University of Genoa

Physics Department

Master Degree Thesis

Calibration of the b-tagging efficiency on jets with charm quark for the ATLAS experiment

Advisors:
Dr. F. Parodi
Dr. C. Schiavi

Co-advisor:
Dr. S. Tosi

Candidate:
A. Lapertosa

29 October 2015

Abstract

The correct identification of jets originated from a beauty quark (b-jets) is of fundamental importance for many physics analyses performed by the ATLAS experiment, operating at the Large Hadron Collider, CERN. The efficiency to mistakenly tag a jet originated from a charm quark (c-jet) as a b-jet has been measured in data with two different methods: a first one, referred as the "D* method", uses a sample of c-jets containing reconstructed D^{*+} mesons (adopted for 7 TeV and 8 TeV data analyses), and a second one, referred as the "W+c method", uses a sample of c-jets produced in association with a W boson (studied on 7 TeV data). This thesis work focuses on some significant improvements made to the D* method, increasing the measurement precision. A study for the improvement of the W+c method and its first application to 13 TeV data is also presented: focusing on the event selection, the W+c signal yield has been considerably increased with respect to the background processes.

Contents

Introduction

The term "b-tagging" collectively describes the tools developed to identify jets generated from the hadronization of the beauty quarks (b-jets). The correct identification is of fundamental importance for the physics analyses involving b-jets in the final state: top quark studies and Higgs searches, as well as the search for new physics fenomena beyond the Standard Model.

The performance of the b-tagging algorithms is calibrated on data, measuring the efficiency to tag a jet originated from a b quark (b-jet tagging efficiency) as a b-jet or to mistakenly tag a jet originated from a charm quark (c-jet tagging efficiency) or from a light quark or a gluon (mistag rate).

The c-jet tagging efficiency was measured by the ATLAS experiment with two different techniques: using a sample of c-jets containing D^{*+} mesons (D* method, 7 TeV [46] and 8 TeV [47]) and using a sample of c-jets produced in association with a W boson (W+c method, 7 TeV [48]). This master degree thesis focuses on the improvements made to the D* method and the W+c method, in view of the measurement of the c-jet tagging efficiency on 13 TeV data.

After a brief introduction on the ATLAS detector (Sec. 1), the role of the b-tagging in the ATLAS physics analyses is discussed (Sec. 2). The methods used to measure the efficiency of the b-tagging algorithms are described, with particular attention to the D* method and the W+c method (Sec. 3).

The D* method uses a sample of c-jets containg D^{*+} mesons, reconstructed within a jet in the $D^{*+} \to D^0(\to K^-\pi^+)\pi^+$ final state. In this thesis work, a study on 8 TeV data was performed in order to improve the precision of this measurement. Different procedures were tested in order to improve the D* mass fit and the beauty flavour fraction determination: the best approaches are discussed along with their results (Sec. 4).

The W+c method uses a sample of c-jets produced in association with a W boson, with the W boson decaying into an electron and a neutrino, and the c-jet identified via a soft muon stemming from a semileptonic c-hadron decay. Exploiting the charge correlation of the two leptons (electron and muon) it is possible to extract an almost pure c-jet sample. In this thesis work, a study for the improvement of the W+c method and its first application to 13 TeV data is presented (Sec. 5).

Chapter 1

The ATLAS experiment at LHC

At the middle of the XX century, 12 European countries founded CERN (originally "Conseil Européen pour la Recherche Nucleaire") to promote international cooperation on scientific research. In the following 60 years, CERN laboratories, based on the particle accelerator technology, addressed many questions and led the scientific community torward the understanding of the atomic, nuclear and sub-nuclear physics. Nowadays, CERN is an intergovernmental organization employing around 2500 people among physicists, engineers, technichians and amministratives in its laboratories builded on the border between France and Switzerland. More than 10000 scientists from all over the world operate with its facilities, making the CERN one of the world largest centers for scientific research.

The energy required to study more and more in deep the fundamental constituents of matter often required the building of a new generation of high-energy accelerators. Currently, these accelerators are the constituents of a unique complex (Fig. 1.1), capable of increasing the speed and consequently the energy of a wide range of particles, such as protons, antiprotons, electrons and ions. At the top stage of this accelerator complex stands the Large Hadron Collider: currently the most powerful accelerator in the world.

1.1 Large Hadron Collider at CERN

The Large Hadron Collider (LHC) [1,2] is a 27 km circular particle accelerator built in the same tunnel once containing the Large Electron-Positron Collider (LEP). LHC accelerates protons or ions to perform collisions at the TeV energy scale, allowing physicists to test the theoretical predictions of many proposed particle physics models.

After the dismantling of LEP, in the period going from 2000 to 2008, the LHC was built in the tunnel at about 100 m underground. The counter-rotating beams collide in four interaction points, located in underground caverns where the equipment of the 4 main experiments stands. After accelerator commissioning and cosmic rays

studies, LHC begun operations on 10 September 2008. An accidental liquid helium leak due to a broken connection between two magnets caused a long stop to repair more than 50 damaged magnets. After 14 months, the LHC was back in November 2009 with particle collisions at the record energy of $\sqrt{s} = 2.36$ TeV in the center of mass. The energy was then increased in 2011 ($\sqrt{s} = 7$ TeV) and again in 2012 ($\sqrt{s}$ = 8 TeV). From September 2013 to March 2015, the LHC underwent the first Long Shutdown: in these months the magnet connections were renewed to get stronger for the new energy level. The LHC resumed operations in April 2015, with both beams circulating at 6.5 GeV energy. In the following months, LHC delivered collisions at the record energy of $\sqrt{s} = 13$ TeV (6.5 TeV + 6.5 TeV) in the center of mass, becoming once again the most powerful accelerator of the world. A new season of physics results is just begun: LHC will operate at 13 TeV allowing experiments to collect precious and interesting data until 2018, when the second Long Shutdown is expected.

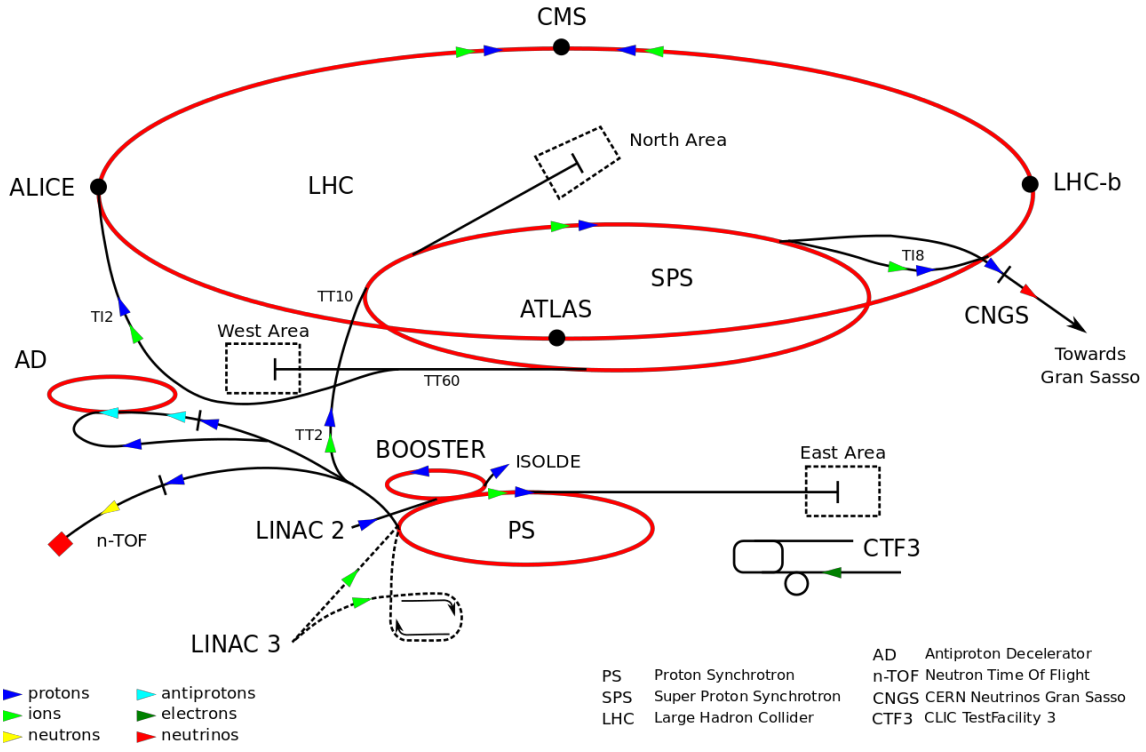

Figure 1.1: CERN accelerator complex: LINAC2, PS BOOSTER, PS and SPS operate in the pre-acceleration chain of the LHC. They also accelerate different beams for a wide range of physics experiments: AD, n-TOF, ISOLDE, CNGS.

Before entering the LHC, the protons extracted from an hydrogen source go through an acceleration chain composed by many stages:

- LINear ACcelerator 2 (LINAC 2): a 36 m long linear accelerator initially increases the energy up to 50 MeV;

- Proton Synchrotron Booster (PSB): four superimposed rings with a 25 m diameter accelerates the protons up to 1.4 GeV;
- Proton Synchrotron (PS): a 628 m long circular accelerator composes bunches of protons and brings them up to 26 GeV;
- Super Proton Synchrotron (SPS): a 2 km diameter ring accelerates the beams up to 450 GeV, before injecting them into LHC.

In future, LINAC 2 will be replaced by LINAC 4, a 80 long linear accelerator that will bring protons up to 160 MeV. PS Booster an PS will be replaced too, by Superconducting Proton Linac and by PS2: these upgrades are the only possible way to reach an higher luminosity level.

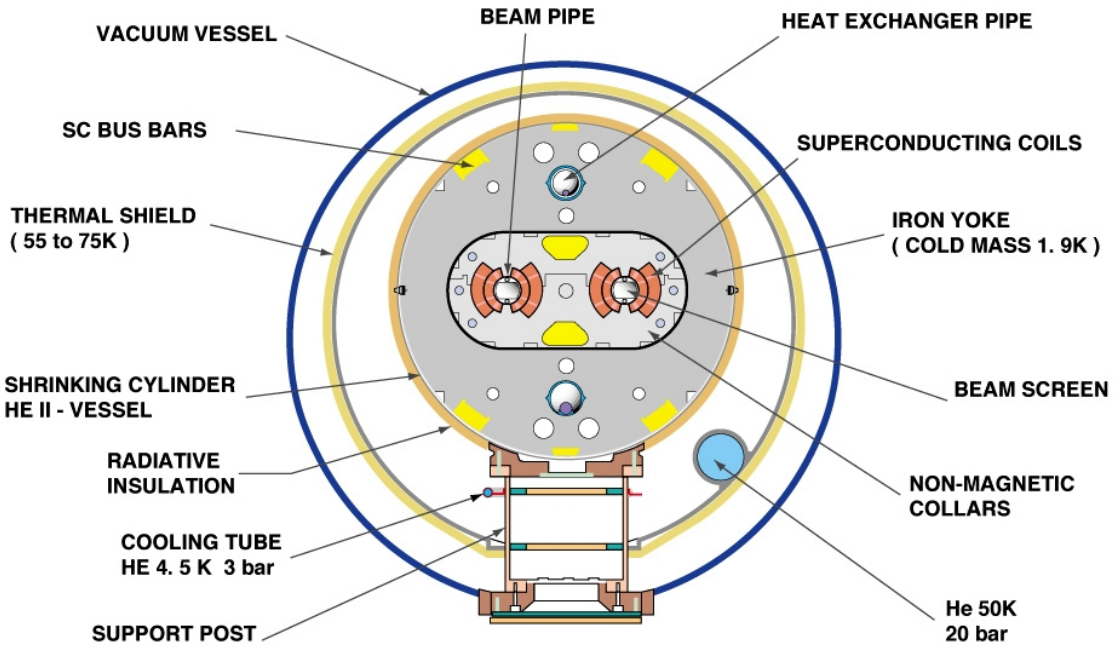

Figure 1.2: Section view of the magnetic dipole of LHC.

Two counter-rotating beams (made up by protons or lead ions) circulate in two separate vacuum chambers contained in a single beam pipe (Fig. 1.2). It takes 4 minutes and 20 seconds to accelerate enough protons to fill each LHC ring, then several minutes are needed to reach the final LHC energy. Beams at the extablished energy circulate for 10 hours or more into LHC, colliding at the 4 interaction points (ATLAS, ALICE, CMS, LHCb) until the number of protons runs out. Then the beams are dumped, the magnets are rumped down, the whole system goes through about $\sim$30 minutes of rest and a new cycle begins.

In order to reach 27 km length, more than 1200 superconducting dipoles are connected together and cooled at 1.9 K by liquid helium. These magnets generate a 8 Tesla magnetic field that bends the beams along their circular track. Almost 400

quadrupole magnets are used to guide and focus the beams, providing the beams with small transverse sections. Radiofrequency cavities are used to accelerate the beams until they reach the desired energy. Each beam is composed by many tiny little bunches of particles (a maximum of 2808 bunches, each one made up by about 10^{11} protons): foreseen by design, the bunches usually collide at 40 MHz rate, resulting in one bunch-crossing event per 25 ns.

In the LHC, the particle beams travel in two adjacent parallel vacuum chambers, rotating respectively in clockwise and counter-clockwise direction. After reaching the final energy, they collide in four experimental areas where the particle detectors of the main LHC experiments stand:

- A Large Ion Collider Experiment (ALICE): an experiment with the specific purpose of studing the quark-gluon plasma physics through ion-ion collisions, in order to gain further knowledge on the state of the matter at extreme energy density;
- A Toroidal LHC ApparatuS (ATLAS): a detailed description of the ATLAS experiment will be provided in Sec. 1.2;
- Compact Muon Solenoid (CMS): a general purpose experiment with many different goals as Higgs boson search, Super Symmetry and Dark Matter searches;
- Large Hadron Collider beauty (LHCb): an experiment focused on beauty physics, aiming toward a better understanding of the rare B hadrons phenomena and a precise measurement of the CP violation parameters.

1.2 ATLAS

ATLAS [4,5] is a multi purpose experiment located at the Interacting Point 1 (IP1) into the LHC tunnel. The ATLAS collaboration is made by more than 3000 physicist coming from 38 different countries. The ATLAS experiment was firstly proposed in 1994, then in 1999 a Tecnical Design Report [4] was defined. The detector assembly was completed in 2008, a first upgrade was performed during Long Shutdown 1 (2013-2014) and two more are expected in 2018 and in 2022.

ATLAS was designed to be a general purpose detector: proton-proton collisions produce a variety of different particles with a broad range of energies. Rather than focusing on a particular physics process, the ATLAS detector can measure an entire range of signals, in order to study whatever physical processes or new particles will emerge from LHC collisions.

The ATLAS detector is 46 m long with a 25 m diameter and its overall weight is over 7000 tons. It is composed by three different sub-detectors: the Inner Detector (Sec. 1.2.1), the Calorimeters (Sec. 1.2.2) and the Muon Spectrometer (Sec. 1.2.3). These components cover the interacting point, embracing it in the shape of

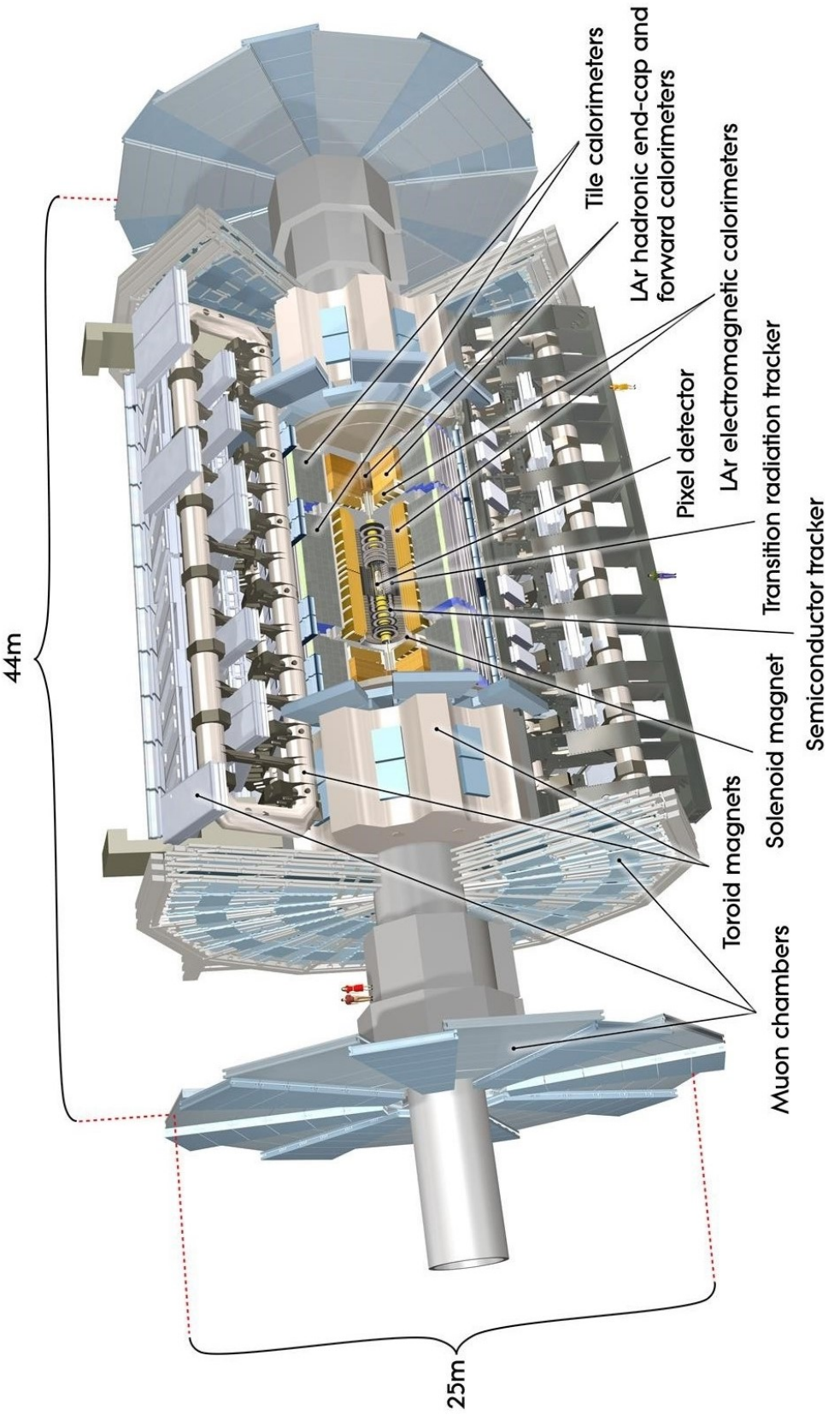

Figure 1.3: Full view image of the ATLAS detector. [3]

a cilinder: each of the three sub-detectors is divided into a barrel section (around the collision point, in the central region) and two end-caps (at the two sides, in the forward region). In particular, the Inner Detector measures charged particles momentum and provides their reconstructed trajectories; the Calorimeters measure the energy of hadrons, electrons, positrons and photons; the Muon Spectrometer identifies muons and measures their momentum. To measure properly the particles momentum and their charge sign, ATLAS is equipped with a Magnet System (Sec. 1.2.4) composed by a superconducting solenoid (around the Inner Detector) and 8 large superconducting toroids (around the Calorimeter). The data acquisition infrastructure (Sec. 1.2.5) manages the huge data flow. Proton-proton interactions occur at a very high rate ($\sim$ 1 GHz): a dedicated trigger system selects interesting physics events, reducing the acquisition rate to a few hundred Hz.

1.2.1 Inner Detector

The Inner Detector (ID) [6] is located all around the interaction point, surrounding the beampipe. It is composed by 4 subdetectors, from the inner to the outer most: the Insertable Beam Layer (IBL), the Pixel Detector, the Semi Conductor Tracker (SCT) and the Transition Radiation Tracker (TRT). Immersed into a 2 T magnetic field provided by the solenoid, the ID provides precise charged particle identification, their trajectories and vertex reconstruction.

Insertable Beam Layer

The Insertable Beam Layer (IBL) [7] has been introduced between the innermost Pixel layer (the B-layer) and the vacuum chamber containing the beams (the beampipe) during the first Long Shutdown occurred between Run I and Run II. A new Beryllium beampipe was built (reducing the radius from 29 mm to 25 mm) in order to obtain enough space for the IBL installation. IBL is composed by 14 staves of modules, mounted directly on the beampipe (Fig. 1.5). Sensors are located longitudinally along the stave, at about 33 mm from the interaction point. Each stave is equipped with 12 Planar sensors (in the central region of the stave) and 8 3D sensors (in the forward region of the stave, 4 for each side). This "mixed scenario" ensures a better z-resolution after heavy irradiation, because of the vertical electrode orientation.

The Planar sensors (active area: $40.8 mm^2$) are bump-bonded to the two Front-End readout chips, while the 3D sensors (active area: 16.8 mm x 20.4 mm) are connected to a single Front-End chip. In particular, Planar sensors are silicon double chip sensors ($n^+ - in - n$) while 3D sensors are single chip sensors ($n - in - p$).

Given the fact that semi condutor detectors are quite sensible to radiation, the main purpose of the IBL is to provide a new additional layer, possibly less sensible to radiation damage, increasing the robustness of the track reconstruction of the detector. Furthermore, the addition of an extra layer placed closer to the interaction point will considerably improve the vertex reconstruction capability of the detector.

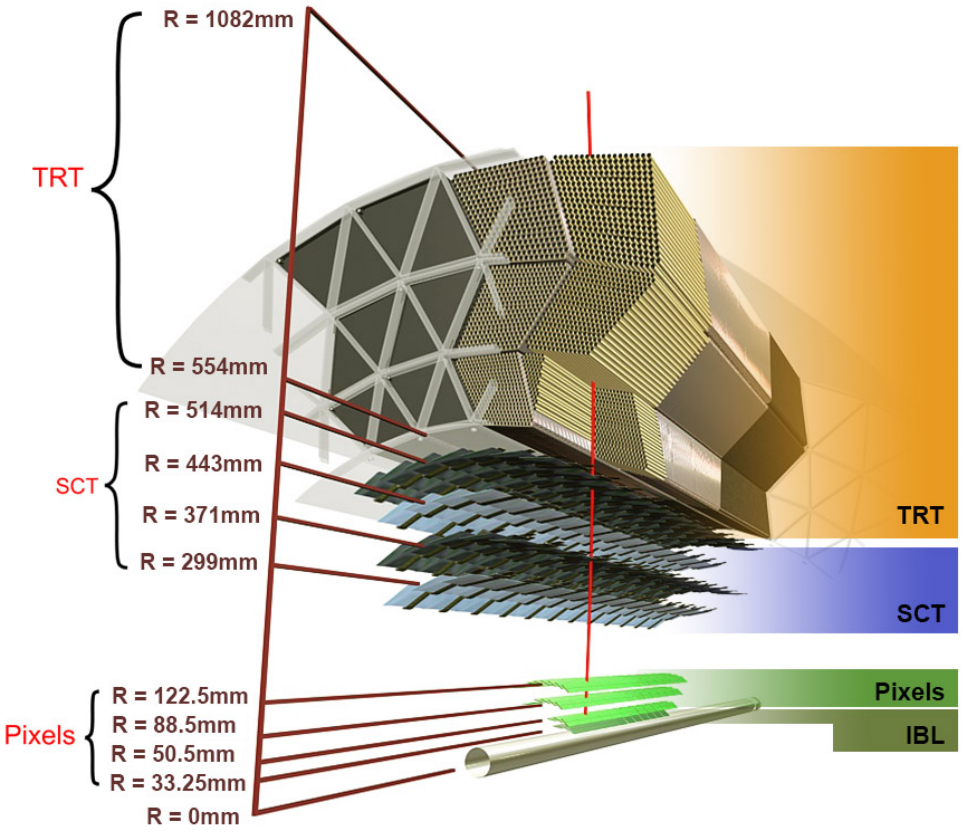

Figure 1.4: Overview of the barrel section of the Inner Detector after IBL insertion (2015). [3]

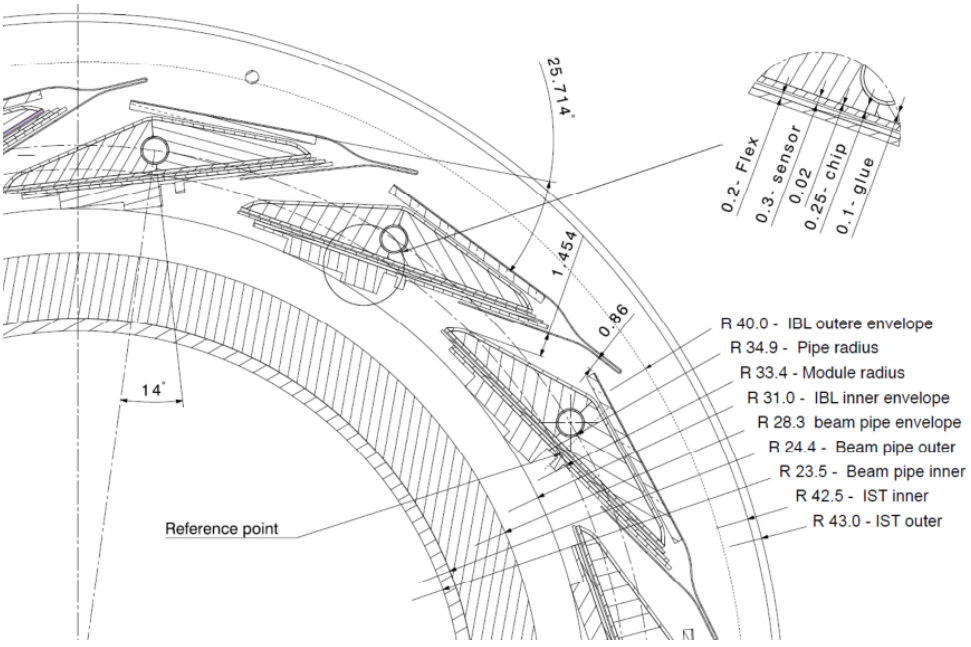

Figure 1.5: Radial view of IBL layout: the staves are mounted between the beampipe and the IBL Support Tube (IST).

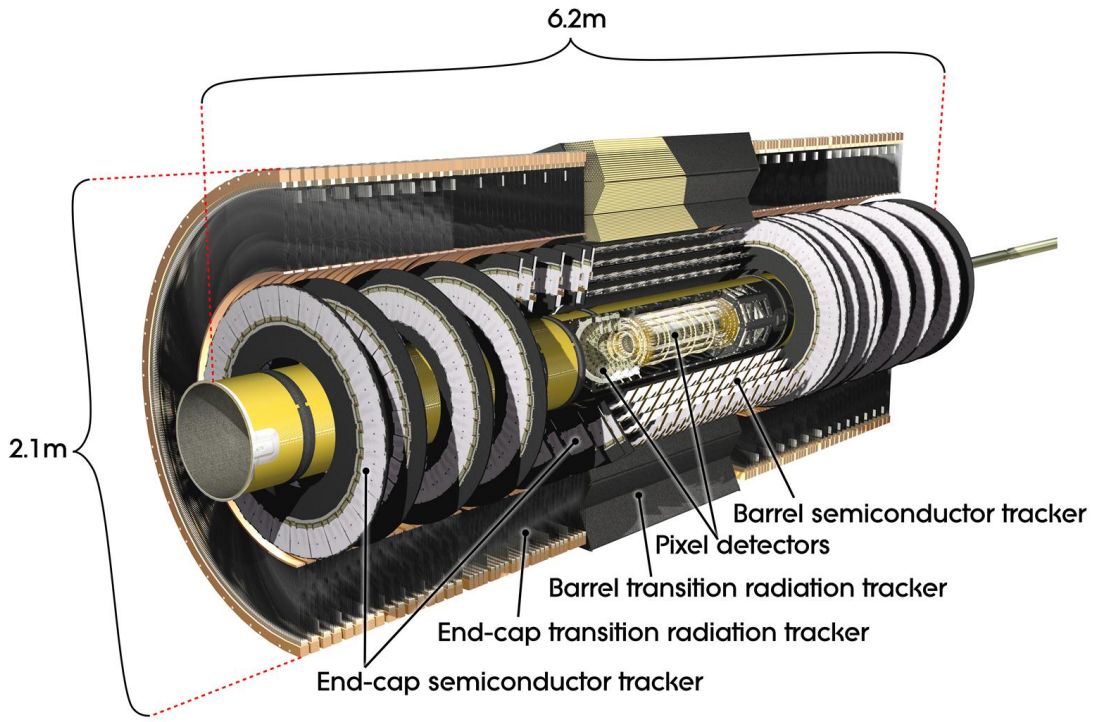

Figure 1.6: Overview of the Inner Detector before introducing IBL: Pixel detector, Semi Conductor Tracker and Transition Radiation Tracker.

Pixel detector

The Pixel detector [8] is located around the interaction point, sorrounding IBL. It is composed by three concentric barrel layers (mean radius: 5.05 cm, 8.85 cm and 12.25 cm; length: 80.1 cm) and two end-caps of three disks each (mean radius: 17 cm; distance from the interaction point: 49.5 cm, 58 cm and 65 cm): this allows to provide three hits for a precise track reconstruction. Thanks to Pixel detector it is possible to identify collision vertices and to measure the impact parameters of tracks with high resolution (impact parameters define the distance of closest approach between the track and the vertex position, see Sec. 1.2.7). With approximately 80 million silicon sensors (50 x 400 μm^2), Pixel detector covers a pseudorapidity range $|\eta| < 2.5$ (pseudorapidity is defined as a function of the polar angle θ, see Sec. 1.2.7) with an extremely high granularity. It reaches an accuracy of 10 μm in R - ϕ and 115 μm in z in the barrel, 10 μm in R - ϕ and 115 in R in the disks. Its role is crucial to perform b-tagging as its high spatial resolution allows to measure with high precision the distance between tracks and the primary interaction vertex.

Semi Conductor Tracker (SCT)

The Semi Conductor Tracker (SCT) [9,10] consists of 4088 modules of silicon strips arranged in four concentric barrels and two end-caps of nine disks each, for a total

surface of 63 m^2. It is designed to provide a minimum of four three-dimensional position measurements per track: it allows precise measurements of track momenta, vertex position and impact parameter with an accuracy of 17 μm in R - ϕ and 580 μm in z.

The barrel region uses two different micro-strips with one set of strips in each layer parallel to the beam direction and a relative angle of 40 mrad. The end-cap region has a set of strips running radially and a set of stereo strips at an angle of 40 mrad.

Transition Radiation Tracker

The Transition Radiation Tracker (TRT) [11] is both a straw drift-tube tracker and a transition radiation detector. It is divided in two barrel sections and two end-caps. The detector consists of about 300000 proportional drift tubes (straws) with 2 mm diameter; each tube is filled with a mixture of $Xe : CO_2 : O_2 = 70 : 27 : 3$, with 5-10 mbar over-pressure. While a charged particle traverses the straws, it produces a trail of electron-ion pairs: electrons drift toward the anode wire creating other electron-ion pairs. The drift time for each straw is used to determine the closest approach distance of each particle to the anode wire, for tracking purposes. In addition, charged particles emits radiation while crossing the interface between media with different dielectric constants (polypropylene and the gas mixture, in this case). The radiation intensity is proportional to the particle relativistic γ factor of the particle, allowing to discriminate between electrons/positrons and hadrons. The barrel sections are composed by 144 long straws disposed parallel to the beam direction; the end-caps are composed by radially disposed 37 cm long straws. The accuracy is not high (about 130 mm per straw in R - ϕ), anyway the high number of hits ($\sim$ 36 per crossing track) compensates the low precision, providing an accurate momentum measurement.

1.2.2 Calorimeter

The Calorimeter system [12] sorrounds the ID and the 2T solenoid: it is composed by an electromagnetic calorimeter designed to measure the energy and provide identification of electrons/positrons and photons and by an hadronic calorimeter for energy measurement of charged and neutral hadrons. With an overall radius of 4.23 m and a length of 13.3 m, with the hadronic calorimeter sorrounding the electromagnetic one, it covers a total range in pseudorapidity $|\eta| < 4.9$. It is consists in a barrel section and two end-caps for both electromagnetic and hadronic calorimeter, plus a forward calorimeter in the forward region (the region at small angle with respect to the beam axis).

In particular the Calorimeter system can measure with high precision the energy, position and shower shape of electrons/positrons, photons and jets (a jet is a tight cone of hadrons emerging from the hadronisation of a quark or a gluon). With

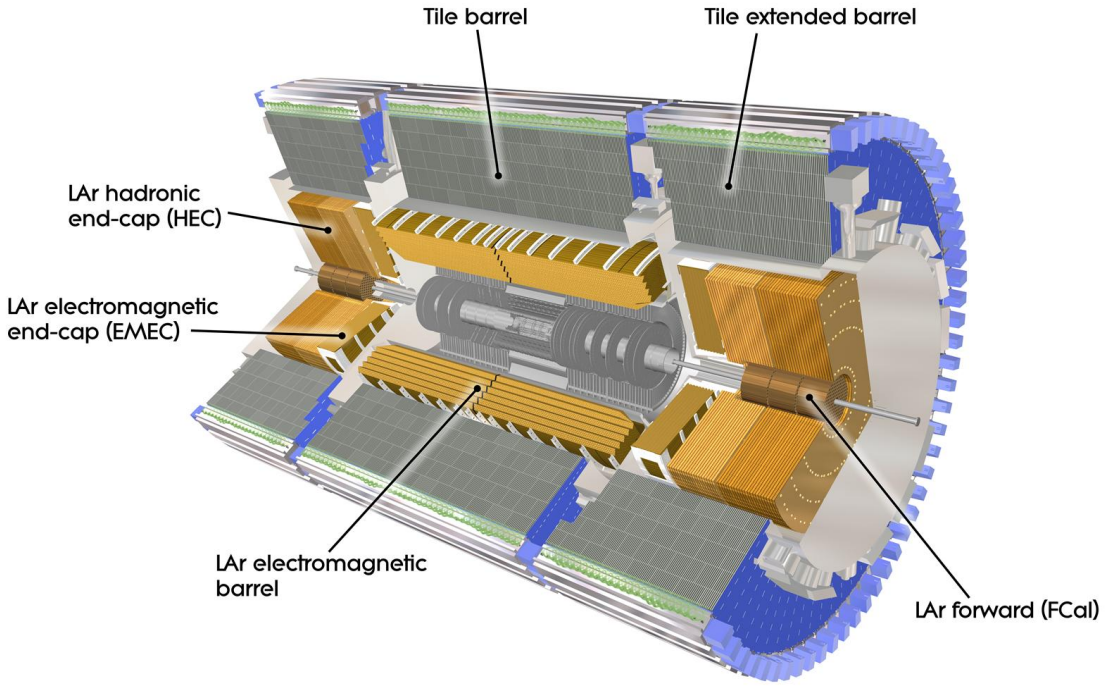

Figure 1.7: A computer generated image of the full ATLAS Calorimeter. [3]

these measurements, it contributes to the definition of missing transverse momentum and to particle identification. It is capable of disentangling electrons/positrons and photons showers from showers coming from hadrons and tau decays and QCD background up to the TeV energy scale.

Electromagnetic Calorimeter

The electromagnetic calorimeter uses Lead layers as absorber material to generate the particle shower and liquid Argon layers as active material to measure energy and position. It is composed by a barrel section ($|\eta| < 1.475$, more than 26 radiation lengths) and two end-caps ($1.375 < |\eta| < 3.2$, more than 24 radiation lengths). The barrel section is divided in two identical half-barrels, while each endcap is divided into two coaxial weels. This geometry enanches the η precision measurements and the γ/π_0 and the e/π separation.

Hadronic Calorimeter

The hadronic calorimeter uses different materials for barrel and end-caps: the barrel section uses Iron as absorber material and scintillating tiles as active material, while the end-caps use Argon technology with Copper as absorber material. The barrel section is composed by the central tile barrel ($|\eta| < 1.0$) and by two lateral tile extended barrel ($0.8 < |\eta| < 1.7$). The hadronic calorimeter end-caps cover

a pseudorapidity range $1.5 < |\eta| < 3.2$. This approach contains the hadrons from propagating further in the muon system and provides good measurements of missing transverse momentum.

Forward Calorimeter

The forward calorimeter uses copper and tungsten as absorber materials and liquid argon as active material, covering the forward region between $3.2 < |\eta| < 4.9$, a region with very high level of radiation. It is composed by one electromagnetic and two hadronic calorimeter layers for each side of the interaction point, longitudinally 1.2 m far from the Inner Detector: this solution avoids the backscattering of neutrons towards the Inner Detector.

1.2.3 Muon Spectrometer

The Muon Spectrometer [13] is the outermost of the ATLAS sub-detectors. Thanks to the magnetic field produced by the air-toroid system (described in Sec. 1.2.4), it identifies muons among all the charged particles exiting from the calorimeter measuring their deflection. It is designed to detect charged particles in the pseudorapidity range $|\eta| < 2.7$, measuring the momentum with small uncertainty; it has also trigger capability in the pseudorapidity range $|\eta| < 2.4$. Due to the energy loss in the dense material of the calorimeter, the muon system can detect muons with momentum larger than 4 GeV. In order to enhance the momentum resolution, the tracks recostructed in the muon spectrometer can be extrapolated back to the internal region of the ATLAS detector, to match the Inner Detector track with the track provided by the muon spectrometer. The Muon system is divided in barrel and end-cap region, employing four different detector technologies: Monitored Drift Tube chambers (MDT), Cathode Strip Chambers (CSC), Resistive Plate Chambers (RPC) and Thin Gap Chambers (TGC). In particular the MDT and the CSC measure with high accuracy the 2D spatial coordinates of the track: they are the tracking core of the muon system because of their high spatial resolution (80 μm). The RPC and the TGC are the trigger core of the muon system: they provide a faster but less precise signal, useful for trigger decision.

1.2.4 Magnetic system

The magnetic system is made of two different solutions: a central superconducting solenoid and 3 air-core superconducting toroids. The solenoid surrounds the Inner Detector and provides a 2 T axial magnetic field that bends particles trajectories. It is 5.3 m long, with a mean radius of 1.25 m; it is cooled at 4.5 K by liquid helium. The toroidal is generated by 3 huge air-core toroids: one in barrel region (1 T magnetic field) and two in the end-cap regions (0.5 T magnetic field). Each air-core toroid consists of 8 coils mounted symmetrically in ϕ, distanced of 45 around the

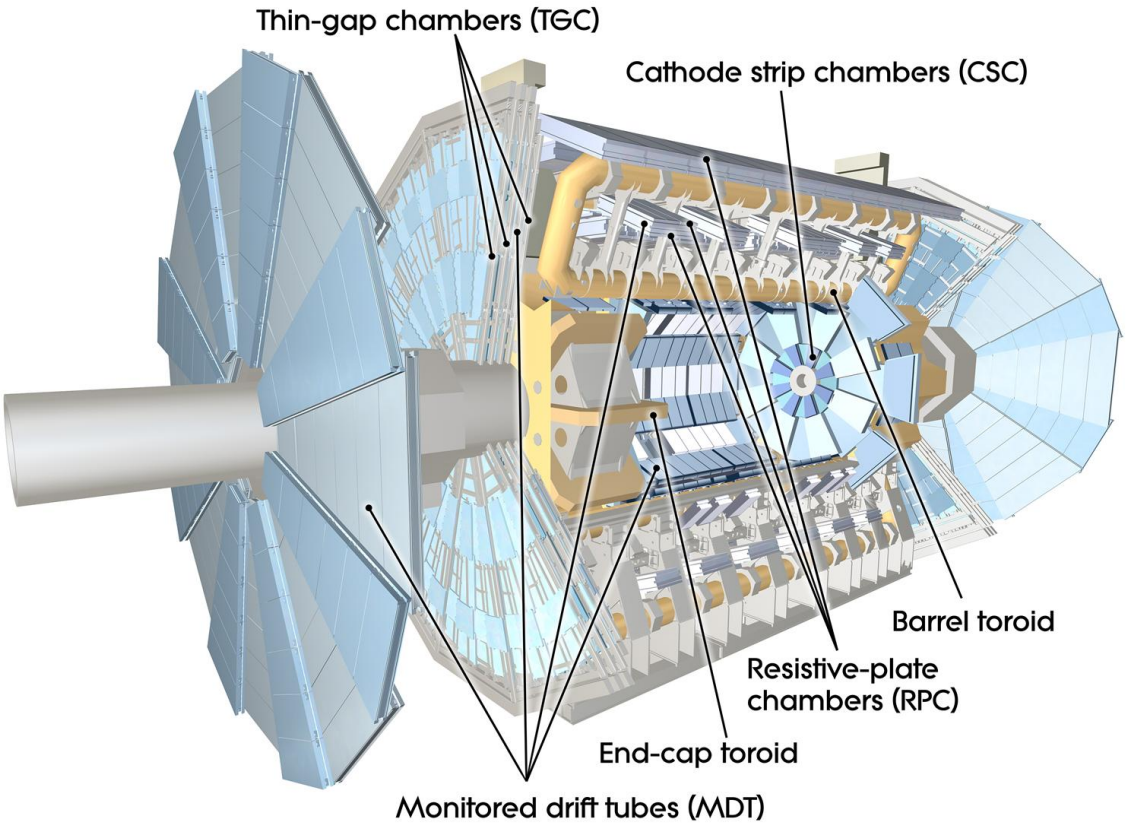

Figure 1.8: Overview of the Muon Spectrometer. [3]

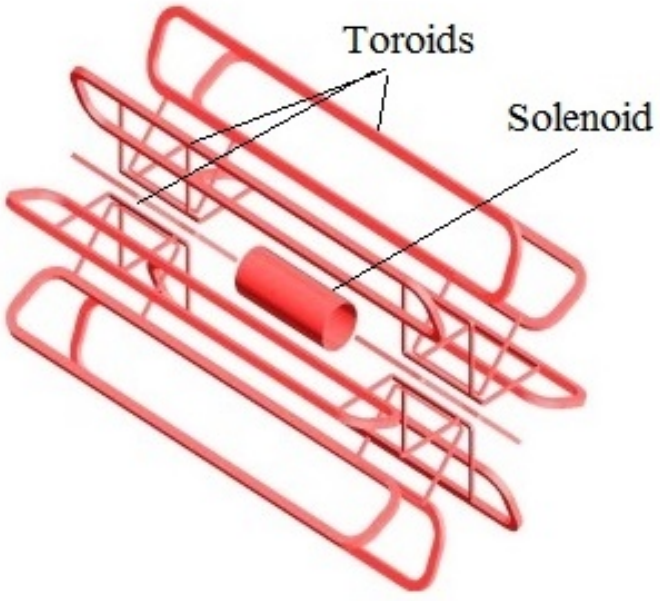

Figure 1.9: The Magnetic System: the superconducting solenoid and the toroids.

beam axis. The toroidal system generate a magnetic field orthogonal to the muon trajectories. The barrel toroids extend for more than 25 m with an inner diameter of 9.4 m and an outer diameter of 20.1 m. The end-cap toroids have a length of about 5 m, with an inner diameter of 1.6 m and an outer diameter of 10.7 m.

1.2.5 Trigger and data acquisition

The high event rate performed by LHC collisions is impossible to handle with any available storage technology, while at the same time, the majority of the events are not interesting. The trigger infrastructure performs a drastic selection of the events, reducing the data flow to an acceptable level before the permanent storage performed by the data acquisition system (DAQ).

Proton-proton collisions are well understood and described as parton interactions within the Quantum Cromo-Dynamic theory (QCD). If the transverse momenta (defined in the transverse plane with respect to the beam direction, see Sec. 1.2.7) involved in the scattering process are higher than ~ 10 GeV, the parton interaction is referred to *hard*, possibly representing an interesting physics event, while if the transverse momenta are lower, the parton interaction is referred to *soft* and no new physics is expected. The soft elastic collisions, in which the momentum transfer between partons is negligible, are the most frequent events at LHC energy scale, the so-callled *minimym bias* events, which will usually be discarded by the selections applied by experiments.

To increase the rate of interesting events, the number of collisions need to be increased; this could be obtained in different ways: increasing the number of bunches in the beam, increasing the number of protons in each bunch, increasing the revolution frequency of the bunches, reducing the transverse sections of the beams. The variable describing the collider ability to produce collisions is the luminosity:

$$\mathcal{L} = f n_p \frac{\mathrm{N}_1 \mathrm{N}_2}{4\pi\sigma_x\sigma_y} \tag{1.1}$$

where σ_x and σ_y represent the gaussian transverse profiles of the beams, N_1 and N_2 represents the number of protons in the bunch, f is the revolution frequence of the bunches and n_p is the number of bunches in the beam.

At the design luminosity ($\mathcal{L} = 10^{34}\ cm^{-2}s^{-1}$), LHC collisions will produce about 23 interactions per bunch crossing: this means a data production of 4000 Gb/s for a bunch crossing rate of 40 MHz (1 bunch crossing event per 25 ns). The ATLAS trigger system [14] is capable to reduce this huge data quantity selecting the interesting events with tight criteria. Until 2012 the trigger system was divided in 3 levels: Level-1, Level-2 and Event Filter, while from 2015 the last two levels are merged to form a unique software based trigger, the High Level Trigger.

The Level 1 trigger selects Regions of Interest (RoI) using the Calorimeter and the Muon Spectrometer information, searching above all for high transverse momentum

particles, the most interesting particles produced in the collisions. Within 2.5 μs, the hardware processors select the events with a roughly 100 kHz data flow.

The High Level Trigger uses the complete information: fine granularity data coming from the Inner Detector, the Calorimeter and the Muon Spectrometer, it performs ID tracks reconstruction and searches among fully-built events within 4 s, reducing the rate to a few hundreds of Hz. Being able to adopt offline reconstruction algorithms, this step can be performed only at the end of the chain, thanks to the previous data reduction.

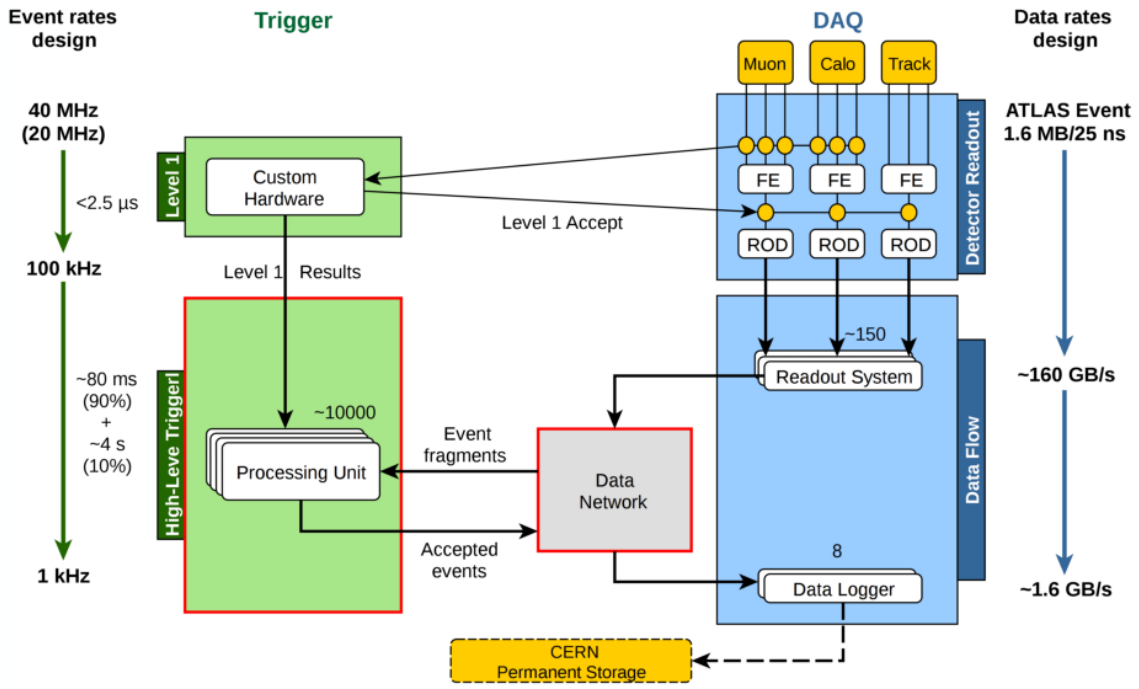

Figure 1.10: Schematic view of the Run II ATLAS trigger and DAQ infrastructure.

During the selection, the trigger levels exchange information in the form of RoI and recostructed events with the data acquisition infrastructure. This information exchange ends with the final permanent data storage of the selected events on the ATLAS software infrastructure.

Since many interactions are produced for each bunch crossing, during the acquisition of a particular triggered event, the information coming from all the bunch crossing interactions is registered: the events collected in this way are defined pile-up events. The ATLAS experiment developed different solutions to reduce the occurred pile-up contamination (see Sec. 2.3).

CERN produces a huge quantity of data each year: to store and manage them, the CERN has its own Data Center. To operate physics analysis on these data, the CERN provided the Worldwide LHC Computing Grid (WLCG) [15]: it is a unique project that provide a global system of more than 170 computing centers linked together to share computing power with the universities and research centers.

1.2.6 Software infrastructure

The ATLAS experiment developed a specific framework for the data analysis, MC simulation and event displaying: the Athena framework [16, 18]. Athena is based on Gaudi architecture [17], previously developed for LHCb experiment, with many specific improvements required for ATLAS data analysis. The framework is written in Python language; it can generate simulated samples and reconstruct both real and simulated events. To perform these tasks, it is possible to run a chain of algorithms in one job or to split the assignment into several jobs.

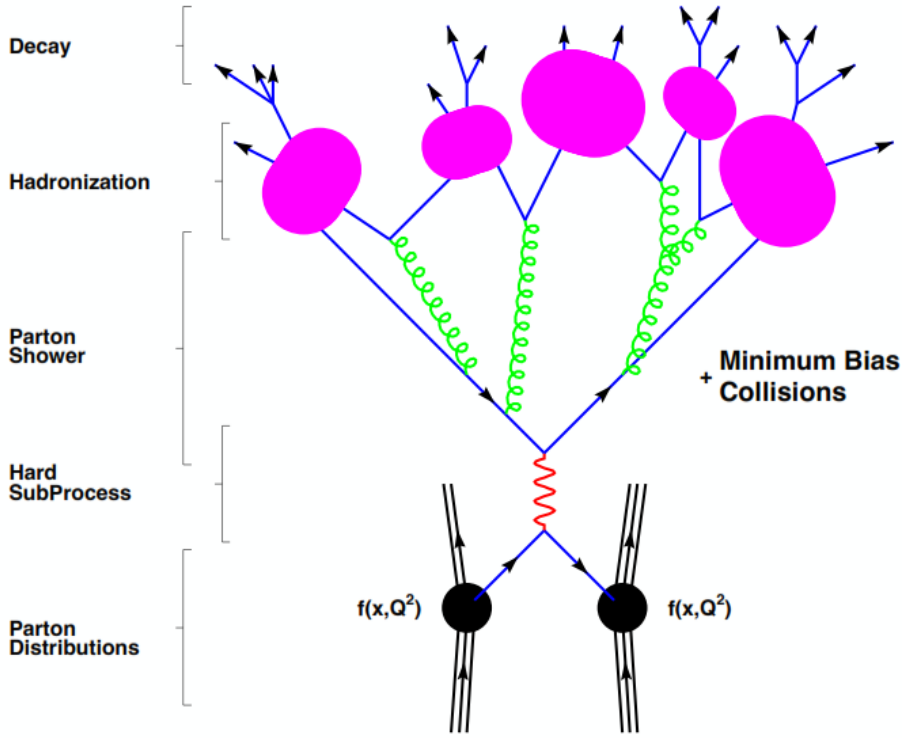

Figure 1.11: Schematic view of the typical hadronic event generation process.

Event generation and detector simulation

The event generation is performed with specific algorithms, such as Pythia [19], Herwig [20] or Sherpa [21]: a more complete simulation includes the use of many algorithms for different tasks, i.e. the background and the signal generation. A typical hadronic process simulation starts from Lagrangian models and matrix elements: the first step is the calculation of the probability for quarks and gluons to produce the parton shower. The next step is the simulation of the hadronization process: colourless hadrons originate from the coloured partons produced in the shower. The last step of this chain is the simulation of the hadron decay chains towards lighter and more stable states.

After the generation, the events are processed by the algorithms that simulate the

interaction with the detector material, such as energy loss and scattering. The most important toolkit for detector geometry simulation is GEANT4 [22]: it provides a detailed description of the detector response to the particle interaction, taking into account also the presence of eventual magnetic fields.

Event reconstruction and analysis

The simulated events generated in this way are completely compatible with real data collected by the detector: the reconstruction algorithms can operate both with simulated and real data. During this step, multiple tasks are performed: pattern recognition, track fitting, vertex determination, energy and momentum measurement. Finally, real data and Monte Carlo simulations are used to perform comparative analysis between experimental results and theoretical prediction. Many different analysis tecniques are used to look at data from different perspectives: usually the analysis consists in software programs in C++ language implemented on ROOT [23,24] software, with dedicated libraries directly linked, such as RooFit [25].

1.2.7 Coordinate system and particle trajectory parameters

The ATLAS coordinate system has the origin in the nominal collision point, with the Y axis pointing vertically upwards, the X axis pointing towards the center of LHC ring and the Z axis pointing towards the direction of the counter-clockwise circulating beam. Actually, the Y axis points upwards with a 0.7° tilt, because of the LHC plane inclination.

Instead of carthesian coordinates, a system of spherical coordinates is preferred for the analysis: the azimuthal angle ϕ lies in the transverse plane and the polar angle θ defines the angle between the particle three-momentum $\mathbf{p}$ and the positive direction of the beam axis.

The coordinate system includes a mix of specific coordinates used in High Energy Physics:

$$(x, y, z) \rightarrow (p_T, \eta, \phi) \tag{1.2}$$

The pseudorapidity η is defined as a function of the polar angle θ:

$$\eta \ = \ -\ln\left(\tan\frac{\theta}{2}\right) \tag{1.3}$$

and the transverse momentum p_T is defined as the component of the three-momentum $\mathbf{p}$ perpendicular to the beam axis:

$$p_T \ = \sqrt{{p_x}^2 \ + \ {p_y}^2} \tag{1.4}$$

Many analyses often require a selection on a precise pseudorapidity range ($|\eta| < 2.5$) corresponding approximately to the interval between θ = 9.39° and θ = -9.39°: in this region there is a better instrumental coverage.

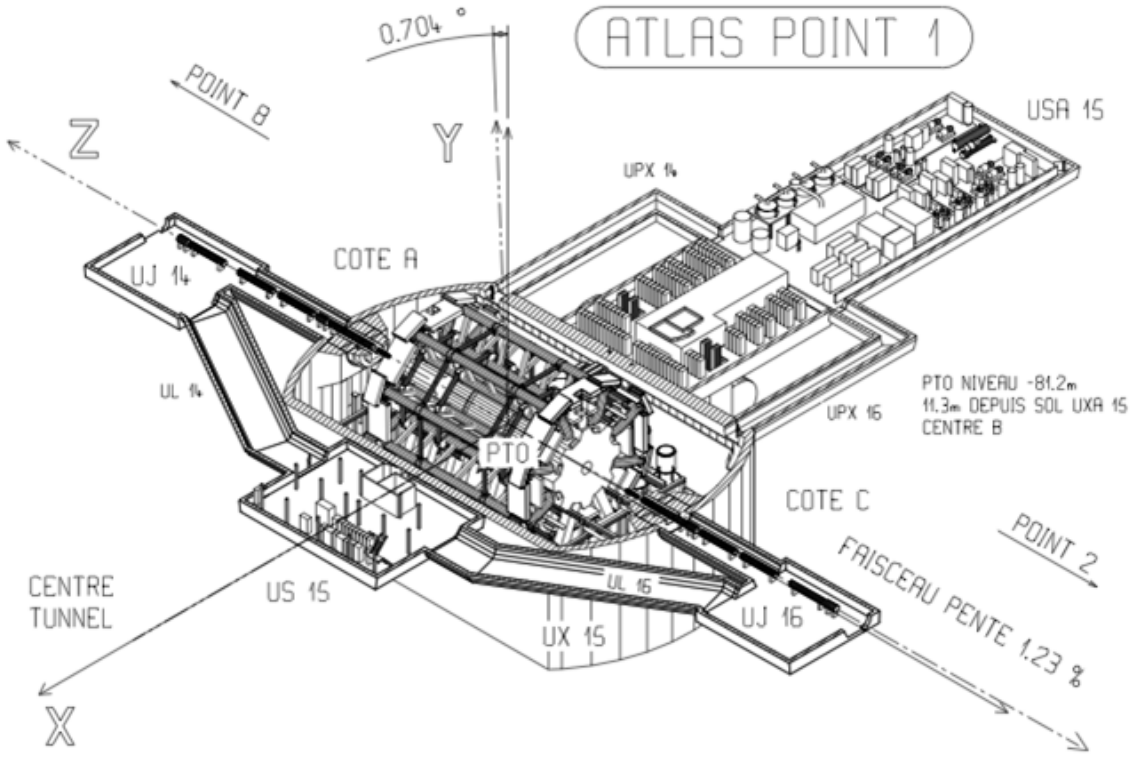

Figure 1.12: Interaction Point 1: ATLAS cavern and coordinate system.

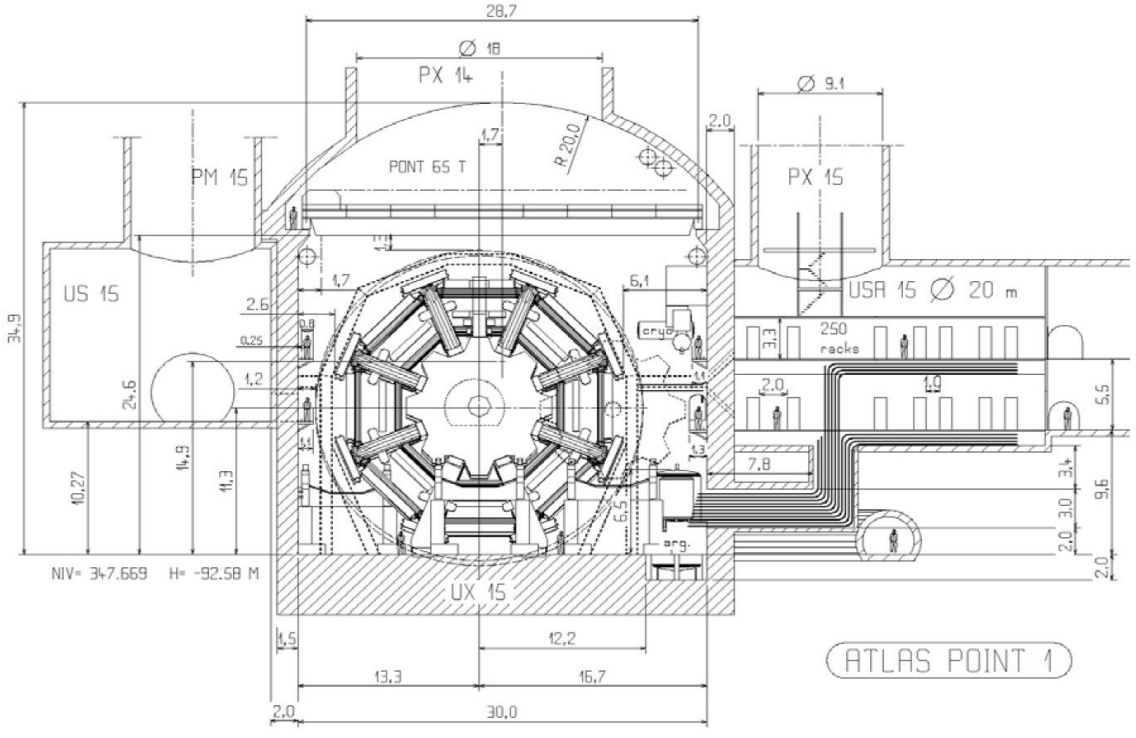

Figure 1.13: Project of the ATLAS cavern.

A charged particle generated at the collision point in the presence of the magnetic field produce a track in the Inner Detector that can be approximated as a helix. The helix can be represented with 5 parameters in the perigee parametrisation, where the perigee of a track is the point of closest approach to the z axis of the coordinate system. The ATLAS perigee parameters are:

- d_0: signed transverse impact parameter, defined as the distance of closest approach of the track to the primary vertex point in the transverse plane, signed according to the reconstructed angular momentum of the track about the axis;
- z_0: longitudinal impact parameter, defined as the difference between the z coordinates of the primary vertex position and of the track at this point of closest approach in the transverse plane;
- ϕ: azimuthal angle of the track at the perigee;
- θ: polar angle of the track at the perigee;
- q/p: charge over momentum of the track.

Chapter 2

ATLAS physics and b-tagging role

In the Standard Model (SM) frame, quarks play an important role: they are the base components of the hadrons (barions and mesons) such as protons, neutrons, pions, kaons and more. At hadronic machines, such as LHC, plenty of quarks stem from each collision and, moreover, many particles generated in the collisions decay in channels containing quarks. According to Quantum Chromo-Dynamics (QCD) theory, the partons (quarks and gluons) are coloured states that can not exist individually. The quarks usually combine to form colourless states: pairs of quarks-antiquarks bound together (mesons) and triplets of differently coloured quarks or antiquarks (barions). This is the reason why free quarks and gluons always undergo a hadronization process: other quarks and gluons are generated in the so-called "parton shower", then a tight cone of hadrons (jet) emerges. It is impossible to detect stand-alone quarks before hadronization (except for indirect measurements of the the top quark, which decays before hadronizing because of his veri large mass), but it is still possible to identify the quark from which the jet originated. The specific goal to identify the beauty content of a jet is pursued by each particle experiment: in particular, the ATLAS Flavour Tagging group manages the task to label a jet coming from a beauty quark (b-jet), from a charm quark (c-jet) or from a gluon or a up, down or strange quark (light jet).

The correct identification of jets containing b hadrons (hence b-tagging) is of capital importance for many physics analyses: for top quark studies and Higgs searches, as well as the search for new physics fenomena beyond Standard Model. ATLAS developed its own algorithms for b-jet identification [29], exploiting the typical properties of the b quarks and the b hadrons emerging from the jet: the long lifetime, the high decay multiplicity and the high invariant mass above all. After the description of the phisics analysis that could benefit from b-tagging, the main ATLAS b-tagging algorithms will be presented in this chapter. A detailed description of the measurement of the tagging effiency on b-jets, c-jets and light jets will be treated in the next chapter.

2.1 Standard Model and beyond

Nowadays, the most satisfying theory describing particle interaction is the Standard Model [31–33], born after a century of deep studies over the fundamental constituents of matter. The SM describes the electromagnetic, weak and strong interactions in terms of fields, following the electro-weak theory (also referred as Glashow-Weinberg-Salam Theory) and the Quantum Chromo-Dynamics. The gravitational interaction is not included in this model: it dominates in the universe on a large scale, but it does not play an important role in particle physics due to its weakeness. From the theoretical point of view, the Standard Model is a gauge theory based on the symmetry group:

$$U(1) \otimes SU(2) \otimes SU(3) \tag{2.1}$$

where $U(1) \otimes SU(2)$ is the group of unified weak and electromagnetic interaction and SU(3) is the group associated to strong interaction. The interactions are exchanged by means of "messenger particles", the gauge bosons: the photons (electromagnetic interaction), the gluons (strong interaction) and $W^{\pm}$ and Z^0 bosons (weak interaction).

Gauge Boson	Mass [GeV]	Interaction	Range [m]
gluon (g)	0	Strong	10^{-15}
photon (γ)	0	Electromagnetic	∞
$W^{\pm}$	80.4	Weak	10^{-17}
Z^0	91.2	Weak	10^{-17}

Table 2.1: Gauge boson are mediators of the Strong, Electromagnetic and Weak Interaction. The measured masses are reported in table, along with the effective ranges of the interactions.

According to the Standard Model, there are few fundamental particles divided in two main classes: 12 fermions with spin 1/2 (6 leptons and 6 quarks) and 5 bosons with spin 1 (actually, 8 different gluons exist). Of the three forces, only the weak force affects all fermions; the electromagnetic force is felt only by the charged particles, while only the quarks interact via the strong force.

	Generation			
	I	II	III	Electric Charge
Leptons	$\begin{pmatrix} \nu_e \\ e^- \end{pmatrix}$	$\begin{pmatrix} \nu_\mu \\ \mu^- \end{pmatrix}$	$\begin{pmatrix} \nu_\tau \\ \tau^- \end{pmatrix}$	$\begin{matrix} 0 \\ -1 \end{matrix}$
Quarks	$\begin{pmatrix} u \\ d \end{pmatrix}$	$\begin{pmatrix} c \\ s \end{pmatrix}$	$\begin{pmatrix} t \\ b \end{pmatrix}$	$\begin{matrix} 2/3 \\ -1/3 \end{matrix}$

Table 2.2: Leptons and quarks: I, II, III generation.

Lepton	Mass		Quark	Mass
ν_e	< 2 eV		u	2.3 MeV
e^-	0.511 MeV		d	4.8 MeV
ν_μ	< 0.19 MeV		c	1.3 GeV
μ^-	105.7 MeV		s	95 MeV
ν_τ	< 18.2 MeV		t	173 GeV
τ^-	1776.8 MeV		b	4.2 GeV

Table 2.3: Leptons and quarks measured masses.

Since the electroweak theory describes massless fermions and gauge bosons, new parts have to be added to the theory to give them mass: the spontaneous symmetry breaking of the electroweak gauge symmetry and the Higgs mechanism. This introduces a new particle into the theory: the Higgs boson. Since the theory can only indirectly and weakly predict favored values for the Higgs boson mass, a long search has been performed at the highest energy colliders: LEP (electron-positron collisions, up to 200 GeV), Tevatron (proton-antiproton collisions, up to 1.96 TeV) and LHC (proton-proton collisions, up to 13 TeV).

To be able to correctly identify b-jets is fundamental for SM analysis, like the Higgs search (for details, see Sec. 2.1.1) and top quark physics (see Sec. 2.1.2), but also for beyond SM physics as Super symmetry particles search (see Sec. 2.1.3).

2.1.1 Higgs search

Higgs boson searches have been carried at LEP [35] and at Tevatron [36], but they did not collect enought data to discover a new particle. LEP contributed to exclude the mass region lower than 114.4 GeV, while Tevatron excluded 147-180 GeV region. At Tevatron an excess of events was found in the region 115-140 GeV, but the statistical significance was too low to declare the discovery. Finally, in 2012 the ATLAS [37] and CMS [38] collaborations discovered a neutral boson compatible with the Higgs boson, with a mass value of ≈ 126 GeV. Further studies reached a better precision on the Higgs mass measurement: the current best estimation of the mass value is 125.7 ± 0.4 GeV [27], combining ATLAS and CMS measurements. With this mass value, it is possible to calculate precisely its production cross sections, the decay rates, the branching ratios and the higgs-fermions couplings.

Two main production mechanisms generate Higgs bosons at LHC: the gluon gluon Fusion (ggF) with a production cross section of about 19 pb and the Vector Boson Fusion (VBF) with a lower cross section of about 1 pb, both calculated at $\sqrt{s} = 8$ TeV. Other important production mechanisms are Higgs-stralung (VH) and the associated production with a top pair.

To study the properties and the couplings of the Higgs boson, it is important to know not only the production rates but also the decay rates. The Higgs boson can decay directly into all possible massive particles, and into photons and gluons

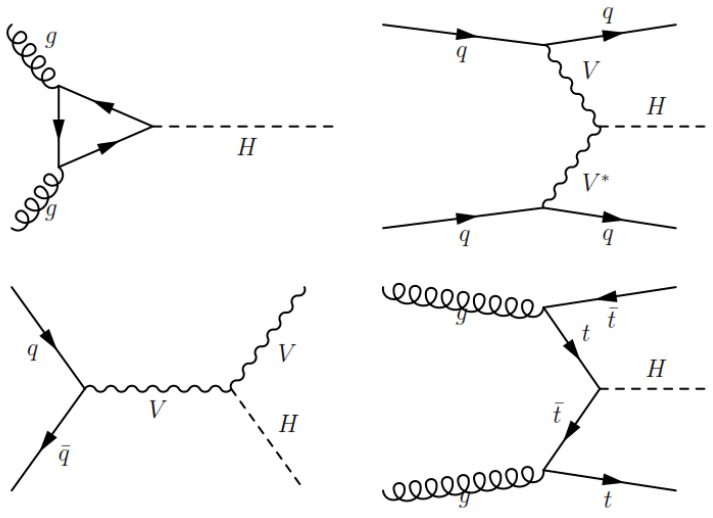

Figure 2.1: Most important Higgs production mechanisms, from top left: gluon-gluon Fusion (ggF), Vector Boson Fusion (VBF), Higgs-stralung (VH), production in association with a top pair.

through loop diagrams. With a 126 GeV mass, the main branching ratio is $H \to bb$ (68%) followed by $H \to WW$ (roughly 20%), gg, $\tau\tau$, ZZ, cc, $\gamma\gamma$.

The Higgs discovery made by the ATLAS and CMS collaborations was performed studying the ZZ and $\gamma\gamma$ decay channels, because these signatures are clearly identified in the detector. On the opposite, at a hadron collider, there is plenty of quark production and this represents a large background for the search in the $H \to bb$ channel. After the discovery, data continues to indicate that it is the SM Higgs boson. In order to measure the Higgs properties, it is essential to measure all the Higgs couplings, included the one with b-quarks. The b-tagging plays a crucially important role in those searches, expecially in the $H \to bb$ channel: this channel has been studied at ATLAS during Run I data analysis and will be performed more deeply with Run II data, with increased energy and luminosity.

2.1.2 Top physics

Top quark is the heaviest particle ever discovered with a measured mass of 173 GeV [27]. It was discovered at Tevatron [39, 40] in 1995, and since then many measurements have been carried out in order to determine its properties. Top quark lifetime is so low ($\sim 10^{-25}$) that it is the only quark decaying before the hadronization process occurs. For this reason, no hadrons containing top quark have been ever observed. The correct identification of a b-jet allows to study in deep the top quark and its properties: the top quark decays almost exclusively into a bottom quark and a W boson via an electro-weak process ($t \to Wb > 99\%$ [27]). Since the top quark has a mass in the range of electro-weak energy scale it could be used to probe electro-weak symmetry breaking and fermion mass generation.

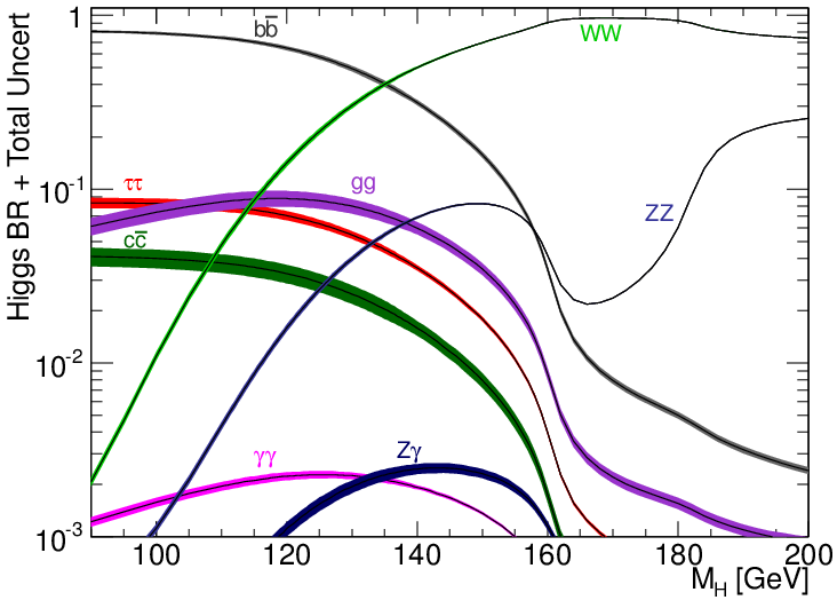

Figure 2.2: Higgs boson decay branching ratios for different mass value: at ≈ 125 GeV the $H \to bb$ is the main channel with about 68%, followed by WW (roughly 20%).

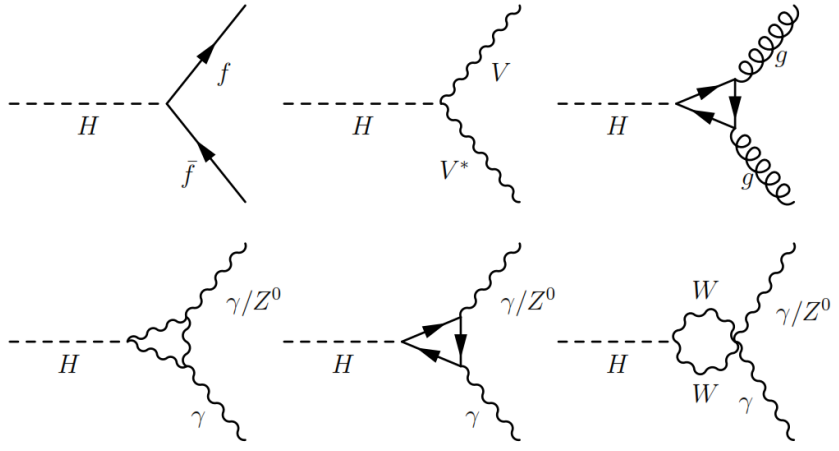

Figure 2.3: Most important Higgs decays channels, from top left: into a fermion pair, a weak vector boson pair, a gluon pair through coupling with a fermion triplet, a $\gamma\gamma$ or a γZ^0 pair through triplets or loop diagrams.

In addition, about 10 million top quark pairs per year are produced at the LHC, the correct understanding of top quark production cross sections and decay channels is important because events having top quarks in the final state are one of the dominant backgrounds for many physics analyses searching for new physics beyond Standard Model at the TeV scale.

2.1.3 Super Symmetry and beyond

Super Symmetry (SUSY) is an extension of the Standard Model which tries to solve some of the SM inconsistencies introducing a new symmetry which relates fermions and bosons. This symmetry is called supersymmetry and gives every lepton, quark and gauge boson a supersymmetric partner. Super symmetric particles are yet to be discovered, they are supposed to be at a higher mass scale with respect to the known particles. The most common SUSY model, addressed as "Minimal SUSY Standard Model" (MSSM) [34], introduces also additionals Higgs bosons to resolve gauge anomalies. The R-parity is the MSSM symmetry, defined as:

$$R = -1^{(3B+2S+L)} \tag{2.2}$$

where B is the Barion number, S is the Spin and L is the Lepton number. Standard Model particles have an even R-parity value, while their superpartners have an odd value. The R-parity conservation has some heavy implications:

- SUSY particles are only produced in pairs;
- SUSY particles decay channels contain an odd number of SUSY particles;
- stable and light SUSY particles exist.

In particular the Lightest Stable Particle (LSP) is electrically neutral and currently is the best candidate to explain the Dark Matter component of the Universe.

Quarks	Leptons	Neutrinos	Gauge Bosons			
(u, d, c, s, t, b)	(e, μ, τ)	(ν_e, ν_μ, ν_τ)	g	$W^\pm$, $H^\pm$	γ, Z^0, h^0, H^0	A^0
($\tilde{u}$, $\tilde{d}$, $\tilde{c}$, $\tilde{s}$, $\tilde{t}$, $\tilde{b}$)	($\tilde{e}$, $\tilde{\mu}$, $\tilde{\tau}$)	($\tilde{\nu}_e$, $\tilde{\nu}_\mu$, $\tilde{\nu}_\tau$)	$\tilde{g}$	$\tilde{W}^\pm$, $\tilde{H}^\pm$	$\tilde{\gamma}$, $\tilde{Z}^0$, $\tilde{h}^0$, $\tilde{H}^0$	$\tilde{A}^0$

Table 2.4: Standard Model particles (upper row) and their supersymmetric partners (lower row) according to the Minimal SUSY Standard Model.

The ATLAS experiment can contribute to SUSY particles searches; the correct identification of quarks either in the signal determination or for background suppression purpose could lead to the discovery of new particles. Decays which involve SUSY particles have some special characteristics: high amount of missing energy due to the lightest SUSY particle which does not interact with the detector, leptons with high transverse momenta, lepton pairs with same charge in the final state.

In addition, the study of heavy quarks and leptons includes the search for a possible fourth generation of quarks and leptons. Other parts of the exciting physics programme of the ATLAS experiment are the search for extra dimensions and magnetic monopoles.

2.2 Jet properties

In this section a description of the characteristics of b quarks and B hadrons is reported: b-jets have a different topology with respect to jets coming from charm quarks, light quarks or gluons and this allows to separate them from the others.

The b quark is the second heaviest known quark, with a mass around 4 GeV [27]. Jets coming from b quarks (b-jets) will contain one or more B hadrons in their final state after hadronisation: the properties of some B hadrons give to b-jets their special topology: the presence of a secondary vertex, quite separated from the primary vertex and reconstructable in most cases, and tracks with large impact parameters (d_0 and z_0, see Sec. 1.2.7).

The most important B hadrons are the $B^{\pm}$, B^0 and B_s mesons which decay via electro-weak interactions and have lifetimes of the order of 1.5 ps. With such a relatively long lifetime, the typical B meson with a 50 GeV momentum would travel about 5 mm in the lab frame before decaying, generating in this way a displaced secondary vertex with a high number of tracks (high multiplicity). The tracks coming from the secondary vertex have also larger impact parameters with respect to the tracks from primary vertex. The b-tagging algorithms exploit all of these properties to label a jet as a b-jet.

The light quarks (up, down and strange) with masses in the MeV scale or the massless gluons are much lighter than the b quark: the so called light jets (l-jets) are quite different with respect to the b-jets in terms of invariant mass and decay length. Lifetime of hadrons containing lighter quarks are several orders of magnitude shorter (10^{-23} s to 10^{-16} s for strong or electromagnetic decays) or longer (10^{-8} s for electro-weak decays). In case they decay instantly, they will not travel a measurable distance, excluding the possibility to identify a secondary vertex due to the hadron decay. In case they decay out of the Inner Detector, meters away from the primary vertex, the possibility to find a secondary vertex in the Inner Detector is excluded. In case they decay in the Inner Detector, the topology of the decay is such that it can be correctly identified as a hadron containing light quarks, also thanks to the different invariant mass.

The c quarks are slightly lighter than the b quarks, with a mass of about 1 GeV [27], so jets coming from c quarks (c-jets) have intermediate properties: since the hadrons containing c quarks have a lifetime similar to that of the B hadrons, they also produce a separated secondary vertex, characterized by a lower track multiplicity and an invariant mass in the middle between b-jets and l-jets. For this reason c-jets are very hard to identify and they are easily mistaken as b-jets. The most important hadrons containing c quarks are $D^{\pm}$ and D^0 mesons which decay via the electro-weak force and have lifetimes compatible with B hadrons (about 1 ps), but slightly briefer. The typical D meson decay produces a secondary vertex displaced from the primary vertex of about 3 mm (estimated for a meson with 20 GeV momentum in the laboratory frame) with a topology very similar to that of b-jet.

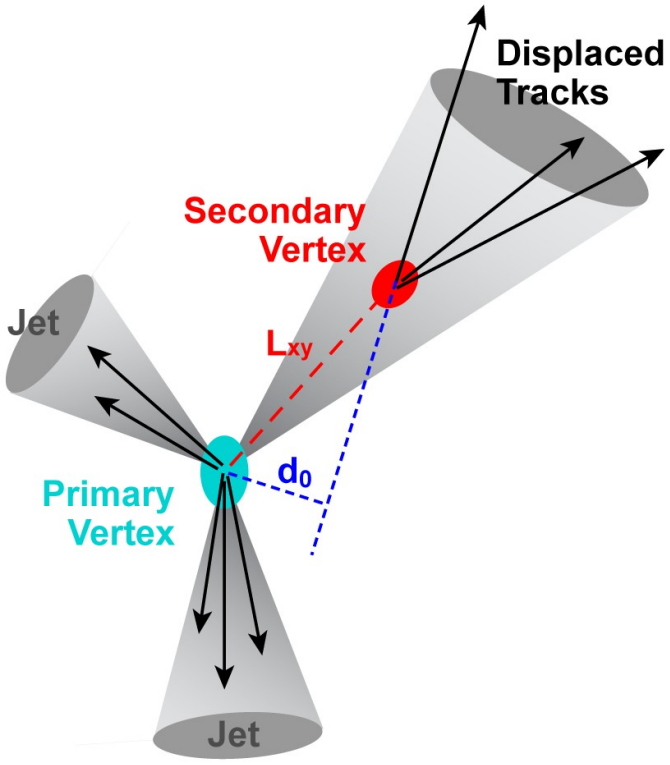

Figure 2.4: Typical structure of a b-jet (or a c-jet) stemming from the primary vertex of the event: while light jets tracks originates in the Primary Vertex (i.e. the collision point), most of the b-jet tracks originates in the Secondary Vertex (i.e. the vertex of the hadron decay). Tracks originated in the Secondary Vertex have higher longitudinal (z_0) and trasversal (d_0) impact parameters.

2.3 Vertices, jets and tracks reconstruction

Since jets are complex physics objects, particular attention is devoted to their reconstruction: starting from the tracks, the primary vertex is identified, then tracks are associated to jets. In this section a brief description of this process is reported, along with the description of pile-up effects reduction.

Due to pile-up interactions, the average number of reconstructed primary vertices is larger than one (about 10 interaction per bunch crossing at 2015 luminosity). Each vertex is required to have two or more tracks; the primary vertex is selected as the vertex that maximises the sum of the associated tracks square transverse momentum p_T^2. The determination of the primary vertex is particularly important for b-tagging, because it is the reference point to determine track impact parameters and the secondary vertex.

For b-tagging purpose, the jets are usually reconstructed using the $anti-k_t$ algorithm [42] with a distance parameter R = 0.4 based on topological clusters of energy deposits in the calorimeters. Two approaches were employed by ATLAS to

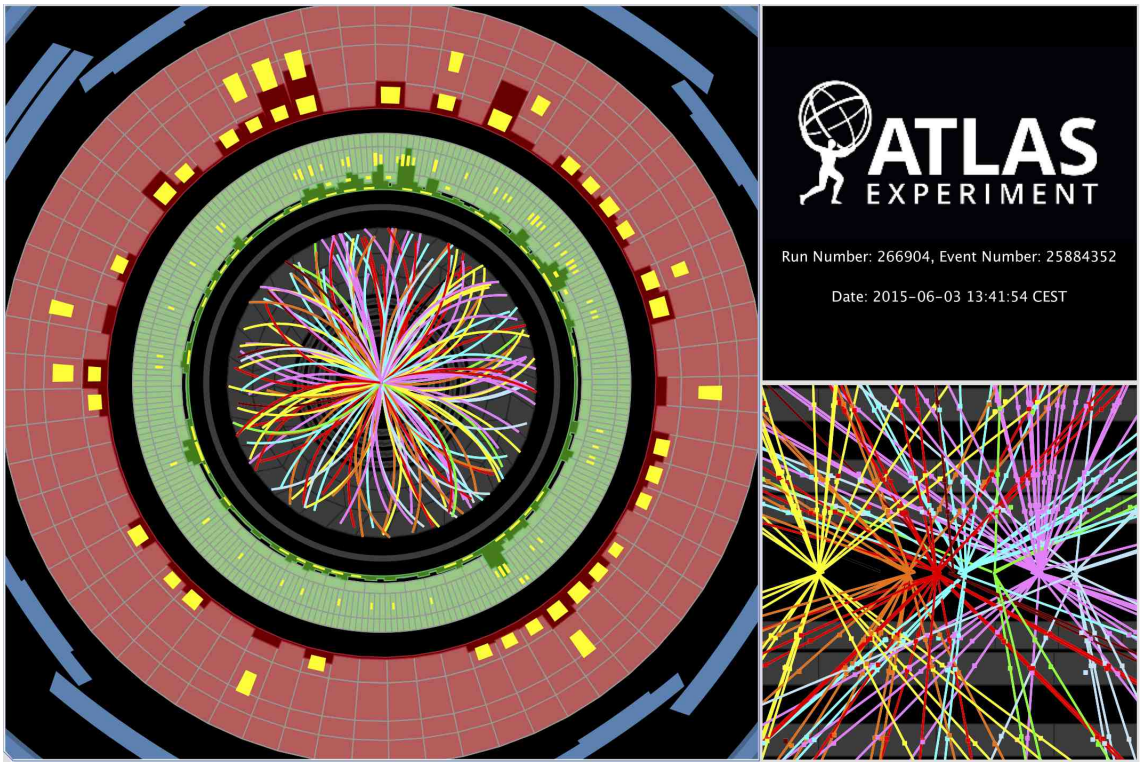

Figure 2.5: Display of a proton-proton collision recorded on 3 June 2015, with the first LHC stable beams at 13 TeV. Transverse plane (left): tracks reconstructed from hits in the Inner Detector are shown as arcs curving in the magnetic field, while the electromagnetic and the hadronic calorimeters are shown in greed and red respectively, with bars indicating energy deposits. Longitudinal plane (right): tracks originate from several vertices, indicating multiple proton-proton interactions, also known as pile-up. [3]

reconstruct clusters for Run I analysis: one starting from energy clusters calibrated at the electromagnetic scale ("EM jets"), and one starting from clusters calibrated using the Local Cluster Weighting (LCW) method ("LC jets") [41].

The tracks associated to the calorimeter jet are required to have small angular separation with respect to the jet axis: $\Delta R(track, jet) = \sqrt{(\Delta\eta)^2 + (\Delta\phi)^2}$, where η is the pseudorapidity and ϕ is the azimuthal angle (see Sec 1.2.7). Greater the jet p_T and narrower is the jet cone, resulting with a smaller cut value of ΔR: a 20 GeV jet has a cone cut value $\Delta R = 0.45$ while a 150 GeV jet has a cone cut value $\Delta R = 0.26$. Tracks ar further required to pass some quality criteria in order to reject fake and badly reconstructed tracks.

The Jet Vertex Fraction (JVF) [43], defined as the summed transverse momentum of the tracks associated with the jet and originated from the primary vertex divided by the summed transverse momentum of all the tracks associated with the jet, is used to reject pile-up jets. JVF measures the fractional p_T from tracks associated with the primary vertex: its value is bound between 0 and 1, a -1 value is assigned to jets with no associated tracks. A small JVF value corresponds to a small pile-up contamination, so, to reduce the contribution from jets arising from pile-up, in Run

I analysis, jets with $p_T < 50\ GeV$ and $|\eta| < 2.4$ were required to satisfy $JVF > 0.5$.

A new discriminant for pile-up suppression has been developed and is currently used by Run II analysis: the Jet Vertex Tagger (JVT) [44], obtained with a 2-dimensional likelihood of R_{p_T}, defined as the p_T sum of the tracks associated with the jet and originated from the primary vertex divided by the fully calibrated jet p_T, and $corrJVF$ defined as JVF, but corrected for the average sum p_T from pile-up tracks associated with the jet. The resulting JVT value is bound between 0 and 1, with -0.1 assigned to jets with no associated tracks. To reduce pile-up contamination, JVT is required to be lower than 0.641 for jets with $p_T < 50\ GeV$ and $|\eta| < 2.4$, while jets with $p_T > 50\ GeV$ have a negligible contamination, so no further selection is applied to them.

2.4 Flavour tagging algorithms

The identification of b-jets is performed with several algorithms, exploiting the long lifetime, high mass and decay multiplicity of b hadrons. The tracks originated from b-hadron decay have large impact parameters with respect to the tracks stemming from the primary vertex. The basic ATLAS b-tagging algorithms [28] are based on the secondary vertex reconstruction (SV1 and JetFitter) and the large impact parameters of their tracks (IP3D). To improve the performance of these algorithms, their output variables are combined through multivariate tecniques to obtain more discriminant variables. This is the case of the most used b-tagging variables: MV1 (for Run I) and MV2c20 (for Run II). During Run I, the online trigger used his own version of SV1 and IP3D b-tagging algorithm to identify on-the-fly b-jets during collisions and store interesting events containing b-jets; during Run II, the MV2c20 algorithm is used.

2.4.1 IP3D

IP3D is the impact-parameter-based algorithm currently used by ATLAS, deriving from JetProb, a previous and simpler algorithm extensively used at LEP and at Tevatron. The impact parameters are computed with respect to the primary vertex. In particular, the transverse impact parameter d_0 is signed to further discriminate the tracks from the b-hadron decay from tracks originating from the primary vertex. The sign is defined as positive if the track intersects the jet axis in front of the primary vertex, and negative if the intersection lies behind the primary vertex. The jet axis is defined from the calorimeter jet direction or, when a compatible secondary vertex is found, the jet axis is defined as the line joining the primary and the secondary vertex. IP3D uses both the transverse (d_0) and the longitudinal (z_0) impact parameters and their correlations. More precisely, IP3D uses d_0 and z_0 significances defined as the ratio between the impact parameter and its uncertainty on the reconstructed impact parameter: $S_{d_0} = d_0/\sigma_{d_0}$ and $S_{z_0} = z_0/\sigma_{z_0}$. For each track, the significances are

compared to the pre-determined 2-dimensional Probability Density Function (PDF) obtained from simulation of the b-jet and light jet. The log-likelihood ratio formalism is used on the jet weight, defined as the sum of the logarithms of the track weight, i.e. the ratio of probabilities of a track.

2.4.2 SV1

SV1 is one of the secondary vertex based algorithms currently used by ATLAS, along with the JetFitter algorithm described later. The secondary-vertex-based algorithms have a much smaller mistag rate (i.e. the probability of mistakenly tagging a light jet as a b-jet) with respect to the impact parameter based algorithms, but they are limited by the relatively small efficiency of finding the secondary vertex ($\approx 60\%$). SV1 searches for the secondary vertex trying to associate every pair of tracks far enought from the primary vertex with a χ^2 fit. Vertices with invariant mass compatible to K_s, Λ and other long lived particles are rejected, as well as vertices with coordinates compatible with a secondary interaction with the detector material. SV1 is based on a likelihood ratio formalism exploiting these properties:

- the invariant mass of all the tracks used to reconstruct the secondary vertex;
- the ratio of the sum of the energies of these tracks to the sum of the energies of all tracks in the jet;
- the number of two-track vertices;
- the ΔR between the jet direction and the line joining the primary and the secondary vertex.

To improve the b-tagging performance, the IP3D and the SV1 algorithm are combined summing their respective weights, since they both use the LLR formalism. This is referred as the IP3D+SV1 algorithm.

2.4.3 JetFitter

Jet Fitter exploits the topological structure of the weak decays of the b-hadron and the c-hadron inside the jet. A Kalman filter is used to find the line where the primary vertex, the b-hadron and the c-hadron vertices lie, as well as their position on this line. The JetFitter algorithm takes six variables as input of a neural network:

- the number of vertices with at least two tracks;
- the total number of tracks at these vertices;
- the number of additional single track vertices on the b-hadron flight axis;
- the invariant mass of all charged particle tracks attached to the decay chain;

- the energy of these charged particles divided by the sum of the energies of all charged particles associated to the jet;
- the weighted average displaced vertex decay length divided by its uncertainty.

The jet p_T and η are used as additional input nodes in the neural network, since the six input variables depend on them. JetFitter has three output nodes, referred as P_b, P_c and P_l, corresponding to the b-jet, c-jet and light jet hypoteses. The discriminating variable is defined as the logarithm of the ratio between P_b and P_l. Another discriminant is obtained by the combination of the IP3D and the JetFitter algorithm, giving the IP3D output weight as an additional input node to the JetFitter neural network. This is referred as IP3D+JetFitter algorithm, but also as JetFitterCombNN.

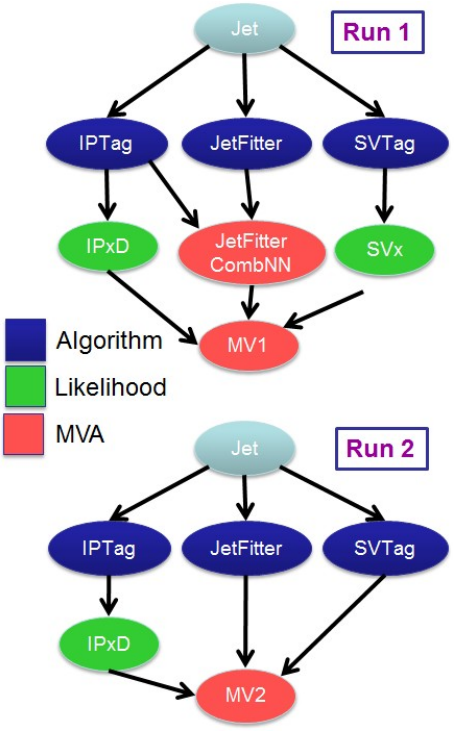

Figure 2.6: Combination of the b-tagging algorithms.

2.4.4 MV1

To better improve the performance, the MV1 algorithm was developed as a combination of three of the previously described algorithms: IP3D, SV1 and IP3D+JetFitter. The MV1 neural network is a perceptron based on the TMVA package [45] with two hidden layers consisting of three and two nodes, respectively, and an output layer with a single node which holds the final discriminant variable. The training relies on a back-propagation algorithm and is based on two simulated samples: b-jets (signal hypothesis) and light jets (background hypothesis).

Jets are divided in ten bins in p_T and four bins in η and the (p_T,η) bin of the jet is used as additional input node in the neural network. The MV1 output weight distribution is shown in Fig. 2.7 for b, c, and light jets in simulated $t\bar{t}$ events: b-jets tend to have high values close to one and light jets low values close to zero. The MV1 discriminant variable was used for b-tagging purpose by many Run I physics analises.

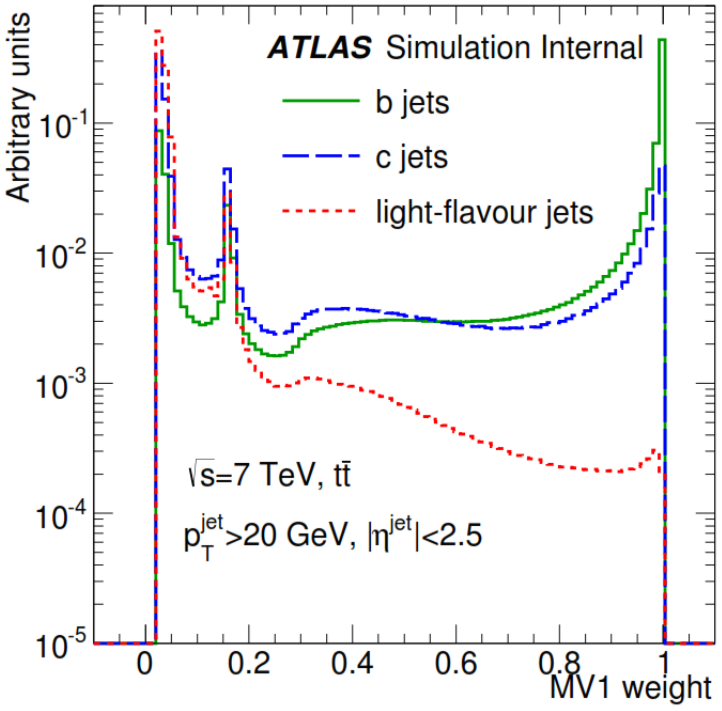

Figure 2.7: Distribution of the tagging weight obtained with the MV1 algorithm, for three different flavours: b-jet, c-jet and light jets. The spike around 0.15 corresponds mostly to jets for which no secondary vertex could be found.

2.4.5 MV2c20

Significant improvements are introduced in the b-tagging performance for the Run II, due to the addition of an extra pixel layer in the detector (the Insertable B-Layer) and from several enhancements to the tracking and b-tagging algorithms. One of the enhancements was to train the MV1 network against also c-jets introducing a c-jet component in the backround hypotesis. The resulting discriminant variable, MV1c, has a better c-jet rejection than MV1 with a small decrease in light jet rejection.

In addition, a new multivariate b-tagging algorithm was developed for Run II: MV2c20 [30]. The input variables obtained from the three basic algorithms are combined using a boosted decision tree (BDT) algorithm to discriminate b-jets from light jets and c-jets. The training is performed on a set of approximately 5 million $t\bar{t}$ events, controlling the fraction of c-jets added in the training: MV2c20 means

training with b-jets as signal and a mixture of 80% light jets and 20% c-jets as background. The 24 input variables used for the training include the jet p_T and η to take advantage of correlations with the other input variables.

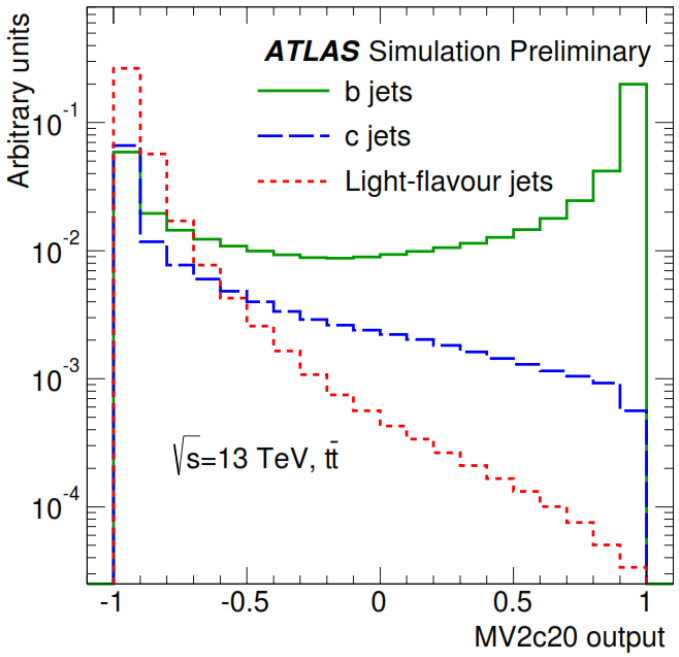

Figure 2.8: Distribution of the tagging weight obtained with the MV2c20 algorithm, for three different flavours: b-jet, c-jet and light jets.

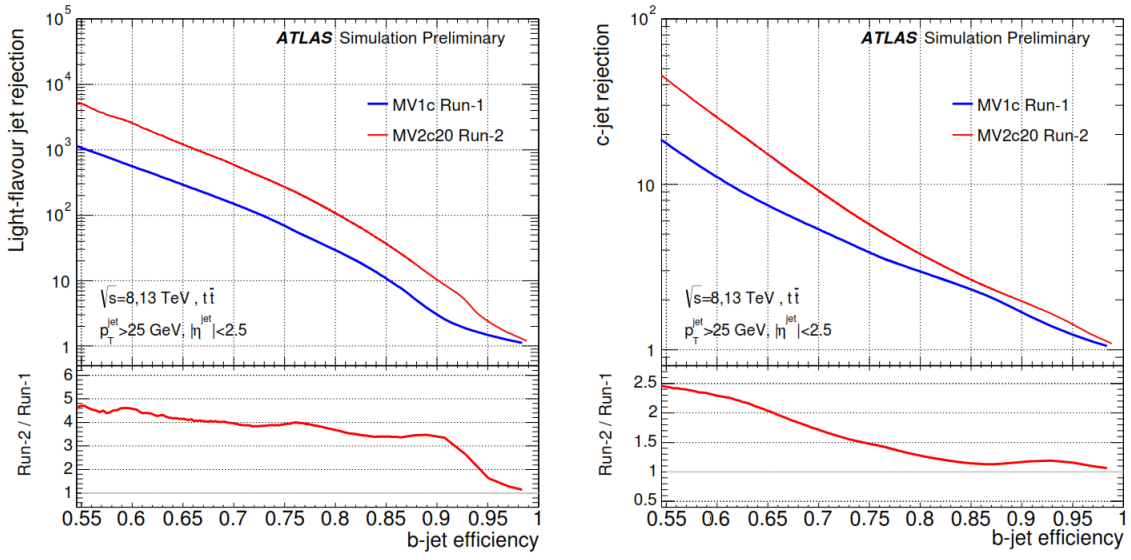

Figure 2.9: The light jet and c-jet rejection versus b-jet efficiency for the MV1c algorithm using the Run I detector and reconstruction software (blue) compared to the MV2c20 algorithm using the Run II setup (red). The samples were generated at different centre of mass energies (8 and 13 TeV) and different pile-up conditions, then the Run II sample was re-weighted in terms of jet p_T, η and average number of interactions in order to get an unbiased comparison with the Run I sample.

The performance for several background mixtures of c-jet and light jets in the training has been compared: the mixture adopted in MV2c20 gave the best c-jet rejection performance, which is considered preferable with respect to MV2c00 trained with only light jets and giving a slightly better light rejection. While MV1 was based on a neural network approach using also intermediate multivariate tools (see Fig. 2.6), the MV2c20 uses a BDT omitting the additional intermediate multivariate tools: this not only improves the performance, but also significantly simplifies the algorithm by directly using the variables from the basic algorithms.

As it can be seen in Fig. 2.9, the addition of the IBL, which results in an extra pixel layer closer to the beampipe, and several improvements to the tracking and b-tagging algorithms allowed to gain roughly a factor of 4 in the light jet rejection and a factor of 2 in the c-jet rejection with respect to the MV1c algorithm.

Chapter 3

Calibration of b-tagging efficiency

Any physics measurement adopts several reconstruction tools (e.g. for trigger selection or particle identification), whose behavior is well known from studies on simulated data. But this needs to be compared to what is seen in real data, to take into account possible differences and to validate the reliability of each tool. This process is usually called calibration and the uncertainty of its results is the best way to quantify how well we control each tool. This applies to each tool and algorithm used in any analysis, b-tagging included.

In order to use the b-tagging in physics analysis, it is requested to measure, on data collected by the detector, the efficiency of the tagger algorithms to correctly identify as a b-jet a jet originated from a b quark (ϵ_b: b-efficiency) or to mistakenly tag a jet originated from a charm quark (ϵ_c: c-efficiency) or from a light quark or a gluon (ϵ_l: l-efficiency or mistag rate). For each b-tagging algorithm, working points are defined as a function of the average b-jet efficiency measured in simulated data. The efficiency is measured for each of the working points with different methods, both on data and on Monte Carlo simulation samples, then the results are presented as data-to-simulation Scale Factors (SF) defined as the ratio of the efficiency in data to that in simulation.

In Tab. 3.1 and 3.2, some MV1 and MV2c20 working points are reported, along with the corresponding c-jet and light jet rejection factors. Rejection is measured as the inverse of the efficiency of selecting quark and gluon jets: a rejection factor of 50 would mean that, if the tagger is applied, the number of background jets is decreased by a factor of 50. Jets originated from τ leptons can be mistakenly tagged as b-jet too: in order to have a good b-tagging performance it is important to take them into account, so τ jets rejection factors are also reported in the tables.

Cut Value	b-jet Efficiency [%]	c-jet Rejection	light jet Rejection	τ jet Rejection
0.992515446	50	13.86	2545.27	46.25
0.9867	60	7.98	651.81	25.21
0.8119	70	4.99	150.00	14.07
0.6065	75	3.96	65.98	9.05
0.3900	80	3.09	27.13	5.64
0.1644	85	2.38	10.29	2.93
0.0616	90	1.75	2.97	1.56

Table 3.1: Working points for the MV1 b-tagging algorithm, including benchmark numbers for the efficiency and rejections rates. The statistical uncertainties are negligible.

Cut Value	b-jet Efficiency [%]	c-jet Rejection	light jet Rejection	τ jet Rejection
0.4496	60	21.33	1846.27	92.89
-0.0436	70	8.05	442.85	26.19
-0.4434	77	4.53	135.78	9.76
-0.7887	85	2.62	27.51	3.83

Table 3.2: Working points for the MV2c20 b-tagging algorithm, including benchmark numbers for the efficiency and rejections rates. The statistical uncertainties are negligible.

The samples used to perform each efficiency measurement must be characterized by a strong predominance of jets having the flavour to calibrate: the first step of each calibration is to select an almost pure sample of the requested flavour, estimating the flavour composition over data.

The c-jet tagging efficiency has been measured on an inclusive sample of jets associated to D* charged mesons (see Sec. 3.1.1) as well as in an inclusive sample of W+c events (see Sec. 3.1.2). The b-jet tagging efficiency has been measured in an inclusive sample of jets containing a muon (see Sec. 3.2.1) and in $t\bar{t}$ events with one or two leptons in the final state (see Sec. 3.2.2), while the mistag rate has been measured in an inclusive jet sample (see Sec. 3.3).

In this chapter a detailed description of the D* and W+c methods for c-jet tagging efficiency determination is reported, along with a brief description of the methods for b-jet tagging efficiency and mistag rate measurement used for the Run II calibrations plans.

3.1 c-jet tagging efficiency measurement

The charm tagging efficiency (i.e. the efficiency to mistakenly tag a jet originated from a charm quark as a b-jet) was measured with two different techniques: using a sample of c-jets containing D^{*+} mesons (from now on, referred as the D* method) [46,47] and using a sample of c-jets produced in association with a W boson (referred as the W+c method) [48], both described in details in two dedicated subsections.

The D* method uses a sample of c-jet containg D^{*+} mesons, reconstructed within a jet in the $D^{*+} \to D^0(\to K^-\pi^+)\pi^+$ final state. Since the c-jet sample is strictly exclusive (contains only c-jets with a D^{*+} meson decaying in a specific channel), the c-efficiency measurement must be extrapolated to an inclusive sample, taking into account both the fragmentation fractions and the branching ratios of the various charm hadron species present in a generic c-jet.

The W+c method used a sample of c-jets produced in association with a W boson, with the W boson decaying into an electron and a neutrino, and the c-jet identified via a soft muon stemming from a semileptonic c-hadron decay. Exploiting the charge correlation of the two leptons (electron and muon) it was possible to extract an almost pure c-jet sample. As seen for the D* method, also the W+c method requires an extrapolation to an inclusive sample of the c-jet tagging efficiency measurement performed over a precise exclusive sample containing a c-jet with a muon coming from the semileptonic decay of the c-hadron.

Due to the extrapolation procedure, the systematic uncertainty of the results increases: the D* method uncertainties vary depending on the p_T bin, from 10% to 40% in the 7 TeV calibration and from 8% to 15% in the 8 TeV calibration, while the W+c method uncertainties for the 7 TeV calibration are between 5% and 13%.

A third method to measure the c-jet tagging efficiency was proposed, but not developed yet: the $t\bar{t}$ method, using a sample of top pair events, both quarks decaying into Wb channel, with one W boson decaying to lepton and neutrino and the other W boson decaying hadronically. In this hadronic decay the interesting channel is $W \to sc$, containing a c-jet. So the events selected would present four jets (two b-jets, one c-jet and one light jet): even if the sample is not composed uniquely by c-jets, the fraction of c-jets is known (i.e. one c-jet on four). The $t\bar{t}$ method does not bias the c-jet selection, resulting in a completely inclusive sample with the advantage of not requiring any extrapolation of the result, significantly reducing the systematic uncertainties.

3.1.1 D* method

The c-jet tagging efficiency measurement with the D* method was performed almost with the same procedure on 4.6 fb^{-1} of 7 TeV proton-proton collision data recorded by ATLAS during 2011 [46] and on 20.3 fb^{-1} of 8 TeV data recorded in 2012 [47]. The D* method was the first method to be developed to measure the c-jet efficiency and was the reference method during Run I.

To perform a charm efficiency calibration, the optimal solution would be to have access to a pure sample of c-jets: the D* method uses events with a $D^{*+} \to D^0(\to K^-\pi^+)\pi^+$ reconstructed within a jet (implicitly including also the charge-conjugate D^{*-}). In this case, the jet sample selected is mostly made up by c-jets, with some contamination of b-jets with $b \to c$ decays, with a modest combinatorial background that can be subtracted by studying the signal sidebands. The contamination from

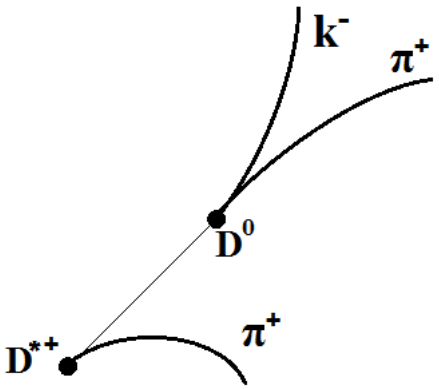

Figure 3.1: D^{*+} decay : $D^{*+} \to D^0\pi^+$, with $D^0 \to K^-\pi^+$

D* mesons that result from b-hadron decays is measured with a fit to the D^0 pseudo-proper time distribution.

After estimating the beauty contamination, the c-jet tagging efficiency is measured comparing the yield of D* mesons before and after applying the b-tagging requirement.

The data samples were collected using a logical "OR" of single jet triggers, allowing to populate the entire jet p_T range used in the calibration (7 TeV: 20 GeV < p_T < 140 GeV; 8 TeV: 20 GeV < p_T < 300 GeV). The Monte Carlo samples were generated with PYTHIA6 [49] (7 TeV) or with PYTHIA8 [19] (8 TeV), then re-weighted in terms of jet p_T and average number of interactions to match the data sample.

Event selection

To reconstruct the D^{*+} decay chain ($D^{*+} \to D^0\pi^+$, with $D^0 \to K^-\pi^+$), pairs of opposite sign charged tracks are selected, assigning them kaon and pion mass hypotheses, in order to reconstruct the D^0 decay. These D^0 candidates are then combined with charged particle tracks with opposite sign to that of the kaon candidate, assigning the pion mass to them. The so reconstructed D^{*+} decay chain is then associated with a reconstructed jet requiring its direction to be within $\Delta R = 0.3$ of the jet direction.

Some kynematic conditions are requested to select good candidates and reduce the combinatorial background:

- all tracks must have at least 5 hits in the silicon tracking detectors (Pixel and SCT), at least one of them in the Pixel detector;
- the transverse momenta of the kaon and pion candidates from the D^0 decay candidate have to fulfill $p_T > 1$ GeV;
- the reconstructed D^0 candidate mass must satisfy the condition $m_{K^-\pi^+}$ −

$m_{D^0} < 40MeV$, where m_{D^0} is the world average D^0 mass (1864.83 ±0.14 MeV [27]);

- the transverse momentum of the D^{*+} candidate has to exceed 4.5 GeV;
- the momentum of the D^{*+} candidate projected along the jet direction has to exceed 30% of the jet energy (50% in the 8 TeV calibration).

The kinematics of the decay cause the D^{*+} to release only a small fraction of energy to the prompt pion, usually called "slow pion"; for this reason the D^{*+} signal is commonly studied as a function of the mass difference Δm between the D^{*+} and D^0 candidates. D^{*+} mesons are expected to form a peak in the Δm distribution around 145.4 MeV (i.e. the difference between D^{*+} and D^0 mass), while the combinatorial background forms a rising distribution, starting at the pion mass (139.57 MeV [27]). The distributions of the mass difference Δm is divided in four p_T bins: [20, 30] GeV, [30, 60] GeV, [60, 90] GeV, [90, 140] GeV (7 TeV calibration) or [20, 30] GeV, [30, 60] GeV, [60, 140] GeV, [140, 300] GeV (8 TeV calibration).

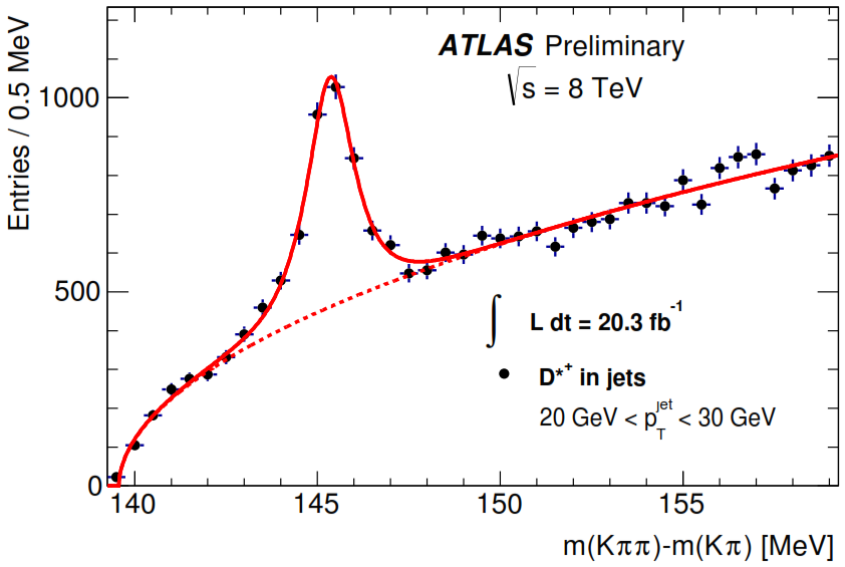

Figure 3.2: Plot of the mass difference distribution Δm of the D^{*+} candidates associated with 20 $GeV < p_T < 30$ GeV: the solid curve shows the total fit to the distribution, the dashed curve shows only the background contribution (8 TeV calibration) [47].

Mass fit

In order to determine the yield of the D^{*+} mesons, the mass difference distribution is fitted with a signal part S and a background part B. The signal is described with a modified Gaussian, giving a better description of the signal tails with respect to a simple Gaussian:

$$S = exp[-0.5 \cdot x^{(1+\frac{1}{1+0.5x})}] \tag{3.1}$$

where $x = |(\Delta m - \Delta m_0)/\sigma|$. Δm_0 and σ are free parameters in the fit and they represent the mean and the width of the Δm peak.

The combinatorial background is fitted with a power function multiplied by an exponential function, with α and β free parameters:

$$B = (\Delta m - m_\pi)^\alpha \cdot e^{-\beta(\Delta m - m_\pi)} \tag{3.2}$$

Background subtraction

Signal and background regions are defined as the region within 3σ around the D^{*+} mass peak and the region between 150 MeV and 168 MeV, respectively. This choice of the intervals aims at including almost all the signal events in the signal region and ensuring a negligible fraction of signal events in the background region.

The background fraction $f_B/(f_B+f_S)$ is computed by integrating both the signal S and background B functions in the Signal region (3σ around the Δm peak).

For the variable of interest, the D^0 pseudo-proper time, the distribution of events from the background region, normalised to the fitted background fraction in the signal region, is subtracted from the corresponding distribution in the signal region. The procedure relies on the assumption (verified on MC samples) that the distribution of the variable of interest is the same for the combinatorial background under the peak and in the sideband.

Flavour composition

As pointed out before, the D^{*+} candidates can come directly from original c-jets, but also from c-jets originated in the $b \to c$ decay in b-jets. The measurement of the flavour composition for the selected D^{*+} sample is of fundamental importance in order to obtain a c-jets pure sample for the c-jet tagging efficiency measurement. The discriminating variable used to identify beauty and charm components is the D^0 pseudo-proper time defined as:

$$t(D^0) = sign(\mathbf{L}_{xy} \cdot \mathbf{p}_T(D^0)) \cdot m_{D^0} \cdot \frac{L_{xy}(D^0)}{p_T(D^0)} \tag{3.3}$$

where m_{D^0} is the D^0 mass, $p_T(D^0)$ is the transverse momentum of the reconstructed D^0 candidate, $L_{xy}(D^0)$ is the distance in the transverse plane between the D^0 decay vertex and the primary vertex of the event.

Firstly, the charm and beauty D^0 pseudo-proper time distributions, referred as $F_c(t)$ and $F_b(t)$, are extracted from simulated samples originated from charm and beauty quarks respectively. The functional form of charm and beauty D^0 pseudo-proper time distribution $F_c(t)$ and $F_b(t)$ is fixed over the simulated distributions,

then their sum is built as follows:

$$F(t) = f_b \cdot F_b(t) + (1 - f_b) \cdot F_c(t) \tag{3.4}$$

where f_b is the beauty fraction of the signal, used subsequently to evaluate the c-jet tagging efficiencies. A binned likelihood fit is performed on the D^0 candidates pseudo-proper time distribution in order to extract the beauty fraction: f_b is a free parameter, while $F(t)$ is normalized to the integral of the fitted histogram. The fit is performed with the variable $ct(D^0)$ in the range [-1,2] mm after the background subtraction: c is the speed of light in vacuum and $t(D^0)$ is the D^0 pseudo-proper time.

The fit to background-subtracted data, in the jet p_T bin [30,60] GeV is shown in Fig. 3.3 and the results in the four bins of jet p_T are summarised in Tab. 3.3 (7 TeV and 8 TeV calibration).

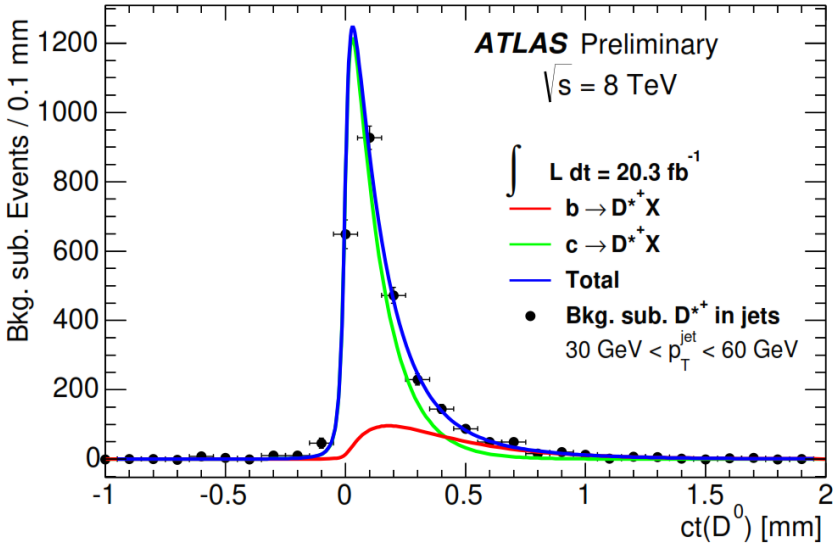

Figure 3.3: Fitted D^0 pseudo-proper time distributions to background-subtracted D^{*+} data samples in the jet p_T bin [30,60] GeV of 8 TeV calibration. The solid lines show the complete fit (blue), the beauty component (red) and the charm component (green).

p_T bin [GeV]	f_b (7 TeV)		p_T bin [GeV]	f_b (8 TeV)
20 - 30	0.212 ± 0.010		20 - 30	0.193 ± 0.017
30 - 60	0.315 ± 0.010		30 - 60	0.197 ± 0.014
60 - 90	0.303 ± 0.015		60 - 140	0.157 ± 0.014
90 - 140	0.315 ± 0.017		140 - 300	0.171 ± 0.023

Table 3.3: Beauty fractions determined by D^0 pseudo-proper time fits to the data in four jet p_T bins for 7 TeV and 8 TeV calibrations.

Tagging efficiency measurement

To measure the c-jet tagging efficiency over the selected sample, a combined fit to the Δm distribution of D^{*+} mesons in jets is performed before and after applying the b-tagging requirement, introducing an extra parameter $\epsilon_{D^{*+}}$ to describe the D^{*+} tagging efficiency.

This inclusive efficiency $\epsilon_{D^{*+}}$ is then decomposed into the efficiency for b and c-jets containing D^{*+}:

$$\epsilon_{D^{*+}} = f_b \epsilon_b(D^{*+}) + (1 - f_b)\epsilon_c(D^{*+}) \tag{3.5}$$

where f_b is the fraction of D^{*+} from beauty determined by the D^0 pseudo-proper time fit before applying the tagging requirement. The efficiency to tag a b-jet, ϵ_b, is taken from simulation and corrected by the data-to-simulation scale factors obtained from the b-jet efficiency calibration analysis (see Sec. 3.2).

The c-jet tagging efficiency determined with a D^{*+} sample is obtained solving the Eq. 3.5 with respect to ϵ_c:

$$\epsilon_c = \frac{\epsilon_{D^{*+}} - f_b \epsilon_b}{1 - f_b} \tag{3.6}$$

Extrapolation to inclusive c-jet sample

The obtained result needs to be extrapolated to a inclusive c-jet sample. The efficiency extrapolation factors for data and simulation X are defined as follows:

$$\epsilon_c^{data} = X^{data} \cdot \epsilon_c^{data}(D^{*+}) \tag{3.7}$$

$$\epsilon_c^{sim} = X^{sim} \cdot \epsilon_c^{sim}(D^{*+}) \tag{3.8}$$

where $\epsilon_c(D^{*+})$ is the c-jet tagging efficiency measured over a D^{*+} sample and ϵ_c is the c-jet tagging efficiency on an inclusive sample of c-jets. The inclusive scale factor $SF_c^{data/sim}$ is given by:

$$SF_c^{data/sim} = \frac{\epsilon_c^{data}}{\epsilon_c^{sim}} = \frac{X^{data}}{X^{sim}} SF_c^{data/sim}(D^{*+}) \tag{3.9}$$

While it is possible to evaluate X^{sim} over simulation, it is not currently possible to measure the corresponding factor in data. The extrapolation factor in data is therefore estimated from simulation, after re-weighting both the production fractions and the branching ratios in the simulation to the best experimental values.

Results

The results of the calibrations are expressed in terms of data-to-simulation scale factors, useful to correct the efficiencies in simulated events to those measured in data. The scale factors are compatible with unity within the uncertainties, while no

significant p_T dependence is observed. The 7 TeV calibration [46] contains results for MV1 60%, 70%, 75% and 85% working points along with some IP3D+SV1 and JetFitter+IP3D working points, while the 8 TeV calibration [47] contains results for MV1 60%, 70% and 80% working points.

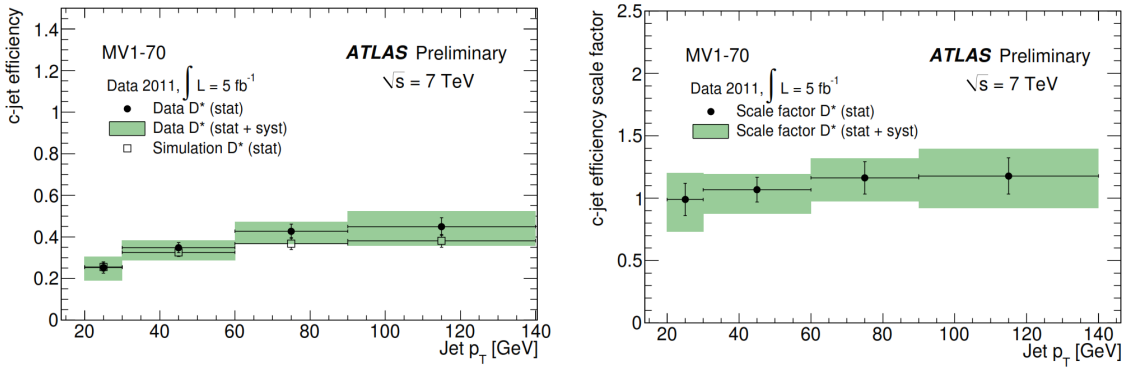

Figure 3.4: The c-jet tagging efficiency measurement results for data and simulation (left) and the corresponding data-to-simulation scale factors (right) for the MV1 tagging algorithm at 70% b-jet effciency working point (7 TeV calibration).

The uncertainties vary depending on the p_T bin, from 10% to 40% in the 7 TeV calibration and from 8% to 15% in the 8 TeV calibration. The most important systematic uncertainties are those related to the mass fit of D^{*+} mesons, to the extractions of the fraction of D^{*+} mesons originating from beauty hadrons and the extrapolation of the scale factor to that of an inclusive c-jet sample. Other minor systematic uncertainties are from the b-tagging efficiency scale factor, the jet energy scale, the jet energy resolution and the pile-up contamination.

3.1.2 W+c method

The c-jet tagging efficiency measurement with the W+c method was performed on 4.6 fb^{-1} of 7 TeV proton-proton collision data recorded by ATLAS during 2011 [48]. With a similar procedure, studies were also performed on 8 TeV data recorded in 2012.

To extract a pure c-jets sample, the W+c method selects events with c-jets produced in association with a W boson. The W boson is reconstructed via its decay into an electron and a neutrino, and the c-jet is identified via a soft muon stemming from a semileptonic c-hadron decay. The dominant production mechanism of the W+c signal is $gs \to W^-c$ and $g\bar{s} \to W^+\bar{c}$ where the W boson is always accompanied by a c quark with an oppositely signed charge, resulting in two opposite sign leptons. Most of the background processes are evenly populated with events where the charges of the decay leptons are of Opposite Sign (OS) or of Same Sign (SS). The charge correlation of the two leptons is exploited to extract a c-jet sample with high purity, as the difference between the number of events with opposite and with same charge

(OS-SS). The same event selection strategy was already exploited in several W+c production cross section measurements [58].

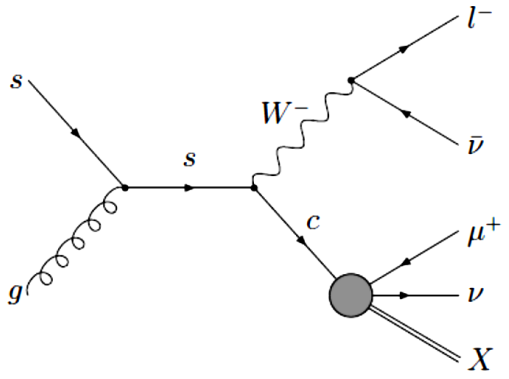

Figure 3.5: The Feynman diagram of the dominant W+c production.

After the determination of the c-jet tagging efficiency using the c-jet sample cointaining a soft muon produced in association with a W boson, an extrapolation to an inclusive c-jet sample is performed. The results are presented as data-to-simulation scale factors, that are applicable to an unbiased sample of inclusive c-jets.

The data samples were collected using single-electron trigger, selecting events with an electron of pseudorapidity $|\eta| < 2.47$ and $p_T > 20\ GeV$. The Monte Carlo samples were generated with PYTHIA [49] or HERWIG [20] with specific configuration for each sample. Besides the W+c sample (i.e. the signal process), many background processes were generated: W+b, W+light, top quark pairs, single-top quarks, dibosons (WW, ZZ and WZ), multijet and Z+jets samples.

Event selection

To select good events, many requests are applied over all the event components: electrons, missing energy, jets and muons.

Firstly, events are required to have at least one primary-vertex candidate: if more than one vertex is found, the vertex with the highest sum of the squared transverse momenta of associated tracks is selected as the primary vertex.

Electrons are required to have a transverse momentum $p_T > 25\ GeV$. The pseudorapidity range allowed for electrons is $|\eta| < 2.47$, excluding the calorimeter transition region $1.37 < |\eta| < 1.52$. Electrons are required to pass the "tight" identification criteria adopted for 2011 data. Only isolated electrons are selected, with a cut on the calorimeter-based quantity that estimates the transverse energy in the calorimeter cells within a cone of radius $\Delta R < 0.3$ around the electron direction (from now on, referred as ETCone). The ETCone of the electron should be lower than 3 GeV, with this request the electrons coming from hadrons decay within a jet are excluded. Exactly one electron with these qualities is allowed in each event.

The neutrino coming from the W decay is represented by the missing transverse momentum (from now on, referred as MET). Each event is requested to have $MET > 25\ GeV$; in addition the W boson transverse mass m_T^W is requested to be greater than 40 GeV:

$$m_T^W = \sqrt{2\ p_T^{el}\ MET(1 - cos(\phi_{el} - \phi_{MET})} \tag{3.10}$$

where p_T^{el} is the transverse momentum of the selected electron, MET is the missing transverse momentum of the event, ϕ_{el} and ϕ_{MET} are the azimuthal angle of the electron and the missing momentum.

Jets with $p_T > 25\ GeV$ and $|\eta| < 2.5$ are selected. To reduce pile-up contribution, jets are further requested to have Jet Vertex Fraction (see Sec. 2.3) $JVF > 0.75$. Exactly one jet fulfilling these conditions is allowed for each event.

The selected jet is required to have a "soft" muon inside, with $p_T > 4\ GeV$ and $|\eta| < 2.5$, coming from the c-hadron decay inside the jet. A specific matching procedure between the jet and the muon (defined Soft Muon Tagging: SMT [59]) provided a χ^2 representing the quality of the jet-muon association: a SMT $\chi^2 < 3.2$ is requested. Only "combined" muons are accepted, i.e. muons with a track both in the Inner Detector and in the Muon Spectrometer. Two impact parameter conditions are added: $|d_0| < 3\ mm$ and $|z_0\ sin(\theta)| < 3\ mm$. A set of ID hit requirements are requested to select high quality tracks, according to 2011 recommendations. The ETCone of the muon is requested to be higher than 3 GeV, resulting in a not isolated muon. Exactly one muon with the described quality associated to the selected jet is allowed per each event.

Signal extraction

In order to exploit the charge correlation of the W+c signal events, the sample is extracted as the difference between the number of events with opposite charged leptons and same signed leptons: $N^{OS-SS} = N^{OS} - N^{SS}$. After the subtraction some background events still remains, predominantly from W+light and multijet. Their contribution to the N^{OS-SS} in terms of OS-SS asymmetry is evaluated with data-driven methods. With this procedure, from the 4.6 fb^{-1} 7 TeV data, about 7445 OS events and 3125 SS events were selected, resulting in $N^{OS-SS} = 4320 \pm 100$. Of these events, the estimated SMT W+c events are $3910 \pm 100(stat) \pm 160(sys)$, resulting in a 90% pure c-jet sample.

Tagging efficiency measurement

The high purity of the selected SMT c-jet sample is well suited to measure the c-jet tagging efficiency $\epsilon_{c(\mu)}^{data}$, defined as the fraction of W+c selected events that pass the b-tagging requirement:

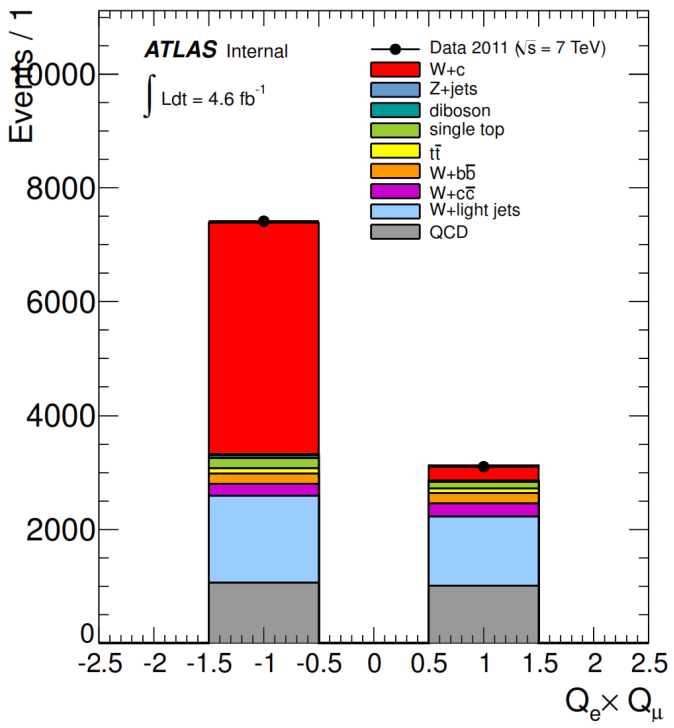

Figure 3.6: Number of Opposite Sign (left) and Same Sign (right) events: the W+c events (red) are dominant in the OS selection, while the most important backgrounds are W+light (blue) and multijet (grey).

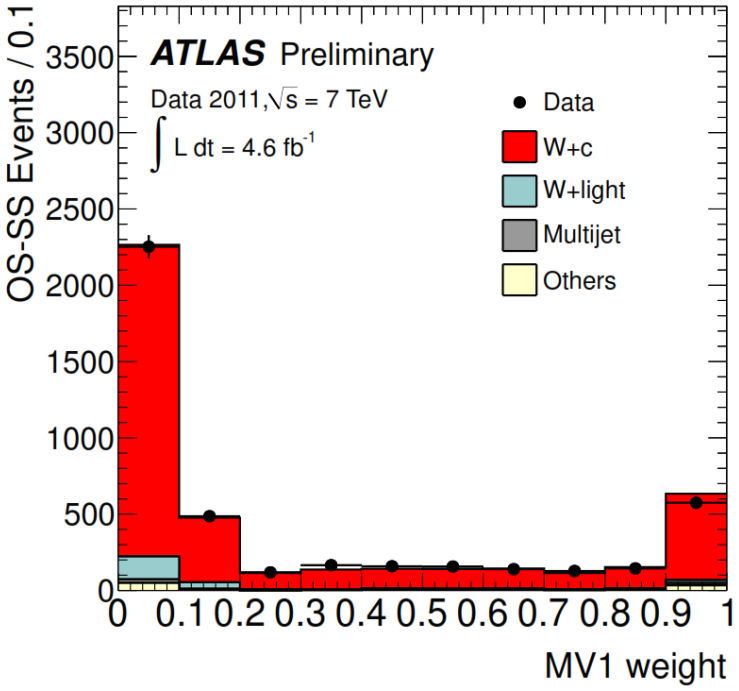

Figure 3.7: MV1 output weight distribution for N^{OS-SS} events: the W+c process is about 90% of the total events, few events from W+light and multijet processes remains.

$$\epsilon_{c(\mu)}^{data} = \frac{N_{W+c}^{b-tag}}{N_{W+c}} \tag{3.11}$$

where N_{W+c} is the number of W+c events before applying the b-tagging requirement (also referred as pretag sample) and N_{W+c}^{b-tag} is the number of W+c events passing the b-tagging requirement. The small background contribution is estimated using MC simulation and subtracted.

The c-jet tagging efficiencies for the MV1 tagging algorithm derived for SMT c-jets measured on 7 TeV data vary between 13% and 50% with the total uncertainty ($3\% - 10\%$) increasing with the tightness of the working point. The systematic uncertainties are dominated by the precision on the W+light and multijet background yields at pretag level, the statistical uncertainties are of the same order as the systematic uncertainties. The c-jet tagging efficiency is measured for a simulated sample too and a data-to-simulation scale factor is presented. Due to several differences between an inclusive sample and a sample of SMT c-jets the derived scale factors $SF_{c(\mu)}^{data/sim}$ need to be extrapolated in order to be applicable to samples of inclusive c-jets.

Extrapolation to inclusive c-jet sample

Selecting a sample of c-jets via the semimuonic decays of c-hadrons results in a different composition of c-hadron types with regard to an inclusive sample due to the different semileptonic branching ratios of the c-hadrons. The c-jet tagging efficiency of an inclusive sample of c-jets ϵ_c can be obtained from the c-jet tagging efficiency of SMT c-jets $\epsilon_{c(\mu)}$ by applying an extrapolation factor α.

$$\epsilon_c = \alpha \cdot \epsilon_{c(\mu)} \tag{3.12}$$

Similarly to the D* method, the efficiency extrapolation factors for data and simulation α are defined as follows:

$$\epsilon_c^{data} = \alpha^{data} \cdot \epsilon_{c(\mu)}^{data} \tag{3.13}$$

$$\epsilon_c^{sim} = \alpha^{sim} \cdot \epsilon_{c(\mu)}^{sim} \tag{3.14}$$

The inclusive scale factor $SF_c^{data/sim}$ is given by:

$$SF_c^{data/sim} = \frac{\epsilon_c^{data}}{\epsilon_c^{sim}} = \frac{\alpha^{data}}{\alpha^{sim}} SF_{c(\mu)}^{data/sim} \tag{3.15}$$

The α^{sim} extrapolation factor is derived from MC samples, resulting in a value of about 0.8 for the different working points. The corresponding extrapolation factor for data α^{data} is estimated using simulated-corrected sample to minimize differences between data and simulation. The fragmentation fraction of the relevant weakly decaying c-hadrons of simulation are re-weighted to the value measured by many

e^+e^- and $e^\pm p$ collision experiments, while the semileptonic branching ratios of c-hadrons are corrected to match the world average values.

Results

The results are presented as data-to-simulation scale factors in Fig. 3.8: the values vary between 0.75 and 0.92 with total relative uncertainties of 13% to 5%. The scale factors decrease and the corresponding uncertainties increase with increasing tightness of the working point. Three main sources of uncertainties are of the same order: the statistical uncertainty, the systematic uncertainty on the measured scale factors for the SMT c-jet sample and the systematic uncertainty due to the extrapolation procedure. The main source for the systematic uncertainties due to the performed extrapolation is the limited knowledge of the charged particle multiplicity of c-hadron decays which has a significant impact on the c-jet tagging efficiency.

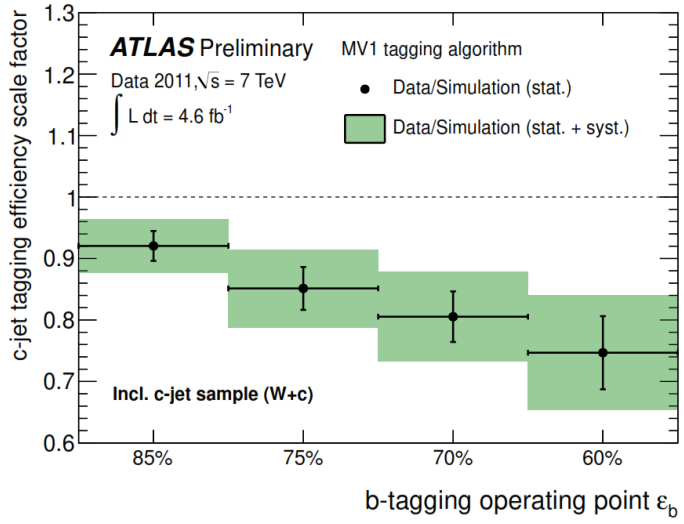

Figure 3.8: The c-jet efficiency data-to-simulation scale factors for different MV1 tagging algorithm working points.

3.2 b-jet tagging efficiency measurement

The beauty efficiency (i.e. the efficiency to tag a jet originated from a beauty quark as a b-jet) has been measured in an inclusive sample of jets with muons inside (p_T^{rel} method) [51] and in samples of $t\bar{t}$ events with one or two leptons in the final state ("Tag counting" and PDF method) [50, 54, 55].

To have different methods is convenient in order to calculate properly the efficiency in different configurations: e.g. the p_T^{rel} calibration method is less precise than the $t\bar{t}$ based methods in high p_T ranges. The p_T^{rel} is a simple method able to give a good estimation of the efficiency with small data samples; the large amount of data collected by LHC allows an alternative measurement of the b-jet tagging

efficiency, based on $t\bar{t}$ events. Compared to the p_T^{rel} method, based on muon di-jet events, the $t\bar{t}$ based methods provide a b-jet tagging efficiency measurement in an inclusive jet sample rather than a sample of semileptonic b-jets and cover a larger range in p_T.

3.2.1 p_T^{rel} method

For the b-jet tagging efficiency calibration with the p_T^{rel} method, the pure b-jet sample is obtained by selecting jets containing a muon: this sample results naturally enriched in b-jets, because of the semileptonic decay of the b-hadrons.

The p_T^{rel} method uses templates of the muon momentum transverse to the jet axis (relative p_T: p_T^{rel}) to measure the flavour composition. The b-jet tagging efficiency is extracted with a fit to the fraction of b-jets before and after b-tagging.

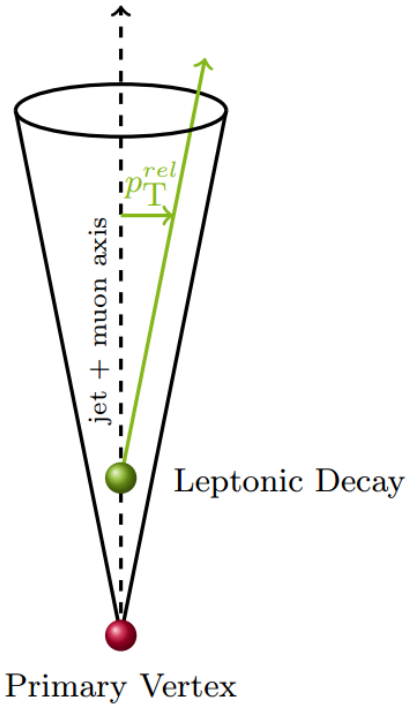

Figure 3.9: Scheme of the p_T^{rel} definition with respect to the jet+muon axis.

The events used in the analyses (about 5 fb^{-1} at 7 TeV) were collected with triggers that require a muon reconstructed from hits in the Muon Spectrometer and spatially matched to a calorimeter jet. To reduce the dependence on the modelling for muons in light-flavour jets, the heavy-flavour content in the sample is increased by requiring that there is at least one jet in each event, other than those used in the measurement, with a reconstructed secondary vertex.

The number of b-jets before and after b-tagging can be obtained for a subset of all b-jets, namely those containing a reconstructed muon, using the p_T^{rel} variable which is defined as the momentum of the muon transverse to the combined muon

plus jet axis. Muons originating from b-hadron decays have a harder p_T^{rel} spectrum than muons in c and light-flavour jets, as can be see in Fig. 3.10. Templates of p_T^{rel} in simulated events are constructed for b, c and light-flavour jets separately, and these are fitted to the p_T spectrum of muons in jets in data to obtain the fraction of b-jets before and after b-tagging.

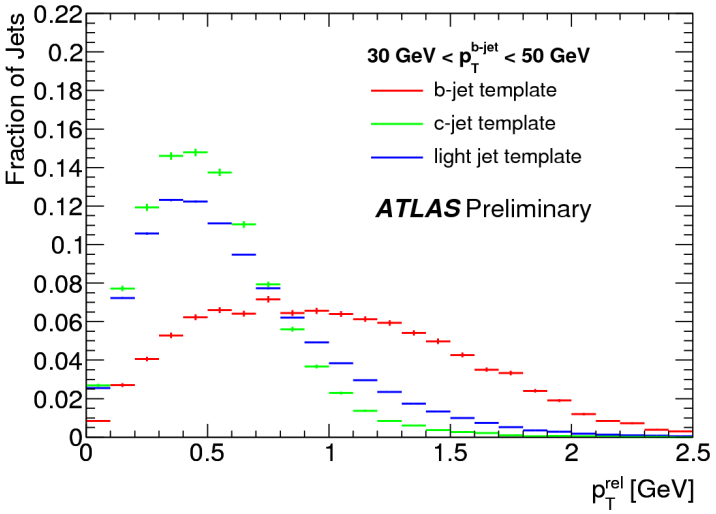

Figure 3.10: The three templates for the p_T^{rel} for jets with 30 GeV < p_T < 50 GeV, normalized to unity. The b and c templates are obtained from the Monte Carlo samples, and the light-jet template is the track-based template from data. [52]

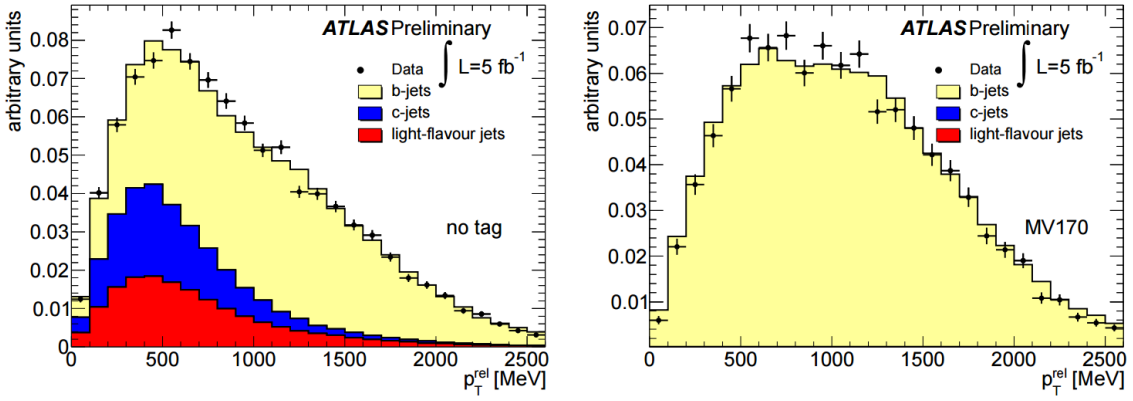

Figure 3.11: Template fits to the p_T^{rel} distribution in data before (left) and after (right) b-tagging, MV1 tagging algorithm, 70% working point, 60-75 GeV jet p_T bin.

As the templates from c and light-flavour jets have a rather similar shape, the fit can only reliably separate the b-jets from non-b-jets. Therefore, the ratio of the c and light flavour fractions is constrained in the fit to the value observed in simulated events.

The efficiency measured for b-jets with a semileptonically decaying b-hadron in data is compared to the efficiency for the same kind of jets in simulated events to

compute the corresponding data-to-simulation efficiency scale factor.

The p_T^{rel} templates for b and c-jets are derived from the simulated sample, using muons associated with b and c-jets. It has been verified that the pre-tagged and tagged template shapes agree within statistical uncertainties. The template for light jets is derived from muons in jets in a light flavour dominated data sample.

As the p_T^{rel} method is directly affected by the calorimeter jet axis determination, a difference in the jet direction resolution between data and simulation or an improper modelling of the angle between the b quark and the b hadron in simulation would cause the p_T^{rel} spectra in simulation and data to disagree, introducing a bias in the measurement. To smear this effect, a new track-based jet axis was defined, formed by the vector addition of the momenta of all tracks in the jet, and then the samples were corrected in agreement with this definition, more compatible with the data profile.

The p_T^{rel} method can only measure the b-jet tagging efficiency in data for b-jets with a semileptonic b-hadron decay. The b-jet tagging efficiency is sligthly different between samples with jets decaying semileptonically and hadronically, because the charged particle multiplicity is different in the two decays channel and because the jets used in the p_T^{rel} analysis always contain a high-momentum muon track, whereas the hadronic b-jets do not.

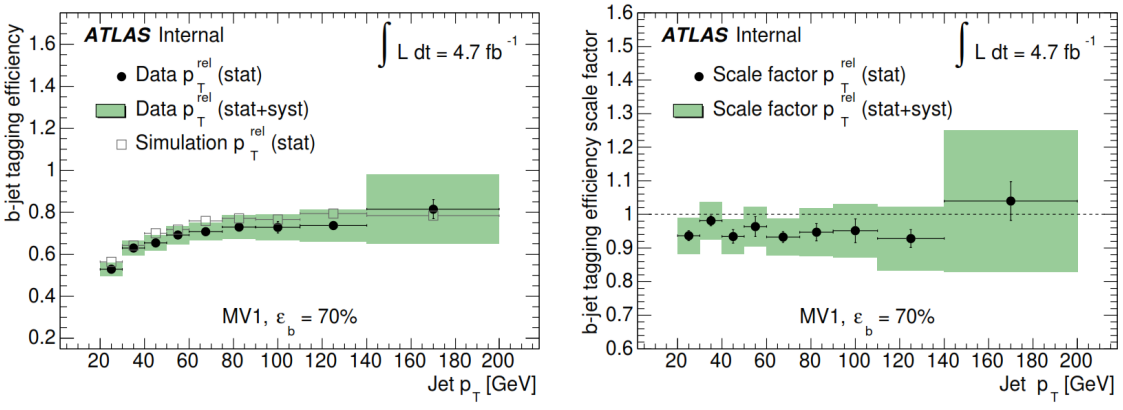

Figure 3.12: The b-jet tagging efficiency in data and simulation (left) and the corresponding data-to-simulation scale factors (right) for the MV1 tagging algorithm at 70% efficiency obtained with the p_T^{rel} method.

However, assuming that the simulation adequately models the relative differences in b-jet tagging efficiencies between semileptonic and hadronic b-jets, the same data-to-simulation scale factor is valid for both types of jets. This assumption was investigated measuring the data-to-simulation scale factor separately for jets with and without muons using a high purity sample of b-jets in $t\bar{t}$ dilepton events. The ratio of the data-to-simulation scale factors for jets with and without muons was found to be consistent with unity for all tagging algorithms and operating points. The uncertainty on the measurement is assigned as a systematic uncertainty on the

scale factors. Other typical systematic uncertainties are due to heavy flavour production, decay and fragmentation provided by simulation models, along with the jet energy scale and the pile-up jets contamination. The total uncertainty (systematic and statistical) ranges from 5 to 20 %, increasing for higher jet p_T bins, resulting that the p_T^{rel} calibration method is less precise than the$t\bar{t}$ based methods in high p_T ranges.

3.2.2 $t\bar{t}$ based methods

In the case of $t\bar{t}$ based b-tagging efficiency measurement, the "Tag counting" method fits the multiplicity of b-tagged jets in $t\bar{t}$ candidate events, while the PDF method exploites the kinematic correlations between the jets in the event. The "Tag counting" selection methods are applied to both the single-lepton and di-lepton decay channels, while the PDF method is applied only to the di-lepton channel.

The events used in $t\bar{t}$ based calibration are selected with single electron or muon trigger, further requiring at least four jets in the single-lepton channel (e or μ) or at least two jets in the di-lepton channel (ee, $e\mu$ or $\mu\mu$). In the single-lepton channel, a treshold for the missing transverse momentum and the missing transverse mass is applied, while in the di-lepton channel a missing transverse momentum and an invariant two-lepton mass selection is applied. Due to W+jets, Z+jets and fake lepton reconstruction, both channels have dedicated background estimation studies to calculate the fraction of the selected events.

The top quark $t \to Wb$ branching ratio is higher than 99% [27], so each $t\bar{t}$ event is expected to contain almost exactly two b-jets in the final state. In reality, not all of the b-jets are well reconstructed or tagged, since the b-jets from top quark decay can be outside the detector acceptance and additional b-jets coming from gluon splitting can contaminate the sample. Moreover, c-jets and light jets, coming from hadronic W boson decay, can be tagged as b-jets.

The agreement in the scale factors among the two methods is very good. The scale factors are close to unity, with an uncertainty ranging from 4% to 30%, depending on the jet p_T bin and on the calibration method.

"Tag counting" method

The "Tag counting" method evaluates the expected fractions of events containing b-jets, c-jets and light jets passing the selection, with Monte Carlo studies. The expected number of events with n b-jets is calculated as the sum of all these contributions. Then the b-jet tagging efficiency is extracted fitting the expected event counts to the observed counts. Different approaches are used to take into account the backgrounds for the single-lepton and the di-lepton channels. The events are divided in p_T bins, a likelihood function is used to extract the beauty, charm and light jet fractions for each of the p_T bin.

PDF method

The PDF method (also referred as Combinatorial likelihood method) uses di-lepton $t\bar{t}$ events, applying a different approach for the four channels considered: events with two or three selected jets, with $e\mu$ leptons or ee and $\mu\mu$ leptons. The b-jet tagging efficiency is calculated for each of the four channels with an unbinned maximum likelihood fit. All different jet flavour combinations are present in the likelihood, with a different PDF (Probability Density Function) for each combination. These PDFs are determinated over simulated samples, except for the b-jet weight PDF, containing the useful information to be extracted from data.

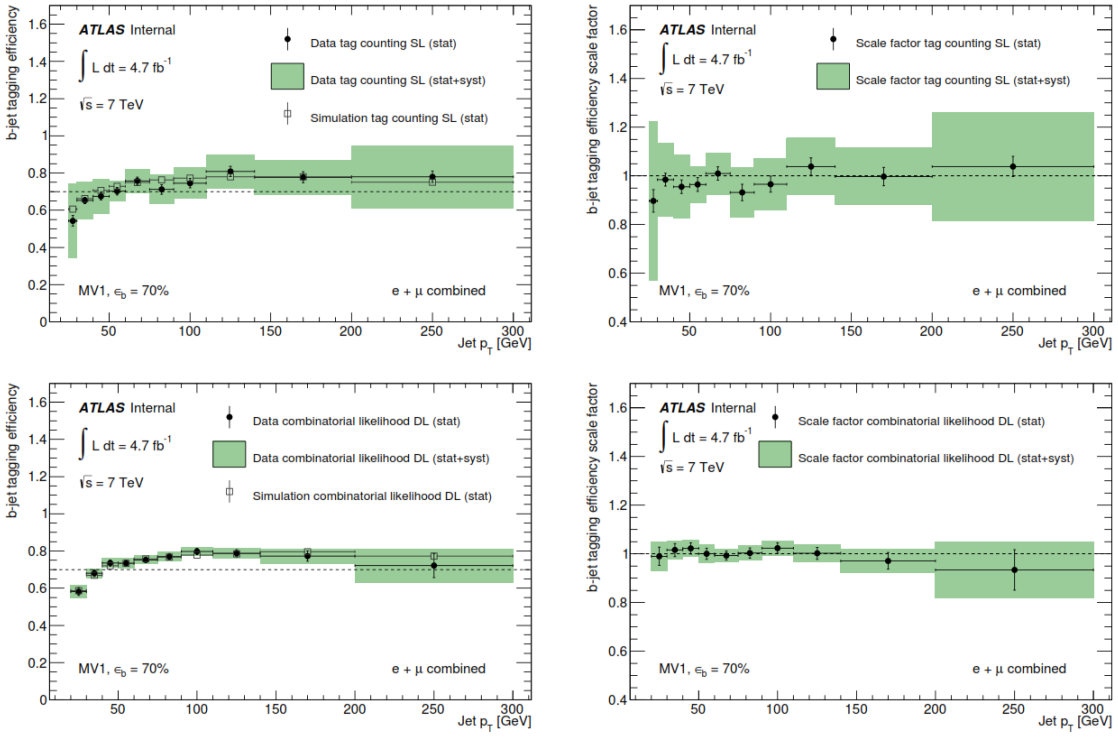

Figure 3.13: The b-jet tagging efficiency in data and simulation (left) and the corresponding data-to-simulation scale factors (right) for the MV1 tagging algorithm at 70% efficiency: "Tag counting" method on single-lepton channel (top) and PDF method on di-lepton channel (bottom).

3.3 Light jet tagging efficiency measurement

The light jet tagging efficiency or mistag rate or (i.e. the efficiency to mistakenly tag a jet originated from a light quark or a gluon as a b-jet) has been measured in an inclusive jet sample with the "Negative tag" method [47, 53].

The sample used for the mistag rate measurement was collected using a logical "OR" of single jet triggers; simulated inclusive jet samples were used too, similarly to the D^{*+} analysis.

Light-flavour jets are tagged as b-jets mainly because of the finite resolution of the Inner Detector and the presence of tracks stemming from displaced vertices from long-lived particles or material interactions. The lifetime-signed impact parameter distribution of these tracks as well as the signed decay length of vertices reconstructed with these tracks are expected to be symmetric. The inclusive tag rate obtained by reversing the impact parameter significance sign of tracks for impact parameter based tagging algorithms, or reversing the decay length significance sign of secondary vertices for secondary vertex based tagging algorithms, is therefore expected to be a good approximation of the mistag rate due to resolution effects.

For advanced tagging algorithms based on likelihood ratios or neural networks, the "Negative tag" rate is computed defining a negative version of the tagging algorithm which internally reverses the impact parameter and the decay length selections. The mistag rate is approximated by the "Negative tag" rate of the inclusive jet sample, multiplied by two correction factors taking into account for the finite resolution of the detector and the tracks stemming from long-lived particles.

Systematic uncertainties are due to jet energy scale, pile-up, trigger bias, heavy flavour efficiency and long lived particle decays. The total uncertainties on the mistag rate scale factors go from 10% to 50%.

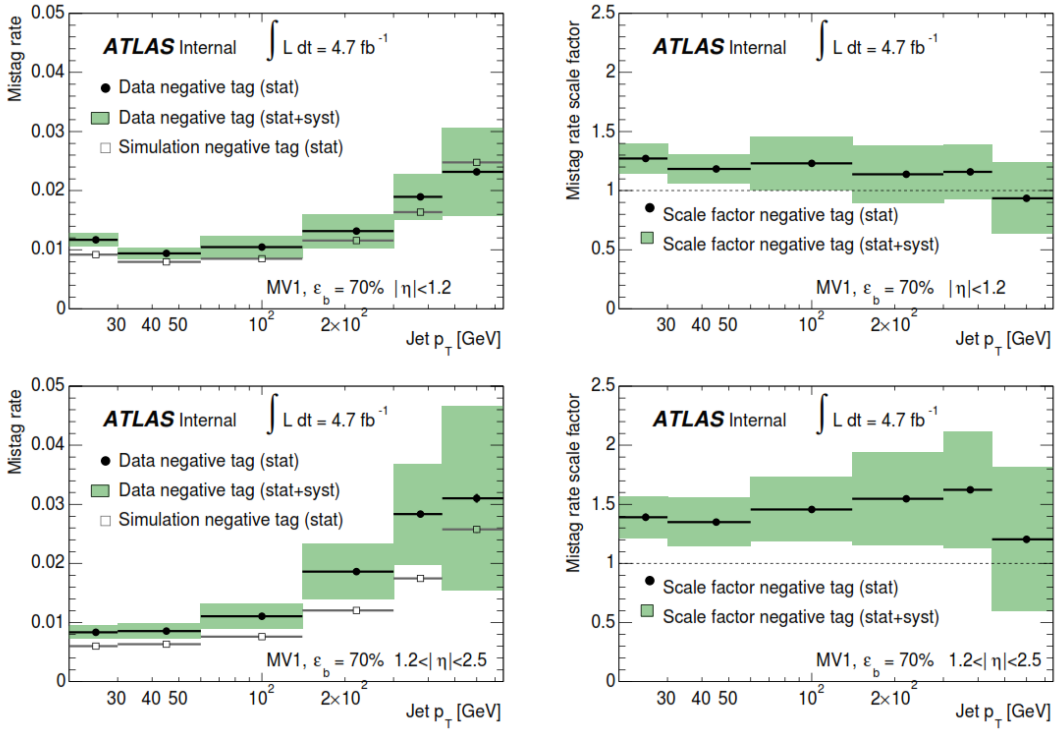

Figure 3.14: The mistag rate measured in data and simulation (left) and the corresponding data-to-simulation scale factors (right) for the MV1 tagging algorithm at 70% efficiency.

Chapter 4

D* calibration

The first part of this thesis work is focused on the measurement of the c-jet tagging efficiency with the D* method (see Sec. 3.1.1): a study on 8 TeV data was performed in order to gain more precision, in view of the measurement to be performed on 13 TeV data.

The D* calibration procedure (almost the same for 7 TeV [46] and 8 TeV [47] measurements) contained a binned fit of the mass plot to extract the signal peak and a binned fit on the D^0 pseudo-proper time distribution, performed after the background subtraction, in order to extract the beauty and charm fraction of the signal.

Different procedures were tested in order to improve the mass fit and the flavour fraction determination: the main aim was to perform a simultaneous fit on both the D* mass and the D^0 pseudo-proper time variable, including the background modelling. The methods developed to perform the combined fits did not gave the expected results, so the D* mass and the D^0 pseudo-proper time were performed separately. The combined fit would have allowed to avoid the background subtraction and all the consequences due to its application, including the possible correlations between the mass fit and the D^0 p.p. time fit.

Another target was to move from a working point calibration (i.e. the calibration of the b-tagging variable for each of the working points, done separately) to a continuous calibration [56], allowing to measure in one single step the efficiency in the whole range of the b-tagging variable (divided in bins).

In this chapter, a brief description of the variables used in the calibration is given (in Sec. 4.1), then the best procedures are discussed along with their results (in Sec. 4.2).

4.1 Variables definition

To perform this test study, the whole 20.3 fb^{-1} data sample collected by ATLAS during 2012 was used. Only jets with $p_T > 20\ GeV$ and $|\eta| < 2.5$ are selected and

then calibrated using simulation based correction factors for p_T and η. To reduce the contamination of pile-up jets, the jets with $p_T < 50\ GeV$ and $|\eta| < 2.4$ are required to satisfy the Jet Vertex Fraction selection: $JVF > 0.5$. Jets are reconstructed with the anti-k_t algorithm from the energy clusters calibrated with the Local Cluster Weighting method. These jets (LC jets) are calibrated to the hadronic energy scale using p_T and η dependent correction factors obtained from simulation. The Jet Energy Scale (JES) is one of the main systematic uncertainties of this measurement, always present in the case of analysis involving jets. Reconstructing the D^{*+} decay chain ($D^{*+} \to D^0\pi^+$, with $D^0 \to K^-\pi^+$) as described in Sec. 3.1.1, it was possible to select about 80000 events compatible with the selection criteria listed before.

For each of these events, 4 useful variables were stored in the Ntuple, as this is the most common ROOT-compatible data format used during the analysis. The considered variables are:

- jet p_T;
- MV1 b-tagging variable;
- Δm;
- D^0 p.p. time.

4.1.1 Jet p_T

The jet transverse momentum p_T is one of the most important variable describing a jet: as for all the physics objects emerging from a collision, the transverse momentum indicates the momentum in the transverse plane of the detector (see Sec: 1.2.7). If the transverse momentum of a physics object is higher than $\sim$ 10 GeV, there is a much higher possibility it is an interesting physics event, while if the transverse momentum is lower, no new physics is expected.

The jets used in the c-jet tagging efficiency measurement with the D* calibration are selected in a specific range of p_T: $20 - 300\ GeV$. The jets are then divided in 4 p_T bins and for each of them the calibration procedure is equally reproduced. The jet p_T bins considered are listed in Tab. 4.1 along with the number of events for each bin: the jet with higher p_T are less frequent, so the bins have a different width in order to include enought events for a suitable statistic treatement (see Fig. 4.1).

Jet p_T bin	Range [GeV]	Number of events
1	20 - 30	43756
2	30 - 60	19515
3	60 - 140	9284
4	140 - 300	6521

Table 4.1: Definition of the 4 jet p_T bins and number of events for each of them.

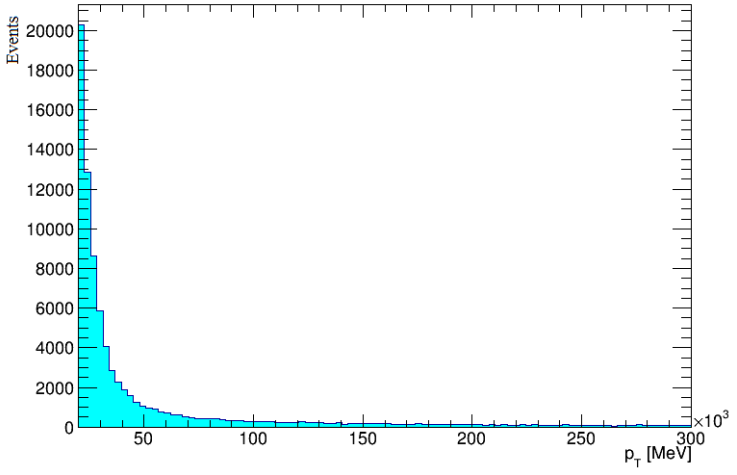

Figure 4.1: Inclusive p_T distribution of the 8 TeV data sample selected jets.

4.1.2 MV1 b-tagging variable

During Run I analysis, the most commonly used b-tagging variable was obtained with the MV1 algorithm (described in Sec. 2.4.4). The MV1 distribution (shown in Fig. 2.7) has high values close to 1 for b-jets and low values close to 0 for light jets. The c-jets have a MV1 distribution that is much similar to that of b-jets, making them very difficult to separate.

In Fig. 4.2 it is shown the inclusive distribution of MV1 b-tagging variable of the 8 TeV data sample. Instead of performing the analysis for each of the MV1 b-tagging working points, in this study the jets (present in a specific jet p_T bin) were divided into 5 tag weight bins, according to the 50%, 60%, 70% and 80% working points, as reported in Tab. 4.2.

Tag weight bin	Initial MV1 value	b-jet efficiency	Final MV1 value	b-jet efficiency
1	1	0%	0.992515446	50%
2	0.992515446	50%	0.9867	60%
3	0.9867	60%	0.8119	70%
4	0.8119	70%	0.39	80%
5	0.39	80%	0	100%

Table 4.2: Definition of the 5 tag weight bins in terms of the initial and final MV1 value; for clarity, the corresponding b-jet efficiency is also reported for each of them.

4.1.3 Δm

The D* decay chain reconstruction allows to determine a useful variable for the analysis, able to discriminate between real D* events and the background: the mass difference variable Δm. It is defined as:

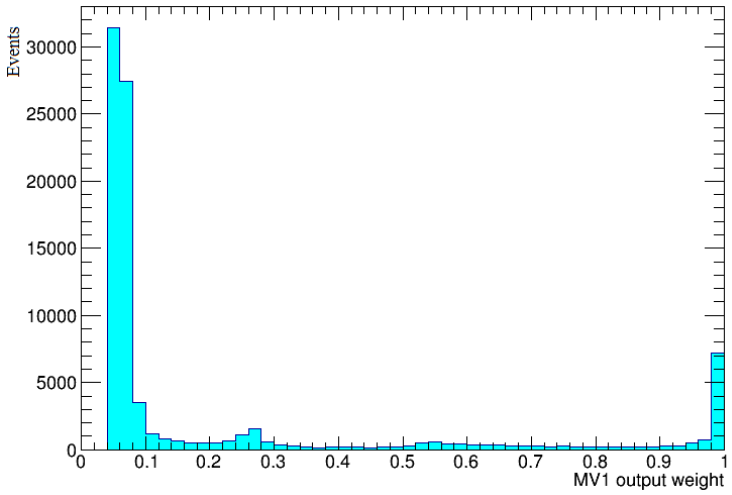

Figure 4.2: Inclusive distribution of the 8 TeV selected jets: MV1 b-tagging variable.

$$\Delta m = m_{K^-\pi^+\pi_s^+} - m_{K^-\pi^+} \tag{4.1}$$

where K^- and π^+ are from the $D^0 \to K^-\pi^+$ decay and the π_s^+ is the slow pion of the $D^{*+} \to D^0\pi_s^+$ decay (charge-conjugate D^{*-} decay chain is implicitly included).

Meson	Measured mass [MeV]	Quarks structure
D^{*+}	2010.28	$c\bar{d}$
D^0	1864.86	$c\bar{u}$
π^+	139.57	$u\bar{d}$

Table 4.3: List of the mesons involved in the D^{*+} decay with their measured mass and quark components. The $D^{*+} \to D^0\pi^+$ decay channel has a 67.7% branching ratio. The extrapolated Δm value between D^{*+} and D^0 is about 145.42 MeV. [27]

The D^{*+} events form a peak in the Δm distribution around 145.4 MeV (i.e. the difference between D^{*+} and D^0 mass), while the combinatorial background forms a rising distribution, starting at the pion mass (139.57 MeV [27]).

4.1.4 D^0 pseudo-proper time

The discriminating variable used to separate the beauty and charm component of the signal in the D^{*+} peak is the D^0 pseudo-proper time $t(D^0)$, defined as:

$$t(D^0) = sign(\mathbf{L}_{xy} \cdot \mathbf{p}_T(D^0)) \cdot m_{D^0} \cdot \frac{L_{xy}(D^0)}{p_T(D^0)} \tag{4.2}$$

where m_{D^0} is the D^0 mass, $p_T(D^0)$ is the transverse momentum of the reconstructed D^0 candidate, $L_{xy}(D^0)$ is the distance in the transverse plane between the D^0 decay vertex and the primary vertex of the event.

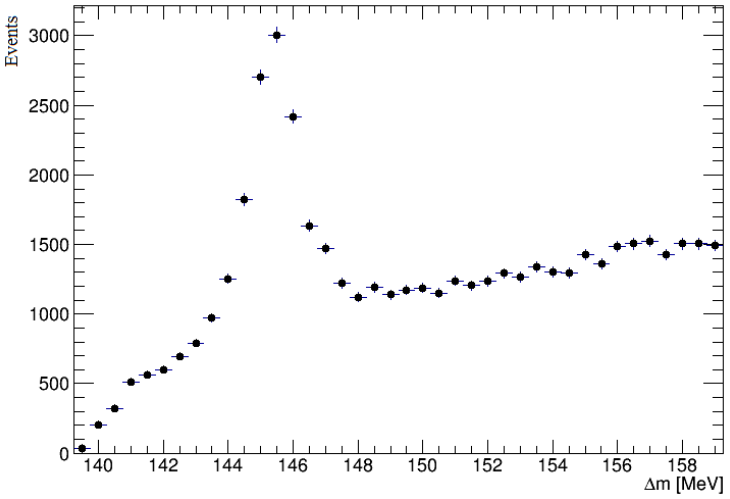

Figure 4.3: Inclusive distribution of the 8 TeV selected jets: Δm variable.

The D^0 p.p. time is multiplied for the light speed c, obtaining the final variable $ct(D^0)$ measured in mm units (see Fig. 4.4).

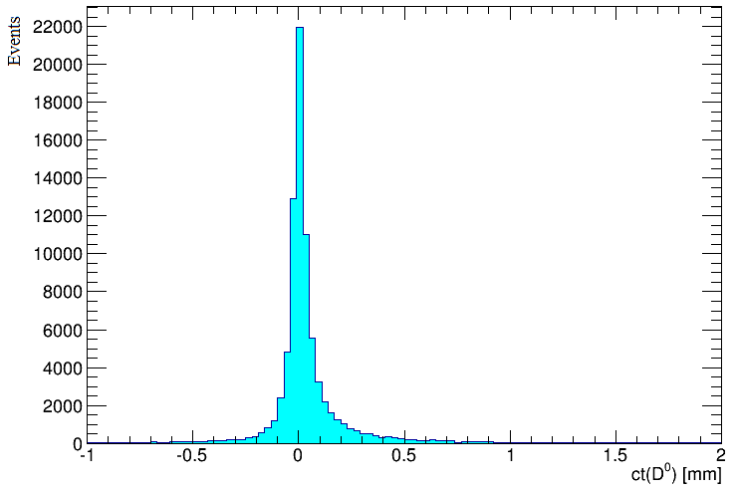

Figure 4.4: Inclusive distribution of the 8 TeV selected jets: D^0 p.p. time variable.

The charm and beauty D^0 pseudo-proper time distributions used to fit the two components over data are extracted from simulated samples originated from charm and beauty quarks respectively. As can be seen in Fig. 4.5, the beauty component (red) has a peak rising from 0 mm with a steady descending slope up to 1 mm, while the charm component (green) begins slightly before 0 mm and has a descending slope ending at about 0.5 mm.

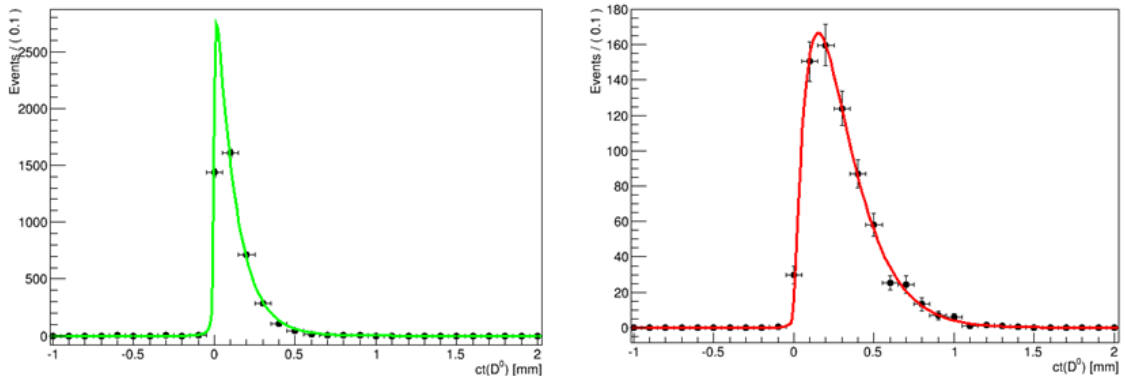

Figure 4.5: Typical distribution of the D^0 p.p. time variable for simulated jets originated from charm (left) and beauty (right) quarks.

4.2 Fitting procedures

The first step of the method to be developed was the implementation of a unbinned likelihood fit on the D* mass peak (see Sec. 4.2.1), identified in the mass difference variable Δm distribution. Among the many procedures tested in order to extract the flavour composition, the one giving the best results was a new version of the precedently used background subtraction (see Sec. 4.2.2)

4.2.1 Unbinned Likelihood on mass

The mass fits performed in the 7 TeV and 8 TeV calibrations were binned fits: the binned fit aims to establish the best fitting function over a set of points, each of them representing the value of the variable in the considered bin range. The fit performed in the calibrations used arbitrary fixed values of the parameters responsible for the center and the width of the D* mass peak. The systematic uncertainty due to the parameters constraints was the most important systematic affecting the measurement.

The unbinned likelihood fit presented in this thesis is performed in a very different way: one of the main target is to perform a fit without constraints on the parameters, in order to remove the related systematic contribution. This fit is performed in bins of tag weight: since the center and the width of the mass peak are p_T dependent, they are set as free parameters. Moreover, the values obtained in the fit are inspected in order to evaluate if a fixed value could be preferred.

Theoretical introduction

Considering n indipendent and identically distributed observations of the x variable:

$$x_1, ...x_n \tag{4.3}$$

the PDF (Probability Density Function) of the variable x_i is a function f that depends on a given set of parameters $\overrightarrow{\theta}$:

$$f(x_i, \overrightarrow{\theta}) \tag{4.4}$$

The likelihood function L is the joint PDF for the whole data sample:

$$L(\overrightarrow{\theta}) \;=\; f(\overrightarrow{x}, \overrightarrow{\theta}) \;=\; \prod_{i=1}^{n} f(x_i, \overrightarrow{\theta}) \tag{4.5}$$

The value for which the likelihood function is maximum also gives the maximum value of its logarithm (the log-likelihood function): maximazing the log-likelihood function is the equivalent of minimizing the χ^2 for a given distribution.

$$ln\ L(\overrightarrow{\theta}) \;= -\frac{1}{2}\chi^2(\overrightarrow{\theta}) \tag{4.6}$$

The poissonian distribution describing the probability of a given number of random and indipendent events is:

$$P(n;\lambda) = \frac{\lambda^n}{n!}e^{-\lambda} \tag{4.7}$$

where λ is the mean number of the expected events and n is the observed number of events.

In the case of a sample containing n events, consisting of n_s events due to the signal process and n_b events due to background processes, it is possible to estimate the number of events contributing to signal and background with the extended log-likelihood function. The poisson distribution of the mean values μ_s and μ_b (of the corresponding random variables n_s and n_b) is therefore:

$$P(n;\mu_s,\mu_b) = \frac{(\mu_s+\mu_b)^n}{n!}e^{-(\mu_s+\mu_b)} \tag{4.8}$$

The PDF for the variable x can be expressed as:

$$f(x;\mu_s,\mu_b) = \frac{\mu_s}{(\mu_s+\mu_b)^n}f_S(x) + \frac{\mu_b}{(\mu_s+\mu_b)}f_B(x) \tag{4.9}$$

where $f_S(x)$ and $f_B(x)$ are the PDFs of the Signal and Background functions already defined in Sec. 3.1.1.

The extended likelihood is defined as:

$$L(x_i;\mu_s,\mu_b) = \frac{(\mu_s+\mu_b)^n}{n!}e^{-(\mu_s+\mu_b)}\prod_{i=1}^{n} f(x_i, \overrightarrow{\theta}) \tag{4.10}$$

Then the extended log-likelihood (without constant terms) is:

$$ln\ L(x_i;\mu_s,\mu_b) = -(\mu_s+\mu_b) + \sum_{i=1}^{n} ln[(\mu_s+\mu_b)f(x_i;\mu_s,\mu_b)] \tag{4.11}$$

Maximizing the $ln\ L$ (or eventually minimizing the opposite sign function $-ln\ L$) allows to extract directly the mean values of the number of signal and background events.

Overall, the fit contains 6 free parameters to be determined for each of the 5 tag weight bin, resulting in a total number of free parameters equal to 30. The free parameters are:

- μ_s and μ_b, the number of signal and background events;
- Δm_0 and σ, the mean and the width of the Δm peak;
- α and β, describing the background shape.

Performed fit and results

One single log-likelihood minimization is performed on all the events belonging to the considered jet p_T bin, fitting at the same time all the 5 tag weight bins mass distributions. The equation used in the fit procedure, expressed in terms of f_S and f_B, is:

$$-\ ln\ L(x_i;\mu_s,\mu_b) = \mu_s+\mu_b - \sum_{i=1}^{n} ln(\mu_s f_S + \mu_b f_B) \tag{4.12}$$

The extended log-likelihood fit on mass is successfully performed on the 4 jet p_T bins for each of the 5 tag weight bins. As example, the mass fit in the 5 tag weight bins of the jet p_T bin [20,30] GeV is shown in Fig. 4.6. The top left plot corresponding to the tag weight bin N.5 is dominated by background events, then the background contamination decreases with the increasing of the b-jet tagging efficiency.

The mean and the width of the Δm peak for the 5 tag weight bins obtained from the unbinned likelihood fit have been evaluated. The tag weight bin N. 4 has a lower number of events with respect to the other bins: the extracted values are not in well agreement with the other bins and have a higher statistical uncertainty. Apart from that bin, there is a good trend of the values, but they are not as much in agreement as expected: it is preferred to keep the mean and the width of the Δm peak as free parameters in the fit.

One of the possible ways to check the validity of the results of the unbined likelihood fit is presented. The Raw Efficiency for a defined MV1 working point is calculated as the ratio between the number of signal events passing the b-tagging requirement (with MV1 weight of the event $<$ MV1 working point weight) and the

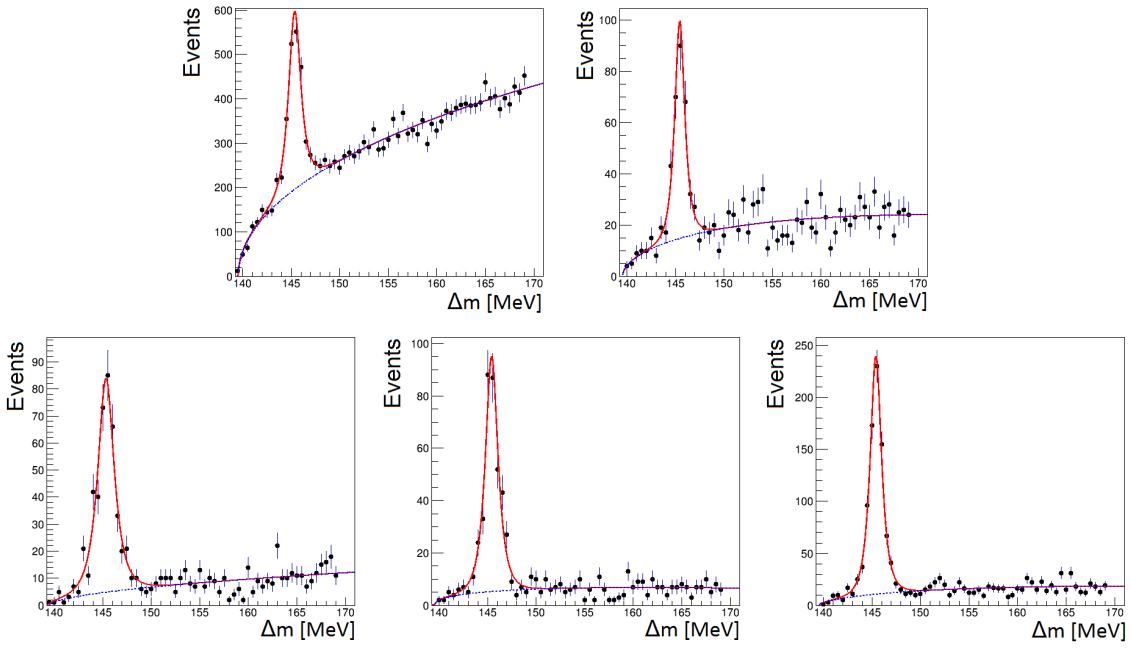

Figure 4.6: Unbinned likelihood fit (extended) on mass performed in the 5 tag weight bins of the jet p_T bin [20,30] GeV: the solid red curve shows the total fit to the distribution, the blue dashed curve shows only the background contribution.

number of all the signal events:

$$\epsilon_{Raw} = \frac{\mu_s \; tagged}{\mu_s \; total} \tag{4.13}$$

For the continuous b-tagging treatment, the number of signal events is extimated in each of the 5 tag weight bins. The number of the signal events selected by a 70% working point requirement is simply obtained summing the number of signal events in the tag weight bins N. 1, 2 and 3 (corresponding to b-jet tagging efficiency of $0\% - 50\%$, $50\% - 60\%$ and $60\% - 70\%$ respectively).

In Tab. 4.4, for each jet p_T bin the Raw Efficiencies at 70% MV1 b-jet efficiency are reported for both the 8 TeV calibration binned fit and the extended unbinned likelihood fit described above. The results are compatible within their statistical uncertainties.

Jet p_T [GeV]	Binned fit	Unbinned fit
20-30	0.402 ± 0.017	0.390 ± 0.018
30-60	0.458 ± 0.011	0.447 ± 0.019
60-140	0.488 ± 0.010	0.488 ± 0.022
140-300	0.510 ± 0.018	0.524 ± 0.037

Table 4.4: The Raw Efficiencies at the 70% MV1 b-jet efficiency: the results obtained with the 8 TeV calibration binned fit and the extended unbinned likelihood fit are compatible.

4.2.2 Background subtraction

The background subtraction procedure described in this section is a reviewed version of the background subtraction tecnique used in the 7 TeV and 8 TeV calibration (see Sec. 3.1.1). The precedent version used an inclusive binned mass fit to extrapolate the background fraction, then the expected number of background events was subtracted bin per bin to the inclusive D^0 p.p. time distribution and the subtracted distribution was fitted with only the beauty and charm components.

The background subtraction tecnique developed in this study contains a two steps procedure: an unbinned likelihood fit on mass (Root based) on each of the 5 bins of tag weight and a binned fit (RooFit based) with the beauty and charm distributions to the background subtracted D^0 p.p. time distribution.

The first step is the unbinned likelihood fit described in Sec. 4.2.1: starting from the mean number of signal μ_s and background μ_b events in the Signal region (within 3σ around the Δm peak), the background fraction α_{bkg} is calculated as follows:

$$\alpha_{bkg} = \frac{\mu_b}{\mu_s + \mu_b} \tag{4.14}$$

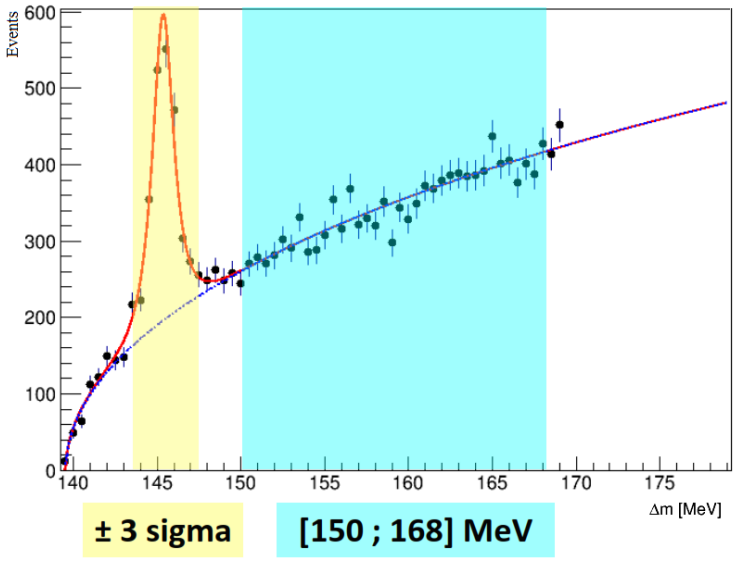

Figure 4.7: Definition of the Signal and the Sideband region on the mass plot.

The Sideband region is defined as the region containing events with the Δm value in the range 150-168 MeV. In this way, the events selected are quite totally from bakground processes, ensuring a negligible fraction of signal events. Using the calculated α_b values, the corresponding fraction of background events is subtracted bin per bin for each of the 5 tag weight bins from the corresponding D^0 p.p. time distribution of the events in the Signal region, using the D^0 p.p. time distribution of the Sideband region to emulate the shape of the background distribution. The D^0 p.p. time distributions of the 5 tag weight bins, after the background subtraction,

are summed together to obtain an inclusive D^0 p.p. time distribution to be fitted.

The second step is the fit on this inclusive distribution performed with RooFit. The beauty and charm distributions derived from simulated samples (see Fig. 4.5) are mixed together to describe the signal:

$$\alpha_b f_b + (1 - \alpha_b) f_c \tag{4.15}$$

where f_b and f_c are the PDFs describing the D^0 p.p. time distribution of the b-jet sample and c-jet sample modeled over MC distribution, and the beauty fraction parameter α_b is the only free parameter in the fit.

The main difference between the previous background subtraction tecnique and the current version, is to treat the D^0 p.p. time distribution differently in the 5 bins of tag weight instead of subtracting the background inclusively. The results of the fits performed with this procedure are shown in Fig. 4.8: the fit reaches convergence for all the 4 jet p_T bins.

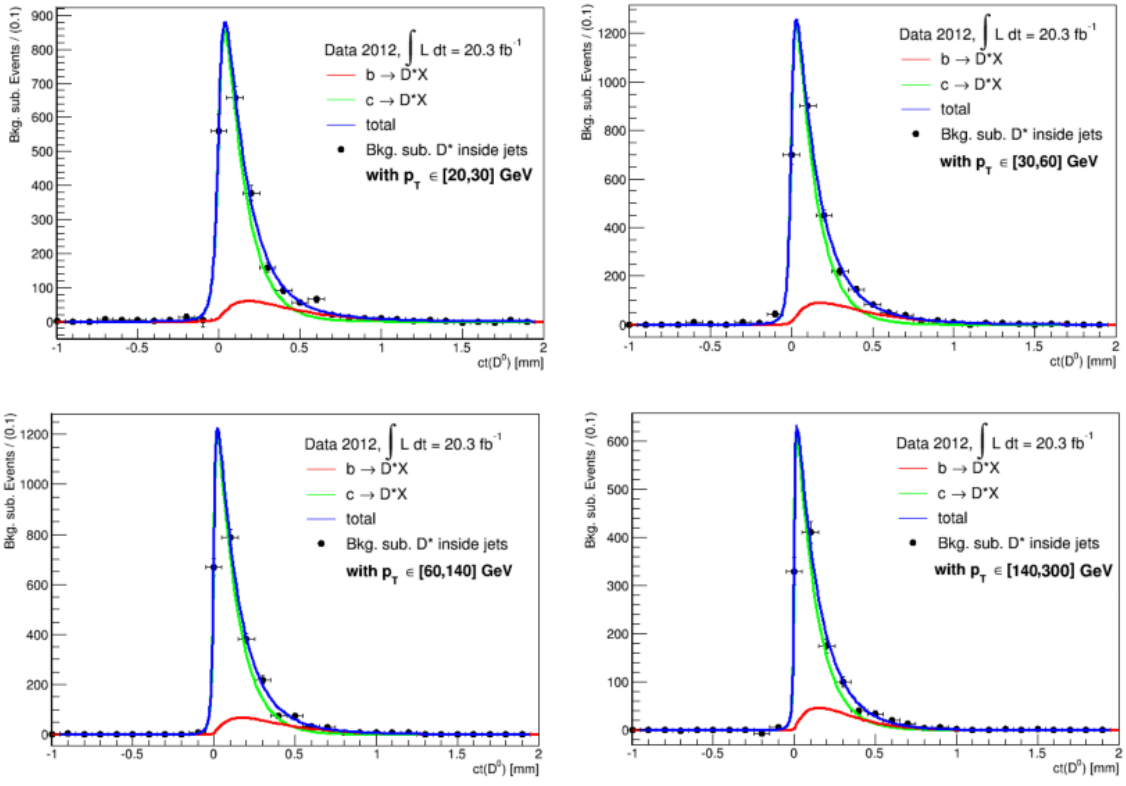

Figure 4.8: Binned fit on the background subtracted D^0 p.p. time distribution of the Signal region events, for the 4 jet p_T bins. The main function (blue) is the sum of the beauty (red) and the charm (green) distributions.

The beauty fraction parameters extracted by the fits are reported along with their uncertainty in Tab. 4.5: the values found with this procedure are compatible within the uncertainties with the values of the 8 TeV calibration [47], with the

exception of the jet p_T bin with range of 20-30 GeV. The uncertainty of the values are very similar to each other for the two procedures.

Jet p_T	8 TeV calibration	This study
20-30 GeV	0.193 ± 0.017	0.166 ± 0.015
30-60 GeV	0.197 ± 0.014	0.183 ± 0.013
60-140 GeV	0.157 ± 0.014	0.150 ± 0.014
140-300 GeV	0.171 ± 0.023	0.168 ± 0.024

Table 4.5: The beauty fraction value with the corresponding uncertainty extracted by the 8 TeV calibration binned fit and by this background subtraction tecnique.

4.3 Conclusions

In this thesis work, a new mass fit procedure was developed: the unbinned likelihood fit allows to extract the needed values without fixing any parameters, as was done previously in the binned fit. This improvement will allow to remove the systematic uncertainty due to the application of constraints on the parameters during the fit.

Moreover, four different fit procedures were developed in order to extract the flavour composition of the signal, each of them with its own advantages and disadvantages. The developed procedures involving a combined fit on mass and D^0 p.p. time distributions did not converge (maybe due to neglected correlations between the tag weight bins and the D^0 p.p. time) or converged with a higher uncertainty on the results (not being an improvement with respect to the previous fit procedure).

The best procedure has resulted to be a new version of the previous background subtraction tecnique, with the subtraction done in bins of tag weight, thanks to the information extracted from the unbinned likelihood fit in bins of tag weight. In the future, an estimation of the validity of the error can be done studying the pull distribution of generated toy samples.

When the 13 TeV data will be available, these procedures could be tested over real data, along with the reference 8 TeV fit procedure. From the comparison of the results provided with all the methods, a cross-check on the α_b value will be possible, then the best method will be selected and used for the 13 TeV calibration.

Chapter 5

W+c calibration

The second part of this thesis work is focused on the developement of the code needed for the measurement of the c-jet tagging efficiency with the W+c method (see Sec. 3.1.2) on Run II data: the first tests on 13 TeV data samples are performed, along with Monte Carlo samples corresponding to the interested processes.

The W+c calibration performed on 7 TeV [48] is based on the exploitation of the charge correlation of the W+c process with respect to the background processes: a tight event selection is applied in order to efficiently separate Opposite Sign and Same Sign events (containing leptons with same and opposite charge, see Sec. 3.1.2).

In this chapter, a brief introduction of the ATLAS Event Data Model is given (in Sec. 5.1), along with the software used (in Sec. 5.2), then there are dedicated sections describing the data and MC samples (in Sec. 5.3 and 5.4), the event selection (in Sec. 5.5) and the results obtained (in Sec. 5.6).

5.1 Event Data Model

The Event Data Model (EDM) is the collection of classes, interfaces, concrete objects and their relationships which, together, provide a representation of the events detected by ATLAS. Thanks to a defined EDM, there is a common definition of the objects (electrons, muons, jets...) and the analysis benefits in terms of software usage and flexibility. Data and Monte Carlo events have a different origin but they are treated in a similar way in order to obtain for both of them the same final format.

The EDM used for Run I analysis was the Analysis Object Data (AOD), and it was designed to be used only in Athena [18], the ATLAS analysis framework. The new EDM developed for Run II is referred as xAOD [60] (also commonly referred as AOD): it was designed to be easily adopted in both full Athena jobs and in very lightweight ROOT [23, 24] analysis jobs. Fig. 5.1 shows the current plan for a standard ATLAS physics analysis, from left to right:

- The first data object in the procedure is the xAOD, a very heavy file ($\sim PB$) structured as a tree-branch structure containing every single useful information

coming from each sub detector and for each kind of reconstructed physics object. The xAOD file can be opened with the ROOT browser in order to be explored (i.e. check the content of the variables), but it is not possible to perform directly the analysis on the xAOD due to its weight and its complexity.

- The Derivation Framework (also named Reduction framework) is the software implemented in Athena that produces a format smaller than the original xAOD, suitable for specific physics analyses, the DAOD (from Derived xAOD, also referred as DxAOD) with the same tree-branch structure of the original xAOD. During the reduction it is possible to remove whole events (skimming), entire physics objects (thinning) or some of their detailed information (slimming). For example, the events without a muon can be removed, along with the number and position of the hits of the Inner Detector tracks and the whole information regarding the MET. Not only the reduction is possible: during this step, some Combined Performance tools (CP) can be applied, smearing and correcting some aspects of the variables. It is also possible to add complex physics objects to the DAOD, for examples variables referred to mesons reconstructed starting from tracks and kinematic information contained in the xAOD. Each physics and combined performance group has his own Derivations, one for each class of analysis.

- The DAODs are much more lightweight than the xAODs ($\sim TB$), but they are still complex objects to handle. The third step in the data reduction is the process that extracts information from the DAOD to build a "flat" ROOT Ntuple (where "flat" refers to the distribution of variables in the file, which does not have the tree-branch structure of DAODs and xAODs). This reduction procedure can be performed both in Athena and in RootCore [26] (a software framework with packages developed in c++), and many CP tools are applied in this step. The Ntuples are much more lightweight ($\sim GB$) and can be easily stored directly on the user laptop or the LXPLUS (Linux Public Login User Service) platform [61] used for the final analysis.

- The last step is commonly represented by the analisis macro, a ROOT-readable c++ file that applies the event selection and produces the histograms of interest for the analysis. The results are then presented in written publications with histograms.

The proton-proton collisions are firstly selected by the ATLAS trigger, then the whole useful information of the selected events is stored in a raw form in the CERN Data Center, representing the World LHC Grid Tier-0 site. The Tier-0 performs also the first fast reconstruction of the raw data (and a second bulk reconstruction about 2 days later): the xAOD produced are stored in the Tier-0 and are identically reproduced in some of the 13 Tier-1 sites (computer centers of Scientific Institutes around the world). Then the Physics groups and the Combined Performance groups (i.e.

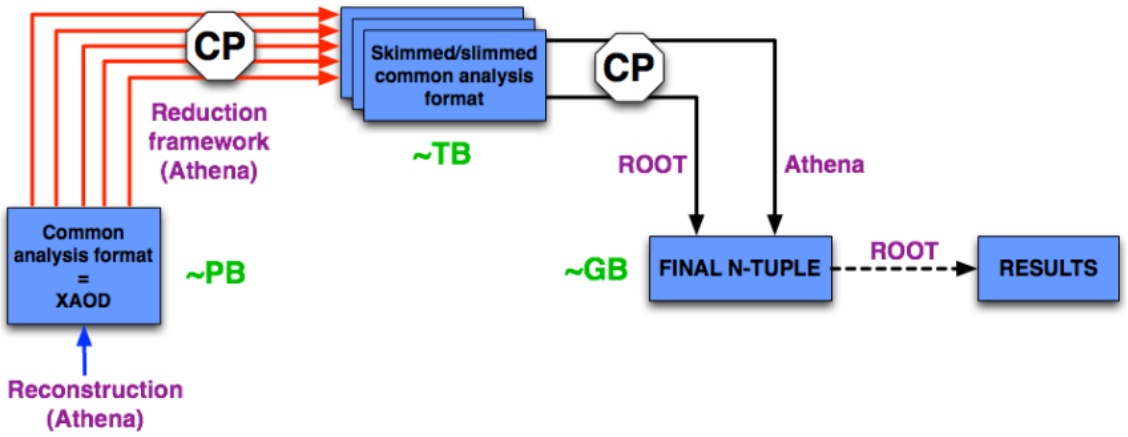

Figure 5.1: The standard procedure commonly used in ATLAS physics analysis: from heavy xAOD to DAOD, then the Ntuple and the results in form of values and histograms.

SUSY, Exotics physics, Higgs physics, Standard Model physics, Muon performance, Flavour Tagging performance) perform the data reduction, producing the DAODs: there are different DAOD formats for different analyses. In particular, the Flavour Tagging group uses 4 different DAOD formats named FTAG1, FTAG2, FTAG3 and FTAG4 (see Tab. 5.1). The DAOD are tipically stored in the Tier-2 sites, about 152 universities and scientific institutions around the world that provide adequate storage and computing power. The final part of the analysis involving Ntuples and hisograms is performed on the Tier-3 sites, represented by local institution computer resources or personal computers.

DAOD format	Calibration analysis
FTAG1	D* method, Negative tag method
FTAG2	$t\bar{t}$ methods (di-lepton)
FTAG3	p_T^{rel} method
FTAG4	W+c method, $t\bar{t}$ methods (single lepton)

Table 5.1: The 4 different DAOD formats used by the Flavour Tagging group.

The Monte Carlo simulated events are generated with dedicated software, then go through the simulation, digitization and reconstruction steps. Each physics process has its own simulation chain: at the end of the chain there are many xAOD representing for example $t\bar{t}$, W+jets, Z+jets. These xAOD are derived and then the Ntuples are extracted from the DAOD.

5.2 Software and tools

The ATLAS experiment developed specific software in order to access the information present in the xAOD and DAOD stored on the Grid. The first step of my thesis

work on the W+c calibration analysis is represented by the search of the data and MC samples.

One of the most useful tools is AMI [62] (Atlas Metadata Interface): it allows to find information on the existing data and MC samples, as number of events or cross section of the MC processes. Besides the metadata information, AMI contains the whole dataset name, including the AMI tags, numbers referring to the specific configuration of the software used during the processing.

The typical DAOD data sample name is:

$$data15\_13TeV.00276262.physics\_Main.merge.DAOD\_FTAG4.f620\_m1480\_p2411$$

where $data15\_13TeV$ indicates it is a data sample collected during 2015 at 13 TeV energy in the center of mass reference, 276262 is the run number, $DAOD\_FTAG4$ is the derivation format, $f620$ is the tag number for bulk reconstruction, $m1840$ refers to the merging and $p2411$ is the derivation framework configuration.

The typical DAOD MC sample name is instead:

$$mc15\_13TeV.361301.Sherpa\_CT10\_Wenu\_Pt0\_70\_CFilterBVeto.merge.$$
$$DAOD\_FTAG4.e3651\_s2586\_s2174\_r6869\_r6282\_p2395$$

where $mc15\_13TeV$ indicates it is a MC sample produced with the 2015 configuration at 13 TeV energy in the center of mass reference, 361301 is a unique identification number referring to that specific MC sample, *Sherpa* [21] is the event generator used, $CT10$ is the specific configuration of the event generator, $Wenu$ indicates it is a sample of W+jets sample with W decaying into electron and neutrino, $Pt0\_70$ is the specific W p_T slice of that sample, $CFilterBVeto$ indicates it is a W+c sample, $DAOD\_FTAG4$ indicates it is a FTAG4 derivation, $e3651$ is the event generation configuration, $s2586$ refers to the full simulation of the detector, $r6869$ refers to the reconstruction, $r6282$ indicates it is a sample reproducing the specific pile-up condition of 25 ns collisions and $p2395$ is the derivation framework configuration.

Specific information on the data acquisition periods, the LHC runs and the integrated luminosity collected can be found on COMA [63] (Conditions and Configuration Metadata for ATLAS). There are specific tools able to calculate with precision the integrated luminosity of a set of data samples, anyway a good estimation could be obtained with COMA or with the online Luminosity calculator tool [64].

The Data Quality information, referring to the performance of the detector during the data acquisition, is used to establish the validity of the collected events: a Good Run List [65] is a list of the usable runs containing only the events registered with a completely working detector. One of the most important tools applied to data makes use of the Good Run List in order to eliminate imperfect events.

Since the xAOD and DAOD are stored on the Grid, they are not directly ac-

cessible. Some tools were developed in order to work with files stored on the Grid: Rucio [66,67] is commonly used to download a small fraction of the xAOD or DAOD sample in order to inspect its content, while Panda [68] is the most used tool to submit jobs that takes as input the files stored on the Grid and produces histograms and Ntuples.

Working with these tools, it was possible to identify the data and MC xAOD samples to be used for the W+c calibration. The Flavour Tagging group provided to produce the FTAG4 derivations of these xAOD samples. Once obtained the derivations, a dedicated software package was needed to extract the useful information and to store it into small flat Ntuples. Since many changes occurred between the 2012 and the 2015 Event Data Model, it was necessary to develope many software components, including the whole analysis code. I developed a complete RootCore package [26] based on EventLoop and SampleHandler packages, able to check out from the DAOD samples stored on the Grid many variables of the ATLAS events: vertices, tracks, hits, electrons, muons, jets, MET. With this package various approaches of the event selection were tested, selecting the most powerful discriminating variables between the W+c signal and the background processes.

After many tests, a suitable event selection was established, then the official production of the Ntuples was launched with the CxAODFramework, a central framework developed by the Flavour Tagging group for the production of the Ntuples for the different calibration analyses. This framework also performs all the smearing and correction of the variables and produce a version of the Ntuples for each standard systematic variation. The RootCore package I developed was a basic contribution to the CxAODFramework, commonly developed by the Flavour Tagging group analysts.

5.3 Data samples

As soon as 13 TeV data were produced, the first tests were performed. The LHC setup for the beginning of the operation delivered collisions at 20 MHz rate (a bunch crossing every 50 ns). The data collected during June and July 2015 with this set up are commonly addressed as 50 ns data. After applying the Good Run List, the 50 ns useful events reached a total of 85 pb^{-1} integrated luminosity: with this low quantity of data it is not possible to perform the W+c calibration. The tight event selection applied in the 7 TeV W+c calibration [48], was able to select only about 10000 events within a total of 4.6 fb^{-1} 7 TeV data.

After two weeks of tecnical stop, LHC resumed the proton-proton collisions at 40 MHz rate (a bunch crossing every 25 ns). During August 2015, a total of 80 pb^{-1} data with the 25 ns setup were collected. The test sample was obtained combining 85 pb^{-1} of 50 ns data with 80 pb^{-1} of 25 ns data. The Ntuples of the test sample were produced with the RootCore package I developed: no correction tools were applied to them and the MET used is from the original fast reconstruction provided during

raw data collection. This test sample was used to develop a new event selection able to correctly identify Opposite Sign events, without the help of the Soft Muon Tagging, not yet developed for 13 TeV data. Even summing up the 50 ns and the 25 ns datasets, it was not possible to perform the W+c calibration: with the tight event selection developed, only about 50 useful events were collected.

During September 2015, LHC increased the number of bunches present in each proton beam, rising in this way the number of collisions. The luminosity reached a peak of $3.23 \cdot 10^{33} cm^{-2} s^{-1}$ and during September about 745 pb^{-1} data were collected, reaching a total of 0.91 fb^{-1} 13 TeV data. At the time of the writing of this thesis work (beginning of October 2015) not all of these data were reviewed by the Data Quality group, since it takes about 10 days to evaluate the validity of the events and produce the Good Run List. Since not all the September data were ready and due to the slight difference between 25 ns and 50 ns data, in this thesis work the dataset used correspond to an integrated luminosity of about 430 pb^{-1}, regarding events collected during September 2015. The Ntuples of the September dataset were produced with the CxAODFramework with all the correction tools implemented, including the proper MET reconstruction.

5.4 MC samples

The first tests of the W+c calibration were performed on the 2014 MC producion, in particular using the HIGG5D2 derivations, produced to study the production of the Higgs in association with a W boson. In this way, it was possible to identify the list of MC samples needed for the W+c calibration: the corresponding original xAOD samples (MC15) were derived using the FTAG4 derivation format.

In principle, the calibration was supposed to be performed on the 50 ns data collected during June and July 2015, but the data collected (85 pb^{-1}) were not enough to study the property of the MV2c20 tagger. For this reason the MC15 samples selected were simulated with pile-up events expected with the 25 ns LHC setup, to be compatible with the 25 ns data that will be used for the calibration.

The MC15 samples are produced with different event generators: Sherpa [21] for the W+b, W+c and W+light jet production, but also for the Diboson and the Z+jets samples, a combination of Powheg [69], Pythia [19, 49] and EvtGen [57] was used to generate the $t\bar{t}$ and single top samples, while for the QCD Multijet samples a combination of Pythia and EvtGen was used.

To correctly match data with MC samples, each selected MC event was weighted with its MC weight and pile-up weight, as follows:

$$(MC_w \ / \ SumMC_w) \cdot PU_w \cdot XSec \cdot kFactor \cdot FilterEff \cdot Lum \qquad (5.1)$$

where MC_w is the MC weight of that selected event, $SumMC_w$ is the sum of the MC weight of all events contained in the original xAOD, PU_w is a correction

for the pile-up contamination of the event, $XSec$ is the total cross section of the sample, $kFactor$ is to correct leading order cross sections to next-to-leading order cross sections, $FilterEff$ is the filter used on MC generated events to select a certain phase space of events and Lum is the total integrated luminosity of the corresponding data sample.

This normalization procedure is performed to each sample independently and then, after each sample is appropriately normalized, they are added together. The $SumMCWeight$ was retrieved with the CutBookKeeper tool, during Ntuple production: some issues were found on the information provided with this tool about some MC samples. This should be further investigated.

5.5 Event selection

To exploit the charge correlation, in order to enhance the W+c signal contribution with respect to the background processes, the events were selected according to the 7 TeV calibration [48] (see Sec. 3.1.2). Since the Soft Muon Tagging [59] tool was not yet developed and tested for the Run II condition, some new requirements were tested in order to replace it.

First of all, a pre-selection is applied during the extraction of the Ntuples from the DAOD, in order to eliminate useless events and obtain smaller Ntuples:

- the events are required to have at least one primary-vertex candidate, made by at least three reconstructed tracks;
- the events are required to have at least one electron, one jet and one muon;
- the missing transverse momentum of the event is required to be higher than 25 GeV, to take into account the neutrino coming from the W decay.

The actual selection is applied at the Ntuple level and contains requirements on all the physics objects present in the process: electrons, missing energy, jets and muons.

Electrons are required to have a transverse momentum $p_T > 25\ GeV$. The pseudorapidity range allowed for electrons is $|\eta| < 2.47$, excluding the calorimeter transition region $1.37 < |\eta| < 1.52$. Electrons are required to pass the "tight" identification criteria established for 2015 data. Only isolated electrons are selected, with a cut on the calorimeter-based quantity (ETCone $< 3GeV$), removing in this way the electrons coming from hadrons decay within a jet. Exactly one electron with these qualities is allowed in each event.

In addition the MET requirement in the pre-selection, the W boson transverse mass m_T^W is requested to be greater than 40 GeV.

Jets with $p_T > 20\ GeV$ and $|\eta| < 2.5$ are selected: the threshold of transverse momentum was lowered from 25 GeV to 20 GeV in order to accept a higher number

of events. To reduce pile-up contribution, jets with $p_T < 50\ GeV$ and $|\eta| < 2.4$ are further requested to have $JVT < 0.641$. Exactly one jet fulfilling these conditions is allowed for each event.

The selected jet is geometrically matched with a muon with $p_T > 4\ GeV$ and $|\eta| < 2.5$, requiring the muon to have a small angular separation with respect to the jet axis: $\Delta R < 0.4$. Only "combined" muons are accepted, moreover two impact parameter conditions are added: $|d_0| < 3\ mm$ and $|z_0\ sin(\theta)| < 3\ mm$. A set of ID hit requirements are requested to select high quality tracks and the ETCone of the muon is requested to be higher than 3 GeV, to select only not isolated muons. To replace the SMT requirement, in addition to the geometrical matching, a new condition was studied: the absolute value of the difference between the longitudinal impact parameters of electrons and muons is required to be lower than 1 mm:

$$|z_0(electron) - z_0(muon)| < 1mm \tag{5.2}$$

Exactly one muon passing the described selection and geometrically associated to the selected jet is allowed per each event.

5.6 Results

The described event selection was developed trying to maximize the yield of the Opposite Sign with respect to the Same Sign events, using the test sample collected during June, July and August 2015.

This newly developed event selection was then applied successfully to about 430 pb^{-1} of events collected during September 2015. The data show the same OS-SS trend seen in 7 TeV data, as it can be seen in Tab. 5.3 and in Fig. 5.2. The MC samples are not fully available yet, so the data-simulation comparison is not shown: the Opposite Sign events abundance is most probably due to the W+c process, since it is the only process with a strong preference in the charge correlation of the two leptons.

Period	Energy	Int. Lum.	OS events	SS events	OS + SS	OS %
2011	7 TeV	4.6 fb^{-1}	7445	3125	10570	70.4
June-August 2015	13 TeV	$85 + 80\ pb^{-1}$	33	14	47	70.2
September 2015	13 TeV	430 pb^{-1}	271	99	370	73.2

Table 5.2: Comparison between the OS-SS yield obtained with 7 TeV data, 13 TeV test sample and available September 13 TeV dataset.

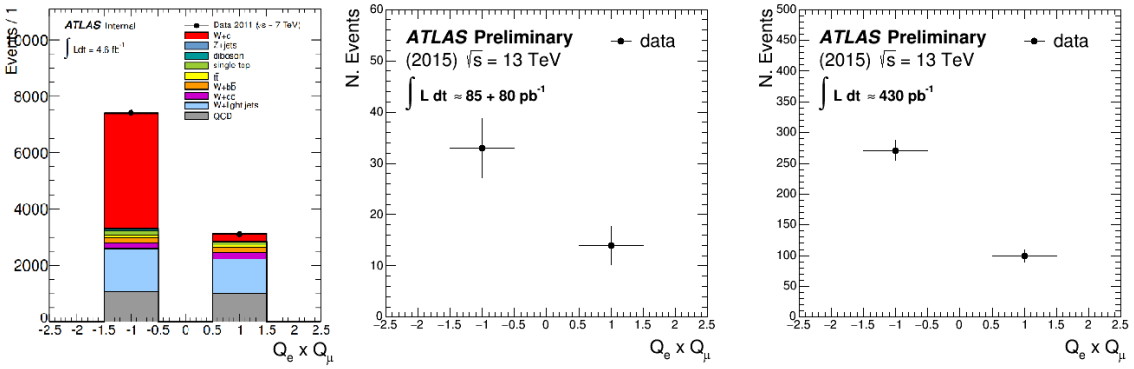

Figure 5.2: Comparison between the OS-SS yield obtained with 7 TeV data (left), 13 TeV test sample (center) and available September 13 TeV dataset (right).

Run Number	Int. Lum. $[pb^{-1}]$	OS events	SS events
276952	6.9	6	0
276954	0.8	2	0
278880	22.7	16	6
278968	10.3	3	5
279169	57.8	37	18
279259	7.6	4	5
279279	17.8	7	5
279284	33.3	27	6
279515	0.4	0	0
279598	71.4	41	17
279764	2.5	1	0
279813	50.4	29	9
279867	31.3	14	5
279928	1.1	0	1
279932	46.9	36	8
279984	68.7	48	14
TOTAL	430.0	271	99

Table 5.3: Estimated integrated luminosity, number of Opposite Sign and Same Sign events selected for each of the Runs used in the 430 pb^{-1} dataset of events registered in September 2015.

With the 7 TeV event selection (including the Soft Muon Tagging) applied to the full 2011 dataset, about 10000 events were selected, with the 70.4 % of events being Opposite Sign.

The test sample (85 pb^{-1} of 50 ns data combined with 80 pb^{-1} of 25 ns data) was used to develope the new event selection. The few events selected (about 50) shows the same OS-SS yield obtained with the 7 TeV selection: the 70.2 % of the events is Opposite Sign.

The available dataset of 13 TeV data collected during September 2015 (430 pb^{-1}), properly reproduced with the CxAODFramework applying the correction tools, show a sensible improvement of the OS-SS yield: the 73.2 % of the events is Opposite Sign. The gain in OS-SS yield could be explained by the suppression of one of the main background processes: the QCD background (shown in grey, Fig. 5.2). With the newly developed event selection, none of the MC events generated to simulate the QCD background processes are selected, so they do not contribute to the OS and SS yield.

Thanks to the high statistics of this dataset, it has been possible to provide the distribution of the MV2c20 b-tagging variable with 13 TeV data (see Fig. 5.3): the presented distribution has the expected flavour composition. The events with a low value of MV2c20 represent events with jets originated from light quarks and gluons, while the events with a high value of MV2c20 contain jets originated from heavy flavour quarks. The OS (red) and the SS (white) distributions have almost the same contribution of light jets, but a very different content in terms of heavy flavour jets: with the OS-SS subtraction, the light jet contribution will be suppressed, while the heavy flavour jet contribution will remain.

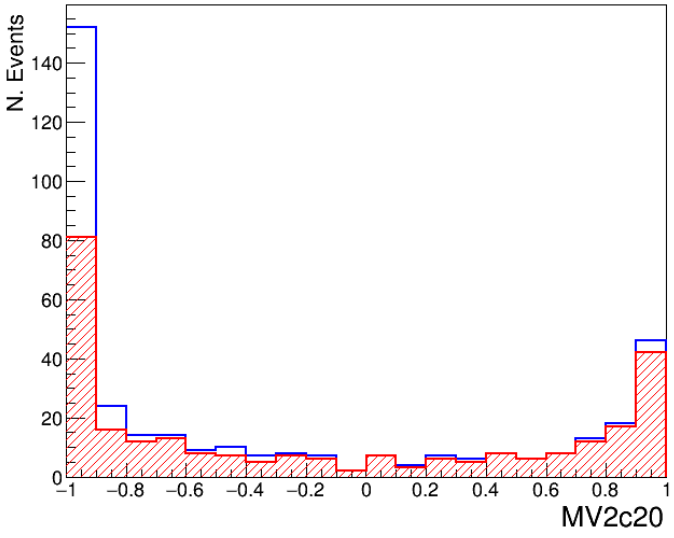

Figure 5.3: MV2c20 b-tagging variable distribution of the selected events: the blue line represents al the selected events (OS+SS), in white the SS events and in red the OS events.

5.7 Conclusions

In this thesis work, the W+c calibration analysis software was developed and the first tests were performed with 13 TeV data. All the useful MC samples were retrieved and the official Ntuples production has been requested. The event selection developed shows the same trend in the OS-SS yield compared to the 7 TeV calibra-

tion, with a slight improvement, despite the Soft Muon Tagging was not yet applied and despite the changes in the Event Data Model. The Flavour Tagging group is currently developing the Soft Muon Tagging algorithm that correctly associates soft muons with jets: the application of this tool will guarantee a even better event selection.

The event selection here discussed is the first step of the calibration procedure: the pure c-jets sample will be extracted as the difference between Opposite Sign and Same Sign events and then the b-tagging will be applied to measure the c-jet tagging efficiency. Currently, the available data are not enough to perform the calibration, but it was anyway possible to test the event selection and to have a first look to the MV2c20 distribution. As soon as 1 fb^{-1} of 13 TeV data will be available, the c-jet tagging efficiency calibration with the W+c method will be provided.

Conclusions

For this master degree thesis, I worked in close contact with the ATLAS Flavour Tagging group, gaining a deep knowledge of the b-tagging algorithms and methods. I tested different solutions for the D* calibration method on 8 TeV data and I developed the W+c calibration on 13 TeV data, becoming one of the experts in the field of the c-jet tagging calibrations. I presented many times the progress and the first results of my contribution at the ATLAS Flavour Tagging meetings.

In the chapter 1, I described in detail the ATLAS experimental apparatus, with a brief introduction on the CERN site and LHC operations.

In the chapter 2, I discussed the importance of the b-tagging for the ATLAS physics programme, I described the properties of b-jets, c-jets and light jets and the b-tagging algorithms developed by the ATLAS experiment to discriminate between them, based on the secondary vertex reconstruction and the large impact parameters of their tracks.

In the chapter 3, I discussed the methods used to measure the performance of the b-tagging algorithms, with a detailed description of the D* and W+c methods used for c-jet tagging efficiency measurement.

In the chapter 4, I described the improvements made to the D* method: I studied different fit procedures for the c-jet tagging efficiency measurement with the D* method using 8 TeV data recorded during 2012 by the ATLAS experiment, improving the precision and moving towards the continuous calibration as a function of the b-tagging variable.

In the chapter 5, I presented the W+c method developed for the c-jet tagging efficiency measurement and I provided its first test on 13 TeV data recorded during 2015 by the ATLAS experiment at LHC. The developed event selection has slightly improved the OS-SS yield, a further gain will be obtained applying the Soft Muon Tagging. When a sufficient amount of 13 TeV data will be available, I plan to perform both the W+c and the D* calibrations on the same sample, allowing to compare the results of the two calibration methods.

Finally, I plan to exploit the experience I gained in the b-tagging field to improve physics analyses that strongly use b-tagging, like the Higgs boson searches: reducing the uncertainties on the b-tagging calibrations, the physics measurements will be considerably more precise. Furthermore, knowing the intimate details of the ATLAS b-tagging, allows me to effectively impact these physics analyses.

Ringraziamenti

In questi quasi 9 mesi sono stato a contatto con molte persone, ognuna di esse ha contribuito a suo modo alla realizzazione di questo (capo)lavoro. Senza di loro non sarei riuscito a partorire questa tesi (sì, è stato un parto) ...per cui diamo il via ai ringraziamenti!i!

Vorrei ringraziare innanzitutto i miei due relatori, Fabrizio P. e Carlo S., e il mio correlatore, Silvano T., perché, nonostante i tantissimi impegni, sono riusciti a dedicarmi un pizzico del loro tempo per guidarmi in questa analisi/avventura che è stata la mia tesi.

Vorrei ringraziare anche Carletto per avermi vivamente consigliato di scegliere il gruppo ATLAS e per i numerosi suggerimenti che mi ha donato lungo la strada.

Vorrei ringraziare inoltre il gruppo ATLAS di Genova, per l'interesse dimostrato, il supporto e i vari consigli. Grazie a loro ho realizzato che il mio lavoro è una piccola parte di un qualcosa di molto grande... in fondo ATLAS è l'esperimento più grande che ci sia!

Un ringraziamento speciale va allo staff del CERN, per aver conservato il mio badge in un cassetto per tutti questi anni e soprattutto per avermelo restituito quando ormai pensavo che fosse perduto.

Inoltre vanno ringraziati tutti i centri di ricerca e le università che fanno parte della World LHC Grid: molti di loro hanno inconsapevolmente contribuito allo svolgimento delle analisi ospitando i miei pacchetti di dati nei loro dischi di memoria sparsi in giro per il mondo.

Vorrei ringraziare *mon amis Danilò* per le ore passate insieme in giro per Genova, ma soprattutto per gli splendidi pranzi della domenica.

Il ringraziamento più grande va alla mia compagna di vita, da molti anni mia costante ispirazione! *Grazie Ester* ♡

Bibliography

[1] L. Evans and P. Bryant, LHC machine, JINST 3: S08001 (2008).

[2] M. Benedict *et al.*, LHC Design Report, CERN-2004-003 (2004).

[3] ATLAS Experiment © 2014 CERN, http://www.atlas.ch/photos/

[4] The ATLAS Collaboration, ATLAS Technical Design Report, volume 1 and 2, CERN-LHCC-99-15 (1999).

[5] The ATLAS Collaboration, The ATLAS Experiment at the CERN LHC, JINST 3: S08003 (2008).

[6] The ATLAS Collaboration, The ATLAS Inner Detector commissioning and calibration, CERN-PH-EP-2010-043 (2010).

[7] The ATLAS Collaboration, ATLAS Insertable B-Layer Technical Design Report, CERN-LHCC-2010-013 (2010).

[8] The ATLAS Collaboration, ATLAS pixel detector electronics and sensors JINST 3: P07007 (2008).

[9] The ATLAS Collaboration, The barrel modules of the ATLAS Semiconductor tracker, Nucl. Inst., 568(2): 642-671 (2006).

[10] The ATLAS Collaboration, The ATLAS Semiconductor tracker end-cap module, Nucl. Inst., 575(3): 353-389 (2007).

[11] The ATLAS Collaboration, The ATLAS Transition Radiation Tracker proportional drift tube: Design and Performance, JINST 3, P02013 (2008).

[12] The ATLAS Collaboration, ATLAS Calorimeter performance: Technical Design Report, CERN-LHCC-96-040 (1996).

[13] The ATLAS Collaboration, ATLAS Muon Spectrometer: Technical Design Report, CERN-LHCC-97-022 (1997).

[14] The ATLAS Collaboration, Performance of the ATLAS Trigger system, JINST 7, C01092 (2012).

[15] I. Bird *et al.*, LHC Computing Grid: Technical Design Report, CERN-LHCC-2005-024 (2005).

[16] The ATLAS Collaboration, ATLAS Computing: Technical Design Report, CERN-LHCC-2005-022 (2005).

[17] G. Barrand *et al.*, Gaudi - a software architecture and framework for building HEP data processing applications, Computer Physics Communications, 140(1): 44-55 (2001).

[18] Athena homepage, http://twiki.cern.ch/twiki/bin/view/Atlas/AthenaFramework

[19] S. Mrenna *et al.*, A brief introduction to PYTHIA 8.1, Comput. Phys. Commun., 178: 852-867 (2008).

[20] M. Bahr *et al.*, HERWIG++ Physics and Manual, Eur. Phys. J., C58: 639-707 (2008).

[21] T. Gleisberg *et al.*, Event generation with SHERPA 1.1, JHEP 02, 007 (2009).

[22] S. Agostinelli *et al.*, GEANT4 - a simulation toolkit, Nucl. Inst., Section A, 506(3): 250-303 (2003).

[23] R. Brun and F. Rademakers, ROOT - an object oriented data analysis framework, Nucl. Inst, Section A, 389 (1): 81-86 (1997).

[24] The ROOT system homepage, http://root.cern.ch/

[25] W. Verkerke *et al.*, The RooFit toolkit for data modeling, CHEP-2003-MOLT007 (2003).

[26] RootCore TWiki, https://twiki.cern.ch/twiki/bin/view/AtlasComputing/RootCore

[27] K.A. Olive *et al.*, Particle Data Group, Chin. Phys. C, 38, 090001 (2014).

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

[29] The ATLAS Collaboration, Performance of b-jet identification in the ATLAS Experiment, JINST DRAFT (2015).

[30] The ATLAS Collaboration, Expected performance of the ATLAS *b*-tagging algorithms in Run II, ATL-PHYS-PUB-2015-022 (2015).

[31] S. L. Glashow, Partial symmetries of weak interactions. Nuclear Physics, 22(4): 579-588 (1961).

[32] A. Salam and J. C. Ward, On a gauge theory of elementary interactions, Il nuovo cimento, 19(1): 165-170 (1961).

[33] S. Weinberg, A model of leptons, Phis. Rev. Lett., 19: 1264-1266 (1967).

[34] H.P. Nilles, Supersymmetry, supergravity and particle physics, Phys. Rep. 110: 1-162 (1984).

[35] ALEPH, DELPHI, L3 and OPAL collaborations, The LEP working group for Higgs boson searches, Search for the Standard Model Higgs boson at LEP, Physics Letters B, 565: 61–75 (2003).

[36] Higgs Working Group, CDF collaboration, D0 Collaboration *et al.*, Updated combination of CDF and D0 searches for standard model Higgs boson production with up to 10 fb^{-1} of data, FERMILAB-CONF-12-318-E, CDF Note 10884, D0 Note 6348 (2012).

[37] The ATLAS Collaboration, Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC, Physics Letters B, 716(1): 1-29 (2012).

[38] The CMS Collaboration Observation of a new boson at a mass of 125 gev with the CMS experiment at the LHC, Physics Letters B, 716(1): 30-61 (2012).

[39] The CDF Collaboration, Observation of Top Quark Production in pp Collisions with the Collider Detector at Fermilab, Phys. Rev. Lett. 74, 2626 (1995).

[40] The D0 Collaboration, Observation of the Top Quark, Phys. Rev. Lett. 74, 2632 (1995).

[41] The ATLAS Collaboration, Hadronic calibration of the ATLAS liquid argon end-cap calorimeter in the pseudorapidity region $1.6 < |\eta| < 1.8$ in beam test. Nucl. Insturm. Meth. A531: 481-514 (2004).

[42] M. Cacciari *et al.*, The anti-kt jet clustering algorithm. JHEP 04, 063 (2008).

[43] The ATLAS Collaboration, Pile-up subtraction and suppression for jets in ATLAS, ATLAS-CONF-2013-083 (2013).

[44] The ATLAS Collaboration, Tagging and suppression of pileup jets with the ATLAS detector, ATLAS-CONF-2014-018 (2014).

[45] A. Hoecker *et al.*, TMVA, Toolkit for Multivariate Data Analysis with ROOT, CERN-OPEN-2007-007 (2007).

[46] The ATLAS Collaboration, b-jet tagging calibration on c-jets containing $D*^{+}$ mesons, ATLAS-CONF-2012-039 (2012).

[47] The ATLAS Collaboration, Calibration of the performance of b-tagging for c and light-flavour jets in the 2012 ATLAS data, ATLAS-CONF-2014-046 (2014).

[48] The ATLAS Collaboration, Calibration of the b-tagging effficiency for c jets with the ATLAS detector using events with a W boson produced in association with a single c quark, ATLAS-CONF-2013-109 (2013).

[49] T. Sjostrand, S. Mrenna, and P. Skands, PYTHIA Generator version 6.418, JHEP 05, 026 (2006).

[50] The ATLAS Collaboration, Measuring the b-tag efficiency in a top-pair sample with 4.7 fb^{-1} of data from the ATLAS detector, ATLAS-CONF-2012-097 (2012).

[51] The ATLAS Collaboration, Measuring the b-tag Efficiency in a Sample of Jets Containing Muons with 5 fb^{-1} of Data from the ATLAS Detector, ATLAS-CONF-2012-043 (2012).

[52] The ATLAS Collaboration, Measurement of the b-jet production cross section using muons in jets with ATLAS in pp Collisions at 7 TeV, ATLAS-CONF-2011-057 (2011).

[53] The ATLAS Collaboration, Measurement of the Mistag Rate with 5 fb^{-1} of Data Collected by the ATLAS Detector, ATLAS-CONF-2012-040 (2012).

[54] The ATLAS Collaboration, Calibration of b-tagging using dileptonic top pair events in a combinatorial likelihood approach with the ATLAS experiment, ATLAS-CONF-2014-004 (2014).

[55] The ATLAS Collaboration, Measurement of the b-tagging efficiency of the MV1 algorithm in pp collisions at 8 TeV using $e\mu$ dilepton $t\bar{t}$ events, ATL-COM-PHYS-2013-381 (2013).

[56] The ATLAS Collaboration, Continuous b-tagging for the ATLAS experiment, ATLAS NOTE DRAFT (2015).

[57] D. J. çange, The EvtGen particle decay simulation package, Nucl. Instrum. Meth. A462:152 (2001).

[58] The ATLAS Collaboration, Measurement of the production of a W boson in association with a charm quark in pp collisions at $\sqrt{7} = 7$ TeV with the ATLAS detector, CERN-PH-EP-2014-007 (2014).

[59] The ATLAS Collaboration, Soft muon tagging and D^*/μ correlations in 7 TeV collisions with ATLAS, ATLAS-CONF-2010-100 (2010).

[60] The ATLAS Collaboration, Report of the xAOD design group, ATL-COM-SOFT-2013-022 (2013).

[61] LXPLUS Service, http://information-technology.web.cern.ch/services/lxplus-service

[62] AMI homepage, http://ami.in2p3.fr/index.php/en/

[63] COMA homepage, https://atlas-tagservices.cern.ch/tagservices/RunBrowser/index.html

[64] ATLAS Luminosity Calculator, https://atlas-lumicalc.cern.ch/

[65] Good Run List, http://atlasdqm.web.cern.ch/atlasdqm/grlgen/All_Good/

[66] Rucio homepage, http://rucio.cern.ch/

[67] The ATLAS Collaboration, Monitoring and controlling ATLAS data management: The Rucio web user interface, ATL-SOFT-PROC-2015-037 (2015).

[68] Panda homepage, http://bigpanda.cern.ch/

[69] S. Frixione *et al.*, Matching NLO QCD computations with parton shower simulations: the POWHEG method, JHEP 11 (2007).