bins. The vertical scale is enlarged in the lower panels to better appreciate all contributions. It can be observed that the quality of the fit is excellent.

The fit results are summarized in Table 7.1 for all p_T bins, together with the theoretical prediction from a PYTHIA MC simulation of QCD b-jet production (Section 2.3). The errors shown are statistical, and they arise from the finite statistics of the data and template samples. The level of agreement

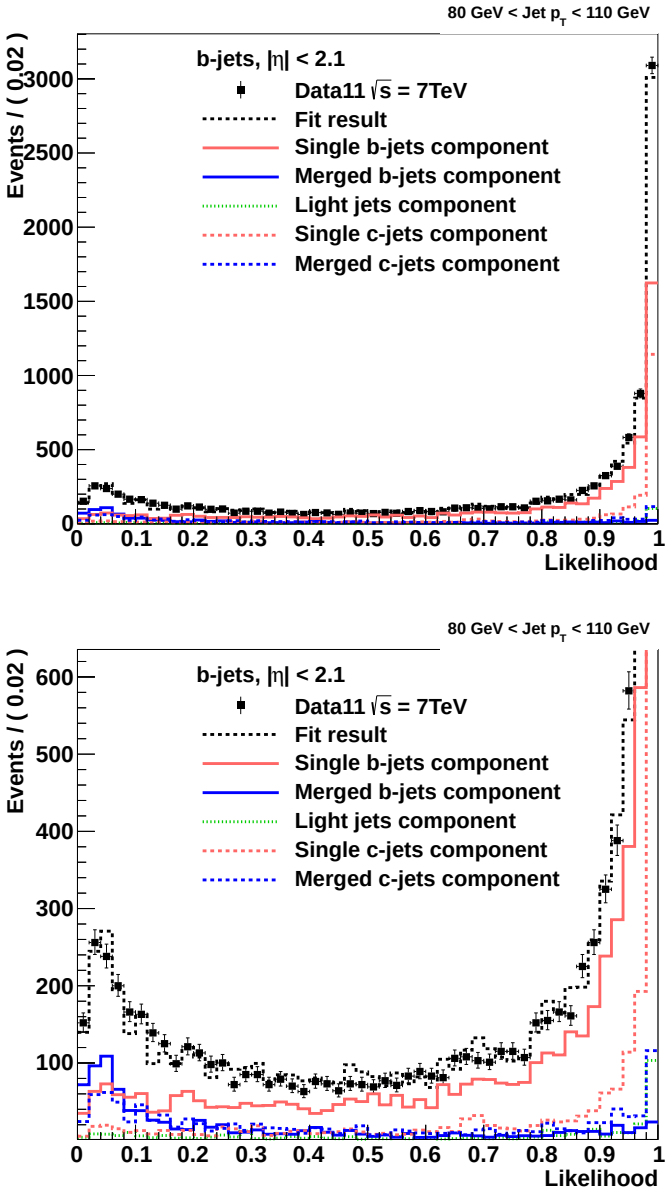

Figure 7.3: Example result of a template fit to the likelihood distribution in data. The fit is shown for jets with p_T between 80 GeV and 110 GeV, in full scale (top) and zooming the vertical scale, to better display the flavour content of the data (bottom). Uncertainties shown are statistical only.

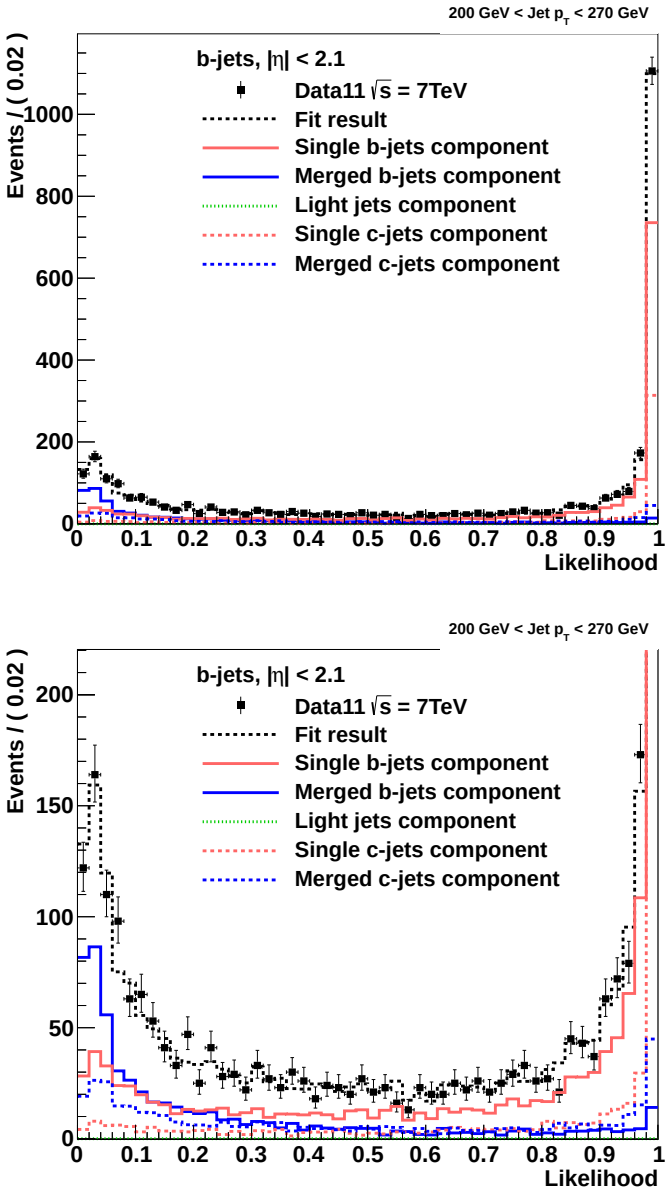

Figure 7.4: Example result of a template fit to the likelihood distribution in data. The result is shown for jets with p_T between 200 GeV and 270 GeV, in full scale (top) and zooming the vertical scale, to better display the flavour content of the data (bottom). Uncertainties shown are statistical only.

is quantified in the Table by the pulls, which correspond to the difference between data and theory normalized to the total uncertainty (pull = (data-theory)/$\sqrt{\text{stat}^2 + \text{syst}^2}$), where the systematic errors are to be discussed in Section 7.4.

The measured composition of b-jets is observed to be in very good agreement with the prediction from the simulation. The single b-jet fraction decreases as expected with p_T, from 63% to 42%, while the merged b-jet fraction increases, from 8% to 20%. The fraction of light jets is the smallest and, except at the higher p_T bins, consistent with zero. The larger fraction of light jets at high p_T is expected from the simulation and is due to the increasing difficulty to efficiently tag b-jets in the boosted regime where the tracks associated to a jet tend to be collimated within a narrow angle.

7.4 Systematic uncertainties

The systematic uncertainties affecting the compostion measurement are mainly those that change the shape of the likelihood templates. The following sources were evaluated:

- track reconstruction efficiency;
- jet transverse momentum resolution
- jet energy scale.

In order to calculate the contribution from the uncertainty in the track reconstruction efficiency a random fraction of the tracks were discarded following the procedure described in Section 6.4. New likelihood templates were produced from the modified events and the fits redone with them.

Jet p_T	single b-jet			merged b-jet			light jet			single c-jet			merged c-jet		
(GeV)	data	theory	pull	data	theory	pull	data	theory	pull	data	theory	pull	data	theory	pull
40 - 60	63±4	56	1.49	8±1	3	2.10	4±4	12	-1.65	31±2	24	2.51	-7±3	5	-2.82
60 - 80	59±2	60	-0.13	5±1	5	0.02	2±2	5	-1.15	26±1	23	1.26	8±2	7	0.30
80 - 110	56±3	58	-0.77	12±1	7	1.80	3±2	5	-0.35	24±2	22	0.66	5±2	8	-0.88
110 - 150	49±5	56	-1.23	12±2	10	0.69	-1±4	5	-1.22	25±4	21	1.01	5±3	9	1.35
150 - 200	47±5	52	-2.64	13±2	12	0.25	-1±4	6	-1.58	31±4	19	2.74	10±3	10	2.49
200 - 270	51±12	49	0.21	19±3	16	0.77	1±7	8	-0.89	19±8	16	0.32	10±5	12	-0.22
270 - 360	51±4	44	1.34	22±1	18	1.45	7±4	11	-0.95	13±3	13	-0.14	8±2	13	-1.64
360 - 480	42±7	36	0.73	19±1	21	-0.64	13±6	19	-0.82	9±5	11	-0.33	17±2	14	1.02

Table 7.1: Measured proportions (in percentage) of the composition of b-jets in QCD production, compared to the theoretical prediction from a MC PYTHIA simulation. Errors shown are only statistical. The pull corresponds to (data-theory)/$\sqrt{\text{stat}^2 + \text{syst}^2}$

The systematic uncertainty originating from the jet p_T resolution is obtained by over-smearing the calorimeter jet p_T in the simulation. The likelihood templates were rederived from this smeared sample, and the likelihood distribution in data fit using these altered samples. The difference between the unsmeared and the smeared scenarios is taken as the systematic error.

The uncertainty originating from the jet energy scale is obtained by scaling the p_T of each jet in the simulation up and down by one standard deviation, according to the uncertainty of the jet energy scale (see Section 6.4), and redoing the likelihood fits on data with the modified templates.

The resulting systematic uncertainties are summarized in Table 7.2. The largest contributions arise from the jet energy scale and jet transverse momentum resolution.

Systematic source	Uncertainty
track reconstruction efficiency	negligible
jet p_T resolution	1%
jet energy scale	2%

Table 7.2: Contributions to the systematic uncertainties affecting the template fitting to experimental data.

7.5 Enriched single and merged *b*-jet samples

The inclusive QCD data sample is ∼50% pure in single b-jets, according to the measurements described in Section 7.3. It is interesting to envisage the use of semi-inclusive samples, defined with adequate selection cuts to have a higher composition in single or in merged b-jets. On the one hand, this

can be used to test the theoretical picture of QCD b-jet production. On the other, it would help validate the universality of the Monte Carlo templates used for fitting.

As discussed in Section 2.5, single b-jets are produced via the Flavour Creation (FCR) and Flavour Excitation (FEX) processes. In the former two heavy quarks are produced in the hard scatter, yielding two hard b-quarks in the final state. The latter can be depicted as an initial state gluon splitting into a $b\bar{b}$ pair, where one of the b-quarks subsequently scatters off a parton in the opposite proton, yielding one hard b-quark and one forward b-quark in the final state. Merged b-jets are produced mostly by Gluon Splitting (GSP). In this process no heavy quarks participate in the hard scatter, but they are produced via the subsequent $g \to b\bar{b}$ branching with both b-quarks clustered in the same jet.

In the FCR process the two b-quarks give rise to a back-to-back pair of single b-jets in the transverse plane[1]. On the contrary, FEX and GSP also give rise to two back-to-back jets, but with only one of them containing heavy flavor. This picture suggests that requiring events with two b-tags is an efficient way to build a sample enriched in single b-jets. On the other hand, requesting events with only one b-tag enriches the FEX and GSP contributions, giving rise to a light jet plus a single or merged b-jet, respectively. These two scenarios are more difficult to disentangle.

Purified sample in single b-jets

A sample enriched in single b-jets is achieved by restricting the data to events with exactly two b-tags, selected with the MV1 tagging algorithm at its

[1] The presence of jets from initial and final state radiation distorts this simplified picture. This sentence should thus be understood as a first order description of the process.

60% working point and satisfying the event and jet selection described in Section 5.3. In order to increase the statistics no requirement on the p_T of the second b-tagged jet is imposed. This sample is expected to be very pure, albeit with some contamination from FEX and GSP due to mistags, where the accompanying jet is incorrectly tagged as a b.

Jet p_T	single b-jets			merged b-jets		
(GeV)	data	theory	pull	data	theory	pull
40 - 60	99±11	84	1.37	-1±1	1	-0.74
60 - 80	82±5	87	-1.01	-3±1	1	-1.60
80 - 110	84±5	88	-0.72	2±1	1	0.37
110 - 150	86±8	85	0.02	4±2	3	0.41
150 - 200	89±9	83	0.67	4±2	3	0.20
200 - 270	95±5	80	1.00	7±2	5	0.61
270 - 360	67±11	81	-1.25	12±2	6	2.21
360 - 480	73±16	73	-0.01	10±1	8	0.98

Table 7.3: Percentage of single and merged b-jets in a QCD sample enriched in single b-jets by requiring events with two b-tags. See details in the text.

The likelihood fits are performed on the purified sample, utilizing the same MC templates as for the inclusive case. The measured fractions of single and merged b-jets, together with their statistical errors and PYTHIA MC predictions for each p_T bin are displayed in Table 7.3. The agreement between theory and experiment is excellent. The fraction of single (merged) b-jets substantially increases (decreases) with respect to the inclusive case, as expected, confirming the intuitive picture discussed in Section 2.5 in terms

of Feynman diagrams.

An example of an enriched template fit is shown in Figure 7.5, comparing for the same p_T bin the inclusive and the semi-inclusive single-b enriched cases. The template description agrees very well within statistics with the data. In fact, even by a simple ocular inspection it can be ascertained that the merged-b fraction decreases, given the evident reduction in the peak at low LL in the lower panel.

Purified sample in merged b-jets

Events with only one b-tagged jet were selected for the purification of the data sample in double b-hadron jets. In order to reinforce this selection, a tight anti-b-tagging requirement on any non-tagged jet in the event was implemented. The anti-b-tagging was performed by imposing, simultaneously, strict cuts on the b-tag weights of the three supported taggers available within ATLAS:

- MV1: $w < 0.07$
- JetFitter: $w < -2$
- IP3D+SV1: $w < -2$

These weight values correspond to a MV1 tagging efficiency of more than 85%, and an efficiency for b-tagging of more than 80% for the JetFitter and IP3D+SV1 algorithms. These high efficiencies are chosen because we want to safely veto on the second jet and do not mind if the price is a high level of fake vetos.

The results are summarized in Table 7.4, which shows the measured fractions of single and merged b-jets, together with their statistical errors and

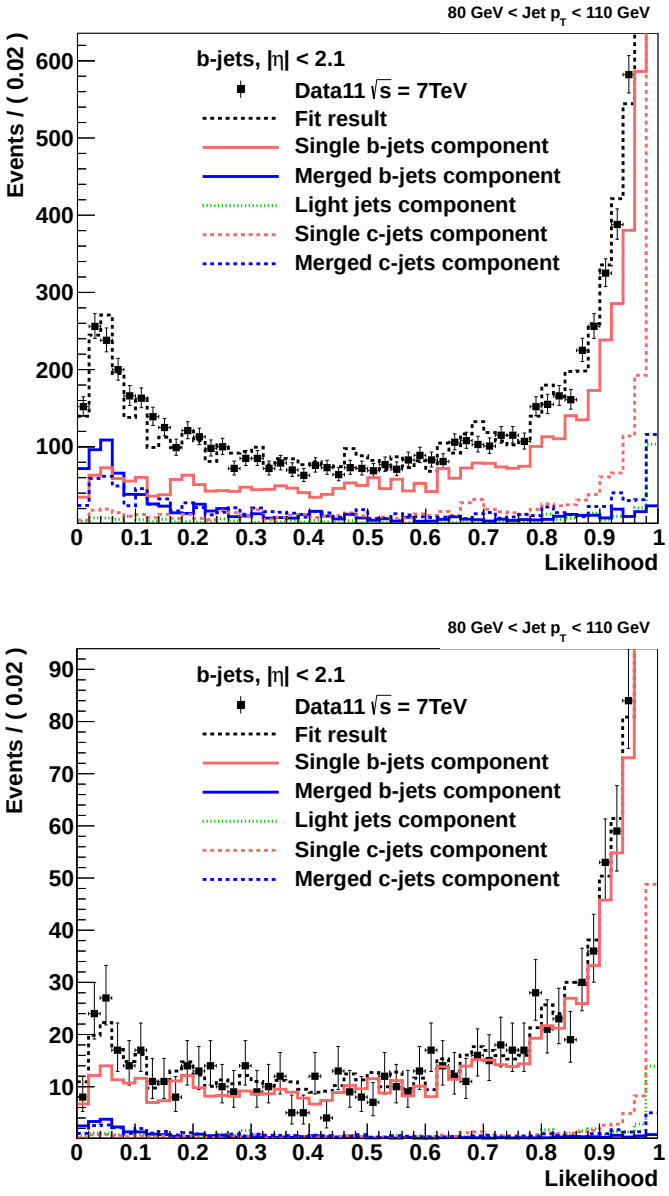

Figure 7.5: Example result of a template fit to the likelihood distribution in the semi-inclusive single-b enriched sample for jets with p_T between 80 GeV and 110 GeV (bottom). The result for the same p_T bin in the inclusive sample is also shown for comparison (top). The vertical scale was enlarged in the two plots to better appreciate all contributions. Uncertainties shown are statistical only.

PYTHIA MC predictions for each p_T bin. Although there is good agreement between data and experiment, it is clear that a much lower level of purification has been achieved than for the previous case of single b-tagged jets. This was expected, given that the selection cuts enhance not only the GSP, but also the FEX component. Some further studies were performed but with only relative success. In particular the b-tagged jet was requested to be back-to-back to the highest (or second highest) p_T jet in the event. Back-to-back jets were defined as those satisfying $\Delta R > 2.8$ and ${p_T}^{\text{tagged}}/{p_T}^{\text{leading}} > 0.7$ (if the b-tagged jets is the leading one, the cut is replaced by ${p_T}^{\text{subleading}}/{p_T}^{\text{tagged}} > 0.7$). No further improvement was found either by restricting to the case where the b-jet is the leading or the subleading one.

Jet p_T	single b-jets			merged b-jets		
(GeV)	data	theory	pull	data	theory	pull
40 - 60	58±3	53	1.50	4±1	3	0.26
60 - 80	58±1	55	1.25	6.1±0.5	5.9	0.09
80 - 110	51±1	52	-0.42	9.4±0.5	8.2	0.55
110 - 150	48±2	48	0.15	14±1	12	1.10
150 - 200	45±3	44	0.38	17±1	15	0.86
200 - 270	44±5	40	0.74	20±1	19	0.24
270 - 360	41±3	34	1.80	21±1	22	-0.30
360 - 480	32±5	29	0.59	23±1	23	-0.34

Table 7.4: Proportion (in percentage) of single and merged b-jets in a QCD sample enriched in merged b-jets by requiring events with only one b-tag. See details in the text.

Chapter 8

Summary and conclusions

In the course of the present thesis a new method was developed to identify b-jets containing two b-hadrons, which do not arise from heavy flavour production at the hard interaction but mainly via a $g \to b\bar{b}$ branching.

The method exploits the expected kinematic differences between double b-hadron ("merged") jets and single b-jets, combining a set of discriminating variables in a multivariate classifier. The differences between single and merged jets originate in the two-subjet structure of merged jets, which tend to have higher multiplicity and larger width. Several jet shape and substructure variables accounting for these envisaged characteristics were investigated in order to obtain the best single-merged discrimination. Due to the noisy environment of the hadron collisions at the LHC track-based variables were preferred over calorimeter variables.

A likelihood ratio estimator was trained using simulated QCD events. Based on discrimination power, correlation and pile-up dependence three input variables were selected for the tagger training: the jet track multiplicity, the track-jet width and the ΔR between the axes of two k_t subjets in the jet. The peformance of the tagger in Monte Carlo events was studied in bins

of the calorimeter jet p_T, achieving a rejection of merged jets of over 95% (90%) for a 50% single b-jet efficiency for jets with $p_T > 150$ GeV ($p_T > 60$ GeV). A comprehensive study of the sources of systematic uncertainties in merged b-jet rejection was performed, the most relevant being the tracking efficiency and the jet energy scale and resolution with a contribution to the uncertainty of 4%, 5% and 5%, respectively. Other sources such as pile-up or the uncertainties in the track momentum resolution and the b-jet tagging efficiency proved to be negligible.

The Monte Carlo distributions of the explored variables were validated using experimental data corresponding to an integrated luminosity of 4.7 fb^{-1} recorded by the ATLAS experiment during 2011. The agreement between data and simulation is excellent.

The tool developed was used to measure the fraction of merged b-jets in QCD b-jet production. The results obtained are in very good agreement with the theoretical prediction from a QCD parton shower simulation of pp collisions.

This tool provides a handle to investigate QCD $b\bar{b}$ production and to reduce backgrounds in Standard Model 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) or associated Higgs production ($WH \to \ell\nu b\bar{b}$ and $ZH \to \nu\nu b\bar{b}$). Jets containing a single b-quark or antiquark also enter in many BSM collider searches, 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.

In order to expand up the results presented here, and to make further advancements in the implementation of the tagger in physics analyses the following improvements should be made: the extension to non-isolated jets

using the concept of ghost-particle matching and active area of a jet for track-to-jet association and labeling and the calibration of the tagger with data. Nontheless, the study presented in this thesis demonstrates that jet substructure variables can provide a good handle for gluon splitting identification in physics searches within ATLAS.

Bibliography

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

[2] CDF Collaboration. Measurements of Bottom Anti-Bottom Azimuthal Production Correlations in Proton-Antiproton Collisions at $\sqrt{s} = 1.8$ TeV. *Phys.Rev.D*, 71:38, 2005.

[3] 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.

[4] 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.

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

[6] 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.

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

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

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

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

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

[12] 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.

[13] 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.

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

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

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

[17] H. David Politzer. Reliable perturbative results for strong interactions? *Phys. Rev. Lett.*, 30:1346–1349, Jun 1973.

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

[19] S. Catani, L. Cieri, D. de Florian, G. Ferrera, and M. Grazzini. Diphoton production at hadron colliders: a fully-differential QCD calculation at NNLO. *Phys.Rev.Lett.*, 108:072001, 2012.

[20] P. Baernreuther, M. Czakon, and A. Mitov. Percent Level Precision Physics at the Tevatron: First Genuine NNLO QCD Corrections to $q\bar{q} \to t\bar{t} + X$. *Phys.Rev.Lett.*, 109:132001, 2012.

[21] 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.

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

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

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

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

[26] 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.

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

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

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

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

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

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

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

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

[35] G. Hanson et al. Evidence for jet structure in hadron production by e^+e^- annihilation. *Phys. Rev. Lett.*, 35:1609–1612, Dec 1975.

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

[37] 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.

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

[39] 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.

[40] 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.

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

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

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

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

[45] 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.

[46] 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.

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

[48] 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.

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

[50] 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.

[51] 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.

[52] 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.

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

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

[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] D. W. Miller. Jet substructure in atlas. *ATL-PHYS-PROC-2011-142*, 2011.

[61] ATLAS Collaboration. Atlas searches for new physics with boosted objects. CERN-LHC Seminar, February 2013.

[62] Jonathan M. Butterworth, Adam R. Davison, Mathieu Rubin, and Gavin P. Salam. Jet substructure as a new Higgs search channel at the LHC. *Phys.Rev.Lett.*, 100:242001, 2008.

[63] Amos Breskin and Rudiger Voss. *The CERN Large Hadron Collider: Accelerator and Experiments.* CERN, Geneva, 2009.

[64] ATLAS Collaboration. Luminosity Determination in pp Collisions at $\sqrt(s) = 7$ TeV using the ATLAS Detector in 2011. *ATLAS-CONF-2011-116*, Aug 2011.

[65] S van der Meer. Calibration of the effective beam height in the isr. *CERN-ISR-PO-68-31. ISR-PO-68-31*, 1968.

[66] Aad, G. and others. The ATLAS Experiment at the CERN Large Hadron Collider. CERN, Geneva, 2008.

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

[68] ATLAS Collaboration. Estimating Track Momentum Resolution in Minimum Bias Events using Simulation and K_s in $\sqrt{s} = 900$ GeV collision data. *ATLAS-CONF-2010-009*, 2010.

[69] ATLAS Collaboration. Tracking Studies for b-tagging with 7 TeV Collision Data with the ATLAS Detector. *ATLAS-CONF-2010-070*, 2010.

[70] Abat, E. and others. Combined performance studies for electrons at the 2004 ATLAS combined test-beam. *Journal of Instrumentation*, 5(11):P11006, 2010.

[71] M. Aharrouche et al. Measurement of the response of the atlas liquid argon barrel calorimeter to electrons at the 2004 combined test-beam. *Nucl.Instrum.Meth.*, A614(3):400 – 432, 2010.

[72] M. Aharrouche et al. Response uniformity of the atlas liquid argon electromagnetic calorimeter. *Nuclear Instruments and Methods in Physics*

Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment, 582(2):429 – 455, 2007.

[73] Aharrouche M. and others. Energy linearity and resolution of the ATLAS electromagnetic barrel calorimeter in an electron test-beam. *Nucl.Instrum.Meth.*, A568(2):601 – 623, 2006.

[74] M. Cojocaru et al. Hadronic calibration of the atlas liquid argon end-cap calorimeter in the pseudorapidity region in beam tests. *Nucl.Instrum.Meth.*, A531(3):481 – 514, 2004.

[75] J. Pinfold et al. Performance of the atlas liquid argon endcap calorimeter in the pseudorapidity region in beam tests. *Nucl.Instrum.Meth.*, A593(3):324 – 342, 2008.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[99] 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.

[100] 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.

[101] R. Snihur. Subjet multiplicity in quark and gluon jets at d0. *Nuclear Physics B - Proceedings Supplements*, 79(1-3):494 – 496, 1999. Proceedings of the 7th International Workshop on Deep Inelastic Scattering and QCD.

[102] 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.

[103] 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.

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

[105] 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.

[106] Scott, D.W. Multivariate Density Estimation: Theory, Practice, and Visualization. *John Wiley and Sons, Inc. United States*, 1992.

[107] T. Carli and B. Koblitz. A Multivariate discrimination technique based on range searching. *Nucl.Instrum.Meth.*, A501:576–588, 2003.

[108] Quinlan, J.R. Simplifying decision trees. *International Journal of Human-Computer Studies*, 51(2):497 – 510, 1999.

[109] A. Hoecker, P. Speckmayer, J. Stelzer, J. Therhaag, E. Von Toerne, and H. Voss. TMVA: Toolkit for Multivariate Data Analysis. *PoS*, ACAT:040, 2007.

[110] Georges Aad et al. Jet energy measurement with the ATLAS detector in proton-proton collisions at $\sqrt{s} = 7$ TeV. 2011.

[111] G. Romeo, A. Schwartzman, R. Piegaia, T. Carli, and R. Teuscher. Jet energy resolution from in-situ techniques with the atlas detector using proton-proton collisions at a center of mass energy $\sqrt{s} = 7$ tev. *ATL-COM-PHYS-2011-240*, 2011.

[112] Roger Barlow and Christine Beeston. Fitting using finite monte carlo samples. *Computer Physics Communications*, 77(2):219 – 228, 1993.

[113] W Verkerke and D. Kirby. Roofit users manual v2.07. 2006.

[114] James, F and Roos, M. Function minimization and error analysis. *CERN Computer Center Program Library D*, 506, 1983.