Data-driven QCD background estimation to W production in association with jets using the ATLAS detector

Lidija Radovanovic
Supervisor: Andrew Hamilton

A submitted to the University of Cape Town for the degree of
A dissertation submitted to the University of Cape Town for the degree of
Master of Science in Physics

Rev: (November 30, 2016)

Abstract

A preliminary analysis was conducted using data from the proton-proton collisions at $\sqrt{s}$ = 8 TeV recorded by the ATLAS experiment, corresponding to an integrated luminosity of approximately 18.9 fb^{-1}. It has been performed to extract the fraction of QCD events in the chosen phase space for the W boson production in association with jets. A data-driven technique was developed for measuring the QCD background. The fraction of QCD events in the signal region was estimated in exclusive jet multiplicities by performing an extended likelihood fit on the E_T^{miss} distribution in data. The fraction of QCD events ranges $2.9 \pm 0.01\%$ for the zero jet bin and increases to $10.74 \pm 0.11\%$ for the two jet bin.The fraction of QCD events decreases with increasing jet multiplicity. Using the results from the fit, some important kinematic distributions have been made. The overall agreement between the data and Monte Carlo simulation is vastly improved by including the data-driven QCD background estimate. This agreement could be further improved by using a data-driven $t\bar{t}$ estimate where the $t\bar{t}$ MC tends to overestimate the amount of events in the higher jet multiplicities.

Declaration

I hereby declare that this thesis has not been submitted, either in the same or different form, to this or any other university for a degree and that it represents my own work.

Lidija Radovanovic

Acknowledgements

I would like to thank my supervisor Andrew Hamilton for introducing me to the world of High Energy Phsics, and I would also like to thank my other supervisor Michael Christian Rammensee who supported me while I was at CERN and continued to supervise me through to completion. Both Michael and Andrew, you supported this young HEP fledgling and were very supportive, compassionate and understanding when I was not able to grasp some things and through other personal difficulties. I am also very grateful to the members of the W+Jets analysis group, especially Valerie Lang, Monica Dunford and Bastian Bernarding who helped with feedback as well as plenty of python, coding and c++ advice. It helped me understand how important collaboration is in the HEP community. Without the aforementioned people, my supervisors and the W+jets team, this thesis would not be possible. I would also like to thank my colleague, Guillermo Hamity for passing on the torch and teaching me the ropes when it comes to Physics Validation for the Tau Physics group. The convenors of Tau CP, Will Davey and Attilio Andrezzea, thank you for letting me be a part of the validation team and for all your comments and helping me and giving prompt feedback on Validation results. Finally, I would like to thank my family, my mom and dad and brother, and my friends. I need to thank my roommates as well for being the best and providing the support and being my cheerleaders when things got difficult. Finally, none of this would have at all been possible if it wasn't for the financial assistance I received from the National Research Foundation during my masters studies as well as SA-CERN who funded my visit to CERN in June 2014.

Contents

List of Figures

Chapter 1

Introduction

Testing the predictions of the Standard Model is one of the main physics goals of the ATLAS [1] experiment. The study of jet production in association with a W boson in proton-proton collisions provides an important test of perturbative QCD (pQCD) [2]. The signature of W production in association with jets in proton-proton collisions is a lepton (e, μ) with a large missing transverse energy accompanied by 1 or more jets. W production in association with jets is a background to some supersymmetry searches as well as a background in certain Higgs boson decay channels. Being able to accurately measure and describe this Standard Model process allows us to constrain this background in other more sensitive searches where the possible signal (SUSY , dark matter) will be overwhelmed by this process. The production of a W boson in association with jets has been studied by the CDF [3] and D0 experiments [4] at the Tevatron Collider. More recently, W production in association with jets studies [5] [6] have been done by the CMS [7] and the ATLAS experiments. The most recent of these is the study performed on the 2011 data collected by ATLAS at a centre-of -mass energy $\sqrt{s}$ = 7 TeV [6].

When measuring a physics process with data collected by ATLAS it is imperative to be able to model the background processes which have the same final state products recorded by the detector. This accurate description of background events will enable us to measure the signal process and from this the cross-section can be extracted and compared to the theoretical predictions provided by pQCD. To separate the signal and background events Monte Carlo (MC) simulations are often used. However, not all the backgrounds can be simulated using MC. The QCD or multijet background is a notoriously difficult background to simulate and there is currently no adequate existing MC technique. This is a fairly significant background in many searches as it has a very large cross section. Given the complexity in modelling

the QCD background, data-driven techniques for estimating the QCD background is one of the primary methods used in studies of W boson production in association with jets. The primary aim of this thesis is to optimize the existing techniques or even discover new methods for determining this background.
The work in this thesis was carried out in collaboration with the W+jets Analysis group, but all results are the work of the author. From a previous study conducted on the 2011 ATLAS data-set, it is known that the isolation requirement on the lepton forms the main discriminating feature between signal and QCD events. In this thesis, the author established the best anti-isolation range for the QCD template, as well as which trigger to use for the QCD template. The author also examined the effect of the different data taking periods on the shape of the E_T^{miss} distribution in the QCD control region used to build the data-driven template and found that this effect was negligible on the template selection. In the muon channel, the effect of the size of the cone around an energy deposition in the track based and calorimeter based variable was investigated. The thesis will conclude with a summary and conclusion, highlighting the merits and pitfalls of this data-driven technique and how it can be improved.

A data-driven estimate for the QCD/ mulitjet background using the 2012 complete data set collected by the ATLAS detector at $\sqrt{s} = 8$ TeV with an integrated luminosity of will form the main objective of this thesis. A chapter is also dedicated the operational work I performed as part of the Tau Physics Validation team. Chapter 2 provides a brief theoretical description of the Standard Model and provides the topology of W+ jet production within the context of proton-proton collisions at the LHC. Chapter 3 gives a description of the ATLAS experiment. Thereafter, the object definitions are given in chapter 4 which will then be used to provide the selection criteria for the W+jets measurement in chapter 6 as well as description of the Physics Validation procedure in chapter 5. Finally, a summary is given at the end of this thesis in chapter 7

Chapter 2

Theoretical Background

"Nature uses only the longest threads to weave her patterns, so that each small piece of her fabric reveals the organization of the entire tapestry."
— Richard P. Feynman

In this chapter an overview of the theoretical description of the Standard Model is given with an emphasis on the most important parts that enhance the understanding of this thesis. In Section 2.1 we give a brief overview and History of the Standard Model and describe each of the components and theoretical milestones which lead to the description of the Standard Model.

2.1 The Standard Model of Particle Physics

At the beginning of the 20^{th} century, only two fundamental forces were known by physicists: the force of gravity and the electromagnetic force. However, neither of these forces could describe the structure of atoms. Physicists endeavoured to describe the interactions at the subatomic scale and asked "What are the constituents of atoms?" at the most fundamental level. Rutherford's identification of the atomic nucleus in 1911, based on the scattering experiments observed by Hans Geiger and Ernest Marsden, led to the discovery of the proton by 1919. By the mid 20th century, a myriad of new particles had been discovered, but an incomplete picture stood and a description of the fundamental building blocks of nature had not fully taken form. In the 1960's, Sheldon Glashow, Steven Weinberg, Abdus Salam

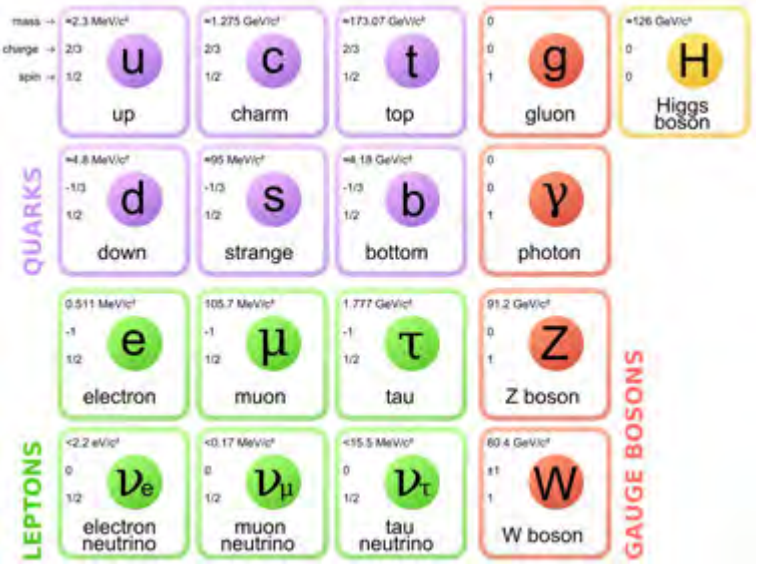


Figure 2.1: The Standard Model Elementary Particles

proved electroweak unification mathematically for which they were awarded the Nobel Prize in Physics in 1979. Their seminal paper describing the interactions of the fundamental particles and their interactions gave the Standard Model its modern form.

It is important to begin an exposition of the Standard Model and the fundamental building blocks of nature by naming them and categorising them. The fundamental particles can be classified into two families; fermions and bosons. Fermions are spin $\frac{1}{2}$ particles which obey Fermi-Dirac statistics while bosons have integer spin and obey Bose-Einstein statistics. The fermions are the matter particles which form two groups of spin - $\frac{1}{2}$ particles: leptons, which interact electromagnetically and weakly. Quarks which interact electromagnetically and strongly. Both the leptons and quarks can be grouped into three generations as can be seen in fig. 2.1, where each generation contains two members. The first generation of fermions is summarised as follows: There are the electrons and neutrinos, (e, ν) and the quarks (u,d) and their anti-particles ($\bar{e}$, $\bar{\nu}$) and ($\bar{u}$, $\bar{d}$) respectively. Therefore, each generation of leptons consists of one charged lepton, with charge -1 in units of electric charge, and one neutral lepton, the neutrino. In the three generations of quarks, as seen in figure 2.1, the up-like quarks all have charge $\frac{2}{3}$ while the down-like quarks have charge $-\frac{1}{3}$. The quarks possess colour charge, and each flavour of quark comes in three colours. Quark do not exist on their own and always form colour-neural bound states known as hadrons. The Standard Model is a quantum field theory which describes the interactions between matter particles and the gauge bosons which mediate the three forces described by the Standard Model.

Each of these three forces is described by a gauge group and they have been unified in the Standard Model. This means that the lagrangian remains invariant under a continuous group of transformations. The product of these three groups is described by the non-abelian symmetry group and is written as

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

The SM has is a very successful, albeit incomplete theory, which has stood up to experimental scrutiny. The theory is incomplete because it does not take the fourth fundamental force, gravity, into account, it also does not describe the dark matter that has been detected in astronomical observations. The theory came about in the 20th century, and is the product and work of many great scientists in the previous century. In order to give a more complete picture, in the following sections I will describe the theoretical milestones which led up to the culmination of the standard model, namely the quantisation of the electromagnetic field, the development of the theory of the strong interaction as well as electroweak unification. Preceding these sections is an overview of the Lagrangian forumlation of the SM. Our best and current understanding of the universe is completely defined by the famous Lagrangian which defines every particle in the universe and its interactions.

2.1.1 The Lagrangian Formulation of the Standard Model

In the study of motion and forces which produce motion, the trajectory of a particle is derived by solving the Lagrange equations. In classical mechanics, the Lagrangian is defined as the difference of kinetic and potential energy.

$$L = T - V \tag{2.2}$$

In equation. 2.2, the equations of motions for particles are derived by solving the Euler-Lagrange equation:

$$\frac{\partial L}{\partial q} - \frac{d}{dt}\frac{\partial L}{\partial \dot{q}} = 0 \tag{2.3}$$

However, since the Standard model is a field theory, and particles are seen as excited states (field quanta) of an underlying field, thus the Lagrangian density $\mathcal{L}$, where $L = \int \mathcal{L} d^3x$, is used to obtain the "equations of motion". The interactions between particles is described by the interaction terms between the underlying quantum fields. The quantum theory of electrodynamics (QED) is a field theory which describes interactions of charged particles.

The QED Lagrangian density (from which everything can be derived) for electrons, photons and the interaction between them is given in equation. 2.4 [8].

$$L = -\frac{1}{4}F_{\mu\nu}F_{\mu\nu} - \bar{\psi}_e\gamma_\mu[\partial_\mu + ieA_\mu]\psi_e - m_e\bar{\psi}_e\psi_e \tag{2.4}$$

Where in this thesis, natural units have been used throughout such that $\hbar$ and c are equal to 1. The first term in equation. 2.4 describes the free electromagnetic field which is defined by the four-potential $A^\mu \equiv (\phi, \mathbf{A})$ containing scalar and vector potentials, ϕ and $\mathbf{A}$ respectively. The electromagentic field tensor $F_{\mu\nu}$ is defined as $F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu$. The last term corresponds to the "mass" term and describes the free electron with mass m_e. The middle term in equation. 2.4 which is grouped together using square brackets, contains the kinetic term describing the kinetic energy of the electron while the second term is the "interaction" term which describes the interaction of the electron with the electromagnetic field.

2.2 QED, QCD and Electroweak symmetry breaking

Since Rutherford's discovery of the nucleus in his backscattering experiment, scientists wished to explain the nuclear force and had considerable difficulty in formulating a theory which would describe these interactions. Physicist Lev Landau expressed this difficulty in his last paper in 1959 entitled "Fundamental Problems" as follows:
'It is well known that theoretical physics is at present almost helpless in dealing with the problem of strong interactions. We are driven to the conclusion that the Hamiltonian method for strong interactions is dead and must be buried, although of course with deserved honour.'
The strong force is the the force that exists between quarks and is mediated by gluons. It

is the force that is responsible for the binding of protons and neutrons in atomic nuclei. Quantum chromodynamics describes the interactions of colored particles where the strength of the chromodynamic force is given by the strong coupling constant

$$g_s = \sqrt{4\pi\alpha_s} \tag{2.5}$$

The word 'constant' is a misnomer is this case since the coupling constant is dependent on the separation distance between the particles that are interacting. At larger distances, the strength of the strong force increases as the separation distance of the interacting particles increases which gives rise to confinement. This means that at higher energies, the constant decreases leading to asymptotic freedom. QCD is an $SU(3)_c$ non-abelian gauge theory. The $SU(3)_c$ symmetry gives rise to self interacting gluons which produces an extremely complex phenomenology. Above the energy scale $\Lambda_{QCD} \approx 200$ MeV, QCD processes can be calculated with perturbation theory since $\alpha_s << 1$. Due to confinement one is not able to observe free quarks. Instead, quarks form bound states which result in colorless objects called hadrons. The bound state of three quarks each with a different color forms a baryon, while quark-antiquark pairs are called mesons.

2.2.1 Electroweak Unification

Just as electricity and magnetism are two seemingly different forces, they are in fact manifestations of the electromagentic force. In 1961, Sheldon Lee Glashow aimed to unify the electromagnetic and weak interactions and to combine them into a single theoretical system where they are different manifestations of the same "electroweak" force [9]. This was achieved by his proposition that the weak interactions were mediated by unstable bosons.

The mediators of weak interactions are the charged $W^{\pm}$ bosons and neautral Z^0. The mediators of the weak force are massive intermediate vector bosons. The weak force occurs between charged and uncharged leptons, as well as quarks. W bosons are copiously produced in hadron collisions.

This is due to the electroweak interaction of a quark in one proton with an anti-quark of different flavour in another proton. For example, a W could get produced when a red up quark (charge +2/3) in one of the protons interacts with a red anti-down quark (charge + 1/3). As this quark anti quark pair annhilate, the W^+ materialises with it's full unit of electric charge. And color vanishes as the red and anti-red combines to zero color. So

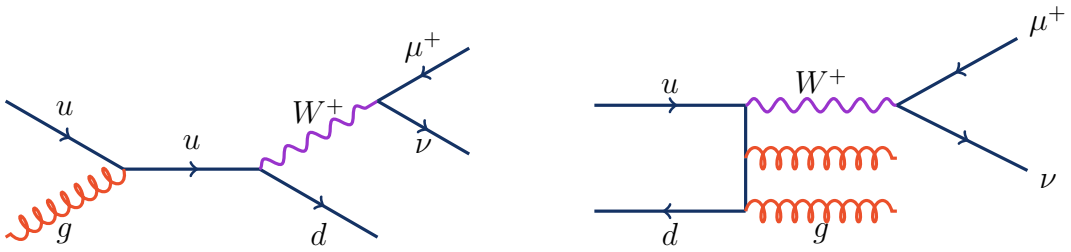


Figure 2.2: Feynman diagrams representing W boson production

these are some of the governing factors with respect to W boson production in a hadron collider. However, in practice, we do not get such an elegant first order production of a W boson. The quarks, before they annihilate to produce a W, emit gluons (seen in Figure 2.2). These gluons are able to radiate more gluons which creates a continuum cascade of energetic particles which hadronise in the detector to form hadronic jets. The LHC can be said to be a "jet making machine" since the scattering of partons is the dominant hard process and the production of QCD jets is the result of these higher order QCD interactions where gluons are radiated. From a theoretical perspective, these numerous interactions are difficult to calculate since the number of feynman diagrams grows exponentially. This poses a problem when trying to compare theoretical calculations/predictions and experimental measurements. In order to allow for a smoother comparison between data and theory, the theorists divide the calculation up into the number of possible energetic final state partons that may be created. In the detector, individual partons are not measured, but the QCD radiation is seen as individual jets of particles.

2.3 Perturbative QCD and W production in association with Jets

At a hadron collider, knowledge of how quarks and gluons scatter and materialise into hadrons is very important for studying any of the physical processes predicted by the Standard Model. Many of the interactions that occur at the Large Hardon Collider (LHC) are QCD hard and soft scattering and many interesting signatures are immersed in the background of pure QCD processes. Given that the Standard Model is renormalisable, perturbative calculations can be performed at higher orders which will predict cross sections for both hard and soft processes. Quarks and gluons bind into colour neutral hadrons which we observe in collider experiments. However, the exact mechanism for hadronisation is not

understood in detail. Therefore, Monte Carlo simulations are used to compare experimental data to to the calculated cross sections. The study of W production in association with hadronic jets provides an important test of perturbative QCD calculations. Since this process contains a large missing E_T signature, it is an important source of background in other SM searches, top quark studies, as well as new physics searches (SUSY searches and dark matter candidate searches). It is therefore essential that we understand the kinematics behind this process and measure its cross section.

Chapter 3

Experimental Setup

""The true method of knowledge is experiment." "
— William Blake

The Large Hadron Collider (LHC) is the world's largest hadron accelerator. The LHC project was approved in December of 1994 by the CERN council. The LHC was first started up on the 10th of September 2008 accelerating the initial beam of protons upto an energy of 7 TeV. There are four main detectors installed at the interaction points around the superconducting ring. The two general purpose detectors ATLAS [1] and CMS [7], a heavy ion experiment ALICE [10] as well as an experiment dedicated to the study of beauty or b-physics, the LHC-B [11]. It is by studying the particle interactions at these high energies that allows physicists to create and observe phenomena which will enable them to answer fundamental questions concerning the nature of the Universe. By recreating conditions which were last seen a few millionths of a second after the Big Bang, physicists are able to observe (and count) rare particle physics processes at a rate which is sufficiently high. The success of the physics program at the LHC was demonstrated when the two general purpose experiments ATLAS [12] and CMS [13] independently reported the discovery of the Higgs boson in July 2012 . In the following sections each element of the detector is described in more detail.

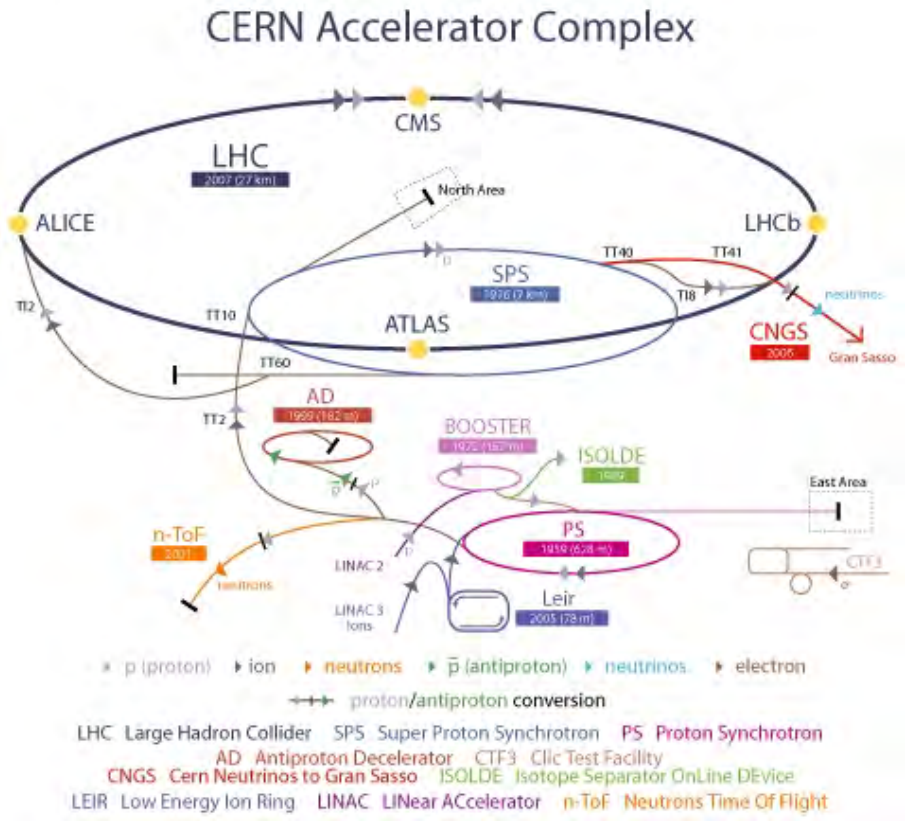


Figure 3.1: The CERN Accelerator complex, Image Copywrite @ CERN [atlas.ch]

3.1 LHC

Before the protons reach the interaction points where the detectors are located, the protons go through the CERN accelerator complex seen in fig. 3.1 which consists of several smaller accelerating rings. First, the protons are obtained from ionising tanks of hydrogen gas. The gas is broken down to atomic hydrogen where it is then stripped of its electron. The complex consists of a linear accelerator and a sequence of 3 synchrotron accelerators. As the protons leave linear accelerator they are traveling at 50 MeV when they are injected into the Proton Synchrotron (PS) Booster. This feeds the protons into the Proton Synchrotron.

Once the proton bunches are injected into the LHC from the Super Proton Synchrotron (SPS) , they have already acquired an energy of 450 GeV. An elaborate configuration of magnetic dipoles and quadrupoles creates a magnetic lens which allows positively charged proton bunches to be accelerated in counter rotating beams. During 2012, the each proton beam was accelerated to 4 TeV. The LHC is expected to operate at full design energy (13 TeV) during Run II of the data taking period which started in 2015. Once the protons

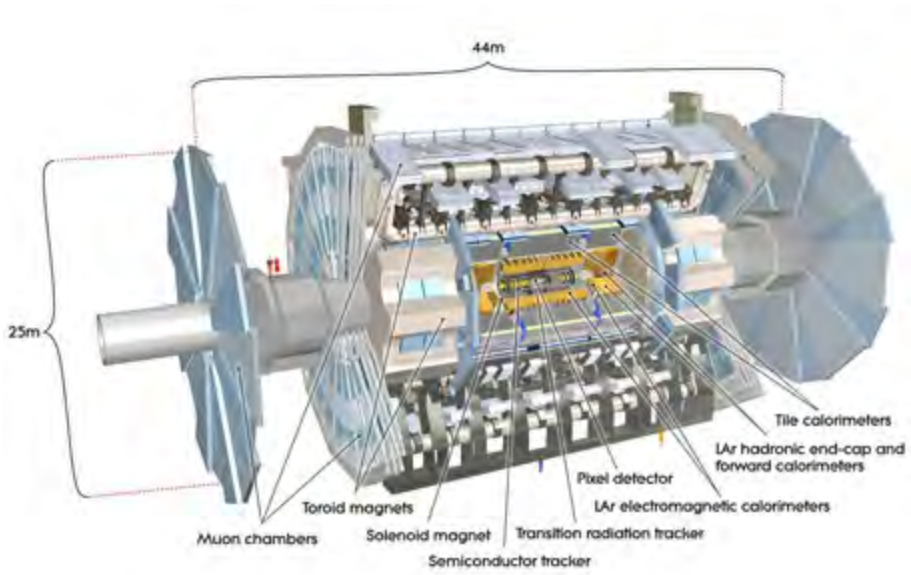


Figure 3.2: The ATLAS detector, Image Copywrite @ CERN [atlas.ch]

bunches have reached the expected centre of mass energy $\sqrt{s} = 8TeV$, they collide in the designated interaction points where the various detectors are located.

3.2 ATLAS Detector

The ATLAS (A Toroidal LHC ApparatuS) [1] detector, seen in fig. 3.2, is one of two general purpose detector which was built to explore physics at the TeV scale. The experimental conditions present at the LHC have set stringent design requirements on the detectors in order for them to cope with the the high interaction rates, radiation exposure and particle multiplicities. For a complete description of the detector, please refer to the official ATLAS Detector documentation.

3.2.1 Psuedorapidity and other variables

The ATLAS detector is the largest particle detector built to date. The 45 m long cylindrical detector has a 25 m diameter which contains three main detector subsystems. Once the proton proton bunches collide, the nominal interaction point defines the origin of the coordinate system. The z axis lies along the beam direction, the x-axis points from the interaction

point to the centre of the LHC ring, while the positive y-axis points upwards away from the interaction point. The psuedorapidity of the particles is defined from the primary vertex as follows:

$$\eta = -\log \tan \frac{\theta}{2} \tag{3.1}$$

where θ gives the polar angle of the particle which is measured from the positive z-axis. The pseudorapidity is often used instead of θ when referring to detector components since it is invariant under Lorentz boosts along the z-direction. In the x-y plane we define the transverse energy E_T, transverse momentum p_T as well as the missing transverse energy (MET) E_T^{miss}. Another important variable is ΔR which is the distance in the pseudo-rapidity azimuthal angle given by the following relationship:

$$\Delta R = \sqrt{(\Delta\eta)^2 + (\Delta\phi)^2} \tag{3.2}$$

3.2.2 Tracks

Tracks which are measured are used to reconstruct the hard scatter interaction point as well as additional collisions that occur in the same bunch crossing due to pile up. A track in ATLAS is parametrised at the point of closest approach with the global Z-axis and the following 5 perigee parameters are used to define the track with respect to the point of closest approach. The 5 perigee parameters are briefly described below with the full definitions given in [14]

- d_0 : transverse impact parameter, closest distance to Z-axis in the transverse plane.
- z_0 : longitudinal impact parameter, which is the Z coordinate of the track closest in distance to the transverse plane.
- ϕ_0 : azimuthal angle of the track at perigee in the range $[-\pi, \pi]$
- θ : polar angle of the track at perigee in the range $[0, \pi]$
- $\frac{q}{p}$: charge over momentum of the track

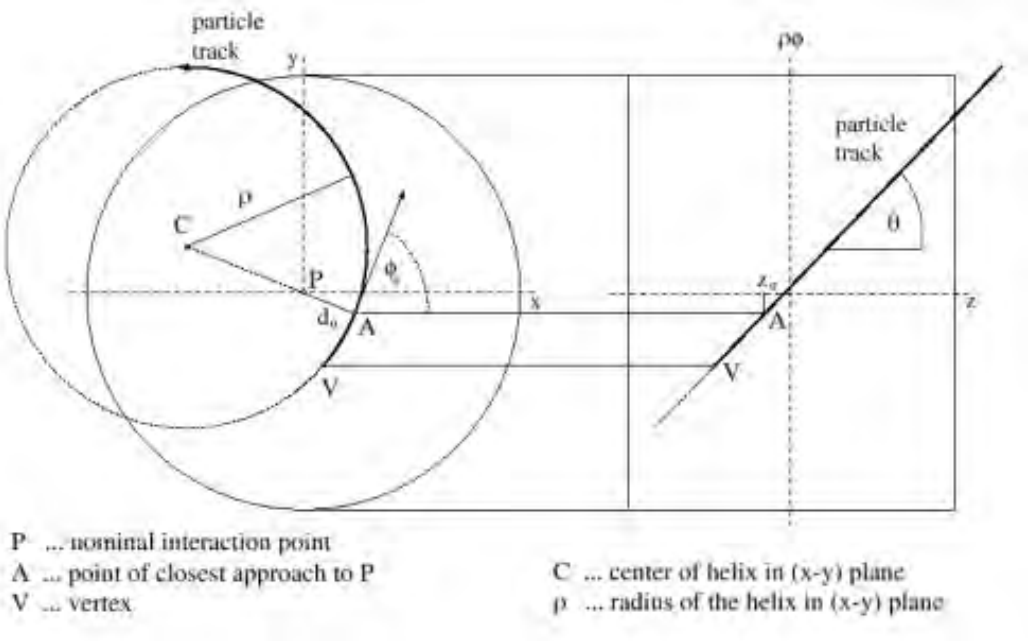


Figure 3.3: The ATLAS perigee parameters, Image Copywrite @ CERN

Figure 3.3 [15] gives a schematic view of the 5 perigee parameters. These 5 parameters contain information about the parent tracks which are used to as physics objects. The tracks can span the entire detector and will have measurements made by many different parts of the detector

3.2.3 Cross section and Luminosity

In collider experiments, in order to describe the number of collisions occurring at a particular interaction point, we use a quantity called instantaneous luminosity L. By integrating the instantaneous luminosity over the period of time that the collisions are recorded, one obtains the total integrated luminosity **L**

$$L = \int dt L(t) \tag{3.3}$$

The unit of total integrated luminosity is then $cm^{-1}s^{-1}$. In this thesis, the integrated luminosity is expressed in units of fb^{-1}, or inverse femto-barn. In 2012, the total recorded integrated luminosity for the 2012 data-taking period was 21.7 fb^{-1} as seen in figure 3.4. The

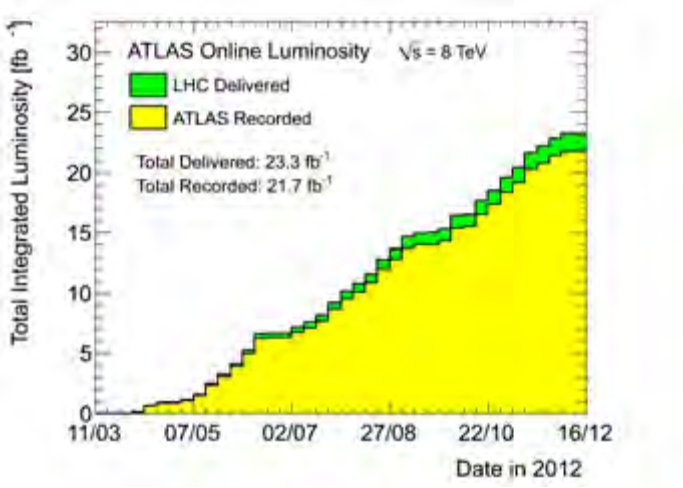


Figure 3.4: Total integrated luminosity

cross section σ_x is proportional to the probability for a given process **x** to occur. Multiplying the cross section by the integrated luminosity gives the expected number of events that should be observed for process **x**.

$$N_x = \sigma_x L \tag{3.4}$$

From this equation, we see that a higher integrated luminosity will yield more data, meaning the that the uncertainties on the measurement are reduced.

3.2.4 Pile Up

In 2012, the LHC was operating at a centre-of-mass energy $\sqrt{s} = 8$ TeV, with an average number of around $< \mu > = 20.7$ collisions per bunch crossing as seen in figure. 3.5.

The effects of pile-up (more than one proton-proton collision per bunch crossing) are quite significant on studies containing hadronic jets and it can be exceedingly difficult to disentangle the effects of pile-up. Not only does in-time pile-up occur, where more than one proton proton collision has occured, but there is also out-of-time pile-up which comes from the following bunch crossing which occurs every 50 ns. Since the triggers have a latency of around $2\mu s$, out-of-time pile-up is observed. To this effect, special corrections to the MC simulations have been implemented to ensure that simulation matches the experimental data. Certain

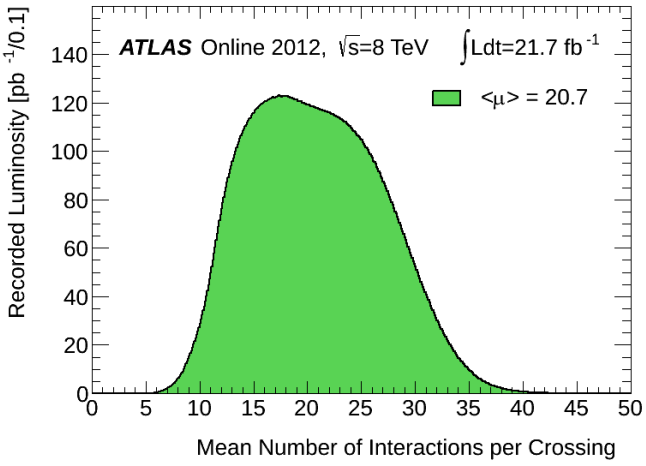


Figure 3.5: Peak average interactions per bunch crossing in 2012

track based variables have also been introduced to deal with the effects of pile-up on analyses containing hadronic jets. In order to reject jets originating from pile-up, a jet selection is applied by ensuring that the variable known as the jet-vertex fraction (JVF) exceeds 0.75.

$$\frac{\Sigma p_T^{jettrks}}{\Sigma p_T^{alltrks}} > 0.75 \tag{3.5}$$

The jet-vertex fraction is the fraction of the total momentum of tracks in the jets which is associated to the primary vertex divided by the scalar sum of of all associated tracks [16].

3.3 Tracking and Magnets

The Inner Detector (ID) is responsible for the precision tracking of charged particles. It is contained within a solenoidal magnet which provides a magnetic field strength of $2T$. The design of the ID is optimised to provide excellent momentum resolution which allows for primary and secondary vertex reconstruction. The p_T threshold of charged tracks in the pseudo-rapidity region $\eta < 2.0$ is $0.5GeV$. The precision tracking detectors cover a pseudo-rapidiy range $\eta < 2.5$ while offering full coverage in ϕ. In the 3512 mm cylindrical envelope which is the ID, there are three independent sub-detector systems. Closest to the beam line is the silicon pixel tracker and the semi-conducting tracker. These work in conjunction with the barrel Transition Radiation Tracker (TRT) modules which consist of gaseous straw tube

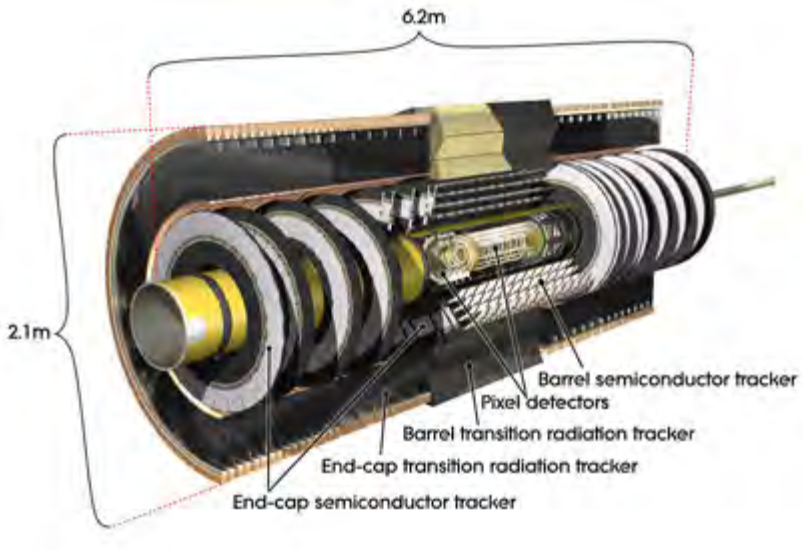


Figure 3.6: The ATLAS Inner detector, Image Copywrite @ CERN [atlas.ch]

elements. At the interaction point, the particle track density within $\eta < 2.5$ is very high due to the 40 000 Hz collision rate where 1000 particles emerge during every collision. This high density requires that the inner detector have very high resolution which will allow for precision tracking measurements.

Discuss the Transition Radiation tracker. Have only mentioned it in this section. Complete this part.

3.3.1 Calorimeters

Calorimeters are very important detectors at the LHC as they provide measurements of the energies of electrons, photons, jets and well as missing transverse energy. These calorimeter measurements are crucial to particle identification and help distinguish between electrons and photons, jets and hadronic tau decays. For this purpose, the calorimeter system in the ATLAS detector can be divided up into two parts. The electromagnetic and hadronic calorimeters. The purpose of the calorimeter is to measure the energy of neutral hadrons and charged particles which interact with and deposit their energy into the material. The calorimeters cover a range of $(\eta) < 4.9$. The η range which is matched to the inner detector is most suited to the fine granular measurements of electrons and photons. In order to distinguish between electrons and photons, tracking and energy information is obtained with the Liquid Argon (LAr) electromagnetic (EM) calorimeter in the barrel region $\eta < 1.475$

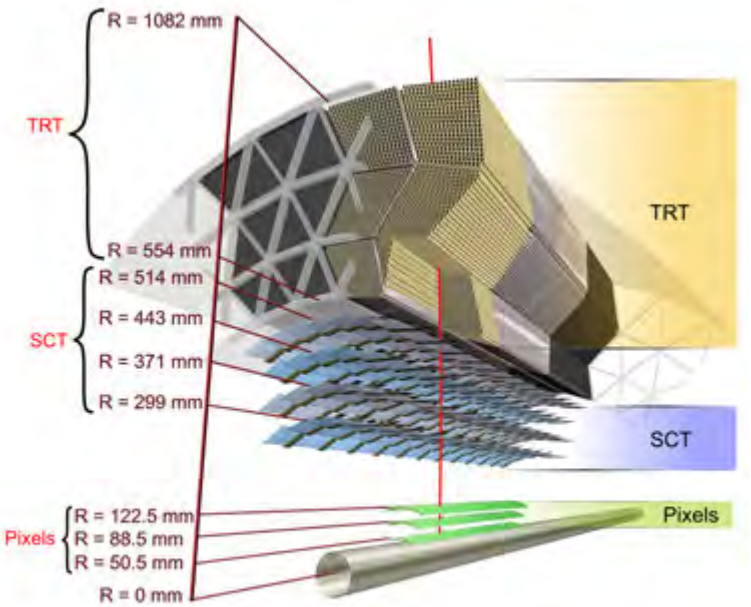


Figure 3.7: The ATLAS Inner detector, Image Copywrite @ CERN [atlas.ch]

and end-cap region within $1.375 < \eta < 3.2$. The EM calorimeter is an accordian shaped detector with alternating layers of of absorber (lead pb) and sampling material (LAr). In order to suppress punch through to the muon system, the thickness of the EM calorimeter is greater than 22 radiation lengths in the barrel region. The accordion shape also provides full ϕ coverage and symmetry without any azimuthal cracks. The granularity of the calorimeter decreases as you move further out into the detector but is still sufficient for measuring E_T^{miss} and reconstructing jets which is important to this analysis.

The barrel and end cap hadronic calorimeters are made up of a tile scintillating hadronic calorimeter in the barrel region and a LAr hadronic end cap calorimeter. The tile calorimeter envelops the EM calorimeter and covers the range $\eta < 1.7$ also consisting of alternating layers of steel (absorber) and scintillating tiles as the active material. Steel is used due to its high density and ferromagnetic properties The tile calorimeter extends from an inner radius of 2.28m up to 4.25 m as the outer radius.

In the ATLAS experiment, noise in the caloirmeters needs to be taken into account. The two primary sources of noise in the ATLAS calorimeters are from the readout electronics while the second source of noise comes from "pile-up" which arises from extra proton proton

interactions during a bunch crossing or from interactions occurring in a crossing shortly after a primary interaction. The response time of the calorimeter is longer than the 50 ns interval between bunch crossings. During an interaction, particles deposit their energy in a copious amount of calorimeter cells in all directions which are measured and various reconstruction algorithms are used to identify the types of particles based on how they have deposited their energy in the calorimeters.

3.3.2 Muon System

The only particles that pass through beyond the calorimeter are muons and neutrinos apart from the small number of punch-though hadrons that reach the Muon Spectrometer. The Muon Spectrometer (MS) surrounds the calorimeters and forms the outer most part of the ATLAS detector. The muon trajectory is bent by a 0.5 T magnetic field in the barrel region by the large superconducting toroidal coils and a 1 T field in the end cap toroids. produced . The MS consists of four subsystems: Monitored Drift Tube (MDT's) chambers and Cathode Strip Chambers (CSC) which provide the precision tracking measurements as well as the Resistive Plate chambers (RPC) and Thin Gap Chambers (TGC) which act as the trigger chambers. The trigger chambers provide the bunch crossing identification as well as the second coordinate in the non-bending plane. Identification of muons for triggering and offline construction is important to this analysis since is it used for reconstructing the W candidates.

3.4 Trigger and Data acquisition

In order to deal with the 25 ns (40 000 Hz) bunch cross rate, a trigger system was introduced to the ATLAS design which reduces the output rate from 40 000 Hz to approximate 200 Hz. The trigger system is divided into three levels. Level 1 (L1) is a hardware based trigger while Level 2 (L2) and the Event Filter (EF) are software based triggers. As seen in figure 3.8, each level of the trigger system is designed to refine the decisions of the previous level by applying additional selection criteria. In order for the L1 trigger to trigger on an event, it will use very limited information from the calorimeter and muon sub detector systems to search for leptons, jets and missing transverse energy. The L1 trigger also defines Regions of Interest (RoI) which are used by the subsequent triggers. The L1 trigger decision is meant to keep the most interesting events and discard the rest is made in less than 2.5 μs thereby

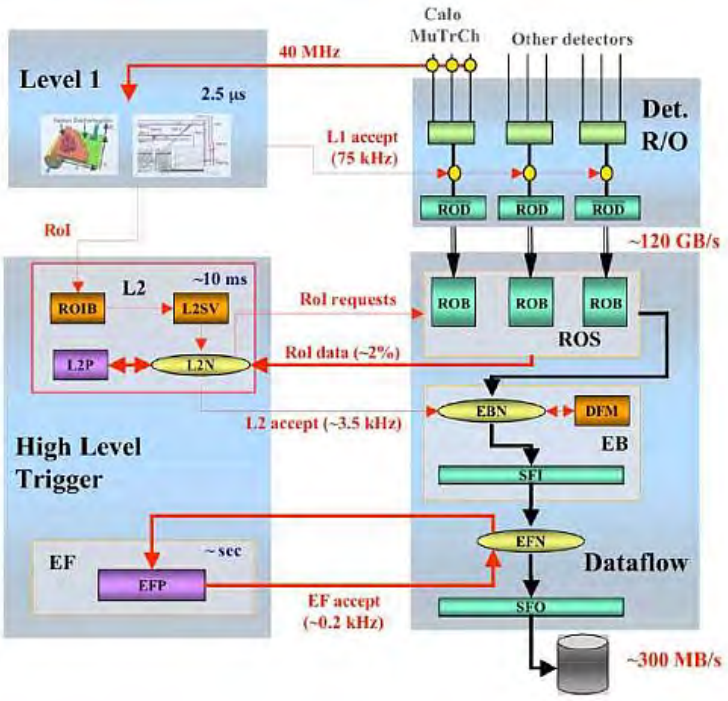


Figure 3.8: The trigger system @ CERN [atlas.ch]

reducing the rate to 75 kHz. Level 2 is a software based trigger and applies a rudimentary analysis on the regions of interest which have been previously identified with the L1 trigger. It has a latency of 10 μs. The Event Filter (EF) is also a software based trigger and which implements the last level of event processing before events are stored to disk. The software based triggers make up the High Level Trigger (HLT

This trigger selection is referred to as the online selection, while the offline selection is performed later during an analysis of the proton-proton collisions.

3.5 ATLAS Simulation

Monte Carlo simulations are a very important part of the experiment and form an important role in providing an understanding of the detector performance and response. By using studies on the simulations, the sensitivity of the experiment to particular physics processes can be optimised. There are also used to continually asses the performance of the detector which enables an accurate interpretation of the data. The simulation program has been integrated into the ATLAS software framework and uses the GEANT4 simulation toolkit [17]

to produce MC simulated events. An *event* is any scattering process, and MC simulations are used to randomly generate a large amount of these events with a detailed simulation of the detector response to that event. The simulation chain is divided into three steps:

- Generation of event and immediate decays
- Simulation of detector and physics interactions
- Digitization of energy deposited in the sensitive regions of the detector into voltages and currents for comparison to the readout of the ATLAS detector.

The output of the simulation chain is the same as that of the ATLAS data acquisition system (DAQ) which enables the real data and simulated data to be run through identical trigger and reconstruction packages. GEANT 4 is used to model the detector geometry as well as the physical conditions present during a data taking period. To do this, very large computing resources are required to accurately model such a complex system. An extensive description of the validation of the software and performance is given in chapter 5 which formed part of the service work I did as part of this thesis.

Chapter 4

Object definitions, identification and reconstruction

A vast amount of data is collected by the ATLAS detector. The different sub-detector systems measure energy deposits, tracks, trigger hits, pixel hits, charge and momentum. This information is fed into reconstruction algorithms which give us a record of the event. This chapter will give an brief outline of how the ATLAS detector turns the raw signals into calibrated physics objects. This section details the reconstruction of leptons, photons and jets as well as the calculation of missing transverse energy and transverse mass used in the measurement of W production in association with jets. A focus is placed on the reconstruction of the muon and properties of the muon object which could be used to discriminate between muons originating from the QCD background and muons originating from a W boson decay.

4.1 How ATLAS sees all the different particles

Figure 4.1 is a wedge cross section of the ATLAS detector and clearly shows all the tracks and energy deposits where are left by the various particles. Electrons and photons leave energy deposits in the Electromagnetic calorimeter. The electrons leave tracks in the inner detector while a photon leaves no track and only a energy deposit is observed. Neutral particles, leave no tracks, and hardrons deposit most of their energy in the hadron calorimeter. Muon's pass straight through, leaving tracks in the inner detector as well as being recorded at the muon spectrometer.

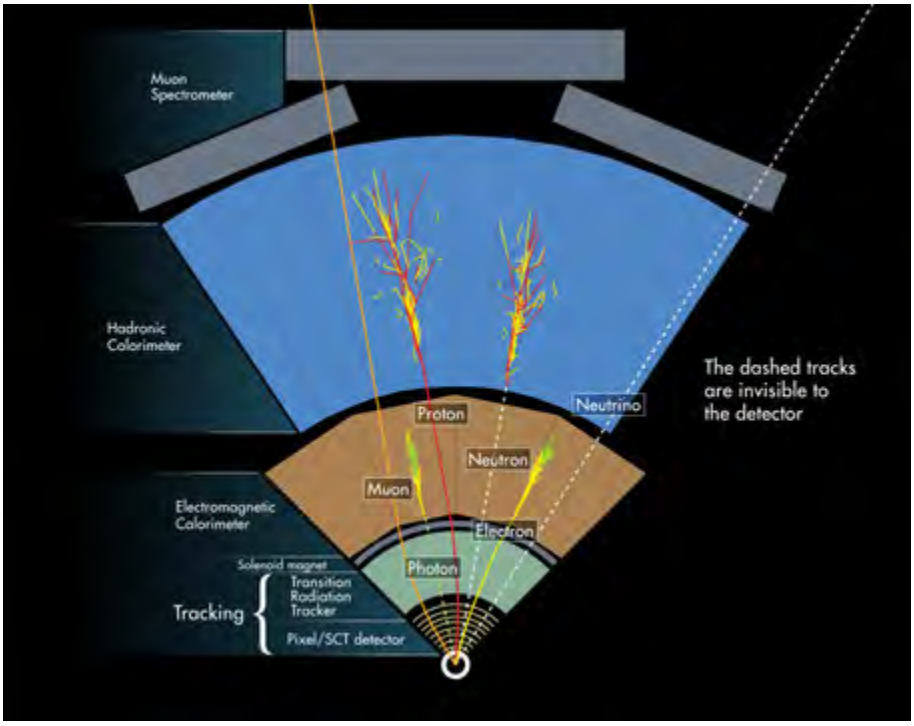


Figure 4.1: The ATLAS wedge, Image Copywrite @ CERN [atlas.ch]

4.2 Muons

4.2.1 Muon Trigger

Muons are heavy and interact very weakly with the detector, typically passing through the inner tracker, calorimeters and muon spectormeter. The muons are charged and hence leave hits in the tracking chambers. They are therefore very useful objects on which to trigger. The ATLAS muon trigger system has been designed to select muons with a high efficiency as well as wide momentum range [18]. Events with muons in them are selected in three stages. The L1 muon trigger system will receive the input from fast muon trigger detectors using resistive plate chambers and well as thin gap chambers (RPC and TGC) in the barrel and end-cap regions respectively. Then, after this step, the Muon Trigger to Central Trgger Processor Interface (MUCTPI) forms the interface betweem the central trigger processor and the muon trigger chambers.

4.2.2 Muon Recontruction

First, it must be noted that there are two types of reconstruction algorithms used at ATLAS. The first one is called Staco. This statistically combines muons that have been measured in the MS and ID. the other algorithm is MUID which performs a track fit to muon hits in the MS and ID. We are able to categorise the muons based on the way that they are reconstructed and these two algorithms produce 4 types of reconstructed muons:

- Stand-Alone muon: This reconstruction is performed by locating the track of the muon in the Muon Spectrometer, it is then extraplolated back to the beam line. By correcting for the energy loss in the calorimeter, the muon momentun is measured at the interaction point.
- Combined Muon: Tracks from the Inner Detector (ID) are statistically combined with tracks from the MS. Energy losses that occur in the calorimeter are again taken in to account by per- forming parametrizations thus providing the most precise momentum and position measure- ments of the muon.
- Calorimeter tagged muons: If tracks from the ID lead to energy depositions in the calorimeter, a particle can be identified as a muon if it satisfies the criteria of a minimum ionising particle. These calorimeter tagged muons are used instead of the ID tracks in order to reduce background in the tag and probe method.
- Segment tagged moun: A track in the ID is extrapolated to the MS. Provided that the track segments in the precision muon chambers are straight, the particle can be identified as a muon.

A full description of the Muon Spectormeter can be found in [19]

4.2.3 Muon criteria for inner detector track

This analysis uses a combined muon. The associated inner deterctor tracks have the following cuts applied as per the recommendatins provided by the Muon Combined Performance group [20]

- Number of pixel hits + number of crossed dead pixel sensors > 0

- Number of SCT hits+number of crossed dead SCT sensors > 4
- Number of pixel holes + number of SCT holes < 3
- A set of TRT hits + number of TRT outliers are given by the following (where n = number of TRT hits + number of TRT outliers: $0.1 < \eta < 1.9$ for n > 5 and $nTRT^{outliers} < 0.9 * n$

4.3 Jets

Jets are collimated bundles of particles produced in proton-proton collisions by the hadronisation of high energy partons emerging from hard scattering processes. Jets are also produced by initial and final state gluon radiation which is associated with this scattering. Jets which have a transverse momentum higher than a few GeV's will interact significantly with with the ATLAS detector. The detector systems responsible for the reconstruction of jets are the calorimeters. As detailed in section 3.3.1, the calorimeters collect of the order of 190,000 cell signals which are used for object reconstruction. In addition to the signals collected by the calorimeters, tracks in the inner detector are also used in the reconstruction of jets. Jet finding algorithms are needed to define jets in proton proton collisions. A topological jet clustering algorithm is used to define jets based on a seed cell in the calorimeters [21]. Jets are built from topo-clusters which are defined by considering the signal to noise (S/N) ratio of the cells of all the the ATLAS calorimeters. If the S/N ratio of a cell is larger than 4, this cell is considered a seed cell. Then, any neighbouring cells which have a S/N greater than 2 are added iteratively to the cluster. Finally, all the neighbouring cells are added to the cluster. If a cell with an energy exceeding 500 MeV is found, this is then used as the seed cell in a second iteration of the clustering. Defining jets using this topo-clustering algorithm means that jets do not have a fixed amount of cells used to define them and they are reconstructed strictly from the calorimeter information. However, any particles depositing energy in the calorimeters will be tagged as jets (such as electrons and muons) and a jet cleaning algorithm is needed during the analysis phase to remove these from the jet collection. The energy of the topological clusters can be calibrated in two ways. First, it can be calibrated to the EM (electromagnetic) scale or a local cluster weighting (LCW) which identifies the cluster as hadronic or electromagnetic based on the shape and density of the cluster. The final cluster energy given at the EM scale is the 'raw' electronic signal from the calorimeter reading and the energy has not been corrected for energy loss arising from non-interacting neutrinos, dead material in the detector or from the topological jet clustering algorithm.

The local cluster weighting energy calibration (which does apply a correction to each cluster) represents the true energy of the particle more closely. Algorithms are very important as they are used to group the hadronic showers and reconstruct them into jet objects which are used in physics analyses. The most common algorithm used by physics analysis groups within ATLAS is the anti-k_T algorithm [22]. It behaves like an an idealised cone algorithm and uses distances d_{ij} between particle pairs i and j. If distances between particle pairs are smaller than the smallest distance d_{ij}, particle i is removed from the list. The distance measure is given by

$$d_{ij} = min(k_{ti}^{2p}, k_{tj}^{2p}) \frac{\Delta_{ij}^2}{R^2} \tag{4.1}$$

where $\Delta_{ij}^2 = (y_i - y_j)^2 + (\phi_i - \phi_j)^2$ and k_{ti}, y_i and ϕ_i are the respective transverse momentum, pseudorapidity and azimuth of the particle i. The usual radius parameter R is the maximum distance allowed between particles. The algorithm keeps adding and combining particles to the list until the distance between particles exceeds R which results in a perfectly conical jet.

4.3.1 Momentum and Missing Transverse Energy

Transverse momentum p_T and missing transverse energy E_T^{miss} are very important and fundamental measurements in particle physics. The transverse momentum is the component of the momentum which is transverse A neutrino will not interact with the detector and will leave no energy behind. This will lead to a missing transverse energy signature in the detector. For a complete description of the Missing Transverse Energy see [23]. The missing transverse energy is defined as the event momentum imbalance in the transverse plane to the beam axis. The momentum is a vector, therefore the sum of all the p_T will be equal to zero is the detector is able to measure all the momenta of the out-going particles.

The cells in the calorimeters are associated with objects as described in the preceeding sections so that the E_T^{miss} is calculated as follows:

$$E_{x(y)}^{miss} = E_{x(y)}^{miss,e} + E_{x(y)}^{miss,\gamma} + E_{x(y)}^{miss,\tau} + E_{x(y)}^{miss,jets} + E_{x(y)}^{miss,softjets} \tag{4.2}$$

$$+E_{x(y)}^{miss,calo,\mu} + E_{x(y)}^{miss,CellOut} + E_{x(y)}^{miss,\mu} \tag{4.3}$$

The MET_Refinal_TRT algorithm is used to calculate the missing transverse energy in this analysis.

4.3.2 Transverse Mass

In order to calculate the mass of the W boson, a quantity called the "transverse mass" is defined in terms of the missing transverse energy and the vectorial information about the lepton. This is important for analyses where a particle (such as a neutrino) iin the final state cannot be detected directly. The transverse mass of the W boson decaying to a lepton with transverse energy $E_{T,1}$ and transverse momentum $p_{T,1}$ and a neutrino with $E_{T,2}$ and $p_{T,2}$ would have the transverse mass defined as:

$$M_T^2 = (E_{T,1} + E_{T,2})^2 - (p_{T,1} + p_{T,2})^2 \tag{4.4}$$

which when we rearrange the equation becomes:

$$M_T^2 = (E_{T,1})^2 + 2E_{T,1}E_{T,2} + (E_{T,2})^2 - (p_{T,1} \cdot p_{T,1} + 2p_{T,1} \cdot p_{T,2} + p_{T,2} \cdot p_{T,2}) \tag{4.5}$$

which simplifies to

$$M_T^2 = ((E_{T,1})^2 - p_{T,1} \cdot p_{T,1}) + ((E_{T,2})^2 - p_{T,2} \cdot p_{T,2}) + 2(E_{T,1}E_{T,2} - p_{T,1} \cdot p_{T,2}) \tag{4.6}$$

By subsituting the relativistic energy-momentum relation $E^2 - p^2 = m^2$ into the first two terms of equation 4.8, the equation simplifies to

$$M_T^2 = (m_1)^2 + (m_2)^2 + 2(E_{T,1}E_{T,2} - |p_{T,1}||p_{T,2}| \cos\phi) \tag{4.7}$$

where ϕ is the angle between the decaying particles. However, for massless particles , $m_{1,2}$ = 0, $|p_T| = E_T$, the transverse mass of the W candidate is calculated as

$$M_T^W = \sqrt{2E_{T,1}E_{T,2}(1 - \cos\phi)} \tag{4.8}$$

where the E_T for the neutrino is obtained from the E_T^{miss}.

Chapter 5

Tau Physics validation

As part of my service work to ATLAS, I performed the Tau Physics Validations for a period of 10 months from August 2014 until June 2015. In this chapter I will briefly describe the process of Physics Validation.

5.1 ATHENA: The ATLAS Software Framework

ATHENA [24] is an evolution of the GAUDI Framework [25] which was developed by the LHCb experiment. Both of these frameworks were built on the component based architecture designed in mind for specific applications to physics data-processing. The component based framework allows flexibility when developing shared components and analysis tools but it also enables the development of tools which are very particular to a specific experiment and physics analysis. The software is used to support code organization and development by providing common functionality, python scripting facilities which allow for individual configuration and interactive use. It is used by almost everyone within the ATLAS collaboration, whether it be used for the production of Monte Carlo samples or for running a physics analysis. The ATHENA framework is organised in a hierarchical way, such that packages within each project only depend upon other packages within the same project or from a lower project. The ATLAS offline software is organised as follows:

- *AtlasCore* This project is the core component of ATHENA and is the core set of packages which are common to all other projects
- *AtlasConditions* This is the set of packages dealing with the ATLAS detector's geom-

etry, calibration, and misalignment.

- *AtlasEvent* This project is dependent on the previous two projects. It contains the set of packages that deal with the Event Data Model (EDM). The EDM defines a number of different data formats such as RAW, AOD, xAOD, D3PD etc.
- *AtlasSimulation* These contain all the tools and services related to simulation and event generation. This includes the Geant4 Simulation , physics event generators, pile-up and digitisation tools.
- *AtlasReconstruction* The contains the packages for event reconstruction as well as on-line monitoring.
- *AtlasTrigger* This is dependent on *AtlasReconstruction*. It contains all packages associated with the high-level trigger (HLT)
- *AtlasAnalysis* This is dependent on the previous package. It provides the tools relevant for physics analysis, monitoring and event display.
- *AtlasOffline* This is the top level project and depends on all the other packages. This final project is dedicated to an official release of ATHENA.

These packages are updated frequently to upgrade the code or to implement bug fixes. For each version of the code, a special tag is created to distinguish between the different versions. The code and tags are all stored using the Subversion (SVN) repository [26]. It is invaluable using versioning control for all the ATLAS offline software which allows developers to revert back to previous working versions of the code, prior to any changes. This allows backwards compatibility with previously released versions of ATHENA. During any given time, the ATLAS software consists of a limited number of open releases. Each release has a tag number associated to is which refers to the major release of ATHENA, the branch of that specific release, and patch number which will contain minor upgrades or changes.

Besides the core tools and packages listed previously, ATHENA also contains some domain specific tools. These include histograming and n-tuple services which allow the user to book, fill and analyse histograms from within the Gaudi Framework. In addition, there are generic tools which allow for performance and resource monitoring, logging and error handling, as well as a job options service. This service contains a catalogue of user-modifiable properties. For example, the value of a cone radius "ConeR" can be set using a job-options file or via the interactive ATHENA prompt.

By allowing scripting functionality within the ATHENA framework, applications can be built that are customised for specific capabilities from a pool of available components and packages. The scripting in ATHENA is primarily done using the open source programming language, Python [27]. The python binding are used to control the packages, specify the input data-sets, the output data-set names, error logging, histrogramming as well as specifying the tools and packages need for running jobs on the ATLAS computing grid. This is the procedure employed for the physics validation described in section 5.2.

5.2 Physics Validation

The validation of the ATLAS simulation chain is performed in two phases. The first phase, the performance of the software has to to be assessed and then the physics performance is tested. This is the part of the validation step that I was involved in. Before the release becomes available for mass MC production, the new release needs to be tested for consistency and to ensure that no bugs have been introduced into the new release. Once a new release of the ATHENA software has been released to the GRID sites, a set of several physics MC samples are produced. The aim of the physics validation is to perform tests and tasks using these MC samples which have been produced with the ATLAS software which has a new bug fix or cache, and checking the quality of the output against a previously validated version of ATHENA. A large number of MC samples are produced, and there are several dedicated groups, each who have a representative from every detector performance (tracking, b-tagging and jet reconstruction) and physics group (Standard Model, supersymmetry, and exotic groups). The experts from these groups liaise with their representative to verify the quality of a single object reconstruction (electrons, muons and taus etc) as well as more complex physics analyses such as the mass reconstruction in $Z \rightarrow \mu\mu$. The new cache will have a new feature implemented, usually for the digitization and reconstruction of Monte Carlo. In order to expose any discrepancies or errors, the validation samples need to be large as these errors will not be discovered in a sample with low statistics. For example, there could be a shift in the number of reconstructed objects or a shift in the reconstructed momentum of the object. A schematic representation of the Physics Validation process is given in figure 5.1.

Once the new samples have been produced, the physics validation convenors will specify a list of tasks for the validation meeting. The detailed procedure involved in the validation is as follows:

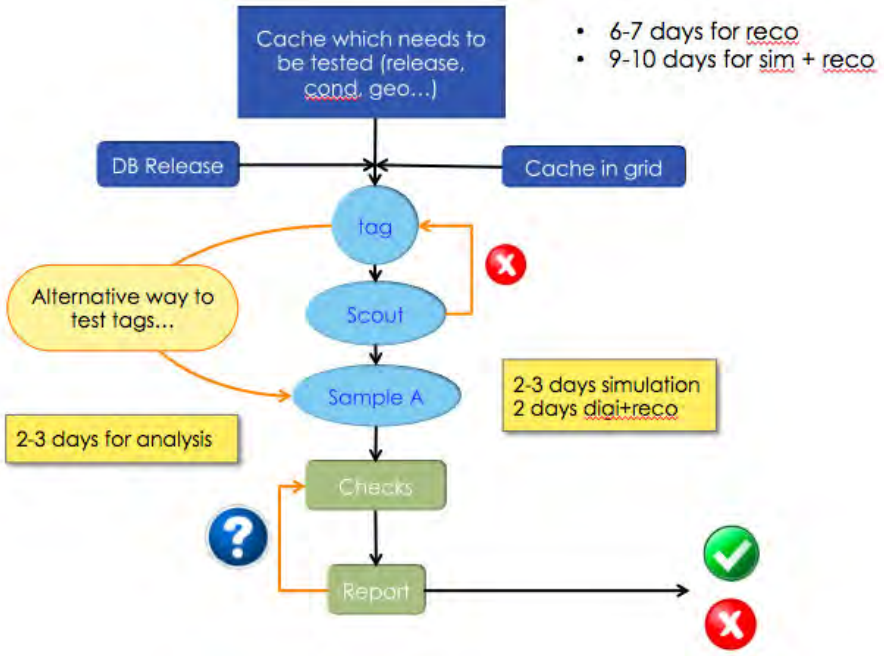


Figure 5.1: Physics validation flow chart

- A request is made for a new cache/configuration by the Simulation, Reconstruction, Data Preparation or MC production convenors.

- The new request gives details pertaining to the geometry of the ATLAS detector, trigger menu options, the reference sample and any addition options needed for the production of the sample. Samples are then produces with the requested configuration that has been requested. For a simulation validation, the entire simulation and reconstruction needs to be performed. If a request is made for digitization and reconstruction only, only the reconstruction of the cache is required.

- The information with the samples and configuration options is then circulated to all the validation groups. This included the details of the newly produced (test) samples and older validated (reference) samples. The convenors also give feedback if differences are expected for a particular validation and then validation contacts are alerted to these differences when performing the physics validation.

- This gives the validation contacts a couple of days to perform the comparisons and to formulate a report indicating any differences seen between objects in the test and reference samples. This report is the presented to the convenors and all the other validation contacts for the various performance and physics groups.

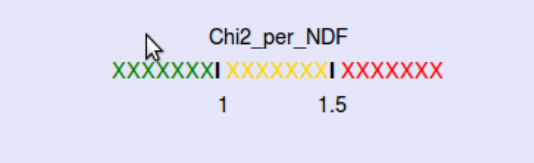


Figure 5.2: squared score attached to each histogram during physics validation

These new samples will be produced using the new software and compared against a previously validated release of ATHENA. It is therefore vital that the software versions maintain backwards compatibility to allow for fair comparisons. As can be seen in figure 5.1, the new cache is released and then it takes 2-3 days for simulation and 2 days for the digitization and reconstruction of the samples. The samples then become available to the validation contacts and each dedicated group will run the necessary validation steps. The new samples are referred to as "test samples" while the previously validated samples are the "reference samples". The samples are centrally generated and each sample has a unique name and tag attached to it. In order to do the physics validation, important discriminating variables are compared. The validation of the tau reconstruction is performed with the tau validation package: TauDQA. The package is dependent on another package called TauHistUtils which produces histograms which are filled by TauDQA for validation. An array of important discriminating variables is plotted for all the tau jet candidates.

When comparing the test and reference samples, the tau validation package overlays the histograms for comparison. In order to quantify the differences between the test and the reference samples, a χ^2 test is used. For the histograms, let T_i and R_i be the number of entries in the i^{th} bins of the test and reference sample respectively. The χ^2 distribution is computed as follows:

$$\chi^2 = \Sigma \frac{(T_i - R_i)^2}{T_i + R_i} \tag{5.1}$$

The histograms are compared and are stored online where all the histograms that have been compared are assigned a RED, YELLOW or GREEN flag using the following criteria depending on their $\frac{\chi^2}{NDF}$ score, where NDF is the number of degrees of freedom and corresponds to the number of bins in the histogram

As can be seen in figure 5.2, for histograms scoring between 0 and 0.5 are flagged GREEN, histograms with a score between 0.5 and 1 are flagged YELLOW. This means that the

histogram should be investigated by the validation contact to ensure that the differences in the histograms are within statistical uncertainties and histograms with a score great than 1 are flagged RED. This indicated the differences are significant and the they need to be examined closely. Depending on the validation task and which changes have been implemented in the new cache, some differences are expected. The flagging of histograms based on the $\frac{\chi^2}{NDF}$ score is not a completely accurate way of quantifying the differences when comparing test and reference samples but allows for a quick reference and inspection of problematic histograms. For the Physics Validation task, a detailed knowledge of tau leptons and their reconstruction within ATLAS was required to give knowledgeable feedback during the validation meeting and to formulate coherent reports and to liase with the Tau Working Group convenors. In section 5.3, the characteristics of the tau lepton are detailed along with the approach used by the Tau Combined Performance Group for tau reconstruction.

5.3 Introduction to Taus

The tau is the heaviest of all the leptons identified in the third lepton generation. Due to its mass, (1776.82 ± 0.16 MeV/c^2), the tau is very unstable and has a very short life time of 2.9×10^{-13} s. It promptly decays to other particles. The decays can be describes by the two following modes: hadronic and leptonic. The decay modes are listed in the table obtained from the Particle Data Group (PDG) [28].

Leading Hadronic decay modes	
Decay Mode	Branching Fraction
$\pi^-\nu_\tau$	10.91 ±0.07%
$\pi^-\pi^0\nu_\tau$	25.52±0.1%
$\pi^-\pi^+\pi^-\nu_\tau$	8.99 ±0.06%
$\pi^-\pi^0\pi^0\nu_\tau$	9.27 ±0.12
$h^-\omega\nu_\tau$	1.99 ±0.08%
$\pi^-\pi^+\pi^0\pi^-\nu_\tau$	2.70 ±0.08%
$\pi^-3\pi^0\nu_\tau$	1.04 ±0.07%
Total Hadronic	64.79 %

Table 5.1: Branching fractions of hadronic tau decays

Almost 64.8% of the decays are hadronic, while approximately 35.2 % are leptonic decays. They are useful objects to study since they are used in many physics searches, namely $H \to \tau\tau$ as well as in W,Z and SUSY searches.

In the table. 5.1, the branching fractions for the leading hadronic tau decays are given. The hadronic τ-lepton decays have a characteristic 1-prong or 3-prong signature. The 1-prong decays, with a charged π, occur in the channels $\tau \to \pi^{\pm}\nu$ 22.4 % of the time and $\tau \to n\pi^0\pi^{\pm}\nu$ 73.5 % of the time. 3-prong decays containing 3 charges πs occur as $\tau \to 3\pi^{\pm}\nu$ at a rate of 61.6 % while $\tau \to n\pi^0 3\pi^{\pm}\nu$ occur at a rate of 33.7 %. These 1- and 3-prong hadronic decays channels are dominated by the charged $\pi^{\pm}$ and neutral π^0. Despite the its short lifetime, it is still long enough to perform a reconstruction by the ATLAS detector in the 3-prong decay vertex.

This is because the neutral pions decay almost immediately into a di-photon pair which are converted inside and measured by the calorimeters. For this reason, the ATLAS reconstruction algorithms only use the hadronic decays to to reconstruct taus. These hadronic tau decays are often confused with QCD multi-jet events. This is because multi-jet events produce tau like tracks in the clusters of the calorimeter as well as tracks in the inner detector. When searching for taus, we search for collimated low multiplicity jets with energy deposits in the calorimeters as well as 1 or 3 tracks in the inner detector. In order to distinguish between QCD multi-jets events and tau decays, since this is not done in the reconstruction phase, a dedicated identification step is required. This identification step is based on variables which allow one to discriminate between QCD jets and a tau lepton.

(a) QCD jet (b) tau hadronic decay

Figure 5.3: A (a) QCD jet and (b) Tau hadronic decay

The charged tracks from a tau decay are collimated in a narrow cone, while tracks from QCD jets are dispersed more widely as can be seen in the figure [29]. Another distinguishing

feature of hadronic tau decays versus QCD jets is that QCD jets have a large number of tracks while tau hadronic decays leave only one (1-prong) or three (3-prong) tracks from pions in the core cone. The core (isolation) cone is defined as:

$$\Delta R = \sqrt{\Delta\eta + \Delta\phi} < 0.2(0.4) \tag{5.2}$$

Exploiting the differences between the shower shapes of tau decays and QCD jets in the calorimeters can also be used as a distinguishing feature for the reconstruction of a tau candidate. Information from the different sub-detector systems along with the particular attributes of hadronic tau decays are used to reconstruct tau objects. However, this information is not adequate to reject background jet events and further discriminating variables need to be used. Apart from the isolation requirement in the tracks and clusters in the calorimeter, certain characteristics of the track systems as well as the shower shapes in the calorimeters provide more information on the reconstructed object. A set of identification variables have been constructed from these criteria which inform how a tau lepton candidate is selected. The criteria are outlined as follows:

- The inner detector provides information pertaining to the charged track or multi-track system. These tracks cannot be matched to any tracks in the muon spectrometer nor should they be characteristic of electron tracks (which have many hits in the TRT).
- The multi-track system need to be well group in either the η or ϕ plane and the invariant mass must be smaller than that of a tau lepton.
- The charge of the tracks must add up to the charge of the resonant tau.
- Calorimeters measure energy deposits and the shapes produced by jets. Tau decays will produce narrow and collimated showers in the EM calirometers.
- Finally, tracking and calorimetry information is collated to ensure that EM showers are matched up with tracks in the ID.

5.4 Example of a Tau Physics Validation

On the 9th of February 2015, a physics validation was carried out consisting of three validation tasks. The details of the validation tasks sent out to the validation contacts consisted

of the following:

- A set of samples were created for a digitization and reconstruction (digi+reco) cache only. No simulation.
- The task was high priority
- The validation consisted of three tasks covering the validation of ATHENA version 20.1 with a 50 ns bunch spacing tested against reference samples made with ATHENA version 19.1.1.5
- The digi+reco configuration has the DC14-like setup as well as MC15-like setup

DC14 was the new computing model implemented and to be tested for RUN-2 at the LHC during 2014 to prepare for RUN-2. MC15 was the 2015 campaign of the ATLAS production group which would provide the geometry and conditions for the 2015 data-taking period during early 2015 data analysis for MC samples

The three task are detailed below and the results of the comparison are given in section 5.4.1.

Task One: Test of digi+reco in ATHENA version 20.1.0.2 with a DC14-like configuration *including* pileup. For this task, the trigger had been turned off. The two reference samples used for comparison were from the last validated ATHENA release 19.1.1.5 samples and the 20.0.0.2 samples both of which included pile-up.

Task Two: Test of digi+reco in 20.1.0.2 in MC15-like configuration *without* pileup. This had the MC15 setup (ATLAS-R2-2015-02-01-00 geometry and OFLCOND-RUN12-SDR-22 conditions) but with the trigger turned off just like in **Task One**. The reference was produced with ATHENA 20.0.0.2.

Task Three: Validation of digi+reco with 50 ns bunch spacing compared to a sample containing a bunch spacing of 25 ns. This sample was updated from a validation meeting preceeding this meeting. The 50 ns bunch spacing options had been correctly set for OFC (Optimal Filter Coefficients) optimization which is used for minimizing electronic noise in the LAr calorimeter.

For all three tasks, the $Z \to \tau\tau$ (PowHegPythia8) sample was used to perform the comparison.

5.4.1 Task One

Due to some central GRID production errors, the first reference sample could not be produced. The comparison of the test sample with the second reference sample (both including pile-up) resulted in some histograms flagged as RED by the simple χ^2 tests.

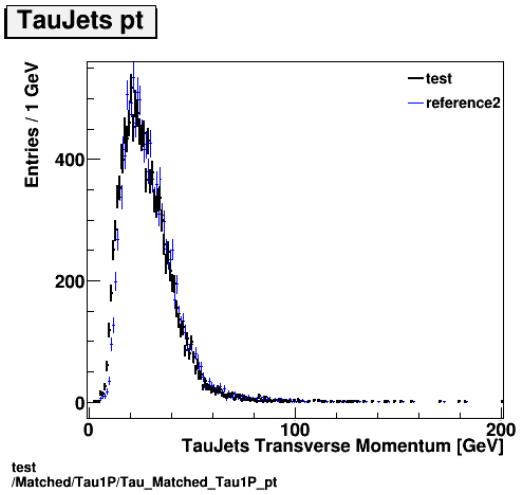
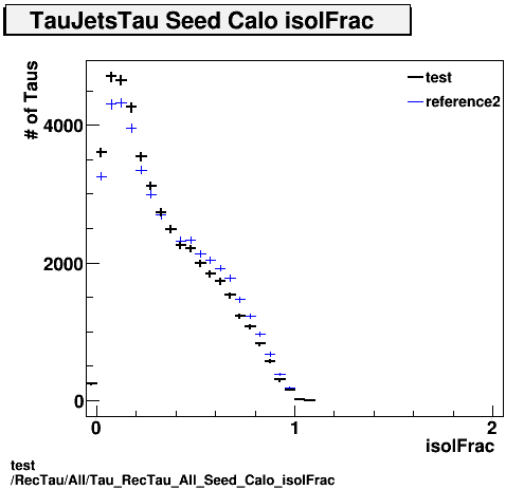


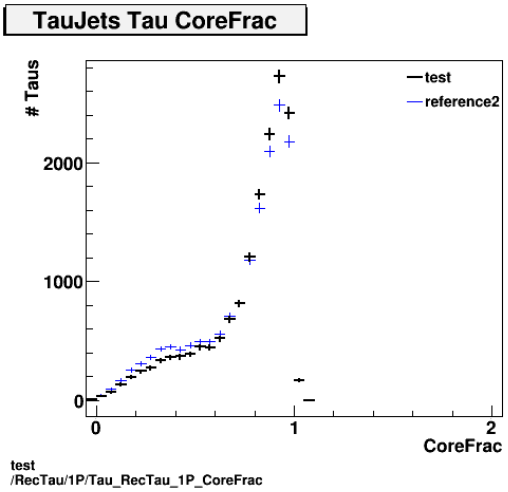
Figure 5.4: Tructh matched p_T histogram for $Z \rightarrow \tau\tau$

For the truth-matched taus, all plots were flagged as GREEN, except for the 1-prong p_T plot were there there was a slight shift in the rising edge of the transverse momentum distribution as seen in figure. 5.4. This difference was not significant and it was concluded, in collaboration with the Tau CP experts that the new test sample had an updated TES (Tau Energy Scale) calibration which would introduce the shift observed in the truth matched p_T distribution. Similarly, all the fake candidates showed excellent agreement and all plots were flagged as GREEN. However, the reconstruction variables, also known as "reco" variables had many histograms flagged as RED. The "reco" tau candidates are mostly comprised of real and fake candidates. It was seen from the truth-matched and fake distributions that there was good agreement between the test and reference sample. The following plots for the reconstructed tau candidates show the isolated calorimeter and core fraction distributions.

In figure. 5.5, the differences seen indicate that the normalisation between the fake and real components was not done properly. However, this discrepancy was indicative of a lower level problem in the new test sample possibly caused by the composition of the sample and the different TES calibrations.



(a) All (1 +3) prong isolated calorimeter fraction $Z \to \tau\tau$

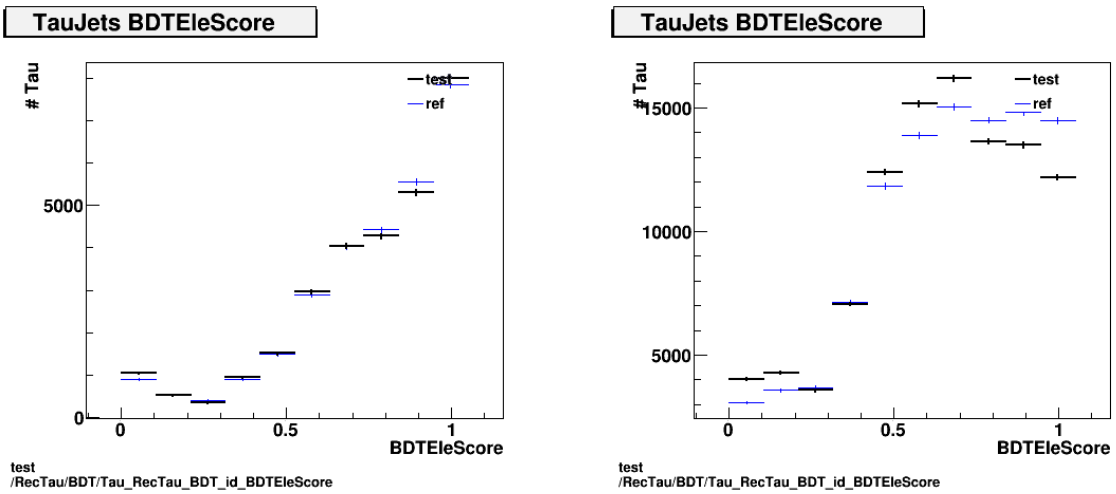


(b) 1-prong core fraction histogram $Z \to \tau\tau$

Figure 5.5: The core fraction distributions of the reconstructed 1 prong matched $Z \to \tau\tau$ sample and the all prong (1 +3) calorimeter isolated fraction

5.4.2 Task Two

In this task, test sample had MC15-like configuration without pile up. The reference sample used was produced using ATHENA 20.0.0.2 which also didn't have pile up. The histograms for this task were all flagged as GREEN except for two histograms including the BDT electron score. BDT scores are used in order to help identify whether a tau object is a true tau. BDT stands for Boosted Decision Trees (BDT). The BDT scores give an indication of how "jet" like or how "electron" like the reconstructed tau object is. These BDT scores are used as an electron veto. Since we don't observe taus in the detector, we reconstruct the taus from a 1-prong or 3-prong hadronic decay [30]. The 1-prong signature can often look like an electron in the detector, however there are ways to distinguish between the 1-prong hadronic tau decay and electron signatures. The cone within which a true tau jet is reconstructed is wider than an electron jet but narrower than a QCD multi-jet as indicated in figure. If we look at the $Z \to \tau\tau$ sample BDT electron score, we see that the distribution is skewed to the right. This is because objects which are more electron like have a score which tends towards 0, while true tau candidates will have distributions skewed to the right towards 1 which is what we would expect from the $Z \to \tau\tau$ sample. The BDT Jet score on the other hand, is not very good at discriminating between tau and electron jets so that is why the BDT electron score provides a useful electron veto for discriminating between tau and electrons.



(a) BDT Electron score for $Z \rightarrow \tau\tau$

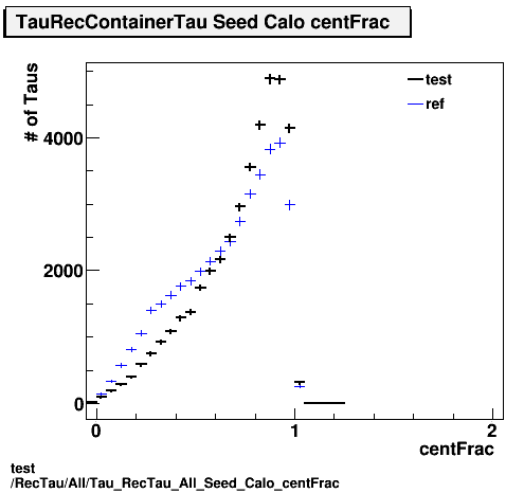
(b) BDT Electron score for $Drell - Yan \rightarrow \tau\tau$

Figure 5.6: BDT electron score distributions for Task Two flagged as RED

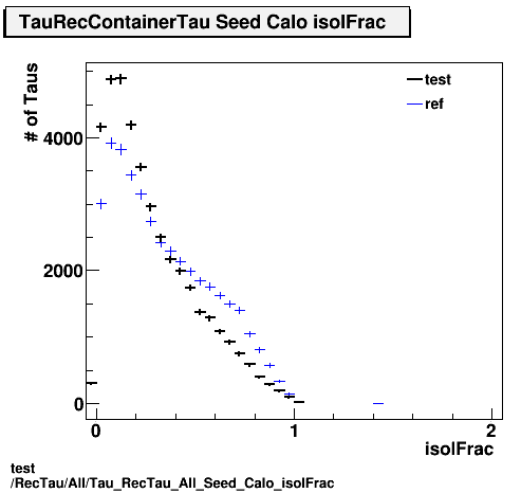
5.4.3 Task Three

Task Three required a comparison of the test sample, which was produced with 50 ns bunch spacing, against a previously validated sample produced using 25 ns bunch spacing. Differences were expected to be observed between the test and reference sample since the sample with 25 ns bunch spacing would have an increase in pile-up. This would affect the reconstructed distributions during the physics validation procedure. As expected, once the comparison had been performed, many variables had been flagged as RED. The pile-up variables were checked for both the test and reference sample.

Looking at figure. 5.7, the reference sample appears to have a lower number of reconstructed events (25 740 events) compared to the test sample which had 40 278 reconstructed events. In the reference sample (always shown in blue) there is a similar bump in the Core Fraction distribution that was observed in **Task One**. This bump indicated an increase in fakes. This means that fewer "real" taus were reconstructed in the reference sample and this is one hypothesis for the counter intuitive difference we observe between the test and the reference sample during this validation. A possible reason for the increase in fake taus could have been caused by masked pixel modules. It was also noted from another validation group, the Jet group, that they has an influx of low p_T jets which could be caused by the configuration in the test sample which used optimized OFC coefficients.



(a) All (1 +3) prong isolated calorimeter fraction $Z \rightarrow \tau\tau$



(b) All (1+3) prong core fraction histogram $Z \rightarrow \tau\tau$

Figure 5.7: The all prong (1 +3) calorimeter isolated fraction and core fraction distributions of the $Z \rightarrow \tau\tau$ sample for Task Three

Chapter 6

The measurement of W production in association with jets

"I never guess. It is a capital mistake to theorize before one has data. Insensibly one begins to twist facts to suit theories, instead of theories to suit facts."
— Sir Arthur Conan Doyle

Measurement of the production of a W boson in association with jets cross section and the QCD background

The production of a W boson in association with jets is an important signature which is used to test the theoretical predictions provided by perturbative QCD. The W boson decays into an electron/positron ($e^{\pm}$) or a muon/anti-muon ($\mu^{\pm}$) and the corresponding neutrino. Both the electron and muons channels are measured. However, this thesis will investigate the muon channel using the 2012 data set collected by the ATLAS detector at $\sqrt{s} = 8$ TeV corresponding to an integrated luminosity of approximately 18.9 fb^{-1}. This thesis focuses on the development of a data-driven QCD estimate using the template method. The optimization studies along with the description of the fit are given in this chapter. The QCD normalization is given in the final section accompanied by detector level distributions.

6.1 Event Selection

We are interested in events containing a W boson decaying into a muon associated with 1 or more high p_T jets. In Tab. 6.1, the main event selection criteria are listed. For the pre-selection, 1 good vertex is required in the event. To trigger on W bosons in the muon channel, the single muon triggers, EF_mu24i_tight and EF_mu36_tight, are employed. The isolated muon trigger EF_mu24i_tight was introduced to deal with the increased center-of-mass energy during the 2012 data taking period. This trigger has a very high efficiency and is very effective at selecting signal events by triggering on muons with a $p_T > 24$ GeV. In order to select isolated muons at trigger level, an isolation algorithm sums the p_T of the tracks in the inner detector of a muon candidate, requiring that the following condition is satisfied :

$$\frac{\Sigma pIDCone20}{p_{T\text{muon}}} < 1.2 \tag{6.1}$$

where $\Sigma pIDcone20$ is the sum of the transverse momentum of the tracks surrounding the candidate muon track in a cone of $\Delta R = \sqrt{\Delta\phi^2 + \Delta\eta^2} < 0.2$ and p_{Tmuon} is the total transverse momentum of the muon candidate. However, for the purpose of this analysis only muon candidates satisfying

$$\frac{\Sigma pIDCone20}{p_{T\text{muon}}} < 0.7 \tag{6.2}$$

passed the event selection.

Exactly one muon is selected with a $p_T > 25$ GeV. There are a number of quality requirements imposed on the muon. The quality requirements are number of hits in the pixel, SCT detectors and the TRT chambers which need to be satisfied in order for the muon to pass this stage of selection. The type of muon used in this selection is a medium plus muon. The selected muon lies in a pseudo-rapidity region above 0.1 until the end of the Muon Spectrometer acceptance which is at a pseudo-rapidity of 2.4. Muons with a transverse momentum of 25 GeV or higher are selected as this will increase the number of signal events. The impact parameter significance $d0Sig$ is a cut on the transverse impact parameter from

Table 6.1: List of Cuts applied to Data and MC for the $W \to \mu\nu$ event selection

	Pre-selection
vertex	>1 good vertex : N_tracks}
	Muon Selection
Triggers	EF_mu24i_tight
	EF_mu36_tight
Muon	medium+
Pseudorapidity	$0.1 < \eta < 1.9$
muon p_T	> 25 GeV
isolation	$\Sigma\frac{ptcone20}{p_T} < 0.07$
	$d0sig > 3mm$
$W \to \mu\nu$ event selection	
Z veto	no second selected muon
MET cleaning	MET cleaning criteria
	jets found with $p_T > 20$ and $\eta < 4.5$
Missing energy	$E_T^{miss} > 25GeV$
Transverse mass	$m_T > 40$ GeV
	Jet selection
Jet p_T	> 30 GeV
Pseudorapidity	$\|\eta\| < 4.4$

the Primary Vertex. The primary vertex is defined as the point on the Z axis where the interaction has been measured to have taken place. We use a cut on the $d0Sig < 3$ since this cut is useful in rejecting heavy flavour (and light quark) backgrounds in electroweak analyses. A tighter isolation cut is required on the signal muon as this will ensure we have a higher background rejection but will still maintain a very high signal efficiency. The cut on the transverse mass is chosen to select W events and the cut on the E_T^{miss} is also for background rejection since the low E_T^{miss} region is dominated by QCD events. Thus these cuts and selections are chosen to select events which have a single high p_T lepton, high p_T jets in a specified pseudo rapidity region as well as a large missing E_T to preserve our signal efficiency whilst also rejecting a large amount of background events, especially the QCD background. As detailed in Chapter 4, jets are reconstructed from topological clusters of calorimeter cells and are calibrated using the $anti - k_T$ algorithm. Jets used in this analysis

are required to have $p_T > 30GeV$, lie within psuedorapidity region $|\eta| < 4.4$ with a distance parameter R = 0.4 on the topological clusters of energy deposited in the calorimeters. The jets are calibrated to the jet energy scale (JES) which is derived from MC. These include corrections on the origin of the jet as well as pile up contributions. Jets originating from b-quarks are b-tagged and rejected from the event selection. Any jets overlapping with a lepton (electron or muon) within $\Delta R < 0.5$ are removed.

The isolation requirement on the signal muon was investigated by testing the suppression of the QCD background as well as the percentage of signal events which are lost when the isolation requirement is tightened. The chosen isolation requirement is $\Sigma\frac{p_T^{cone20}}{p_T} < 0.07$.

6.2 Modeling Backgrounds to W boson production in association with Jets

Dedicated Monte Carlo samples are used in order to estimate the signal and background whilst a data-driven approach is used to estimate the QCD background which will be the focus of this thesis. The same W event selection, as specified in table 6.1 , is applied to all the MC samples. The following processes are estimated using MC: $W \rightarrow \mu\nu$ in association with jets, diboson production, $Z \rightarrow \mu\mu$ plus jets, $W \rightarrow \tau\nu$ + jets, $Z \rightarrow \tau\tau$, single top production, Wbb and Wcc production as well as $t\bar{t}$ production.

6.2.1 MC background estimation

The background processes can be divided into three categories, the electroweak background originating from diboson events and Z bosons decaying to τ and muon leptons. These backgrounds contribute a small fraction to the overall background. The next category of background events originate from the decay of top quarks. This background is very dominant in the higher jet multiplicities and constitutes over 40 % with events containing more than 5 jets.

The number of background events is determined by applying a scaling factor to each MC sample which has been calculated using the theoretical cross section, the k-factor (which is the ratio of the NLO to LO cross section for a given process) and then multiplied by the integrated luminosity of the data. A complete list of all the MC samples used in the analysis

Table 6.2: List of MC samples

process	Generator	dataset	cross section (nb)	k-fac
$W \to \mu\nu$ Np0	AlpgenPythia	147033	8127.2	1.143
$W \to \mu\nu$ Np1	AlpgenPythia	147034	1792.8	1.143
$W \to \mu\nu$ Np2	AlpgenPythia	147035	542.42	1.143
$W \to \mu\nu$ Np3	AlpgenPythia	147036	147.68	1.143
$W \to \mu\nu$ Np4	AlpgenPythia	147037	37.76	1.143
$W \to \mu\nu$ Np5 incl	AlpgenPythia	147038	11.934	1.143
WW	Herwig	105985	12.41928212	1.68
ZZ	Herwig	105986	0.99293481	1.55
W/Z	Herwig	105987	3.66826914	1.9
$W \to \tau\nu$ Np0	AlpgenPythia	117700	8135.7	1.143
$W \to \tau\nu$ Np1	AlpgenPythia	117701	1793.7	1.143
$W \to \tau\nu$ Np2	AlpgenPythia	117702	541.24	1.143
$W \to \tau\nu$ Np3	AlpgenPythia	117703	146.48	1.143
$W \to \tau\nu$ Np4	AlpgenPythia	117704	37.264	1.143
$W \to \tau\nu$ Np5 incl	AlpgenPythia	117705	11.537	1.143
Single Top	MC@NLO Jimmy	108344	0.56426	1.074
$Z \to \mu\mu$ Np0	AlpgenPythia	147113	719.16	1.18
$Z \to \mu\mu$ Np1	AlpgenPythia	147114	175.74	1.18
$Z \to \mu\mu$ Np2	AlpgenPythia	147115	58.882	1.18
$Z \to \mu\mu$ Np3	AlpgenPythia	147116	15.673	1.18
$Z \to \mu\mu$ Np4	AlpgenPythia	147117	4.0057	1.18
$Z \to \mu\mu$ Np5 incl	AlpgenPythia	147118	1.2544	1.18
$Z \to \tau\tau$ Np0	AlpgenJimmy	107670	711.81	1.23
$Z \to \tau\tau$ Np1	AlpgenJimmy	107670	155.13	1.23
$Z \to \tau\tau$ Np2	AlpgenJimmy	107670	48.804	1.23
$Z \to \tau\tau$ Np3	AlpgenJimmy	107670	14.16	1.23
$Z \to \tau\tau$ Np4	AlpgenJimmy	107670	3.7744	1.23

along with their theoretical cross sections is listed in table 6.2 and table 6.3. However, not all the distributions are modelled by MC and the multi-jets background is not derived from MC due to the inherent difficulties in modelling this background. Although it is known MC

Table 6.3: List of MC samples continued

process	Generator	dataset	cross section (nb)	k-fac
Wbb Np0	AlpgenPythia	110801	52.255	1.143
Wbb Np1	AlpgenPythia	110802	45.54	1.143
Wbb Np2	AlpgenPythia	110803	23.671	1.143
Wbb Np3	AlpgenPythia	110804	12.525	1.143
Wcc Np0	AlpgenPythia	126606	143.07	1.143
Wcc Np1	AlpgenPythia	126607	143.68	1.143
Wcc Np2	AlpgenPythia	126608	80.762	1.143
Wcc Np3	AlpgenPythia	126609	35.932	1.143
Wc Np0	AlpgenPythia	126601	758.93	1.143
Wc Np1	AlpgenPythia	126602	274.24	1.143
Wc Np2	AlpgenPythia	126603	71.634	1.143
Wc Np3	AlpgenPythia	126604	16.425	1.143
Wc Np4	AlpgenPythia	126605	4.7468	1.143
ttbar	PowhegPythia	117050	137.379	1.0
ttbar tuned	PowhegPythia	110404	137.36129	1.0

simulation can model the interaction of particles with the ATLAS detector and the detector response, it is not known how often jets can be misidentified as leptons. The probability is small, which means that the MC samples would have to be extremely large to achieve an adequate amount of statistics with events that have faked leptons. This makes the existing MC samples insufficient to describe the QCD background and justifies the need for a data-driven QCD background estimate.

6.2.2 Data-driven measurement on the QCD multijet background

The focus of this thesis is to develop a data-driven estimate for the QCD background which dominates at lower jet multiplicities in W production is association with jets. In order to use a data-driven technique an understanding of the underlying topology of the signal events (W production in association with jets) and QCD events is needed. An understanding of the detector response to a signal event and to that of a QCD event will provide the main differences between these signatures and will enable us to develop a data-driven QCD template using appropriate discriminating variables. As detailed in the previous sections,

events with very large E_T^{miss} mostly originate from signal events. In the first order production of W boson, in the primary interaction between the incoming quark and anti-quark which annihilate to form the W boson. When it decays, the neutrino (which is not detected) will carry off a large part of the p_T resulting in a very large E_T^{miss} signature. However, in multi-jet events resulting from semileptonic heavy flavour decays (eg. $b\bar{b}$), the W bosons which decay have lower p_T and therefore multi-jets events will have lower missing transverse energy. So the E_T^{miss} is an important discriminating variable since the shapes of these distributions is very different. In the following sections, the selection and methodology of developing a data-driven QCD template is detailed.

6.2.3 Defining Control Regions for the QCD template

In order to build a data-driven template, one needs to carefully select a Control Region (CR) where the majority of events will originate from the multi-jet background and will not have a high contamination of signal events. The differences in topology between signal and QCD events indicate that the E_T^{miss} cut needs to be removed and the cut on the transverse mass m_T needs to be relaxed to select a QCD rich CR. Two important cuts which are used for the SR to reduce background contamination are is the isolation requirement on the muon as well as the cut on the impact parameter significance. Inverting the isolation requirement on the muon is the crucial step in defining the QCD CR. An initial trigger study was performed using the non-isolated muon triggers which would be the starting point of the analysis. The results of this initial study found that the most appropriate trigger for the QCD template estimate would be the EF_mu24_tight trigger. Muons passing this trigger would have the most similar kinematic properties to the QCD events which could pass the signal selection. This trigger triggers on muons with a p_T greater than 24 GeV, using the "tight" selection for muons but not having any restriction on the isolation on the relative track p_T. The isolation requirement on the muon is the most discriminating feature for selecting QCD events. It is important to chose the cuts in such a way as to minimise the contamination of signal events in the template, while having sufficient statistics. The CR is selected by rejecting events which pass the isolation requirement on the muon, in other words, the isolation requirement is inverted. Since the CR is not completely orthogonal and free of signal events, the same selection is run on the MC simulation (W+Jets, $t\bar{t}$, etc) and then this contamination is subtracted from the template. Theoretically, the shape of the QCD multi-jet background for the chosen discriminating variable should be the same in both regions. Therefore the shape of the QCD multi-jet background can be extrapolated to the signal region using events from

the control region. Several templates were investigated. Firstly, the anti-isolation range on the muon had to be decided so various templates with six different anti-isolation ranges were investigated. The templates are listed in Tab. 6.4 .

Table 6.4: Table of data-driven templates created by varying the anti-isolation range

Name of template	Anti-isolation min	Anti-siolation max
mu_QCDtempl007	$p_T^{cone20} = 0.07$	$p_T^{cone20} = 0.5$
mu_QCDtempl009	$p_T^{cone20} = 0.09$	$p_T^{cone20} = 0.5$
mu_QCDtempl05	$p_T^{cone20} = 0.1$	$p_T^{cone20} = 0.5$
mu_QCDtempl1	$p_T^{cone20} = 0.1$	$p_T^{cone20} = 1.0$
mu_QCDtempl100	$p_T^{cone20} = 0.1$	$p_T^{cone20} = 100.0$
mu_QCDtempl1000	$p_T^{cone20} = 0.1$	$p_T^{cone20} = 1000.0$

The effect of varying the cone size of the relative track-backed isolation ($p_T^{\text{cone20/30}}$) and the relative calorimeter-based ($E_T^{\text{cone20/30/40}}$) isolation was investigated using the templates listed in table. 6.5. For each of these templates, the $E_T^{cone(n)}min$ and $E_T^{cone(n)}max$ had fixed values of 0.1 and 0.5 respectively. Similarly, the minimum and maximum range for the p_T^{cone} was between 0.1 and 0.5 across all templates. In table. 6.4 , the template names are listed with their corresponding minimum and maximum anti-isolation ranges. The **mu_QCDtempl007** has the minimum (p_T^{cone20} set at 0.07, since this is the maximum isolation from the signal region and with a restircted isolation maximum set at 0.5. **mu_QCDtempl009** has the minimum (p_T^{cone20} set at 0.09 as well as a maximum of 0.5. The last four templates have a minimum p_T^{cone20} set at 0.1 and we expect these to have fewer statistics but also not as much contamination so it is important to test which template will give us adequate statistics but also not be contaminated with too many signal events. It is also important to note that while a template may be less contaminated with signal events, the final distributions for the muon p_T and leading jet p_T are also compared to make a final decision. The agreement between the data and the MC and data-driven QCD distributions is another motivation for selecting a particular template.

Table 6.5: QCD templates investigated using various combinations of relative track and calorimeter-based isolation

Name of template	$p_T^{cone(n)} size$	$E_T^{cone(n)} size$
muon_QCDtempl20	20	20
muon_QCDtempl30	30	20
muon_QCDtempl40	40	20
muon_QCDtempl30Et	20	30
muon_QCDtempl40Et	20	40
muon_QCDtempl30EtPt	30	30
muon_QCDtempl40EtPt	40	40

6.3 The data driven QCD estimate

In order to obtain the normalization for the data-driven QCD template, the data-driven QCD templates, along with the other MC samples, are fit to the data using the E_T^{miss} distribution. In order to do this, the signal and background MC samples are summed together, which will form the electroweak (EWK) template. Before the two templates are fit to the data, it is important to note here that the contamination in the data-driven QCD template is subtracted from the template prior to the fit. In order to extract the constituent fractions of each template, a kinematic variable that behaves differently in each of the templates is used to do the fit. In the following figure, the discriminating variable is chosen to do the fit.

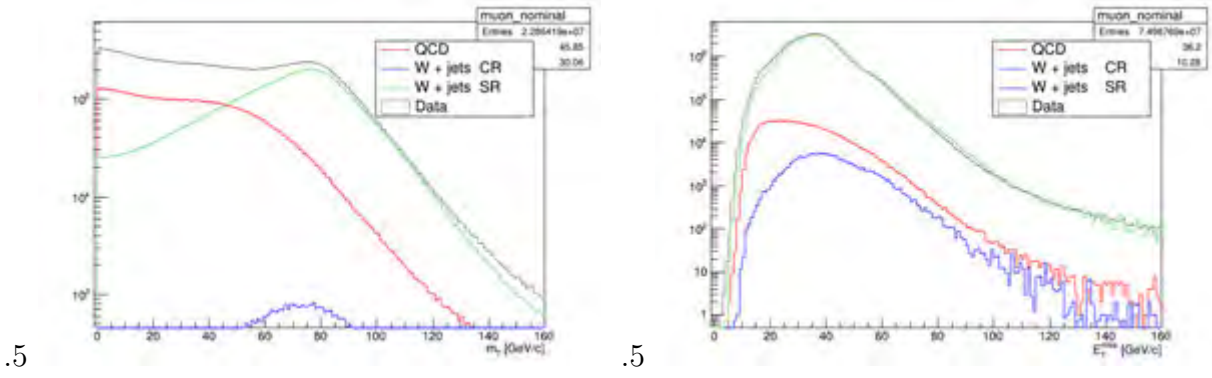


Figure 6.1: Distributions of the E_T^{miss} and m_T which were compared in order to choose the best discriminating variable

In figure 6.1, we see that there shape of the signal W+jets template is similar in the SR and

CR, while the shape of the QCD template in the CR is significantly different to the shape of the electroweak template. Two tools were used to estimate the contribution of each template, RooFit [31] and TFractionFitter [32] which are classes within the data analysis framework ROOT. In order to estimate the constituent fractions of the QCD and electroweak template, an extended maximum likelihood fit is performed using binned templates. Firstly, for each exclusive jet multiplicity, the integrals of the original histograms obtained from the nominal (signal event selection) and QCD template selections in a specified fit range are calculated. This will give us the number of (unscaled) events to which the scaling factors obtained from the fit will be applied. This number of events is the starting value for the fit. The integral (or total) number of QCD events is also calculated to obtain the statistics of each template. This integral is calculated before and after the $E_T^{miss} > 25GeV$ cut is applied to the QCD template.

The QCD template and EWK template are fit to data before the $E_T^{miss} > 25$ GeV cut is applied, however, the the cut on the missing transverse energy is applied again after the fit to get the fraction of QCD events in the SR. All other cuts used for the SR are used in order to preserve as many QCD events which have similar kinematics and could therefore pass the selection requirements for the signal region.

$$N_{SR}^{QCD} = \sum_{i}^{n} (N_{CR,i}^{Data} - N_{CR,i}^{EWKMC}) \times R_i^{fit} \tag{6.3}$$

where the scaling factor SF is calculated as follows

$$QCDSF = \frac{N_{fit}^{QCD}}{N_{CR,fitrange}^{QCD}} \tag{6.4}$$

and similarly, the scaling factor for the EWK template is obtained from

$$EWKSF = \frac{N_{fit}^{EWK}}{N_{SR,fitrange}^{EWK}} \tag{6.5}$$

The contamination in the QCD template is calculated according to

$$Contamination = \frac{N_{CR}^{Data} - N_{CR}^{EWK}}{N_{CR}^{Data}} \tag{6.6}$$

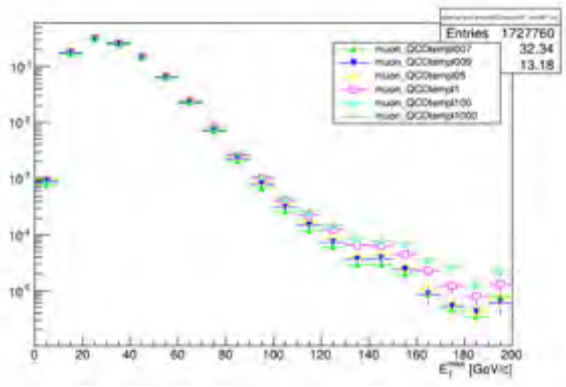
6.4 Determining which template to use

The following section shows the results of the RooFit fit results applied to all the templates. The as seen in figure. 6.2, the variation in the E_T^{miss} distribution when using different anti-isolation ranges is significant for the 0 jet bin. Here there is a significant variation in the tails of the E_T^{miss} distribution, but this is expected since this bin also has the highest contamination of signal events in the QCD CR. However, these variations lie outside of the chosen fitting range (15 GeV and 75 GeV).

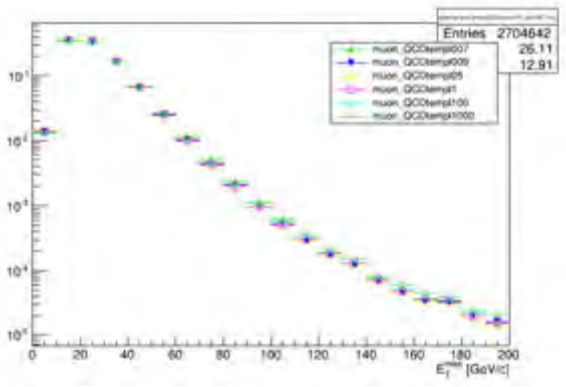
Certain features are desirable in our chosen template and we can use the following criteria to assess which template will be suitable for our data-driven QCD estimate.

- The EWK Scaling Factor. If this factor is close to 1, it ensures that the normalisation performed on the MC samples was done properly.
- The QCD Scaling Factor. The smaller this factor is, the higher the statistics in the template.
- Contamination signal events in the QCD data-driven template.

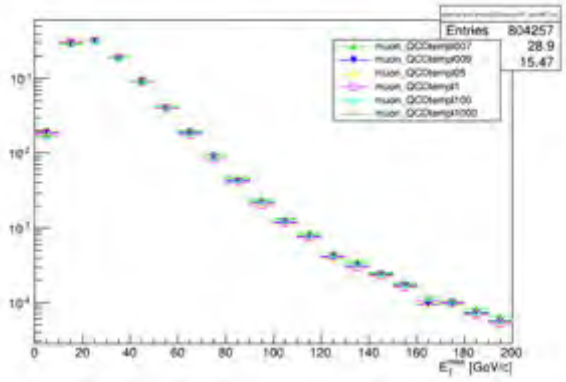
The plots in figure. 6.3 depict the percentage of QCD events in the SR for each template as well as the associated QCD and EWK scaling factors, and contamination. We favour a template with a low contamination contamination. In figure. 6.3 (b), it is observed that the **mu_QCDtempl007** (shown in neon green) has the highest contamination of signal events in the QCD CR while **mu_QCDtempl1000** (shown in yellow) has the lowest contamination. The remaining four templates look have contamination percentages lying between these two templates as expected. **mu_QCDtempl1000** which also has the smallest scaling factor indicating that it has high statistics which is favourable when selecting a data-driven QCD template. However, the anti-isolation on this template means that the kinematics of the muons passing this CR selection will be less similar to the kinematics of the muons passing SR selection. It is expected that **mu_QCDtempl007** will provide a better agreement with the signal region and this was observed especially for the inclusive leading jet p_T distribution as seen in ratio plots shown in figure 6.4.



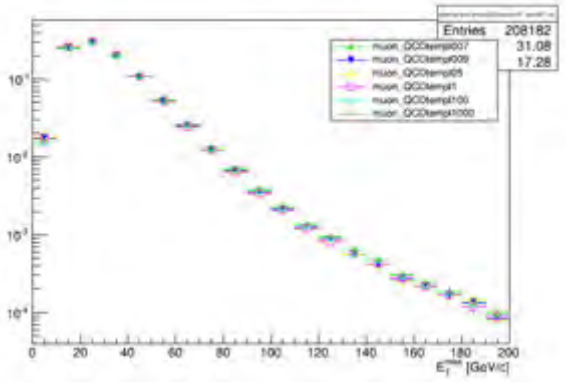
(a) $N_{jets} = 0$



(b) $N_{jets} = 1$



(c) $N_{jets} = 2$



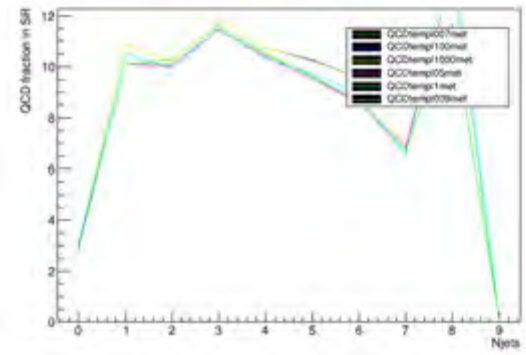
(d) $N_{jets} = 3$

Figure 6.2: Comparison of the E_T^{miss} shape in the different QCD templates when using different anti-isolation ranges

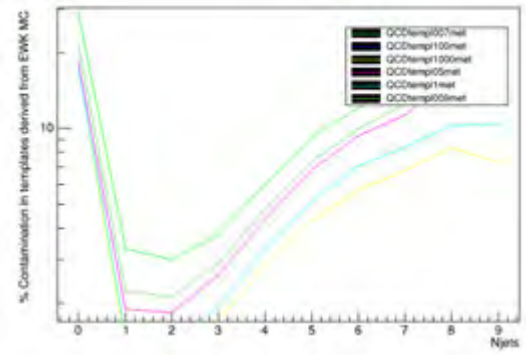
6.5 Data Driven results

From the investigation and optimization studies, the chosen QCD CR region is defined in table. 6.6.

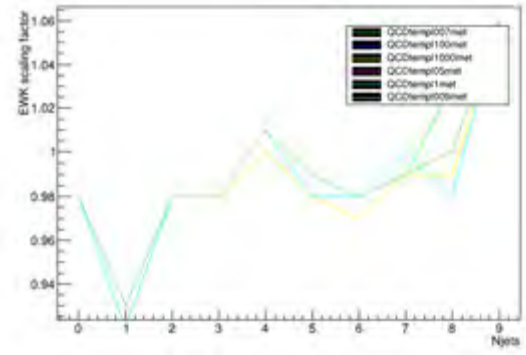
This control region uses a prescaled trigger which corresponds to an integrated luminosity of 2.48 fb^{-1}. The results for the QCD fit for the chosen CR in the fit range 15 to 75 GeV are shown in table 6.7 for the W selection. The table shows that the amount of contamination from the signal and the electroweak backgrounds is highest in the zero jet bin but it fractional in the 1 jet to 5 jets bins. This behaviour is not well understood and would need additional



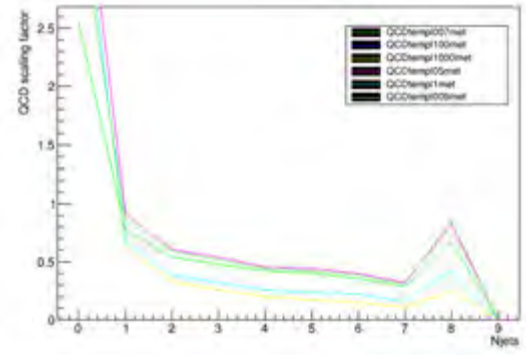
(a) % QCD in SR



(b) % Contamination in templates derived from MC



(c) % EWK Scaling Factor



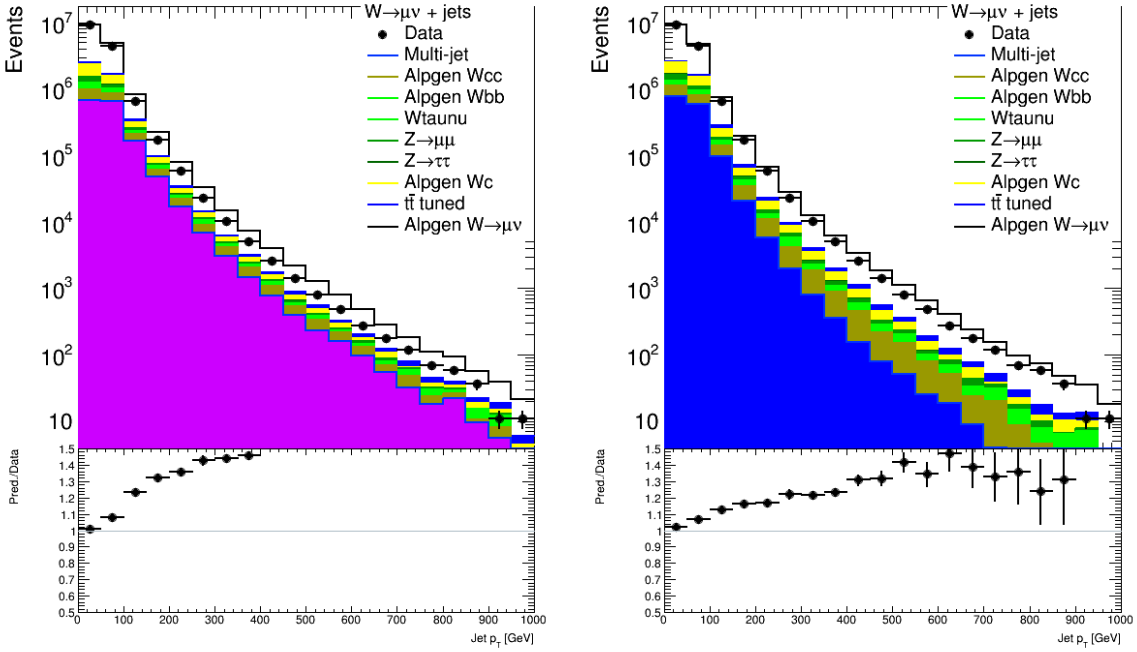
(d) % QCD Scaling Factor

Figure 6.3: Figures of merit used for selecting QCD template in muon W channel

Table 6.6: QCD-CR requirements

Triggers	EF_mu24_tight
Muon	medium+
anti-isolation	$\frac{p_T^{cone20}}{p_T} > 0.1$

studies on the MC based simulations. The scale factors for the normalization of the EWK and QCD templates are also calculated by the fit shown in equation 6.5 and equation 6.4 respectively. The scale factors for the EWK and signal template are approximately 1 which



(a) N_{jets} (b) N_{jets}

Figure 6.4: The detector level plots comparing the leading jet p_T using (a) **mu_QCDtempl1000** and (b) **mu_QCDtempl007**

indicated that the MC samples where well normalised to the integrated luminosity 18.9 fb^{-1} of data.

The results of the QCD fit for the fit range 15 - 75 GeV with relative anti-isolation $\frac{p_T^{cone20}}{p_T} > 0.07$ are shown in figure 6.5 for exclusive jet multiplicities 0 - 3 jets. In the plots in figure 6.5, the data-driven QCD background estimate is shown in blue, while the signal and electroweak backgrounds are shown in yellow. The fit range is indicated along with the percentage of QCD events in the SR.

6.6 Detector Level Results

One the signal region event selection criteria have been applied to the data and MC, and a data-driven QCD template has been obtained, the number of events measured in data

Table 6.7: Exclusive jet bin results of the QCD fit in the fit range 15 GeV to 75 GeV. The colums shows the fraction of QCD events in the fit range $f_{QCD,SR}$, the contamination in the QCD control region, the scale factors for the EWK and QCD templates c_{QCD-CR}, the number of QCD events in the control region after subtraction

n_{jet}	$f_{QCD,FR}$[%]	c^{CR}_{QCD}[%]	$SF_{Sig+EWK}$	SF_{QCD}	n^{CR}_{QCD}	data	$n^{events}_{Sig+EWK,SR}$
0jets	2.9 ±0.01	28.24	0.98	2.51	1866137	64260830	62433168 ±20609
1jets	10.58 ±0.02	3.2	0.93	0.74	900632	8513358	7613552 ±5114
2jets	10.7 ±0.05	2.91	0.99	0.51	241272	2255210	1981571 ±2698
3jets	10.74 ±0.11	3.69	1.0	0.45	64339	598963	522805 ±1645
4jets	7.76 ±0.2	5.83	1.0	0.39	14936	192362	171978 ±361
5jets	5.83 ±0.37	8.95	0.98	0.37	3553	60991	55139 ±165
6jets	4.13 ± 0.42	11.72	0.98	0.32	738	17878	16693 ±75
7jets	2.17 ± 1.42	14.45	1.02	0.19	102	4728	4743 ±35
8jets	1.33 ± 2.4	22.3	1.07	0.17	15	1182	1195 ±16
9jets	0.0 ± 5.58	17.73	1.04	0.0	0.0	257.0	264 ±6
10jets	0.0 ± 1.6	16.93	1.14	0.0	0.0	60.0	62 ±3

and the number of events estimated from the Electroweak MC background and data-driven QCD background can be determined. The jet multiplicity distribution, exclusive in the number of jets, is shown in figure 6.6. It is clear from the table 6.7 and figure 6.6 with the exclusive jet distribution that multi-jet production is the dominant background up to 4 jets jet multiplicities and then the $t\bar{t}$ background begins to dominate at higher jet multiplicities. There is an unusual feature in the 1 jet bin, the leading order jet production bin, where the MC over estimates the data by almost 10 %. This bin is particularly interesting for PDF constraints and this effect would need to be investigated further. The likely cause of this difference could be due to the increased pile-up conditions which were present during the 2012 data-taking period.

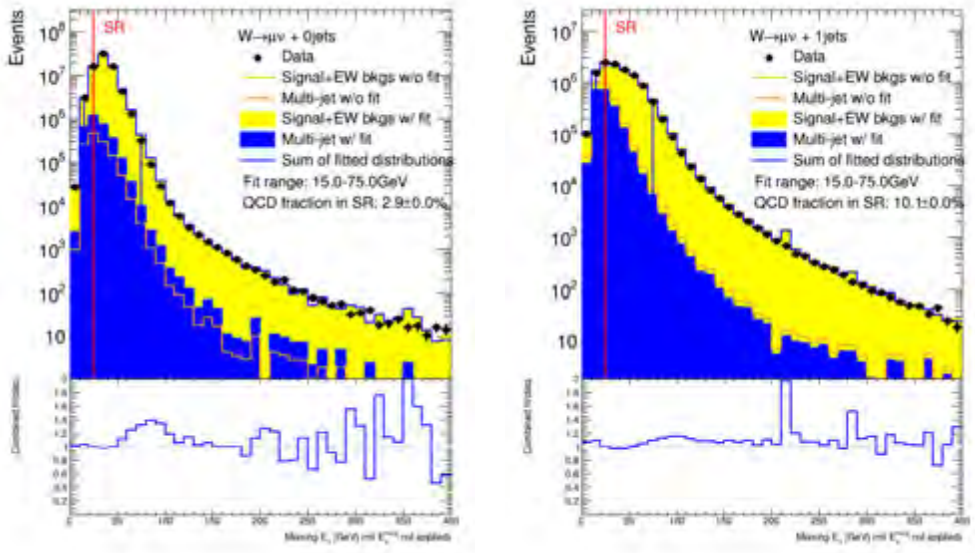
The following section shows several kinematic distributions depicting the agreement between data and the total SM prediction at detector level. The measured detector level distributions are distorted and smeared by the limited resolution and acceptance of the ATLAS detector. In order to uncover the underlying true distributions (particle level distributions), the measured results would need to be unfolded. There are many systematic uncertainties which are introduced by the detector but the actual estimation of them is beyond the scope of this thesis. The reconstruction of the jets, muon and the missing transverse energy E_T^{miss} involve

calibrations which introduce specific uncertainties. A few of the systematic uncertainties encountered are listed below.

- Jet energy scale uncertainties
- Jet energy resolution uncertainties
- Jet vertex fraction
- E_T^{miss}
- Muon scale factor uncertainties in trigger, reconstruction and isolation
- Data-driven QCD background systematics (fit range, choice of anti-isolation)

Figures 6.7 show the muon and W related kinematic distributions for exactly 1 jet produced in association with the W boson. Figure 6.8 shows inclusive plots of the same kinematic distributions for *at least* 1 jet. Figure 6.9 and figure 6.10 show the m_T and E_T^{miss} distributions as well as the jet-related kinematic distributions with exactly 1 jet produced. It is seen that over all, there is a fair agreement between data and MC and the data-driven QCD background estimate. The Alpgen generator used to produce the MC seems to describe the data quite well but there is some over estimation of events in tails of jet and muon p_T distribution. However, the slope of the muon p_T is described and modelled well by Alpgen. The ratio plots indicate the 10 % over estimation by the MC for the 1 jet bin. These plots would need to be unfolded to particle level to remove the systematic effects introduced by the detector. The distributions at particle level would then be used compare the cross section for the production of the W boson to the theoretical predictions provided by pQCD. This however is beyond the scope of this thesis.

It is shown that the data-driven QCD background estimate improves the agreement between data and the MC predictions. This preliminary analysis has helped inform how the rest of the cross section measurement can be improved and has provided a methodology for estimating the QCD background using a data-driven approach.



(a) $Njets = 0$ (b) $Njets = 1$

(c) $Njets = 2$ (d) $Njets = 3$

Figure 6.5: The results of the RooFit template fits which determined the percentage of QCD in the SR. The resuls for N_{jet} = 0 - 3 are shown in this figure. In the plot, the fitting range is indicated and was chosen to be between 15.0 and 75.0 GeV.

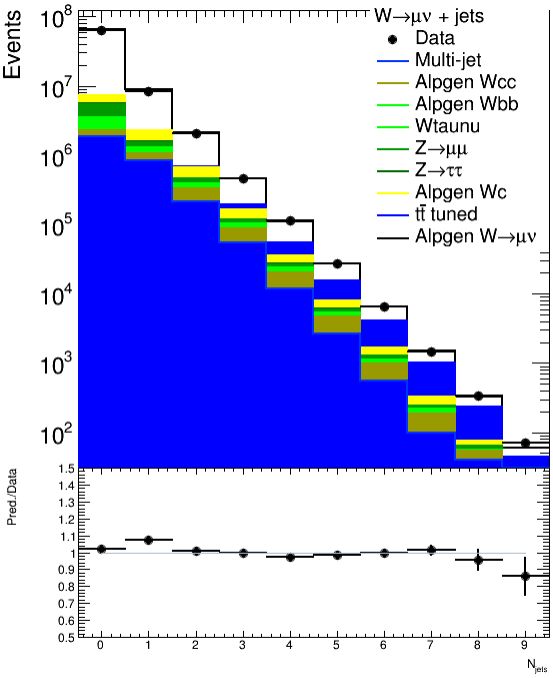
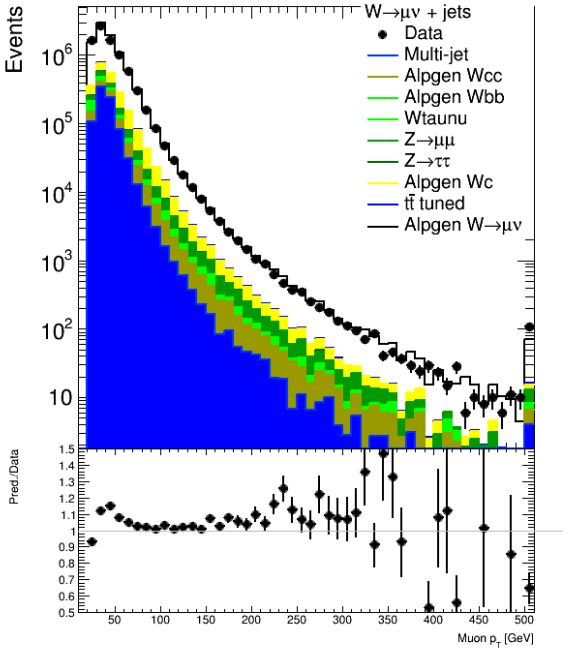
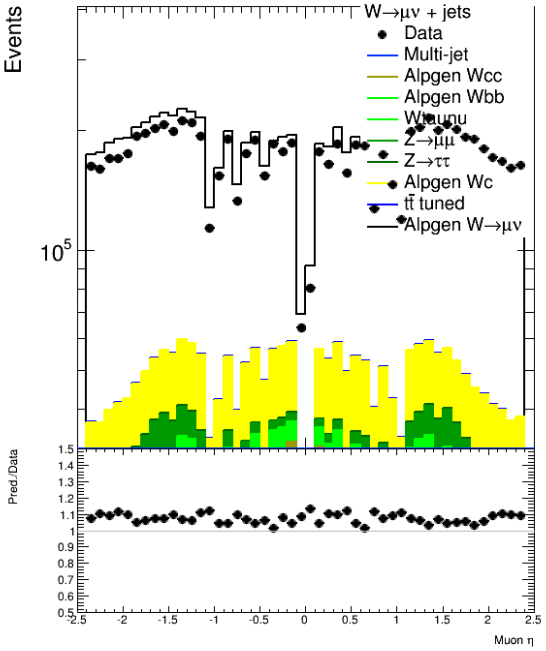


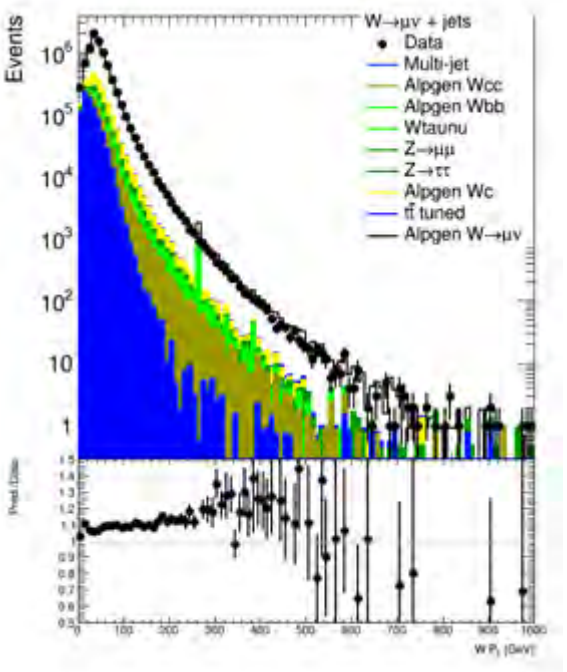
Figure 6.6: The exclusive jet multiplicity distribution in the W + jets signal region



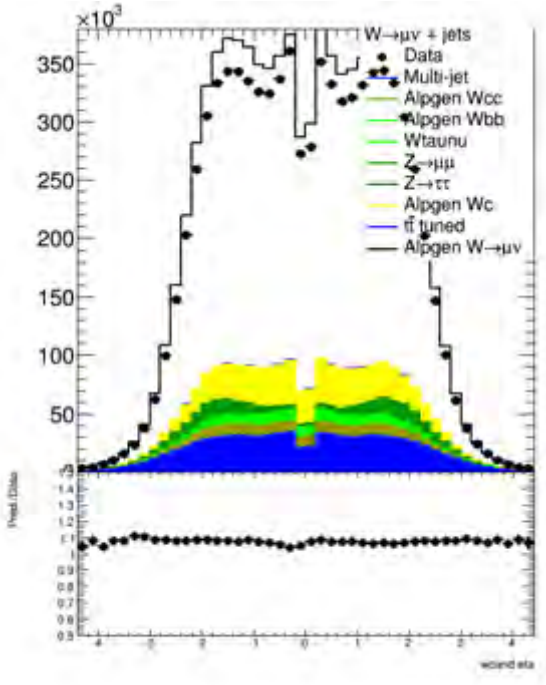
(a) muon p_T



(b) muon η

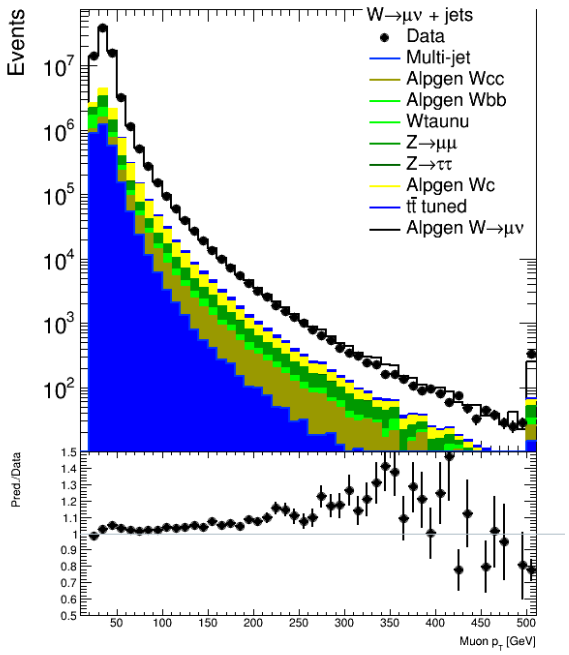


(c) Wp_T

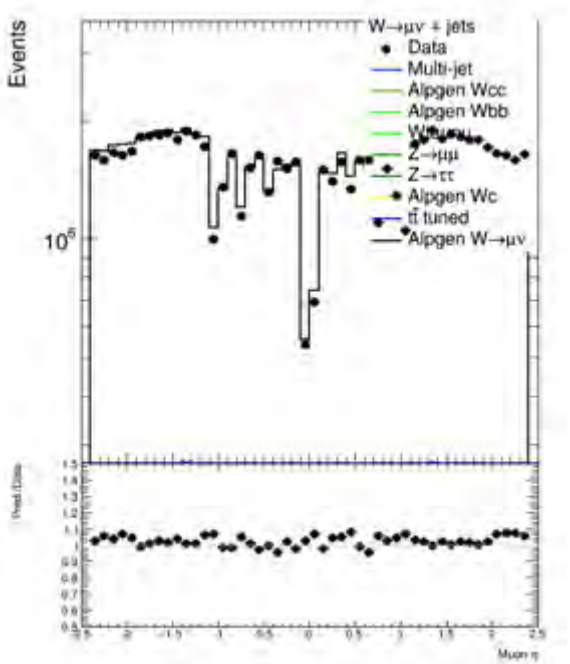


(d) $W\eta$

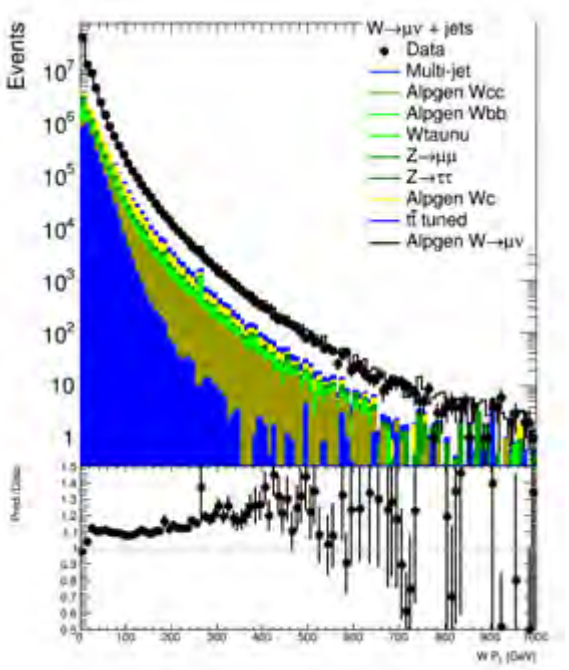
Figure 6.7: The detector level plots for exactly 1 jet (a)



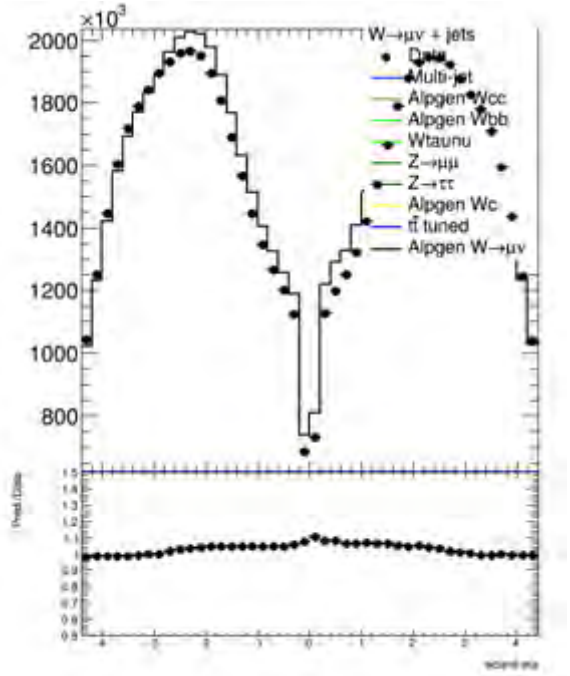
(a) muon p_T



(b) muon η

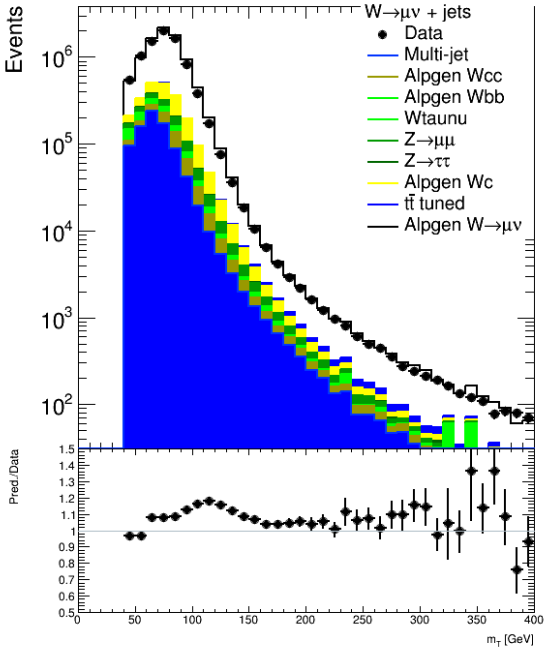


(c) Wp_T

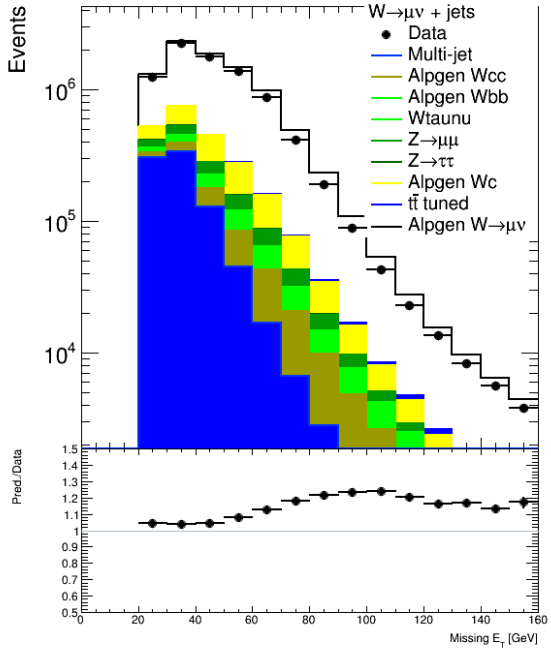


(d) $W\eta$

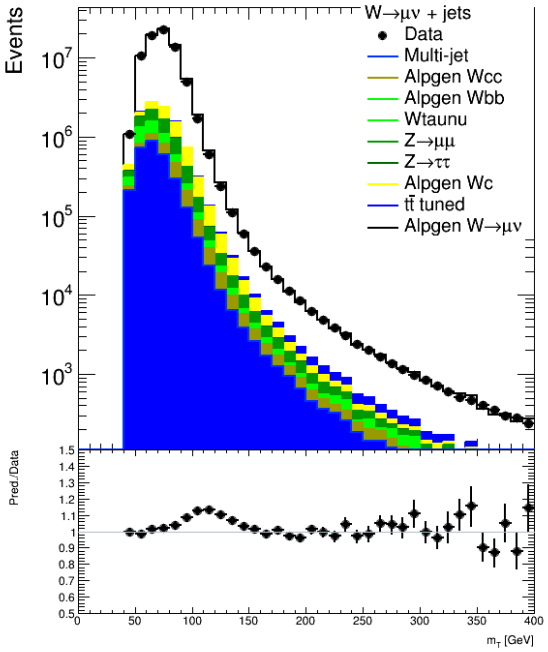
Figure 6.8: The inclusive detector level plots with at least 1 jet



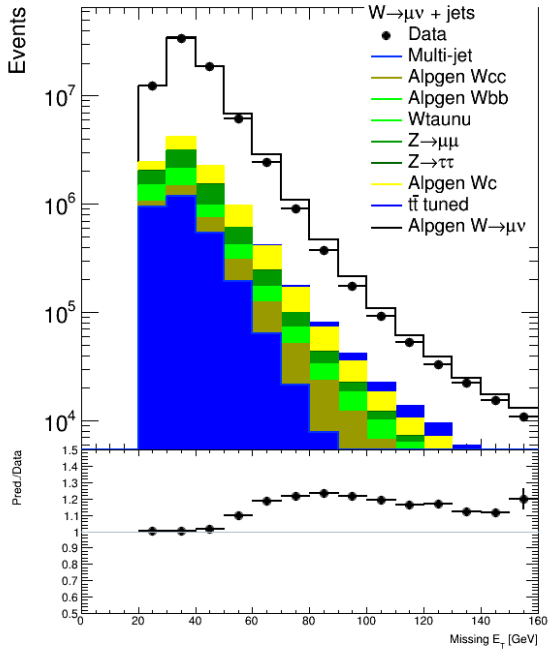
(a) m_T



(b) E_T^{miss}

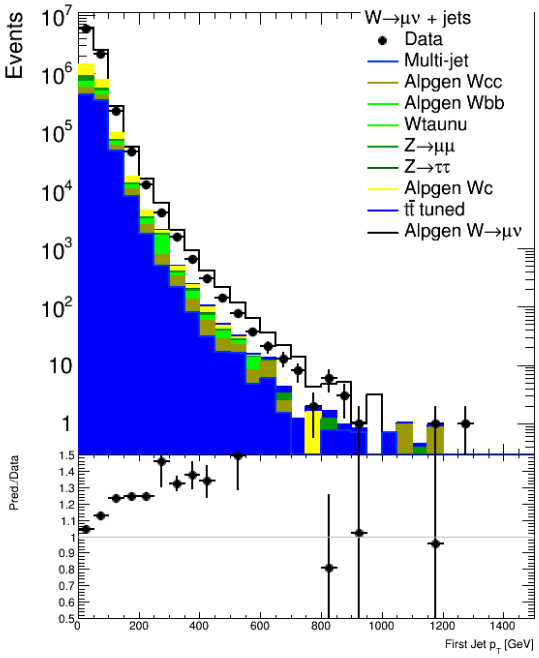


(c) m_T

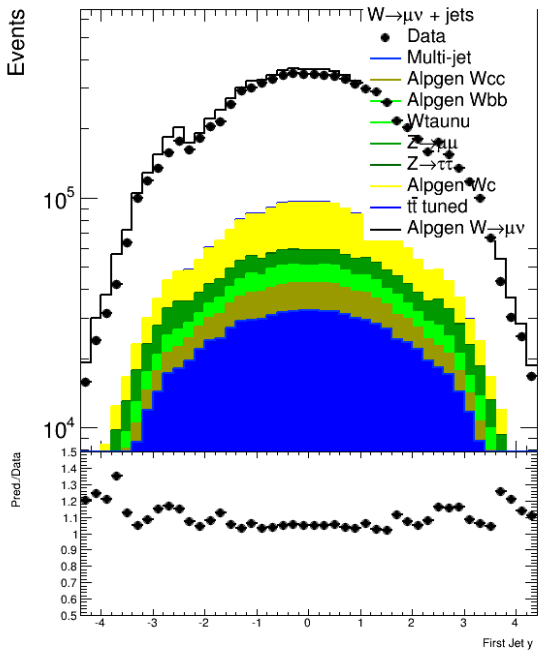


(d) N_{jets}

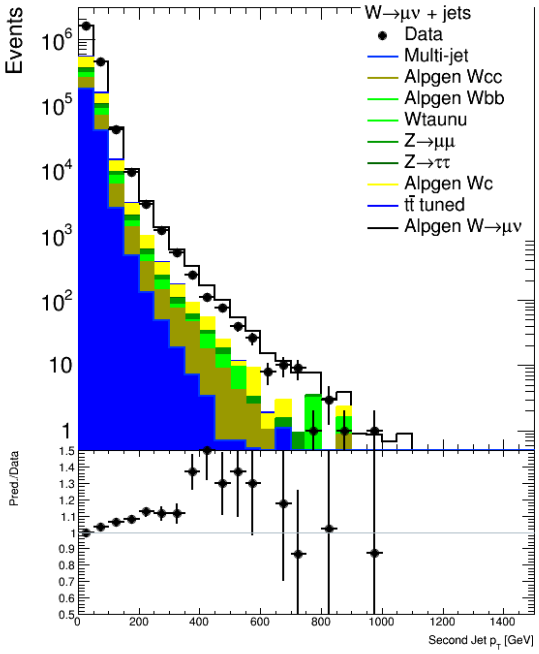
Figure 6.9: Comparison of the detector level plots with exactly 1 jet and inclusive detector level distributions for m_T and E_T^{miss}



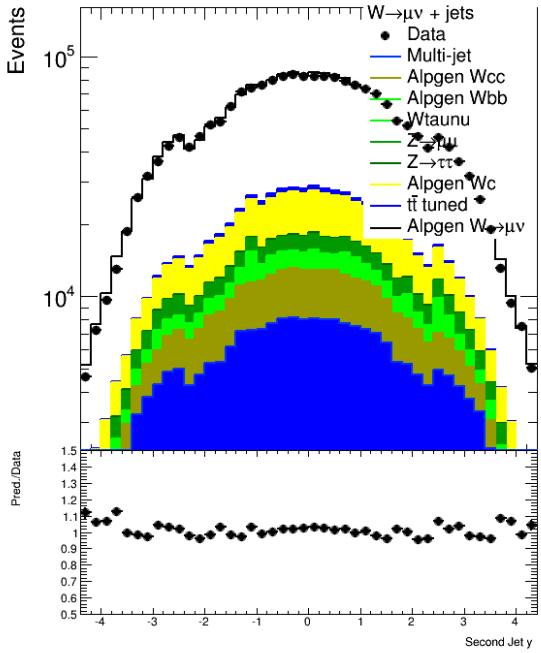
(a) N_{jets}



(b) N_{jets}



(c) N_{jets}



(d) N_{jets}

Figure 6.10: The detector level plots with for the jet related kinematic distributions. Leading jet p_T (a) and rapidity y (b) and second leading jet p_T (c) and rapidity y (d).

Chapter 7

Summary and Outlook

The QCD background, also referred to as multi-jet events in this thesis, is a significant source of background to W production in association with jets analyses. The QCD background arises from multijet events and are due to a misidentified muon which could pass the signal muon selection. These muons mainly come from heavy flavour decays (mainly c and b quarks decaying semileptonically) or pions and kaons which decay within the tracking region of the detector. Being able to accurately measure and describe this Standard Model process allows us to constrain this background in other more sensitive searches where the possible signal (SUSY , dark matter) will be overwhelmed by this process.

This preliminary analysis using data from the proton-proton collisions at $\sqrt{s} = 8$ TeV recorded by the ATLAS experiment, which corresponds to an integrated luminosity of approximately 18.9 fb^{-1} has been performed to extract the fraction of QCD events. This thesis has shown that the data-driven QCD background is well estimated when inverting the isolation requirement on the muon. The fraction of QCD events in the signal region was estimated in exclusive jet multiplicities by performing an extended likelihood fit using RooFit on the E_T^{miss} distribution in data. The overall agreement between the data and MC simulation is vastly improved by including the data-driven QCD estimate.

However, by using a data-driven $t\bar{t}$ estimate where the MC tends to overestimate the amount of $t\bar{t}$ in the higher jet multiplicities since there are high uncertainties in the simulated cross-sections. It was also found that the MC estimate for the 1 jet bin was over estimated by 10 % and this could be due to the increased pile-up conditions which have not been adequately modeled by the MC. In an attempt to improve the techniques for estimating the QCD multi-jets background to searches involving the production of a W boson in association with jets,

several data-driven QCD samples were created to . Different techniques and ideas (varying isolation on the signal muon, using different triggers for the data-driven QCD estimate) for improving the data-driven QCD background were investigated and provided useful input to the WJets groups as ideas for improving the analysis. More studies could be done to investigate the effect of the calorimeter anti-isolation parameters in combination with the track based isolation parameters.

The fraction of QCD events in the signal region is reported and found to be $2.9 \pm 0.01\%$ for the zero jet bin and increases to $10.74 \pm 0.11\%$ for the two jet bin and then decreasing to 7.77 ± 0.2, 5.83 ± 0.37 for higher jet multiplicities.

This analysis would also be vastly improved by obtaining the systematic uncertainties on the data-driven QCD background. The only uncertainties taken into account are the statistical uncertainties obtained from the fit. The effect of the fitting range could be investigated and these systematic uncertainties would help us constrain the data-driven background more accurately.

Bibliography

[1] **ATLAS** Collaboration, G. Aad *et. al.*, *The ATLAS Experiment at the CERN Large Hadron Collider*, *JINST* **3** (2008) S08003.

[2] B. Blok, Y. Dokshitser, L. Frankfurt, and M. Strikman, *pQCD physics of multiparton interactions*, *Eur.Phys.J.* **C72** (2012) 1963, [arXiv:1106.5533].

[3] A. Messina, *Measurement of the W + jet cross section at CDF*, *Brazilian Journal of Physics* **37** (06, 2007) 840 – 842.

[4] **D0** Collaboration, V. M. Abazov *et. al.*, *Studies of W boson plus jets production in $p\bar{p}$ collisions at $\sqrt{s} = 1.96$ TeV*, *Phys.Rev.* **D88** (2013), no. 9 092001, [arXiv:1302.6508].

[5] **CMS** Collaboration, V. Khachatryan *et. al.*, *Differential cross section measurements for the production of a W boson in association with jets in proton?proton collisions at $\sqrt{s} = 7$ TeV*, *Phys.Lett.* **B741** (2015) 12–37, [arXiv:1406.7533].

[6] **ATLAS** Collaboration, G. Aad *et. al.*, *Measurements of the W production cross sections in association with jets with the ATLAS detector*, *Eur.Phys.J.* **C75** (2015), no. 2 82, [arXiv:1409.8639].

[7] **CMS** Collaboration, S. Chatrchyan *et. al.*, *The CMS experiment at the CERN LHC*, *JINST* **3** (2008) S08004.

[8] M. E. Peskin and D. V. Schroeder, *An introduction to quantum field theory*. Advanced book program. Westview Press Reading (Mass.), Boulder (Colo.), 1995. Autre tirage : 1997.

[9] S. Glashow, *Partial Symmetries of Weak Interactions*, *Nucl.Phys.* **22** (1961) 579–588.

[10] **ALICE** Collaboration, K. Aamodt *et. al.*, *The ALICE experiment at the CERN LHC*, *JINST* **3** (2008) S08002.

[11] **LHCb** Collaboration, A. A. Alves, Jr. *et. al.*, *The LHCb Detector at the LHC*, *JINST* **3** (2008) S08005.

[12] **ATLAS Collaboration** Collaboration, G. Aad *et. al.*, *Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC*, *Phys.Lett.* **B716** (2012) 1–29, [arXiv:1207.7214].

[13] **CMS Collaboration** Collaboration, S. Chatrchyan *et. al.*, *Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC*, *Phys.Lett.* **B716** (2012) 30–61, [arXiv:1207.7235].

[14] P. F. esson, T. Atkinson, M. J. Costa, M. Elsing, S. Fleischmann, A. N. Gaponenko, W. Liebig, E. Moyse, A. Salzburger, and M. Siebel, *ATLAS Tracking Event Data Model*, .

[15] E. Moyse, "The TrkParameters package, howpublished = "http://atlas-computing.web.cern.ch/atlas-computing/links/buildDirectory/AtlasOffline/14.0.0/InstallArea/doc/TrkParameters/html/index.html/", year = 2015, note = "[online accessed 17-november-2015]"."

[16] **ATLAS** Collaboration, T. A. collaboration, *Tagging and suppression of pileup jets*, .

[17] **GEANT4** Collaboration, S. Agostinelli *et. al.*, *GEANT4: A Simulation toolkit*, *Nucl. Instrum. Meth.* **A506** (2003) 250–303.

[18] **ATLAS** Collaboration, G. Aad *et. al.*, *Performance of the ATLAS muon trigger in pp collisions at* $\sqrt{s} = 8$ *TeV*, *Eur.Phys.J.* **C75** (2015), no. 3 120, [arXiv:1408.3179].

[19] **ATLAS** Collaboration, *ATLAS muon spectrometer: Technical design report*, .

[20] M. Corradi, "ATLAS Muon Combined Performance Guidelines." https://twiki.cern.ch/twiki/bin/view/AtlasProtected/MCPAnalysisGuidelinesData2012Muon/, 2015. [Online; accessed 27-May-2015].

[21] W. Lampl, S. Laplace, D. Lelas, P. Loch, H. Ma, *et. al.*, *Calorimeter clustering algorithms: Description and performance*, .

[22] M. Cacciari, G. P. Salam, and G. Soyez, *The Anti-k(t) jet clustering algorithm*, *JHEP* **04** (2008) 063, [arXiv:0802.1189].

[23] **ATLAS** Collaboration, G. Aad *et. al.*, *Performance of Missing Transverse Momentum Reconstruction in Proton-Proton Collisions at 7 TeV with ATLAS*, *Eur. Phys. J.* **C72** (2012) 1844, [arXiv:1108.5602].

[24] **ATLAS Collaboration** Collaboration, G. Duckeck *et. al.*, *ATLAS computing: Technical design report*, .

[25] G. Barrand, I. Belyaev, P. Binko, M. Cattaneo, R. Chytracek, *et. al.*, *GAUDI - A software architecture and framework for building HEP data processing applications*, *Comput.Phys.Commun.* **140** (2001) 45–55.

[26] "https://subversion.apache.org/." SVN Version Control.

[27] "https://www.python.org/." The official home of the Python Programming Language.

[28] **Particle Data Group** Collaboration, K. Olive *et. al.*, *Review of Particle Physics*, *Chin.Phys.* **C38** (2014) 090001.

[29] Y. Sakurai, *The ATLAS Tau Trigger Performance during LHC Run 1 and Prospects for Run 2*, arXiv:1409.2699.

[30] Z. Zinonos, *Reconstruction and identification of hadronic decays of tau leptons in ATLAS*, arXiv:1409.0343.

[31] W. Verkerke and D. Kirkby, *The RooFit toolkit for data modeling*, *ArXiv Physics e-prints* (June, 2003) [physics/0306116].

[32] R. Barlow, "Tfraction fitter, a root class." https://root.cern.ch/doc/master/classTFractionFitter.html. [Online; accessed 18-February-2015].