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

María Laura González Silva

Doctoral Thesis in Physics

Physics Department

University of Buenos Aires

November 2012

UNIVERSIDAD DE BUENOS AIRES

Facultad de Ciencias Exactas y Naturales

Departamento de Física

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

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

por **María Laura González Silva**

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

Buenos Aires, Noviembre 2012

Agradecimientos

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

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

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

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

Resumen

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

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

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

Abstract

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

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

Contents

Chapter 1

Theoretical framework

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

1.1 The Standard Model

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

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

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

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

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

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

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

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

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

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

.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

The Quantum Chromodynamics (QCD) Lagrangian density is given by

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

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

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

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

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

1.2 Perturbative QCD

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

1.3 Monte Carlo tools

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

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

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

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

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

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

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

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

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

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

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

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

Pythia

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

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

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

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

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

Herwig++

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

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

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

Detector simulation

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

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

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

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

1.4 Jet physics

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

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

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

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

1.4.1 Jet algorithms

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

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

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

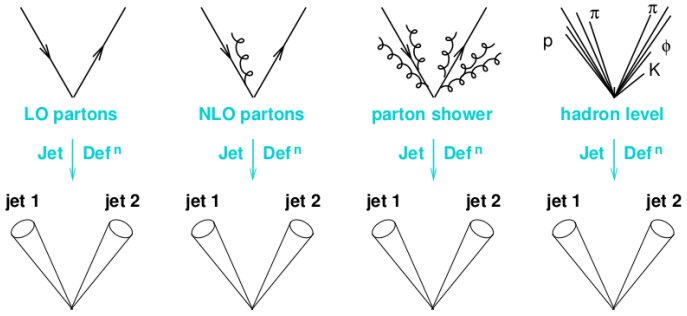

Figure 1.1: The application of a jet definition to a variety of events that differ just through soft/collinear branching and hadronization should give identical jets in all cases [35].

Fixed cone jet finder in ATLAS

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

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

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

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

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

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

Sequential recombination algorithms

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

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

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

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

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

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

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

and the beam distance,

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

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

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

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

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

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

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

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

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

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

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

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

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

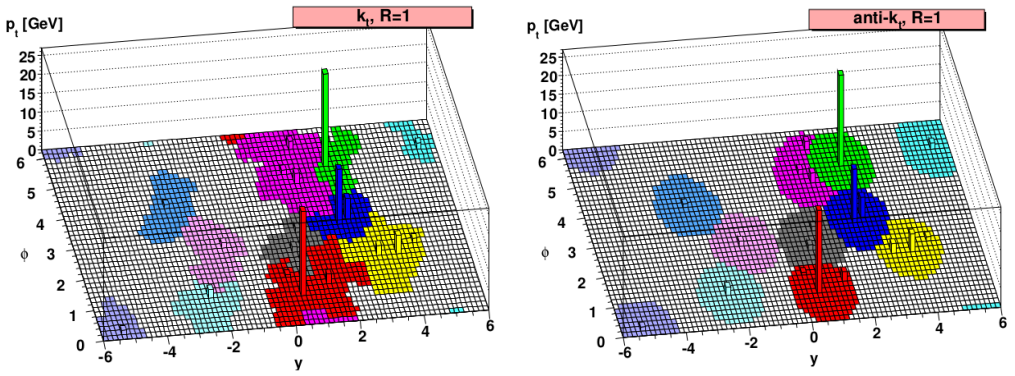

Figure 1.2: A sample parton-level event, generated with HERWIG, clustered with the k_t and anti-k_t algorithms, illustrating the active area of the resulting jets [41].

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

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

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

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

1.4.2 Jet substructure

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

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

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

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

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

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

1.5 Heavy flavor jet production

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

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

initial state (IS).

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

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

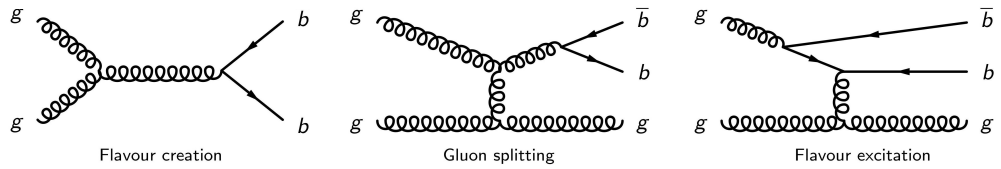

Figure 1.3: Representative diagrams of the three channels contributing to QCD b-quark production up to NLO. The flavour creation channel (left) is the only one present at LO. At NLO, two new channels open up, referred to as gluon splitting (center) and flavour excitation (right).

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

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

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

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

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

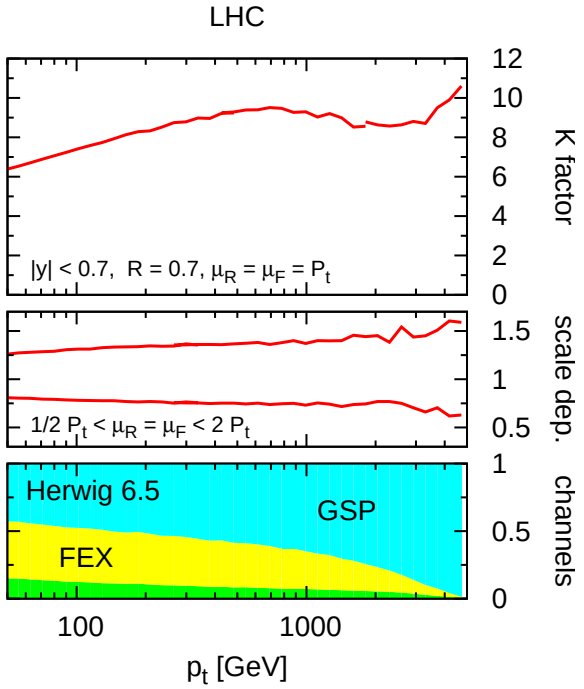

Figure 1.4: Top: K-factor for inclusive b-jet spectrum taken from [54], clustering particles into jets using the k_t jet-algorithm [38] with R=0.7, and selecting jets in the central rapidity region ($|y| < 0.7$). Middle: scale dependence obtained by simultaneously varying the renormalisation and factorisation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

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

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

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

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

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

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

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

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

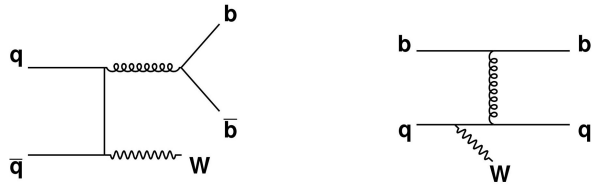

Figure 1.5: Feynman diagrams for W production in association with b quarks.

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

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

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

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

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

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

Chapter 2

The Multivariate Analysis

After the evaluation of the best discriminating variables, a tagging algorithm, capable of efficiently identifying single b-jets while rejecting merged b-jets, can be constructed using several different approaches. In this chapter we present the results of a multivariate likelihood method. Compared to single variable or 2-dimensional cuts analyses, with this technique the sensitivity is largely improved because several variables are combined to achieve the maximum separation power.

2.1 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 discribes a decision boundary. The following multivariate methods were explored:

- Likelihood ratio estimators (LLR)
- Neural Networks (NN)

- Boosted decision Trees (BDTs)

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

$$y_L(i) = \frac{L_S(i)}{L_S(i) + L_B(i)}, \tag{2.1}$$

In the case of poorly correlated variables, the likelihood for being of signal type is obtained by multiplying the signal probability densities of all input variables and normalising this by the sum of the signal and background likelihoods,

$$L_{S(B)}(i) = \prod_{k=1}^{n_{var}} p_{S(B),k}(x_k(i)), \tag{2.2}$$

Because correlations among the variables are ignored, this PDE approach is also called âĂIJnaive Bayes estimatorâĂİ. 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 idea behind the latter approach is to estimate the shape of a

[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.

PDF by the sum over smeared training events. For a PDF $p(x)$ of a variable x, one finds [60].

$$p(x) = \frac{1}{Nh} \sum_{i=i}^{N} K(\frac{x - x_i}{h}) = \frac{1}{N} \sum_{i=i}^{N} K_h(x - x_i), \tag{2.3}$$

where N is the number of training events, $K_h(t) = K(t/h)h$ is the kernel function, and h is the bandwidth of the kernel (also termed the *smoothing parameter*). For the present implementation a Gaussian form of K is used.

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.

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} .

A decision tree (BDT) is a binary tree structured classifier similar to the one sketched in Fig. 2.1. Repeated yes/no decisions are taken on one single variable at a time until a stop criterion is fulfilled. 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 [61].

Training and testing the MVA methods

A sub-set of the dijet Monte Carlo sample was used for training the three methods in the context of the Toolkit for Multivariate Data Analysis, TMVA [62], written in C++ language. After the event and jet selections were performed, the b-tagged jets with $|\eta| < 2.1$ were classified as signal (single b-jets) or background (merged b).

Based on discrimination power, correlation and pile-up dependence three variables were selected for the training: the jet track multiplicity, the track-jet width and the ΔR between the axes of two k_t subjets in the jet. Distributions of the these variables for single and merged b-jets were given to the multivariate methods as input. Other variables such as τ_2 or $\max\{\Delta R(trk, trk)\}$ were also tested leading to no gain in performance.

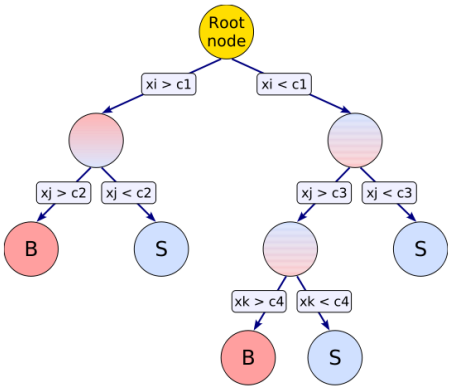

Figure 2.1: Schematic view of a decision tree. Starting from the *root node*, a sequence of binary splits using the discriminating variable x_i is applied to the data. The leaf nodes at the bottom end of the tree are labeled "S" for signal and "B" for background depending on the majority of events that end up in the respective nodes. Image taken from Reference [62].

Different training options were evaluated for the likelihood and Neural Network classifiers. The final configuration for the likelihood estimator uses the KDE approach for PDF estimation of the input variables. The NN trained is a Multi-layer perceptron (MLP) with two hidden layer of n_{var} and $n_{var} - 1$ neurons respectively, using a sigmoidal neuron activation function. The Boosted Decision Tree approach was implemented with 400 trees in the BDT forest. Distributions of the final output for the three methods, evaluated in an orthogonal sample of simulated dijet events, are shown in Fig. 2.3 for a medium p_T bin. The testing sample satisfies the same selection used for the training sample.

The outputs of the MVA discriminants explored 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

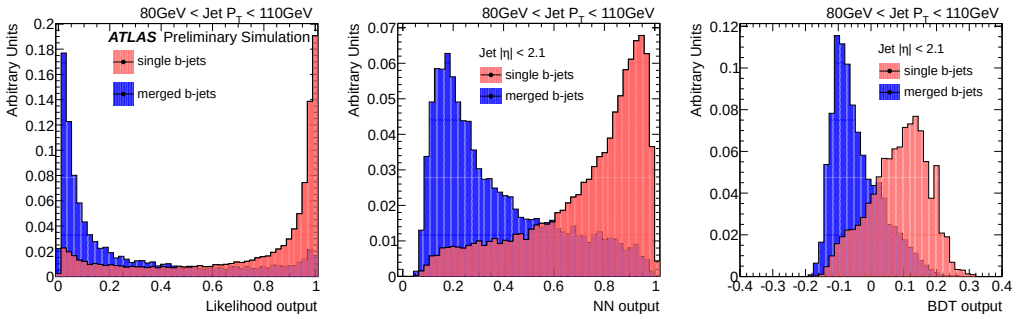

Figure 2.2: Distribution of the MVA discriminant outputs for the Likelihood (a), Neural Network (b) and Boosted Decision Trees (c) classifiers, for single and merged b-jets between 80 GeV and 110 GeV.

of the different methods, expressed in rejection of merged b-jets as a function of single b-jet efficiency (see next section), agrees within statistics, see Fig. 2.3.

As opposed to Neural Network 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 [62]. It is for these reasons that the likelihood classifier is the selected method for our tagger.

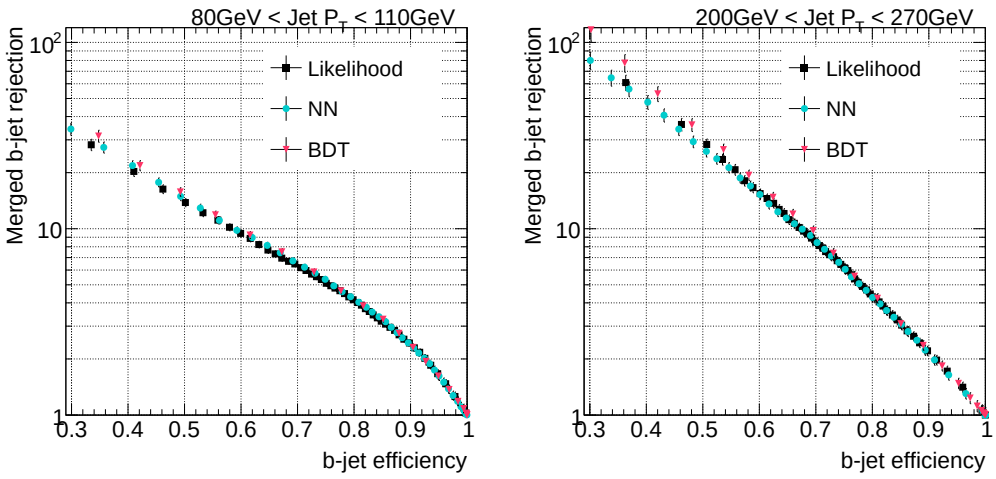

Figure 2.3: Rejection of merged b-jets as a function of single b-jet efficiency for the the different MVA methods evaluated for low and high jet p_T.

2.2 Likelihood training and performance

As indicated in section 2.1, a discriminant between single and merged b-jets was built by training a simple likelihood estimator, with the following three variables as input,

1. Jet track multiplicity
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. 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, and thus increase the statistics of merged jets in the low p_T bins. For the evaluation of the method the same procedure was followed.

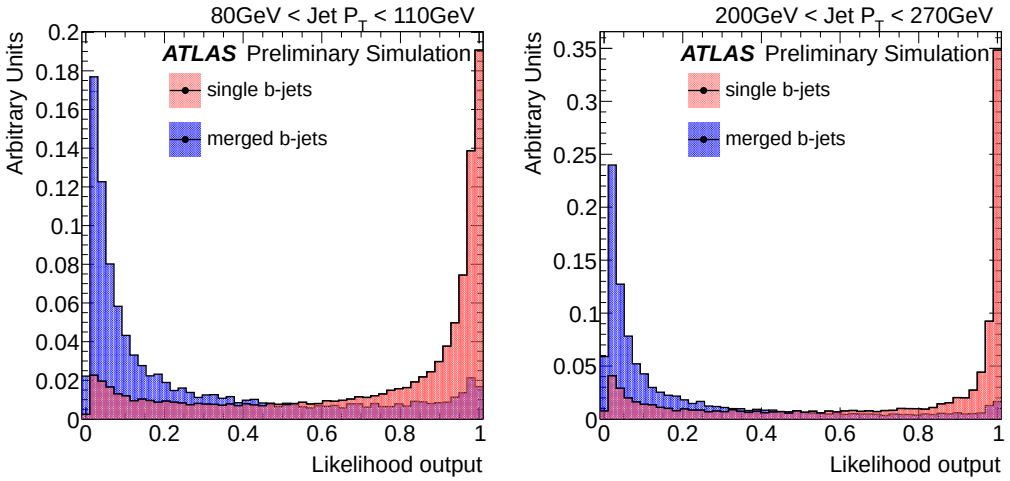

Figure 2.4: Distribution of the likelihood output for single and merged b-jets for medium and high p_T jets.

The distribution of the likelihood output for single and merged b-jets is shown in Fig. 2.4 for low, medium and high transverse momentum jets.

The performance of the tagger in the simulation can be displayed in a plot 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. This is shown in Fig. 2.5 for the eight bins of jet p_T mentioned in section **??**. The performance 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 2.1, together with their relative statistical error. These are propagated from the Poisson fluctuations of the number of events in the merged and single b

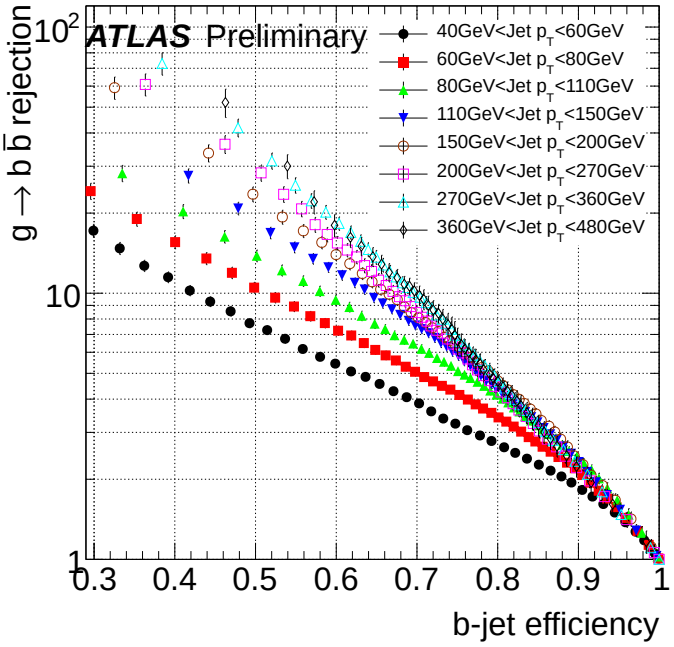

Figure 2.5: Rejection of merged b-jets as a function of single b-jet efficiency for dijet events in 8 jet p_T bins.

distributions. The error is slightly lower for the 60% efficiency working point 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 2.1: The merged b-jet rejection for the 50% and 60% efficiency working points in bins of p_T.

2.3 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 **??** 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);
- 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. 2.6 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 the 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.

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

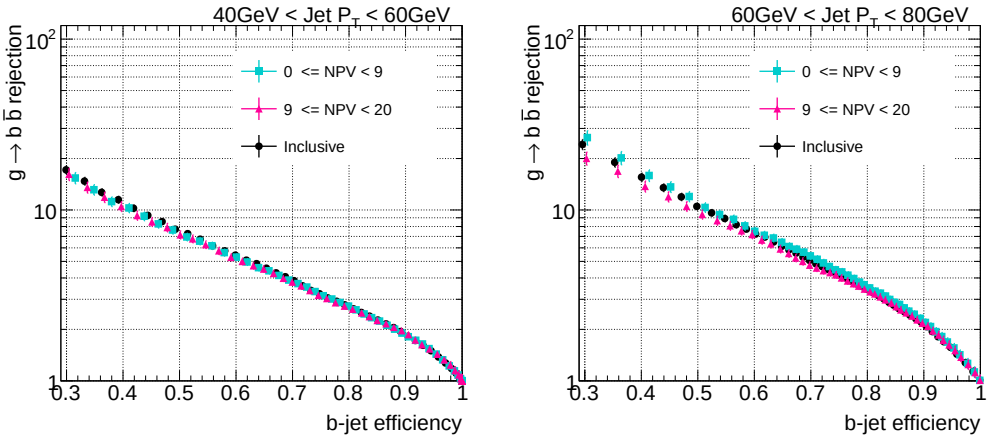

Figure 2.6: Rejection of merged b-jets as a function of single b-jet efficiency in bins of NPV for two low jet p_T bins.

by the varied SFs show no major deviations from the nominal, see Fig. 2.7. In the same manner, the effect of the b-tagging calibration uncertainty on the likelihood peformance, shown in Fig. 2.8, 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:

$$\epsilon_{trk} = \frac{N_{rec}^{matched}(p_T, \eta)}{N_{gen}(p_T, \eta)} \tag{2.4}$$

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

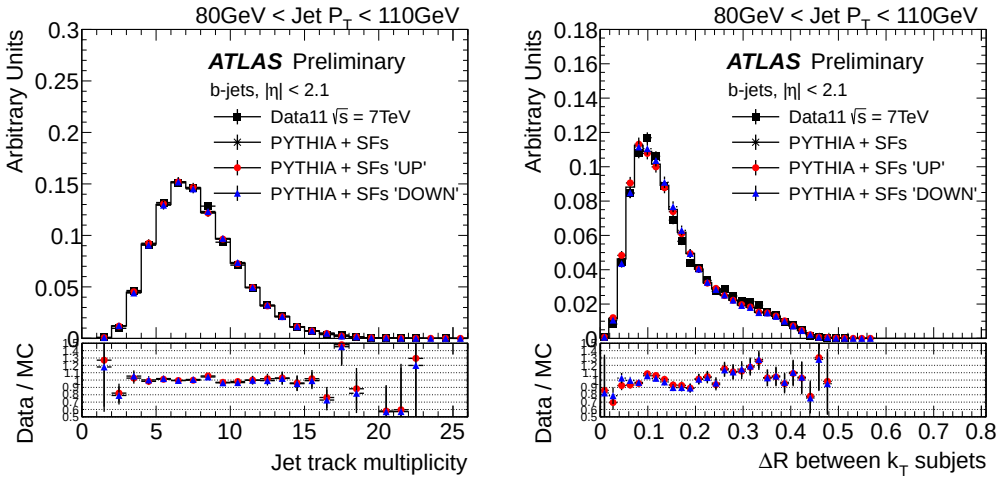

Figure 2.7: The effect of a variation in the b-tagging Scale Factors on the tracking variables distributions. Scale Factors were varied up (down) by 1-sigma to evaluate the systematic uncertainty from this source. The ratio data over MC is shown for MC PYTHIA with SFs varied up (circles) and down (triangles).

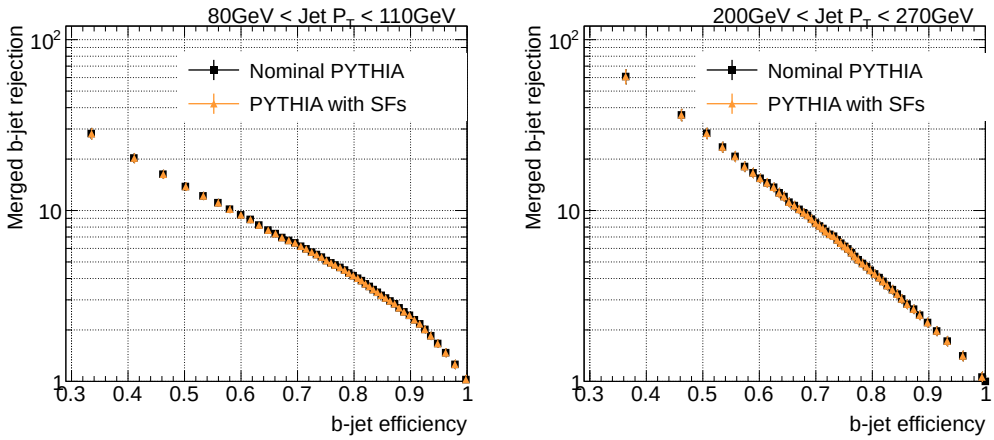

Figure 2.8: Rejection of merged b-jets as a function of single b-jet efficiency with and without scale factors as weights.

in that bin[2]. 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. The uncertain knowledge of the material in the inner detector is the main source of systematic uncertainties in the tracking efficiency [63]. 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 tracking efficiency systematics 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 $2.3 < |\eta^{\mathrm{track}}| < 2.5$ [63]. 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 systematic uncertainty of 4% both for the 50% and 60% efficiency working

[2]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.

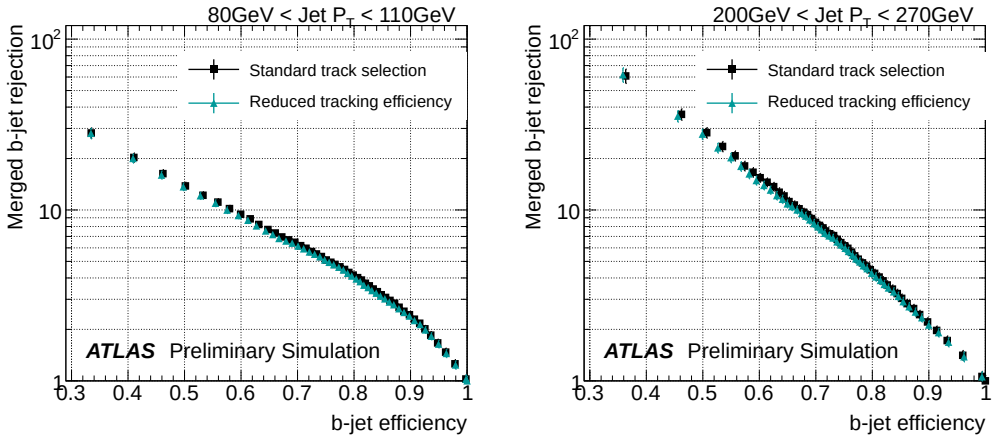

Figure 2.9: Rejection of merged b-jets as a function of the single b-jet efficiency showing shift in likelihood performance caused by a reduction in the tracking efficiency.

points. The performance comparison is shown in Fig. 2.9 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 $1/p_T$, with an upper bound to the resolution uncertainty given by $\sigma(1/p_T) = 0.02/p_T$ [64]. The effect is found to be negligible, see Fig 2.10.

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

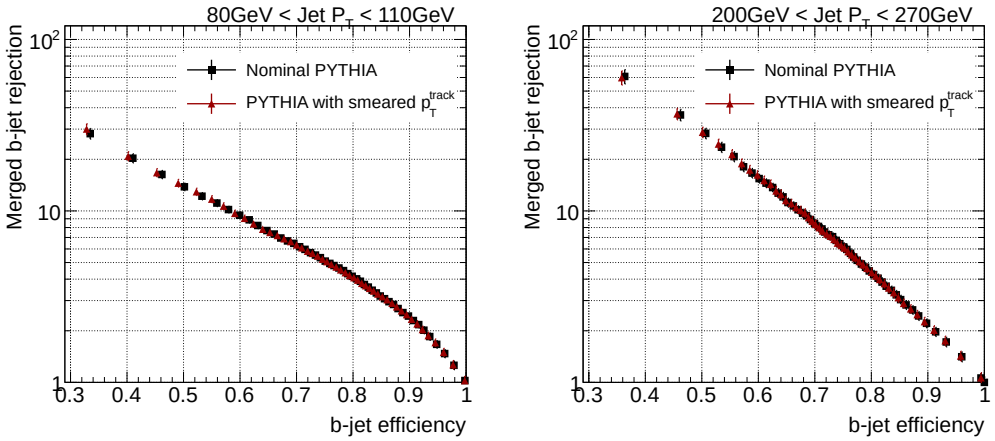

Figure 2.10: Rejection of merged b-jets as a function of the single b-jet efficiency showing the effect of the track momentum resolution uncertainty. It is found to be negligible with respect to the statistical uncertainty.

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 [65]. In the case of b-jets, and 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. 2.11a 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.

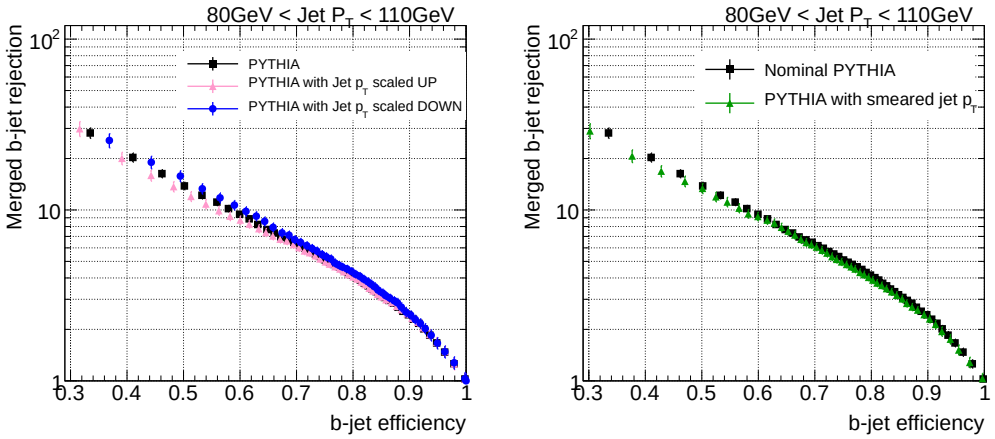

Figure 2.11: Rejection of merged b-jets as a function of single b-jet efficiency for (a) jets with smeared p_T and (b) for jets with varied energy scale compared to nominal.

The jet momentum resolution was measured for 2011 data and found to be in agreement with the predictions from the PYTHIA-based simulation [66]. 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. 2.11b, 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 2.2.

Although the likelihood training was peformed in EM+JES calibrated jets, the performance of the tagger was also evaluated in jets calibrated with the LC+JES scheme, described in Section **??**. A small degradation of the performance is observed, but comparable with the statistical uncertanties. A comparison of the performances is shown in Fig. 2.12 for two p_T bins, representative of the jet momentum range covered.

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

2.4 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 1.3).

Fig. 2.13 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,

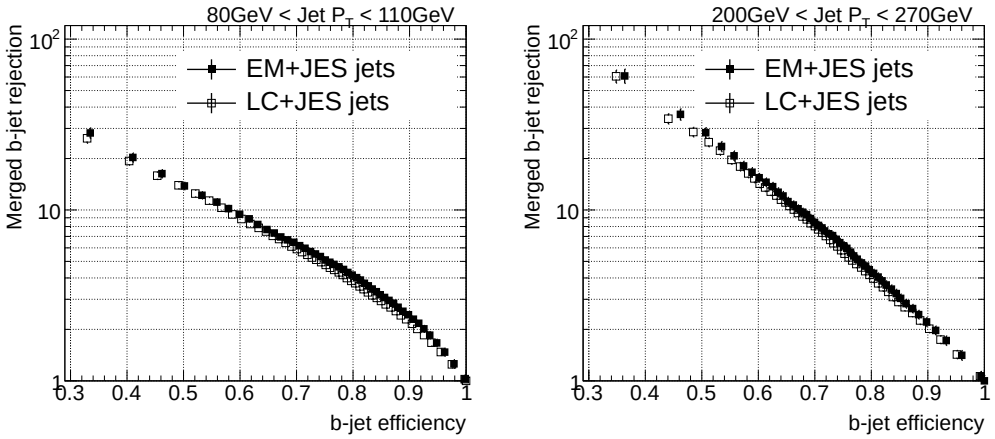

Figure 2.12: Rejection of merged b-jets as a function of single b-jet efficiency for jets calibrated to the EM+JES (LC+JES) scale, between 80 GeV and 110 GeV and 200 GeV and 270 GeV.

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. 2.14 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.

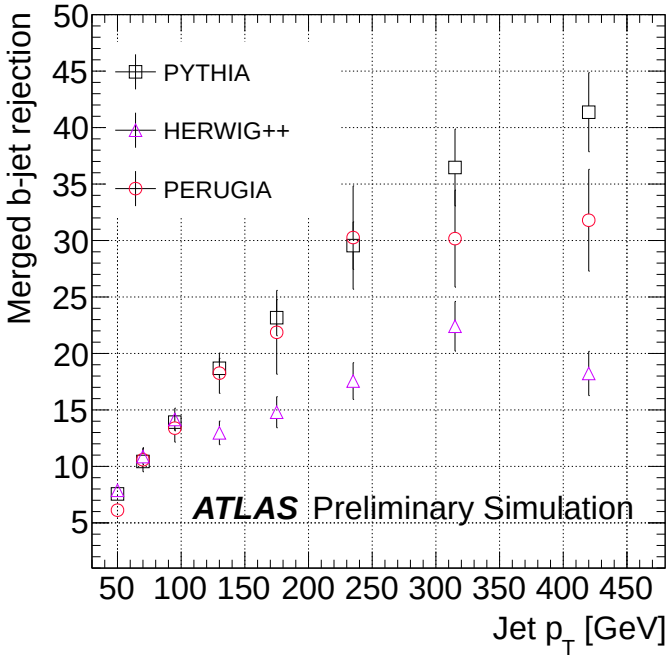

Figure 2.13: Rejection of merged b-jets as a function of jet p_T for different Monte Carlo generators, at the 50% efficiency working point.

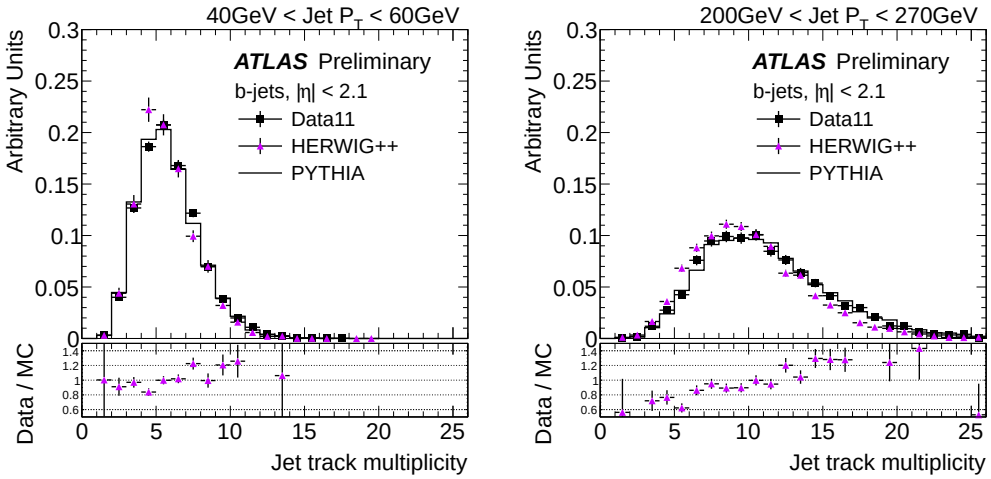

Figure 2.14: Distribution of the jet track multiplicity in 2 different jet p_T bins, for experimental data collected during 2011 (solid black points) and HERWIG++ events (solid violet triangules). The ratio data over HERWIG++ simulation is shown at the bottom of the plot. PYTHIA distribution is also shown for reference.

Bibliography

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[64] 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.

[65] Aad, G. and others. Jet energy measurement with the ATLAS detector in proton-proton collisions at $\sqrt{s} = 7$ TeV. 2011.

[66] 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.