School of Physics and Astronomy

University of Glasgow

Glasgow

G12 8QQ

PhD Thesis

$H \to b\bar{b}$ decays in associated VH production using data from the ATLAS detector

Johnny Friend

October 17, 2024

Submitted in fulfilment of the requirements for the

Degree of Doctor of Philosophy

Abstract

This thesis covers my contribution to the $VH(H \to b\bar{b})$ analysis conducted using $139\,\text{fb}^{-1}$ of data collected by the ATLAS experiment at the LHC. I primarily focused on the boosted 2-lepton MVA analysis, training on new variables and optimizing hyperparameters and contributing to truth-flavour tagging studies. A set of analysis improvements, including those described in this thesis, improved the observed (expected) significance from 2.6 (3.4) to 4.1 (4.5) standard deviations in the 2-lepton region. My development work on the online trigger tool (TTWeb) is also outlined, including my contribution to preparing TTWeb for the start of Run 3 of the LHC.

Declaration

I certify that the thesis presented here for examination for a PhD degree of the University of Glasgow is solely my own work other than where I have clearly indicated that it is the work of others (in which case the extent of any work carried out jointly by me and any other person is clearly identified in it) and that the thesis has not been edited by a third party beyond what is permitted by the University's PGR Code of Practice.

I declare that the thesis does not include work forming part of a thesis presented successfully for another degree.

I declare that this thesis has been produced in accordance with the University of Glasgow's Code of Good Practice in Research.

I acknowledge that if any issues are raised regarding good research practice based on review of the thesis, the examination may be postponed pending the outcome of any investigation of the issues.

Contents

1 Introduction **5**

2 Higgs Boson Theory **7**

2.1 The Standard Model 7

2.2 The Higgs Boson 10

2.2.1 $VH(\rightarrow b\bar{b})$ 12

3 CERN **14**

3.1 The Large Hadron Collider 16

3.2 Collider Complex 17

3.3 Higgs Production 18

4 ATLAS **20**

4.1 The ATLAS detector 20

4.2 ATLAS Co-ordinate system 23

4.3 Inner Detector 23

4.3.1 Silicon Pixels 23

4.3.2 SCT 24

4.3.3 TRT 24

4.4 Calorimeters 25

4.4.1 Electromagnetic Calorimeter 25

4.4.2 Hadronic Calorimeter 26

4.4.3 Forward Calorimeter 27

4.5 Muon Spectrometer 27

4.5.1 Muon Tracking 28

4.5.2 Muon Identification 29

4.6 Magnet system 29

4.6.1 Toroid Magnets . . . 30
4.6.2 Central Solenoid . . . 30

5 ATLAS trigger **32**
5.1 ATLAS qualification task . . . 33
5.2 TriggerToolWeb . . . 33
5.2.1 Recent Developments . . . 37
5.3 VHbb lepton triggering . . . 38

6 VHbb analysis **42**
6.1 Analysis overview . . . 42
6.1.1 Data and simulated samples . . . 43
6.1.2 Object and event selection . . . 45
6.1.3 Systematic uncertainties . . . 49
6.1.4 Multivariate analysis . . . 50
6.1.5 Statistical analysis . . . 52
6.2 Overlap and inclusive Higgs studies . . . 55
6.2.1 Overlap study . . . 55
6.2.2 Inclusive Higgs samples study . . . 57
6.3 Boosted 2L MVA study . . . 60
6.3.1 Significance of MVA output . . . 60
6.3.2 Initial variable tests . . . 61
6.3.3 New vs Old MC samples . . . 62
6.3.4 Variable distribution plots . . . 63
6.3.5 Overtraining issues . . . 65
6.3.6 Baseline variables . . . 67
6.3.7 Truth flavour tagging vs direct tagging . . . 67
6.3.8 Hyperparameter optimisation . . . 72
6.3.9 Colour ring variable testing . . . 72
6.3.10 Colour ring results . . . 75
6.4 Modelling . . . 77
6.5 Summary of Run 2 boosted $VH(\rightarrow b\bar{b})$ results . . . 89
6.5.1 Previous Run 2 boosted $VH(\rightarrow b\bar{b})$ results . . . 89
6.5.2 Run 2 boosted legacy analysis results . . . 90

6.5.3 Run 2 combined $VHbb(cc)$ analysis results 90

7 Conclusion **95**

Chapter 1

Introduction

Particle physics asks one question: what is the universe made of? A simple question with an almost infinitely complex answer, one could say the universe is made of stars and planets but what are they made of? The real question is what are the smallest indivisible building blocks from which everything is made. The best current answer to this question is the Standard Model which describes the fundamental particles of our universe and how they interact but this model is not the final answer. The Standard Model has its shortcomings, there are remaining unexplained phenomena which are observed in our universe such as the existence of dark matter, dark energy (driving the expansion of the universe) and the observed baryon-asymmetry of the universe all suggest that there is more to learn. So particle physicists consider modifications to the Standard Model or even a new theory entirely to uncover the underlying structure of the universe. One method to further our understanding is to build cutting edge research experiments to precisely measure particle interactions predicted by the Standard Model and compare values to the theoretical result. If the experimental and theoretical results don't match then we have an opportunity to learn and adapt our theories to better match the observed results. The Higgs boson is the most recently discovered fundamental particle in the Standard Model, found by the ATLAS and CMS collaborations in 2012. Measuring all of the Higgs properties as precisely as possible remains a massive goal for the particle physics community. This thesis begins by discussing Higgs boson theory and the Standard Model of particle physics in chapter 2 focusing on the $VH(\rightarrow b\bar{b})$ process which is the main physics analysis process in this thesis. I then go on to discuss the CERN facilities, the Large Hadron Collider and the accelerator complex in chapter 3. This is followed by a detailed look at the ATLAS experiment in chapter 4 describing the many subsystems that make up this behemoth detector. We then look at the ATLAS trigger in chapter 5, which

is responsible for collecting interesting data from the many collisions that are recorded at ATLAS; I also helped develop the new online trigger tool app which is used to maintain trigger configurations for the new period of data collection. Finally, this thesis will cover the $VH(\rightarrow b\bar{b})$ analysis in detail in chapter 6, where I will describe the studies I have conducted in order to achieve improved results for the analysis.

Chapter 2

Higgs Boson Theory

Scientific understanding can best make progress when experiment and theory develop in parallel and experimental results can test theoretical models. In the case of particle physics the most complete and robust theory that has withstood confrontation with experimental results is the Standard Model (SM). The SM theory describes the basic building blocks of the universe and how they interact with each other at the most fundamental level. SM theory lays out how matter interacts at the sub-atomic scale. Experimental particle physics results should be compared with the SM theoretical prediction, in order to test the underlying theory.

The SM does have shortcomings. It does not account for gravity, nor does it include a dark matter particle - yet there is evidence for the existence of dark matter in the universe. The SM also does not account for the baryon asymmetry observed in the universe or the hierarchy problem - related to difference in scales of the Planck mass and the SM particles.

2.1 The Standard Model

The theory of the Standard Model was elaborated in the 1970s after experiments had discovered the existence of quarks. Since then the theory has been relentlessly tested by particle physics experiments around the globe. The top quark was discovered in 1995 at Fermilab (CDF and DØ experiments) [1, 2], then later in 2000 the tau neutrino was discovered at Fermilab (DONUT experiment) [3] and finally, the last SM particle to be discovered, the Higgs Boson was discovered in 2012 at the Large Hadron Collider (ATLAS and CMS experiments) [4].

The SM provides a comprehensive theory of all known elementary particles and the interactions between them described using the mathematical framework of quantum field theory

(QFT). The interactions between each particle take place via three of the four fundamental forces (strong force, weak force, and electromagnetic force). There isn't a quantum theory of gravity and so the SM cannot make any predictions about how quantum particles interact gravitationally. Above a certain energy threshold the electromagnetic force and the weak force unify into the electroweak force (EW). Quantum electrodynamics (QED) describes how matter and light interact in the universe - specifically charged particles interacting through emission/absorption of photons (quanta of light). QED continued on from classical electromagnetism using the basis of a relativistic QFT.

The SM Lagrangian ($\mathcal{L}_{SM}$) is invariant under gauge transformation - the theory acts the same way under translational symmetry, rotational symmetry and for different inertial reference frames. The internal symmetries of the SM are represented by the set of Lie groups: $SU(3)_C \bigotimes SU(2)_L \bigotimes U(1)_Y$. Here, $SU(3)_C$ denotes the strong interaction governed by the mathematical framework quantum chromodynamics (QCD) with the conserved quantity colour charge, C, $SU(2)_L$ and $U(1)_Y$ represent the unified EW interaction with conserved quantities of weak isospin, L, and hypercharge, Y.

The SM theory contains particles from two separate categories: fermions and bosons. Fermions are point-like particles that make up all the visible matter in the universe and comprise quarks and leptons. Bosons can be divided into gauge-bosons and scalar bosons. Gauge-bosons are force-mediating particles such as the photon (electromagnetism), gluon (strong force) and the W and Z bosons (weak force). The only scalar boson in the SM is the Higgs boson which acts as a propagator of the Higgs field - responsible for giving all particles mass (discussed in more detail in section 2.2) [5, 6]. Fermions have half-integer spin values, whereas bosons have integer values of spin. Quarks have electric charge values of $\frac{2}{3}$ or $-\frac{1}{3}$; they combine to make hadrons (group of closely bound quarks) with integer values of charge and form almost all of the mass in our universe. Leptons, in contrast to quarks, can stably exist alone and have small mass and integer values of charge (the tau lepton is an outlier with a mass greater than 1GeV which leads to rapid decays when a tau is produced).

Figure 2.1 shows all the fundamental particles in the Standard Model displaying their mass, electric charge and spin values provided by the Particle Data Group [7]. It also maps all the fermions through colour bands to the force-mediators they interact with, and splits the particles into three generations each of which contains two quarks a charged lepton and neutrino, each generation of particles have the same quantum numbers but with orders of magnitude larger mass than the previous generation. All particles in the SM have

an associated anti-particle (with a few exceptions) which have the same quantum numbers excluding electric charge which takes the opposite sign value for anti-particles. The gluon, photon, Z-boson and Higgs boson are all their own anti-particle, the positively and negatively charged W-bosons are each others anti-particle and the neutrinos, although they are neutrally charged, have anti-particles that are differentiated by their chirality (left-handedness or right-handedness).

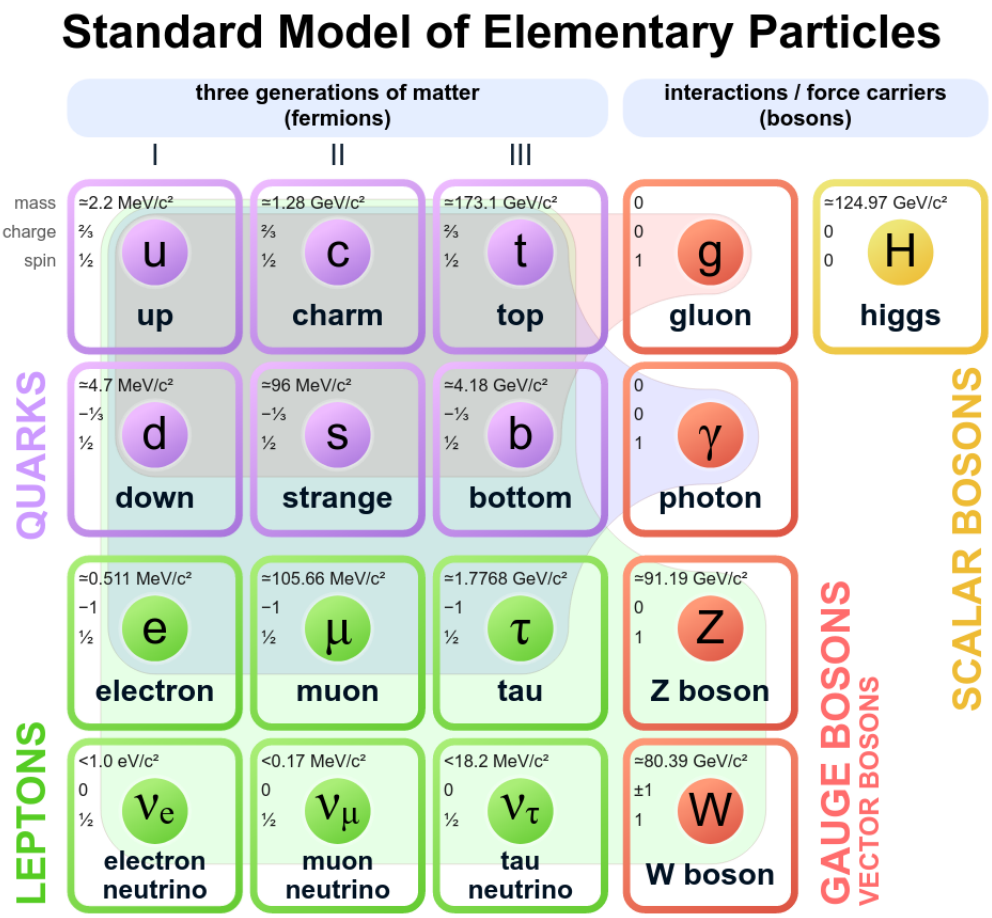


Figure 2.1: A diagram of all the particles that make up the Standard Model, displaying each particle's mass, charge and spin. All numbers are taken from the Particle Data Group [7], image taken from [8].

The Standard Model quarks and gluons all have colour charge; this is described by the 'three triplet model' which denotes the three charges as red, green and blue. Each colour also has the opposite charge equivalent, denoted by anti-red, anti-green and anti-blue. The primary colours are used as an allegory for this charge that comes in three types; the particles themselves are not coloured. Colour confinement is a rule of QCD stating that any colour charged particle can not exist in isolation, therefore, quarks and gluons must group together in a process known as hadronization to form colour neutral stable particles. The most common types of hadron are mesons and baryons. Mesons are made up of a quark anti-quark pair

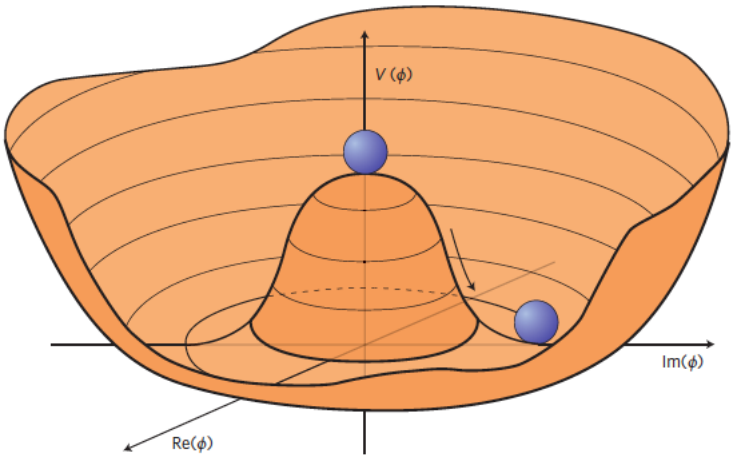


Figure 2.2: The Higgs potential field displayed on imaginary axes, showing the distinctive 'Mexican Hat' shape where the local minimum is offset from the origin [9].

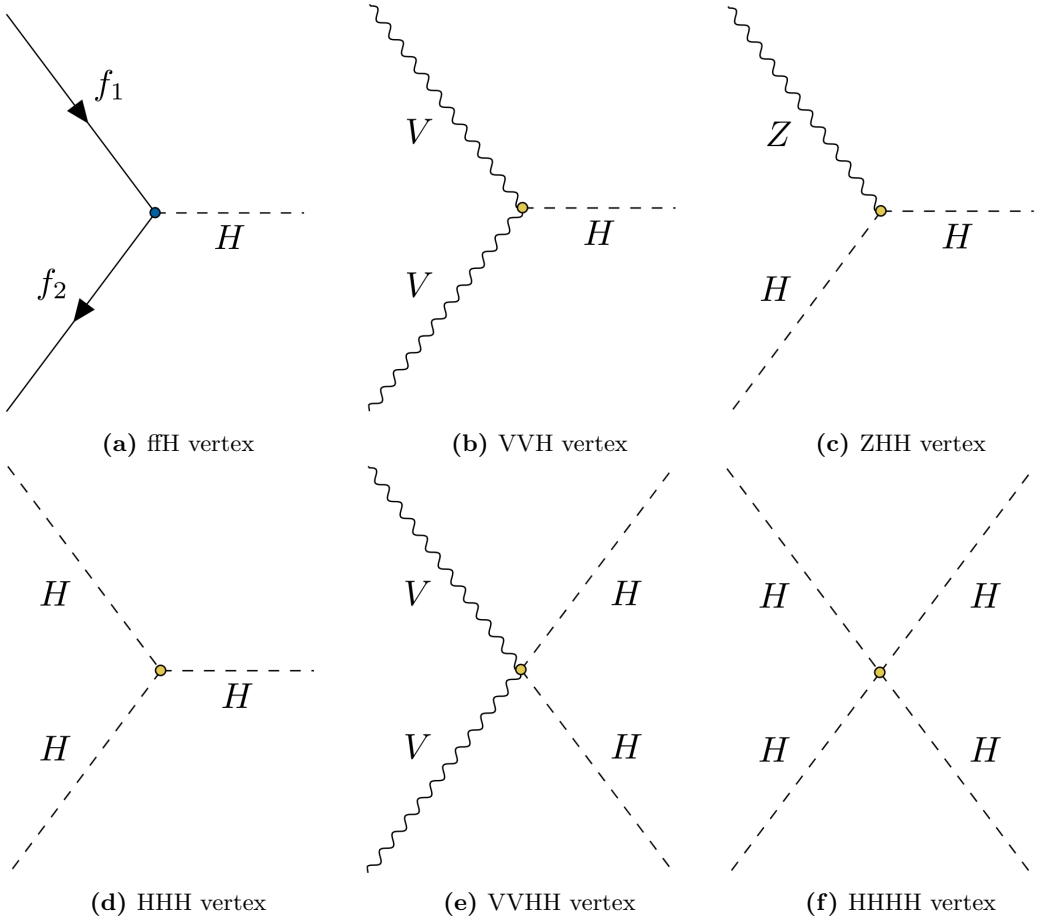
with opposite colour charges (e.g. red with anti-red), baryons are formed from three quarks or three anti-quarks contains one of each colour charge (e.g. one red quark, one green quark and one blue quark).

2.2 The Higgs Boson

The Higgs boson, named after theoretical physicist Peter Higgs, is the most recently discovered particle in the SM. It was theorized in the 1960s by Peter Higgs, Francois Englert and a number of other theorists independently. After the discovery of the Higgs boson in 2012, by the ATLAS and CMS collaborations at CERN, Higgs and Englert were awarded the 2013 Nobel Prize in Physics. The Higgs mechanism was proposed as a solution to the electroweak hierarchy problem, specifically associated to the high masses of W and Z bosons when compared to the mass-less photon. The Higgs field is the reason massive fundamental particles in the SM have inertia (rest mass), and the Higgs boson is an excitation of the Higgs field in the same way that the photon is an excitation of the electromagnetic field. The mass of a fundamental particle is proportional to the strength of its interaction with the Higgs field.

In the early universe (roughly the first 10 seconds after the big bang) all elementary particles were mass-less and travelled at the speed of light [10]. After the universe had sufficiently cooled the Higgs field started interacting through the Yukawa coupling with elementary particles giving them their rest masses. Figure 2.2 displays the Higgs potential field with the first blue ball (positioned at the origin of the real and imaginary plane) representing the early high energy universe and the second blue ball (positioned at the local minima in the potential field) the state after the universe had cooled and the Higgs field began to interact

with particles. The minimum of the Higgs potential is offset from zero due to the shape of the potential field, and this offset is the reason why fundamental particles are massive in the SM.



(a) ffH vertex **(b)** VVH vertex **(c)** ZHH vertex

(d) HHH vertex **(e)** VVHH vertex **(f)** HHHH vertex

Figure 2.3: The Feynman diagrams for all Higgs boson interactions in the Standard Model. The Higgs boson and the fermions interact through the (a) ffH vertex, the Higgs boson interacts with the vector bosons, V, through the (b) VVH vertex and the (e) VVHH vertex, it also interacts with the Z boson through the (c) ZHH vertex. The Higgs boson can interact with itself through the (d) HHH vertex and the (f) HHHH vertex [11].

The Higgs boson has a rest mass of $125.1\,\mathrm{GeV/c^2}$. It interacts with the fermions through the Yukawa interaction giving them mass. The mass of a fermion is proportional to its Higgs coupling (ffH vertex displayed in figure 2.3(a)). The Higgs boson itself is very unstable and decays almost immediately after production (lifetime of $1.6 \times 10^{-22}\,\mathrm{s}$ [12]). The most probable decay is to a pair of b-quarks; figure 2.4 displays a pie chart of all the possible decay products of a Higgs boson and their branching ratios.

The Higgs interaction vertices are displayed in figure 2.3. These Feynman diagrams show all the possible ways a Higgs boson can interact with the other particles in the SM theory. Higgs boson self-interaction is the reason the Higgs boson itself has mass. The Higgs boson interacts with the vector bosons, labeled V in the Feynman diagrams, which leads

to electroweak symmetry breaking giving rise to the massive W and Z bosons. It does not directly interact with photons or gluons, and hence they are mass-less. The vector boson type (either W or Z) is consistent for each interaction. Figure 2.3(a) is the ffH vertex which is responsible for giving the fermions mass. From figure 2.4 the largest branching ratio for a decaying Higgs is to a pair of b-quarks at around 58%. Experimental measurements of cross-sections times branching ratios can be compared to theoretical predictions. If the experimental and theoretical results don't match it could be that the Higgs is decaying to non-SM particles - potential evidence of new physics.

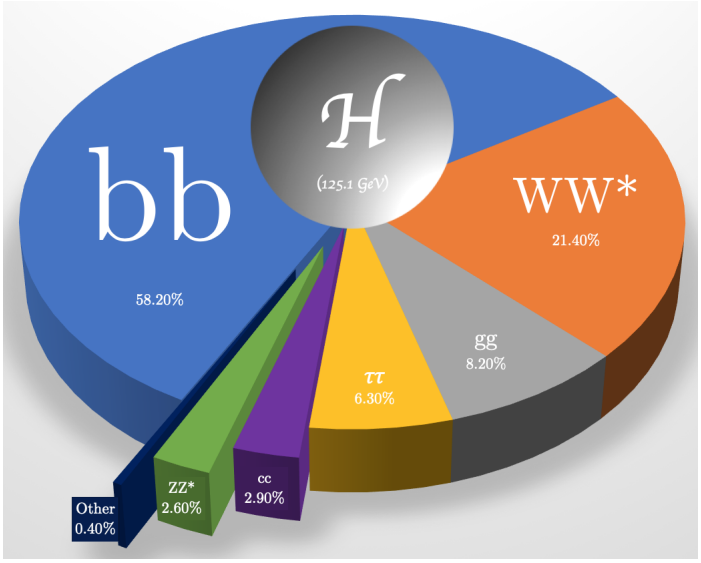


Figure 2.4: Pie chart showing the decay ratios of the Higgs boson to other particles. These ratios are given for a Higgs mass of 125.1 GeVc^2. [11].

2.2.1 $VH(\rightarrow b\bar{b})$

This thesis will discuss a specific Higgs process which is the basis of the analysis in section 6, namely the $VH(\rightarrow b\bar{b})$ interaction which begins with VH production. This can be seen from looking at the VVH vertex from figure 2.3(b) and rotating it to have one vector boson emitting a Higgs boson. For this decay to happen the incoming particle energy must be a virtual vector boson, labelled V^* (temporarily carrying more energy governed by the uncertainty principle). The analysis focuses on the $H \rightarrow b\bar{b}$ decay since this has the largest branching ratio; this can be viewed as the ffH vertex from figure 2.3(a), flipped horizontally to produce the quark pair from a decaying Higgs. The remaining vector boson, produced concurrently with the Higgs, decays leptonically - W boson decaying to a lepton-neutrino or antilepton-antineutrino pair, or Z boson decaying to a lepton-antilepton pair or neutrino-antineutrino pair. The leptonic decays are preferred for analysis as they produce a cleaner

signature in the detector than hadronic decays.

Figure 2.5 displays how the $VH(\rightarrow b\bar{b})$ process could take place at a proton-proton collider (e.g. the LHC at CERN). From the left of the figure the two incident protons are seen exchanging quarks, a quark-antiquark pair annihilate to produce a high energy virtual W/Z boson which then decays into a Higgs boson plus the same type of EW boson as the virtual boson, after which the Higgs decays to the b-quark pair and the EW boson decays leptonically. The final state particles in this interaction are the b-quark pair and the leptons or the neutrinos (dependent on the W/Z decay).

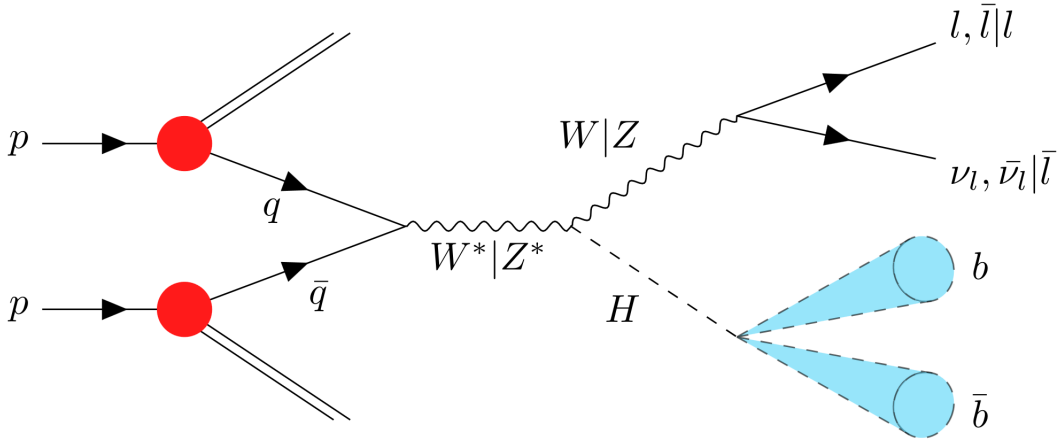


Figure 2.5: Feynman diagram of colliding protons producing a Higgs boson and vector boson with the Higgs decaying to a pair of b-quarks and the vector boson decaying leptonically [11].

Chapter 3

CERN

Organisation Européenne pour la Recherche Nucléaire, also known as CERN, is a cornerstone of nuclear and particle physics research around the world and has designed and built some of the most advanced research experiments of the last century. CERN aims to push the limits of science and human knowledge through large scale collaboration and currently comprises 24 member states.

Starting with the Synchro-cyclotron accelerator opened in 1957, CERN has built hundreds of experimental facilities over the decades. Currently there are 25 active CERN experiments ranging from the giant particle detectors around the Large Hadron Collider (LHC): ATLAS [14], CMS [15], LHCb [16] and ALICE; [17] to the Alpha Magnetic Spectrometer (AMS) mounted on the International Space Station and many others such as NA64, a fixed target experiment (searching for new particles in the dark sector, such as axions and dark photons), and GBAR, an experiment probing the gravitational effects of anti-matter. Figure 3.1 displays the current layout of all the accelerators and experiments located at the CERN campus and surrounding area. The particle accelerator complex delivers beams of particles to experiments.

The particle accelerator program at CERN has been at the forefront of accelerator science for over 50 years. The Proton Synchrotron (PS) completed in 1959 with a circumference of 628 m was the first in a long line of synchrotrons built at CERN. In 1976 the Super Proton Synchrotron (SPS) was built expanding the horizons of circular colliders with a 7 km circumference. Paving the way forward for collider physics, upgrades to the SPS allowed the first proton-antiproton ($p\bar{p}$) collisions in 1981, which led to the discoveries of the W and Z bosons. After the success of the SPS another massive circular collider was proposed, the Large Electron Positron collider (LEP) which sat in a tunnel on average 100 m below the Earth's surface and 27 km in circumference. LEP collided electrons with positrons and

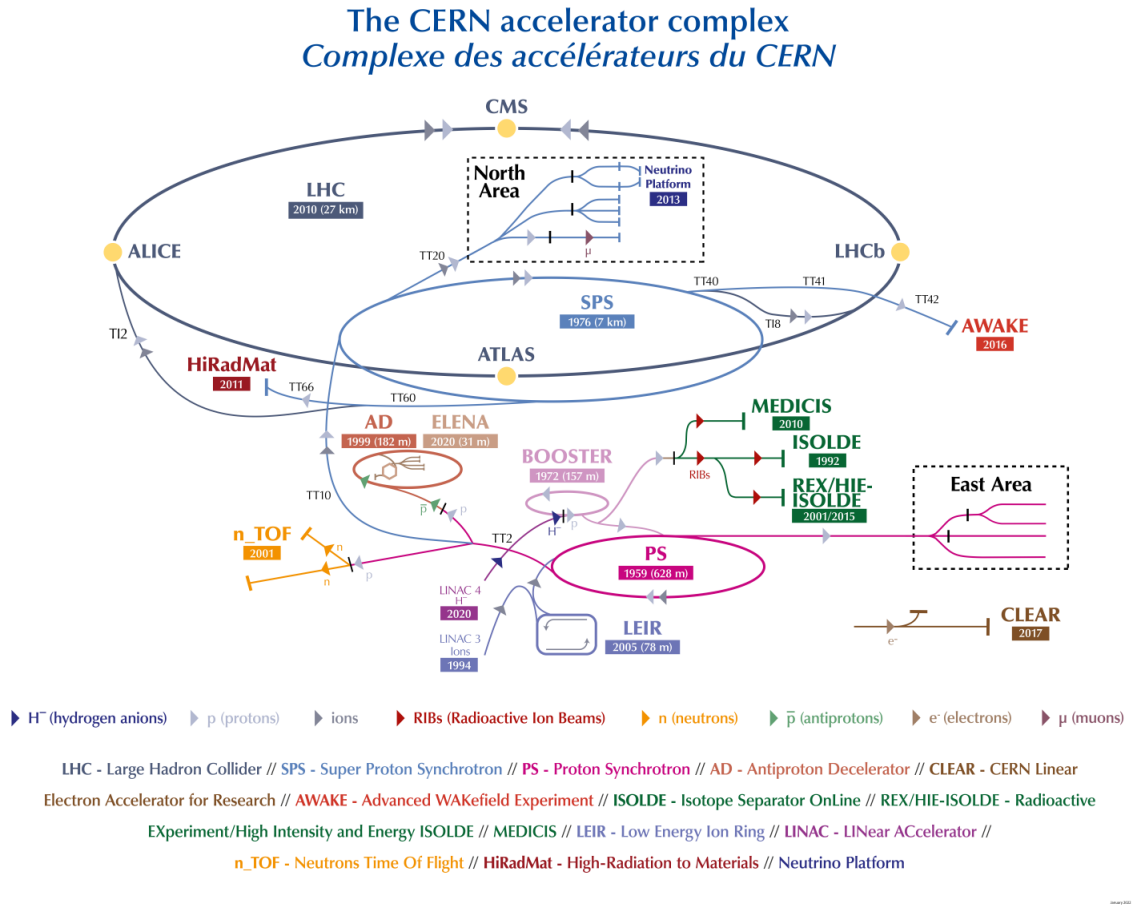


Figure 3.1: CERN Accelerator Complex layout graphic containing all of the physics experiments and accelerators on the CERN campus and surrounding area. [13]

Figure 3.2: Aerial view of CERN campus and surrounding area. The 27km long LHC collider and its four collision points shown by the circle and the four dots around it. The French-Swiss border is displayed by the dotted line. [18]

operated around the mass of the Z boson and was used primarily to observe and measure the properties of the newly discovered W and Z bosons. After the first run of LEP, it was upgraded to increase the collision energy to almost double the initial energy to produce and study pairs of W and Z bosons.

The tunnel of LEP was then used again to host the next-generation circular collider: the Large Hadron Collider (LHC).

3.1 The Large Hadron Collider

The LHC is currently the the world's largest and most powerful particle accelerator. Opened in September 2008 it was the first CERN-built collider to use superconducting electromagnets to accelerate particles around its 27 km ring (figure 3.2 shows an aerial view of the LHC). The electromagnets are made from coils of special electric cable and are cooled to $-271.3\,^{\circ}\mathrm{C}$ (1.85 K). At this temperature the electromagnets can conduct electricity very efficiently and can operate without loss of energy as heat in the cables. Almost all the magnets are connected to liquid helium cooling systems to maintain this optimal operating temperature. The LHC uses thousands of magnets to precisely direct and accelerate bunches of protons through the machine; clusters of magnets form magnet lattices [19]. The LHC has 1232 main dipole

magnets which apply a force bending the particles around the LHC ring; these dipole magnets are 15 m long and weigh 35 tonnes each. 392 quadrupole magnet lattices apply a constricting force on the beam squeezing the particles within the bunch closer together, this creates a very precise beam which is required for collisions at the detectors. Due to the energy dependency of this squeezing process the beam will spread out along its own direction of travel, smearing the well-separated bunches of particles within the beam. To solve this issue there are 688 sextupole magnet lattices and 168 octupole magnet lattices which apply a force along the beamline to maintain the particle bunch integrity (these lattice magnets are also used at the extremities of the main dipoles to deal with any magnetic field inconsistencies).

Around the LHC ring are four particle detectors; these are all situated at collision points where the particle beams are focused to a very small point and the opposing beams are aimed at each other. This beam focusing is achieved by using three quadrupole magnets in a system called inner triplets. Each collision point has two inner triplets on either side of the detector, which narrow the beam down to just 16 microns wide from the usual beam width of 0.2 mm. In total there are 8 inner triplets around the LHC [20].

3.2 Collider Complex

The LHC protons come from a bottle of hydrogen gas (not much bigger than a water bottle) which holds enough hydrogen to refill the machine 100,000 times. The molecular hydrogen (H_2) gas passes through a duoplasmatron which bombards the hydrogen with free electrons to produces hydrogen ions (H^-), which are protons with 2 loosely bound electrons (e^-). The low energy protons cannot be fed directly into the LHC, because the magnets only work for high energy protons, so they must be accelerated in smaller facilities. It is designed like this to produce the most precise beam possible, the beams can be monitored at every stage as the protons pass from one accelerator to the next. These protons start their journey into the LHC in LINAC4 (which recently replaced the LINAC2 during the second long shutdown, known as LS2), a linear accelerator that boosts the hydrogen ions from rest up to 160 MeV. The weakly bound electrons are removed from the protons by passing the ion beam through a very thin carbon foil, which allows the high energy protons to pass while capturing the loosely bound electrons. These fast protons are then fed into the Proton Synchrotron Booster (PSB) which is made up of 4 stacked synchrotron rings that continue to accelerate the protons up to 2 GeV. The PS is the next stop for the protons, which has been a workhorse of the CERN

accelerator complex for over 50 years. Operating up to an energy of 26 GeV, the PS has accelerated many different types of particles such as electrons and positrons, alpha particles (helium nuclei), oxygen, sulphur, argon, xenon, lead nuclei and antiprotons [21]. The SPS is the final stop for the protons before entering the LHC; the SPS accelerates the protons up to an energy of 450 GeV. After the protons leave the SPS they are fed into the LHC in both directions around the ring, ramped up to the required collision energy and repeatedly smashed into each other by the bunch (around 10^{11} protons per bunch) every 25 ns at the 4 major experiments located on the ring. Provided nothing goes awry, the beam is then dumped after 10 to 20 hours of continual data taking, and the process of accelerating the protons begins again.

The upgrade that introduced LINAC4 during LS2 is the one of the first developments aimed at increasing the luminosity of the LHC. The luminosity boost will provide more collisions and therefore more data for the physics community. CERN is planning on rolling out the high-luminosity (HL) upgrade within the next decade and move onto the next phase of LHC data taking with the HL-LHC. The plan is for the LHC to operate for at least another 30 years, with plans for LS3 to begin in 2026 which will complete the HL-LHC construction and detector upgrades. This final phase of the LHC will ramp up the collision energy and continue to produce Higgs bosons at unprecedented scales.

3.3 Higgs Production

The LHC was built with the goal of exploring new physics at unprecedented energy scales for colliders. In 2012, the Higgs boson was discovered giving way to many new analyses and precise measurements centered around this new particle. After many upgrades the LHC now produces more Higgs bosons than ever before. The rate of Higgs production (R_H) in a collider is calculated using equation 3.1.

$$R_H = \mathcal{L}_{inst} \times \sigma_{pp \to H}. \tag{3.1}$$

Instantaneous luminosity ($\mathcal{L}_{inst}$) is the measure of the number of potential collisions within a given area over a period of time, and is measured in $\mathrm{cm}^{-2}\mathrm{s}^{-1}$, and the cross section of proton-proton to Higgs production ($\sigma_{pp \to H}$) is a measure of the probability that a proton-proton collision will produce a Higgs boson. The LHC was designed to nominally run with a $\mathcal{L}_{inst} = 10^{34}\mathrm{cm}^{-2}\mathrm{s}^{-1}$ [22]. Integrating the instantaneous luminosity with respect to time

is a measure of total amount of data taken. A helpful unit that physicists use for this is the femtobarn (equal to $10^{-39}\,\mathrm{cm}^2$) which is an absolutely minuscule area but these are the tiny scales needed to describe particle accelerator collisions. The total data taken during a run is measured in inverse femtobarns (fb^{-1}). When two protons collide, they can interact in several different ways producing many types of particles, so the probability of creating a Higgs boson from one of these collisions is very low. Roughly one in every 10^{11} 13 TeV proton-proton collisions produces a Higgs candidate event, so to analyse precise properties of the Higgs boson we need a colossal amount of data. Currently the LHC is the only accelerator in the world capable of producing enough high energy events to be able to study the Higgs boson in detail. The theoretical predicted cross section for the proton-proton to Higgs boson production is around 5×10^4 fb from [23]. Using this and the nominal LHC $\mathcal{L}_{inst}$ of $10^{-5}\,\mathrm{fb}^{-1}\mathrm{s}^{-1}$ (changed from $\mathrm{cm}^{-2}\mathrm{s}^{-1}$ to $\mathrm{fb}^{-1}\mathrm{s}^{-1}$) we can calculate the expected rate of Higgs production events at the LHC from equation 3.1 which comes out at $0.5\,\mathrm{s}^{-1}$. So, on average the LHC produces a Higgs event every 2 seconds when running at full luminosity.

Chapter 4

ATLAS

The ATLAS detector is situated at point 1 on the LHC and is the largest detector (by volume) used for particle physics with thousands of researchers, engineers, technicians, and students from all across the globe working together in this enormous international scientific collaboration. ATLAS is roughly shaped as a cylinder 25 m in diameter and 46 m long and is situated in a cavern 100 m below ground. The LHC was first turned on in 2008 and ATLAS began recording collisions in 2010 at 7 TeV centre-of-mass energy, ramped up to 8 TeV centre-of-mass energy, with a total integrated luminosity of $26.4\,\text{fb}^{-1}$ collected during Run 1 (2010-2012) [24,25](shown in figures 4.1 and 4.2). After the first long shutdown another $147\,\text{fb}^{-1}$ of data was recorded at 13 TeV centre-of-mass energy during Run 2 (2015-2018) of which $139\,\text{fb}^{-1}$ was able to be used for physics analysis [26] (as seen in figure 4.3). The analysis in this thesis uses data collected during Run 2.

4.1 The ATLAS detector

The ATLAS detector could be considered one of the largest cameras in the world, taking billions of snapshots per second of each proton-proton collision at the nominal interaction point (IP). When a collision event takes place at the IP, a multitude of particles are produced and emitted in various directions. The detector's primary task is to measure and identify these particles with high precision. It accomplishes this goal by generating a magnetic field across the detector volume that causes the paths of charged particles (for example, electrons or positrons) to curve, thereby permitting the precise determination of their momenta. Additionally, the detector records the energies of particles by absorbing them, allowing for the calculation of their masses. The ATLAS detector is composed of three principal sections

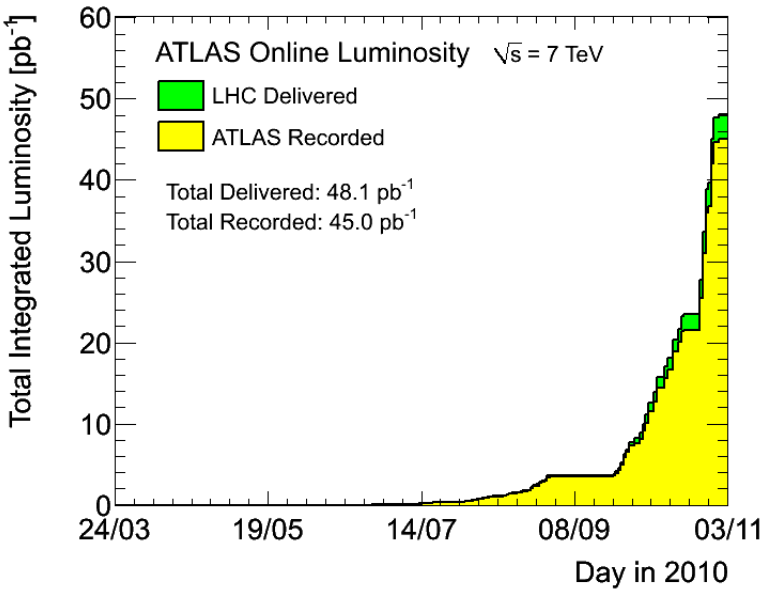


Figure 4.1: Plot of the cumulative luminosity versus day delivered to (green), and recorded by (yellow) ATLAS during stable beams and for pp collisions at 7 TeV centre-of-mass energy [24].

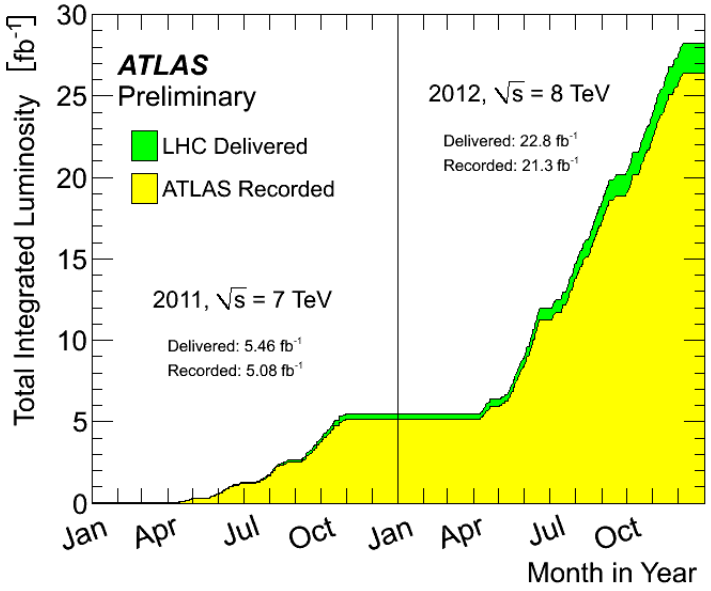


Figure 4.2: Plot of the cumulative luminosity versus month delivered to (green), and recorded by (yellow) ATLAS during stable beams and for pp collisions at 8 TeV centre-of-mass energy [25].

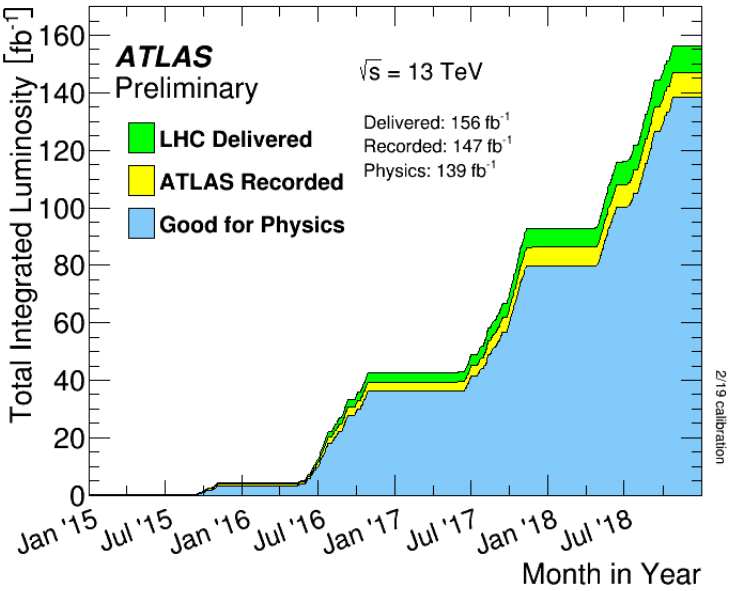


Figure 4.3: Plot of the cumulative luminosity versus month delivered to (green), recorded by (yellow) ATLAS, and good for physics (blue) during stable beams and for pp collisions at 13 TeV centre-of-mass energy [26].

surrounded by a system of magnets, namely the barrel and two end-caps, each equipped with multiple sub-detectors: including the Inner Tracker, Electromagnetic Calorimeter, Hadronic Calorimeter, and Muon Spectrometer. The magnet system is comprised of a central solenoid and toroidal magnets in the barrel, with the end-caps possessing their own toroid magnets(all sub-systems are shown in figure 4.4). The ATLAS detector, through the collaboration of these sub-detectors, is capable of producing a highly detailed picture of a high-energy proton-proton collision.

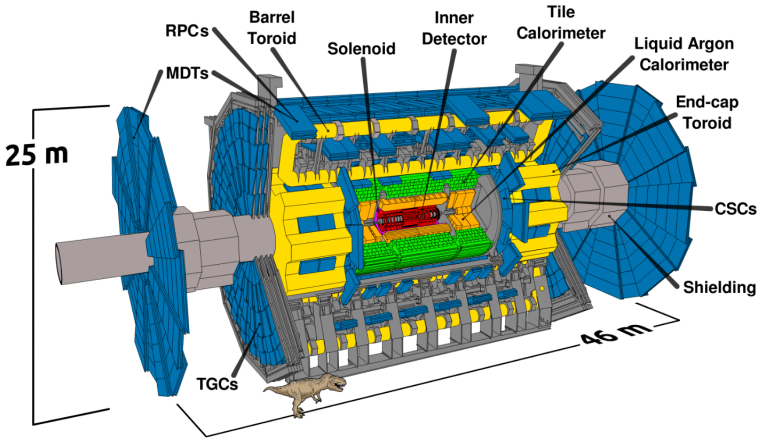


Figure 4.4: Layout of the ATLAS detector to scale, showing the Inner Detector (red), Electromagnetic Calorimeter (orange), Hadronic Calorimeter (green), Magnets (yellow) and Muon Spectrometer (blue) [27].

4.2 ATLAS Co-ordinate system

The detector geometry can be described using a Cartesian coordinate system with the origin located at the IP. The x-y plane is transverse to the beam pipe, with the positive x-direction defined as pointing towards the center of the LHC ring, and the positive y-direction defined as pointing upwards towards the surface of the Earth. The z-axis is parallel to the beam pipe. The most accurate information is collected perpendicular to the beam-line, and so due to the cylindrical nature of the detector, the most natural coordinate system is polar coordinates, where r and ϕ are used to represent the radial distance from the z-axis and the azimuthal angle around the IP in the x-y plane, respectively. The angle between the beam-line and a particle's trajectory is denoted as θ, but this angle is not Lorentz-invariant. Instead, pseudo-rapidity is used (η), defined as $\eta = -\ln(\tan(\theta/2))$. The barrel covers the region $\eta < 1.05$, and the end-cap covers the region $1.05 < \eta < 4.9$.

4.3 Inner Detector

Starting with the detectors closest to the beam pipe, we have the ATLAS inner detector (ID) which is used primarily for tracking. The ID is a cylindrical shape with length of 3.5 m and radius of around 1 m and sits within a solenoidal magnetic field with a strength of 2 T [28] [29]. As can be seen in figure 4.5, there are three sub-detectors that make up the ID: the silicon pixels layer, the Semiconductor Tracker (SCT) and the Transition Radiation Tracker (TRT). The ID has barrel and end-cap regions to capture particles from a large range of pseudo-rapidities.

4.3.1 Silicon Pixels

The silicon pixel layer is the closest detector to the beam-line. It uses a silicon semiconductor wafer that provides non-destructive tracking [30]. As a charged particle passes through the pixel, it creates many electron/hole pairs which are accelerated using an electric field to an electrode, which then provides an electric current to the anode of the device. This signal is then read out by the electronics and can be measured such that you can get the signal accurate to a single pixel. The dimensions of each silicon pixel are $50 \times 400\,\mu m^2$, which leads to very high resolution.

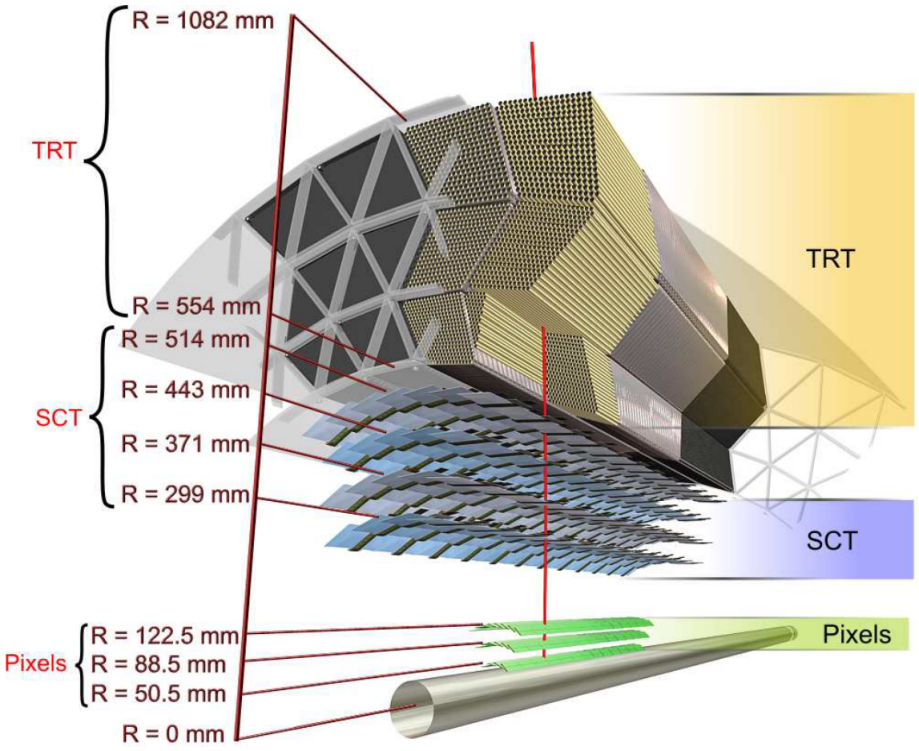


Figure 4.5: Drawing showing the track of a charged particle as it traverses the inner detector through the three layers of silicon pixels, four layers of silicon-microstrip sensors (SCT) and the transition radiation tracker (TRT) [14].

4.3.2 SCT

The SCT is a silicon-microstrip detector that provides precise tracking information of charged particles much like the silicon pixels of the previous layer. It is positioned radially outwards from the silicon pixels [31]. It consists of sets of silicon-microstrip sensors arranged in concentric cylinders in the barrel and in annular structures placed one after another in the two end-cap structures, which cover a pseudorapidity range of $|\eta| < 2.5$. The SCT's high spatial resolution allows for accurate reconstruction of particle trajectories and precise identification of particle vertices, it is slightly less accurate than the previous silicon pixel layer but more cost effective and so covers a larger volume in the detector. The SCT and silicon pixels combine efforts to track particle vertices, which are important for determining the initial starting location of each particle to infer whether it came from the primary interaction point.

4.3.3 TRT

The transition radiation tracker is the furthest sub-detector from the IP in the Inner Detector. It comprises thin gaseous straw tubes with a diameter of 4 mm arranged parallel to the beam in the barrel whereas in the end-caps they are perpendicular to the beam stacked in radial planes [32]. The gas mixture in these tubes is 70% xenon, 27% CO2, and 3% oxygen. These

tubes hold a positively charged gold-plated tungsten wire running down the centre.

When a charged particle passes through the TRT, it ionises the gas in the tube, which then sends an electric signal down the central wire to record a hit. On average, there are 36 hits across the TRT in the barrel and around 22 in the end-caps. These hits provide precise tracking for charged particles emanating from the IP. The magnetic field that is applied across the Inner Detector bends the path of charged particles. The amount of bend of these tracks is used to calculate the momentum of the particle (more bend equates to less momentum).

The TRT is also useful for electron identification. The minuscule gaps between the tubes in the TRT have a different refractive index than the gas mixture within the tubes, so transition radiation is released when charged particles passes through these boundaries. The total energy lost by a relativistic charged particle on the transition between two dielectric media is dependent on that particles Lorentz factor $\gamma = E/mc^2$. Since electrons have very low mass they have a very high Lorentz factor, γ, and give off a lot of energy in this region and can therefore be detected.

4.4 Calorimeters

Moving out from the ID we get to the calorimetry system [33] which is designed to absorb almost all of the particles coming from the IP (everything apart from muons and neutrinos) and is used primarily to measure their energies. The calorimeters are again split into barrel and end-cap regions. There is the electromagnetic calorimeter (ECal) which captures electrons/positrons and photons, the hadronic calorimeter (HCal) to capture mesons and baryons, and the forward calorimeter (FCal) which sits very close to the beam pipe to measure particles travelling out at large pseudo-rapidities. Figure 4.6 shows the different sub-systems that make up the calorimeter in ATLAS.

4.4.1 Electromagnetic Calorimeter

The liquid argon (LAr) Electromagnetic Calorimeter is designed to capture electrons/positrons and photons, it is the inner-most sub-detector within the barrel calorimeter system and within the end cap system it is situated radially outside of the FCal which surrounds the beam at the extremities of the detector [34]. The ECal contains a layer of liquid argon sampler material which provides a medium for the charged particle to emit photons and for photons to pair-produce and so-on giving a showering effect. A lead plate absorbs the showering particles and

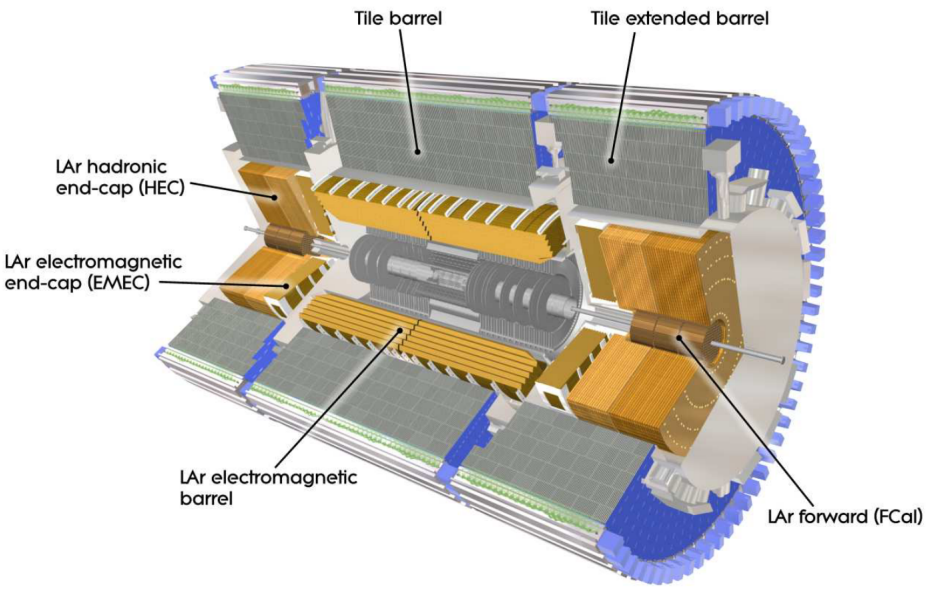


Figure 4.6: 3D Cut away of the calorimetry system in ATLAS [14].

determines their energy, the amount of energy incident on the plate is proportional to the particles energy entering the calorimeter. The lead plates and layers of LAr are arranged in a "zig-zag" formation known as the accordion structure (seen in figure 4.7) this reduces the dead-time of the calorimeter - the time between detecting a hit and being ready to register the next hit. The ECal is made up of barrel and end-cap regions spanning $|\eta| < 3.2$ and due to its accordion structure it covers the full range of ϕ angles. The calorimeter is segmented into modules. Figure 4.7 shows an end-cap module with an example of how the cells are divided for the trigger (the trigger records events for analysis, discussed in more detail in section 5). Energy is lost by the electron or photon as it passes through the ID; this is accounted for by using a LAr pre-sampler but only in the range $|\eta| < 1.8$. The pre-sampler provides an energy sampling in a thin LAr layer (11mm thick) before the calorimeter.

4.4.2 Hadronic Calorimeter

The Hadronic Calorimeter contains two sub-systems, namely the Tile Calorimeter and the LAr Hadronic Calorimeter [36]. The HCal is designed to capture all hadrons (charged and neutral) passing through the detector and does this in a similar way to the ECal by inducing particles to shower and absorbing the decay products. The barrel contains the Tile Calorimeter which uses tile scintillators as the sampler material and steel as the absorbing medium. In the end-caps this is achieved using copper plates as the absorbing medium and LAr as the sampler material. The combined barrel and end-cap regions cover the same η range ($|\eta| < 3.2$) as the ECal.

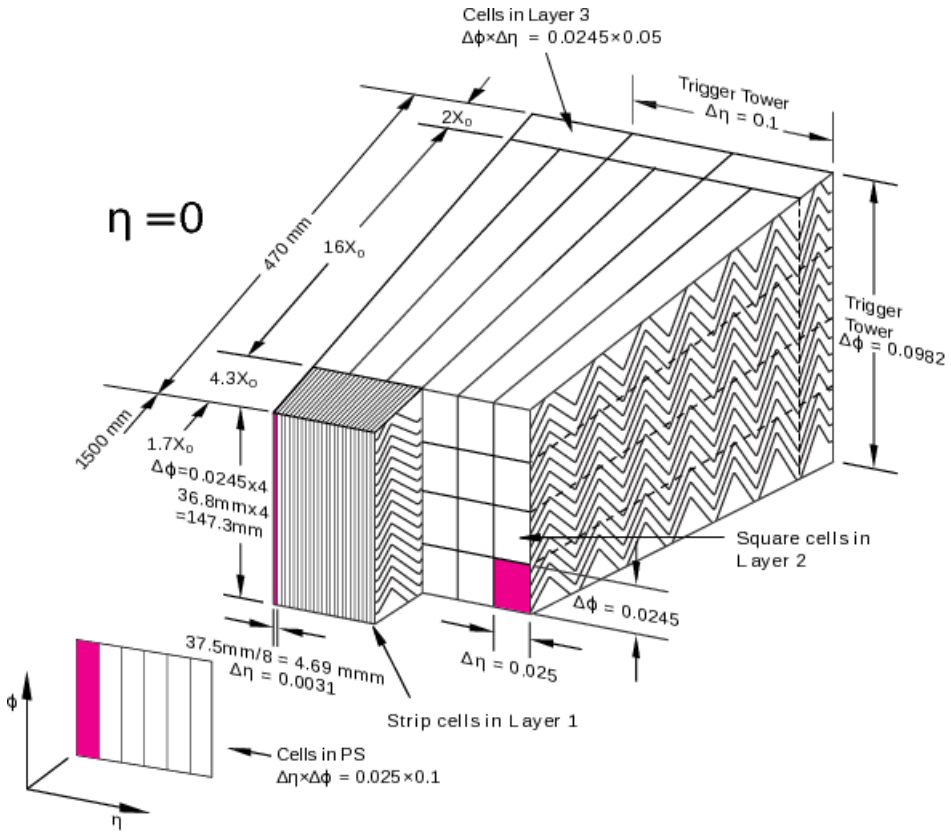


Figure 4.7: Diagram of a calorimeter module where the separate layers are visible. The granularity in η and ϕ of the cells of each of the three layers and of the trigger towers is also shown. The accordion structure is visible within the layers. [35].

4.4.3 Forward Calorimeter

Finally, the FCal is located only in the end-caps very close to the beam, covering a range $3.1 < |\eta| < 4.9$. This detector captures both hadrons and electromagnetic particles using a combination of copper and tungsten as the absorbing medium and again using LAr as the sampler material. The FCal is placed 1.2 m from the ECal to avoid the neutrons that scatter backwards into the detector (in a process called neutron albedo) damaging it.

4.5 Muon Spectrometer

Muons have a long lifetime, are minimum ionising and therefore can penetrate the calorimeter. They travel through the calorimeter layer and into the muon spectrometer (MS) which is the furthest sub-detector from the IP. The MS is responsible for tracking and identifying muons. The path of muons is bent using the massive barrel and end-cap toriodal magnets. The bent paths are used to infer the momentum of the muons as they pass through the MS. The muon vertices are tracked using the MS. There are four sub-systems that make up the MS: the Monitored Drift Tubes (MDT) and Cathode Strip Chambers (CSC) which are responsible

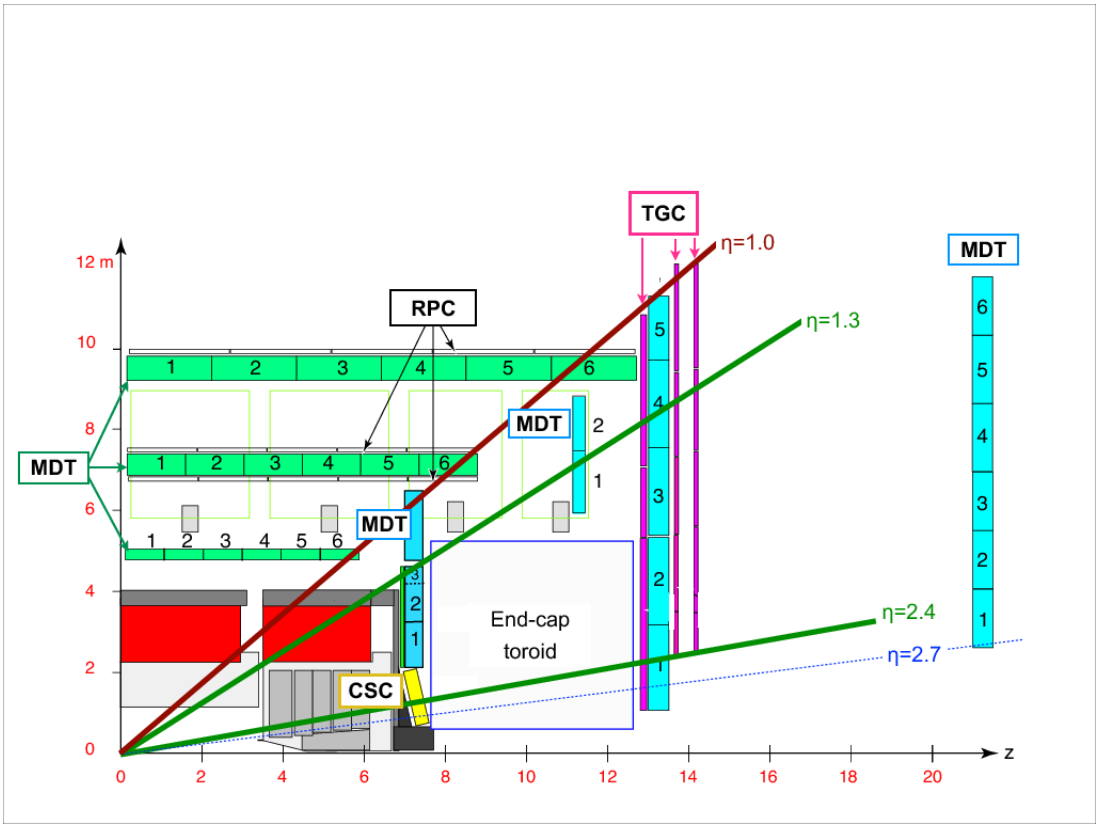


Figure 4.8: Schematic of a quarter of the Muon Spectrometer looking in the z-plane showing some example particle trajectories at different pseudo-rapidity values (η) [37].

for the tracking; and the Resistive Plate Chambers (RPC) and the Thin Gap Chambers (TGC) which are used for rapid triggering. Figure 4.8 displays the combination of all these sub-systems as they are arranged within the ATLAS detector.

4.5.1 Muon Tracking

The MS sits further away from the beam-line than any other sub-detector therefore it doesn't have to withstand as much radiation. So drift tubes while being less radiation resistant, are an excellent solution and a relatively inexpensive way of tracking muons. The MDT chambers contain drift tubes made of aluminium with a diameter of 3 cm filled with a gas mixture of 93% argon and 7% CO2 [38]. These tubes contain a 50 µm anode wire running through the centre to collect the drift electrons that are released as the muons pass by. The MDT has barrel and end-cap components covering a range $|\eta| < 2.7$ and provide a tracking precision of roughly 35 µm in the $r - z$ plane. At small angles to the beam line there is a much larger volume of muons, so to aid in the tracking of increased traffic there is the CSC which covers a range $2 < |\eta| < 2.7$. The CSC contains multiple anode wires and is known as a multi-wire proportional chamber. The extra wires in this detector provide increased granularity. From

figure 4.8 it is shown that the CSC is placed in the inner-most end-cap plate where better resolution is much needed.

The MDT and CSC provide great accuracy on the location of the muons, but the speed at which this information is read out is far too slow for fast-triggering of muons. So, other faster read-out detectors are needed to complete the MS system.

4.5.2 Muon Identification

The RPC and TGC are tasked with quickly identifying muons in the ATLAS detector as well as providing vertex location information in the $r - \phi$ plane. The RPC contains gaseous detector segments with pairs of parallel electrode plates with a large voltage applied across them; this induces large avalanches of particles as a charged particle passes through. The signal response after a muon traverses an RPC segment is almost instantaneous providing the required fast response for triggering. The TGC works similar to the CSC discussed in the previous section, it is also a multi-wire proportional chamber detector but with more focus on fast read-out rates and accurate time stamping. The RPC is only in the barrel and covers a range $|\eta| < 1.05$, whereas the TGC is located in the end-caps with a detecting range of $1.05 < |\eta| < 2.4$.

4.6 Magnet system

The magnet system is very important at ATLAS; it provides a magnetic field across the detector which allows us to measure the momentum of charged particles by bending their travel path through the detector. The magnet system at ATLAS consists of four superconducting magnets:

- the 2 T central solenoid (CS)
- the air-core barrel toroid (BT)
- the two air-cored end-cap toroids (ECT)

The CS provides a magnetic field across the inner detector, and the BT and ECTs provide a magnetic field across the calorimeter systems and the muon spectrometer. The layout of the ATLAS magnets are displayed in figure 4.9. The same toroidal magnetic field could be generated by a single magnet but this would restrict access into the central region of the detector, so, the ECT's have been designed so that they can be easily removed. The

ECTs slide into the BT on a rail system and are perfectly aligned with the CS, the overall dimensions of the magnet system are 26 m in length with a diameter of 20 m [39].

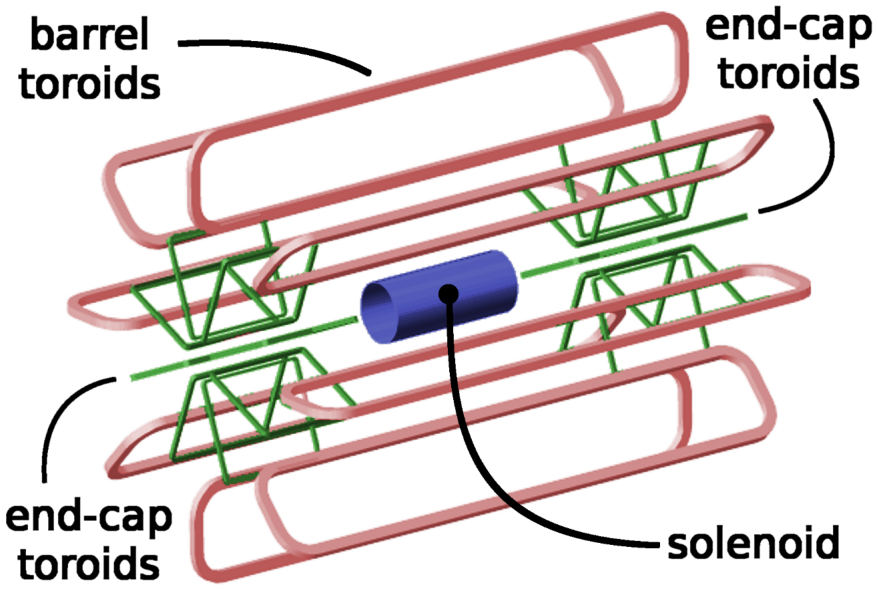


Figure 4.9: Schematic of the magnet system in ATLAS [39].

4.6.1 Toroid Magnets

Both the BT and ECT have eight coils placed symmetrically around the beam axis in the radial plane. The peak magnetic field in the BT and ECT are 3.9 and 4.1 T, respectively. The coils are made up of aluminium stabilized NbTi (niobium-titanium) superconducting wire wound up and held in an aluminium alloy casing. These superconducting magnets have an operating temperature close to absolute zero and must be constantly cooled using a forced flow of helium at 4.5 K. The helium runs through pipes welded onto the casing of the toroid magnets, they also have Dewar flasks (vacuum insulated containers) coupled to an external refrigeration unit to maintain this very low temperature. In the transition region ($1.4 < |\eta| < 1.6$) the magnetic field is weaker due to the overlap between the BT and ECT, the ECT has been rotated 22.5° with respect to the BT to reduce this effect.

4.6.2 Central Solenoid

The CS lies between the ID and the calorimeters, and is made up of a single winding coil supported in a large cylinder. The superconducting wire is the same as the one used in the BT and ECT, with the housing made from aluminium. The cooling of the CS is achieved by using the same Dewar flasks system as the toroidal magnets, but due to the location of the CS it is housed within the barrel calorimeter cryostat which reduces the number of interior

walls needed for the cooling systems. This helps reduce the amount of material in this crucial area of the detector - as we are still inside of the calorimeter any particles that get absorbed or deflected by non-active media can throw off not just the tracking and identification but also energy/missing energy calculations.

Chapter 5

ATLAS trigger

The ATLAS trigger system selects data with the most interesting characteristics for physics analysis [40]. The LHC produces 1.7 billion proton-proton collisions per second at the interaction point of ATLAS; this amounts to a combined data volume of 60 Terabytes per second [41]. It would be near impossible and totally impractical to record every event produced in ATLAS, therefore, the trigger system is implemented to select events with the most interesting characteristics for analysis. In ATLAS, these "interesting" events can be triggered by many different criteria, for example a number of high-energy leptons, photons, jets (stemming from quark decays) or a large amount of missing transverse energy,E_T^{miss}.

The ATLAS trigger consists of two parts; the Level-1 hardware Trigger (L1) and the High-Level software Trigger (HLT). The L1 Trigger is a hardware-level trigger which uses custom-made electronics and firmware to select events using data from the calorimeters, forward detectors and the muon spectrometer. It decides whether to keep or discard an event in 2.5 µs; if an event passes selection it is held in buffer storage. The L1 trigger is where the first event selection processes occur such as requiring the minimum p_T of leptons, or the number of jets in an event. If an event passes the L1 trigger it will then move on to the HLT, which can accept up to 100, 000 events per second [41].

The HLT runs on a computing farm of around 40, 000 CPUs. It operates by running algorithms that can check information from specific sub-detectors, making a more informed decision on whether to record an event. This is where b-tagging algorithms and other more computationally intensive processes are completed. If the event passes all requirements of the HLT then it is put into permanent storage. The HLT can store up to 1000 events per second [41].

The $VH(\rightarrow b\bar{b})$ analysis described in this thesis (in chapter 6) is searching for very spe-

cific decay modes, and these can be found using the ATLAS trigger. The analysis requires events with two b-jets coming from the Higgs decaying into two bottom quarks and 0/1/2 leptons decaying from the vector boson depending on the analysis channel. So the trigger configuration can be set, to look for these particles/objects and save the events that pass the checklist set by the analysis. All ATLAS analyses require different trigger configurations, therefore the online TriggerTool app (TTWeb) was developed to house these configuration files. TTWeb is an essential 'cog' in the ATLAS analysis machine that allows user friendly and simple access to the Trigger Database.

5.1 ATLAS qualification task

My ATLAS qualification task was to develop features for the Run 3 TriggerTool which was redesigned as an online application. The goal of the qualification task was to develop features for the menu experts to upload and compare trigger configurations. The trigger menu defines a list of selection criteria for every physics object that can be detected using ATLAS. For example, the 2-lepton $VH(\rightarrow b\bar{b})$ analysis requires either 2 electrons or 2 muons in the final state, so the analysis team will view the trigger menu and select the 2 electron and 2 muon triggers. TTWeb acts as the front end to the Trigger Database; it is used by menu experts to set up specific configurations that search for rare and interesting physics processes and also used by trigger shifters in the ATLAS control room to view trigger keys.

5.2 TriggerToolWeb

The trigger tool web application was developed as an upgrade to the previous trigger tool app which was coded in java, which made it difficult to upgrade for the purposes needed during Run 2. So the decision was made to create a new trigger tool app as an online application using the React JavaSript library. This allowed us to make a much more powerful tool to assist in the running of the ATLAS experiment. The work in this qualification task was presented in a poster at an international ATLAS conference in Lisbon in the autumn of 2022.

The new trigger tool app designed for Run 3 was initially created by two masters students who I worked closely with during the development of the app. The app was designed to be easy to use and look and feel simple but with all the necessary functionality. As a user enters the app there is a table of Super Master Keys (SMK). A Super Master Key is a unique combination of a particular set of L1 triggers and a particular set of HLT triggers. SMKs

are designed for specific types of runs, some are designed for use during Physics runs and others for Cosmic runs. Within an SMK is a list of L1 prescale keys and HLT prescale keys with the main differences between the prescale keys being their rates and the included sub-detectors (as sometimes one will be taken out for upgrades or tests). A prescale key holds a list of individual trigger items with their configured settings, including prescale value, enabled/disabled status and cut value. The L1 trigger rate is kept at or below 100 kHz and the HLT trigger rate should be 100 Hz or less. Prescale values work as follows: prescale value < 0 means the trigger is disabled; prescale value $= 1$ means all events that pass the trigger are recorded; and prescale value > 1 means only 1 in N events are recorded after passing the trigger, where N is the prescale value. The cut value, in addition to the prescale value, is used to control the rate at which events pass the trigger selection. After selecting a SMK there will be two tables displaying the L1 prescale keys and the HLT prescale keys, one key for each L1 and HLT will be used at a time during a run and these configure the trigger for the specific run conditions. A user can view (and edit) each of these prescale keys on the L1 (figure 5.1) and HLT (figure 5.2) prescale pages, the L1 Prescale page columns are:

- L1 item
- Prescale value
- Enabled/Disabled status
- Cut value

The HLT Prescale page columns are as follows:

- HLT chain
- Prescale value
- Enabled/Disabled status
- Express stream Prescale value
- Express stream Enabled/Disabled status
- Groups
- Streams
- L1 item

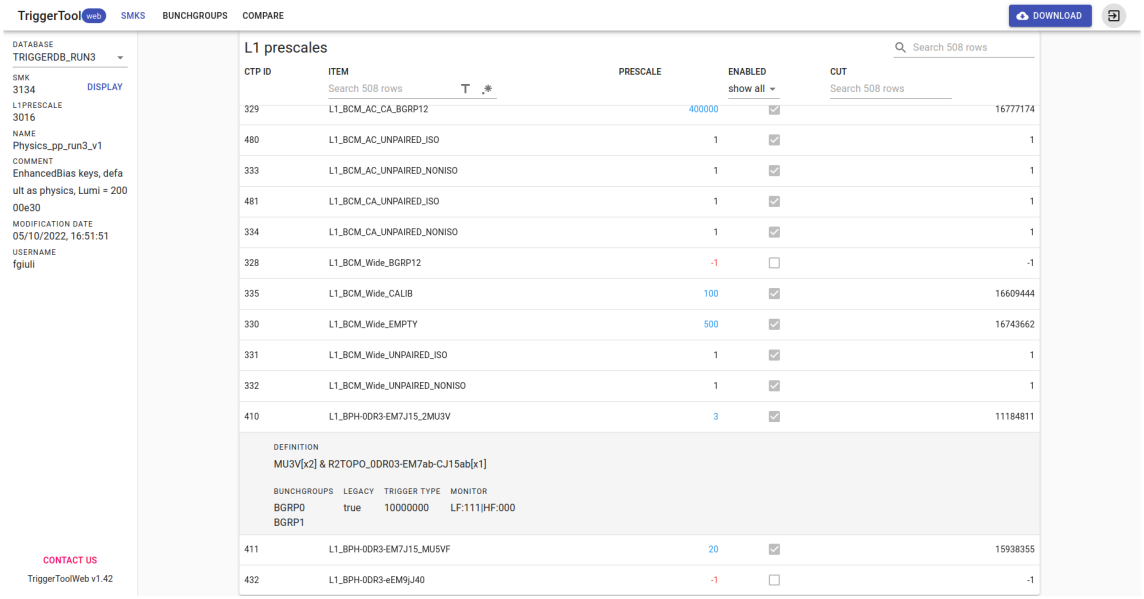


Figure 5.1: L1 Prescales Page displaying the according trigger configuration for a L1 Prescale Key, with extra information displayed for the selected L1 item.

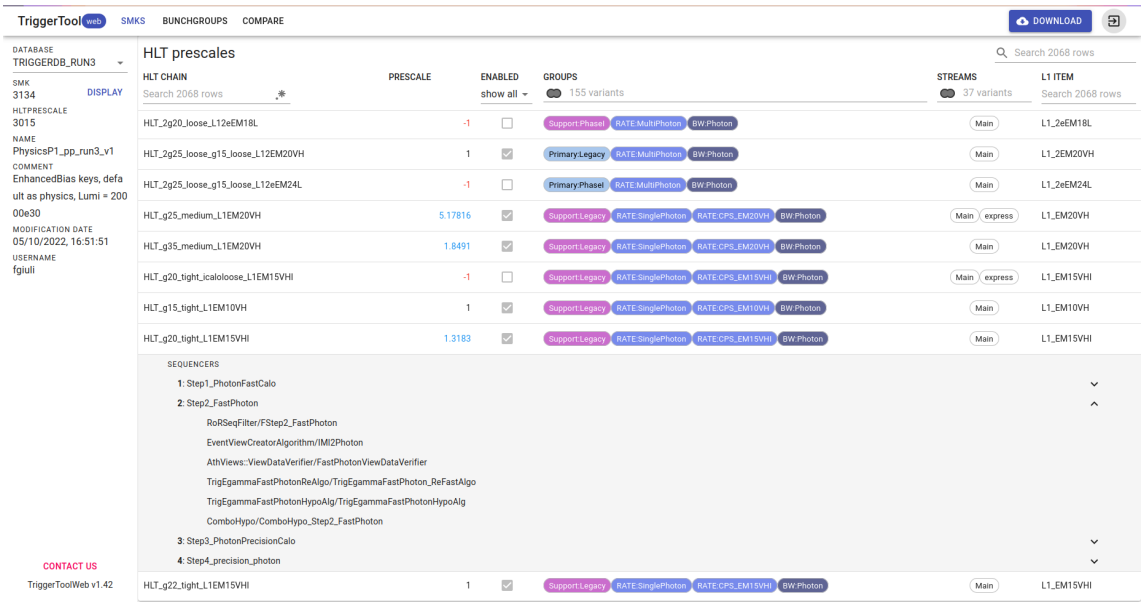


Figure 5.2: HLT Prescales Page displaying the according trigger configuration for a HLT Prescale Key.

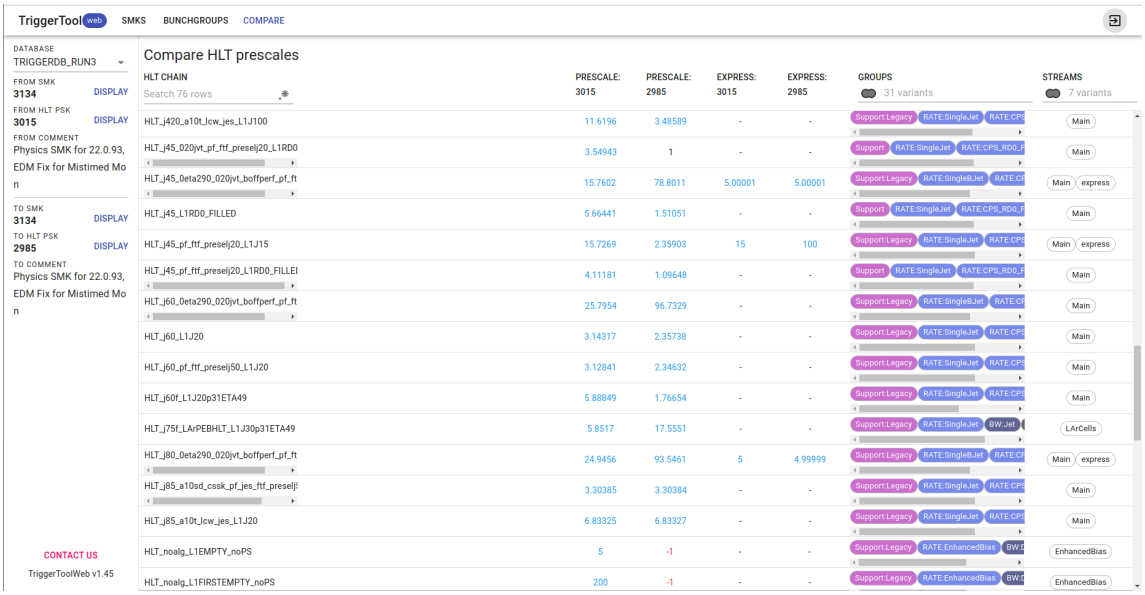


Figure 5.3: Compare HLT Prescales Page displaying two HLT trigger configurations.

Figure 5.4: Bunch Group Set information shown on the Trigger Tool Web app.

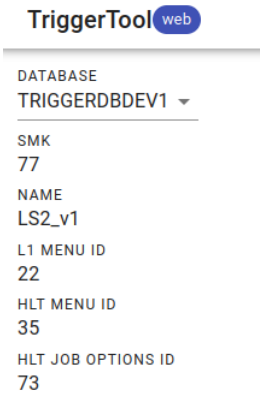


Figure 5.5: Left panel upgrades displaying the L1 menu ID, HLT menu ID and HLT job options ID.

The compare prescales page allows the user to look at two different (L1 or HLT) prescales and compare them all in one page, shown in figure 5.3. Another feature is the Bunch Group Set page shown in figure 5.4. This allows the user to see detailed information about any valid bunch groups set. A bunch group set details the pattern of bunches of protons within the LHC ring. The LHC can hold up to 3564 bunches of protons at once but not all of the available bunches are always filled. The bunch group set records the empty of filled status of each bunch travelling around the LHC. When two bunches cross at an interaction point they can be paired (both bunches contain protons), unpaired (only one bunch contains protons) or empty (neither bunch contains protons). For collisions to occur we require paired bunch crossings.

5.2.1 Recent Developments

Initial developments on the trigger tool included adding L1 menu ID, HLT menu ID and HLT job options ID to the left panel of the SMK page (shown in figure 5.5) and adding a bulk change feature (top middle enable/disable buttons seen in figure 5.6). The left panel information is very useful to the viewer, displaying information about the current page, such as name, comment and the key numbers, as well as last modification date and the user that uploaded the keys. This panel now also can be used to navigate around the app (this can be seen with the addition of the display buttons in the left panel of figure 5.7). A user can edit all the information shown in the L1 and HLT pages by clicking the blue edit button in the bottom right. While in edit mode the user can use the bulk change feature to enable or disable all shown items on the page, this can be refined by using the column filters.

A large task was developing the combined prescales page that displays both L1 and HLT

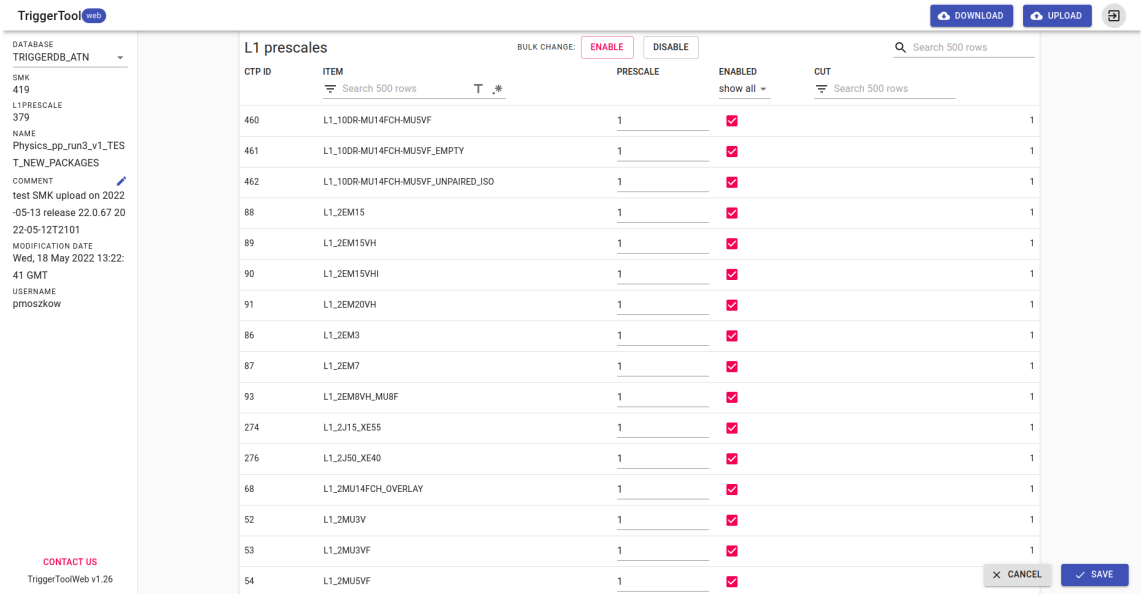


Figure 5.6: Trigger tool web app initial developments on L1 prescales page showing the left panel updates and bulk change options at the top.

prescale information (figure 5.7). This is useful for shifters and experts to view both L1 and HLT trigger configurations together on one page. It was designed to work like the other two prescale pages (HLT and L1) but without the editing capability as this was intended to be a view only page. The user can only view L1 prescale and HLT prescale keys from the same SMK. Some difficulties came from matching the HLT chains with their respective L1 items, as some HLT chains have multiple corresponding L1 items and others have none, each of these cases had to be handled separately when designing the page.

Another required development was to add express stream information to the HLT prescales page (express prescale value and express prescale enabled status). The Express stream contains a minimal rate (around 20 Hz) of events as required to perform a reliable first-validation of the detector subsystems and reconstructed physics objects. So, for the beginning of Run 3 when validation and calibration was needed it was critical that the HLT page contained this express stream information. The trigger tool was successfully deployed, with the additional features that I developed, for the start of Run 3.

5.3 VHbb lepton triggering

I will discuss how the triggering process works for the VHbb 2-lepton analysis. The analysis requires two triggered leptons. The L1 trigger items include a number of single electron and single muon triggers requiring two separate leptons of the same flavour to pass from a single event. The L1 trigger utilises information from the calorimeters and muon spectrometer.

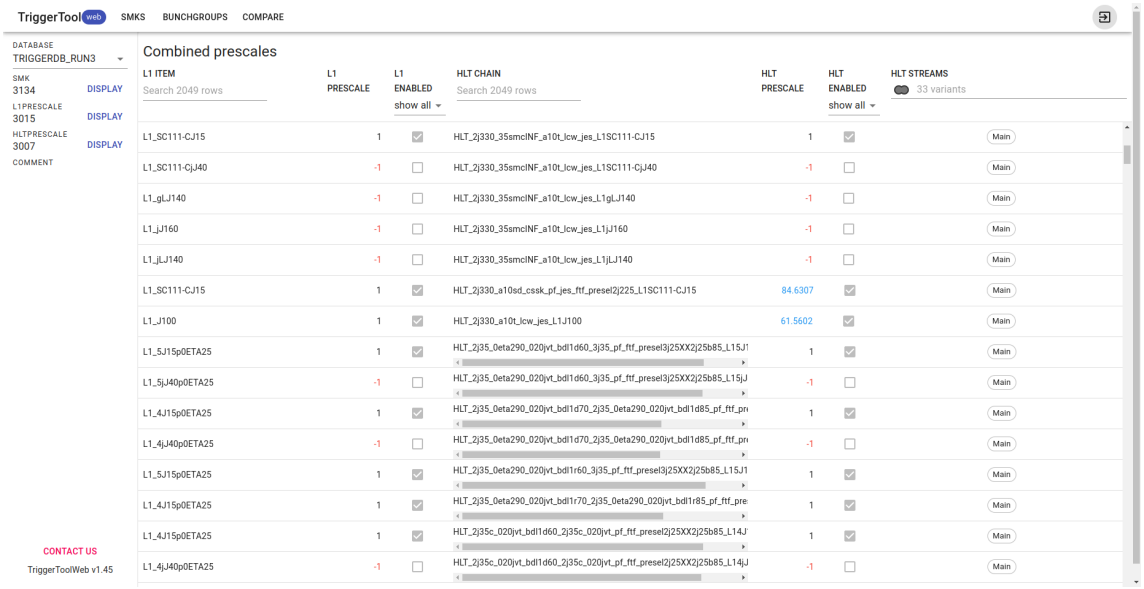


Figure 5.7: Combined Prescales view page recently developed for viewing prescale information from the L1 and HLT prescale pages.

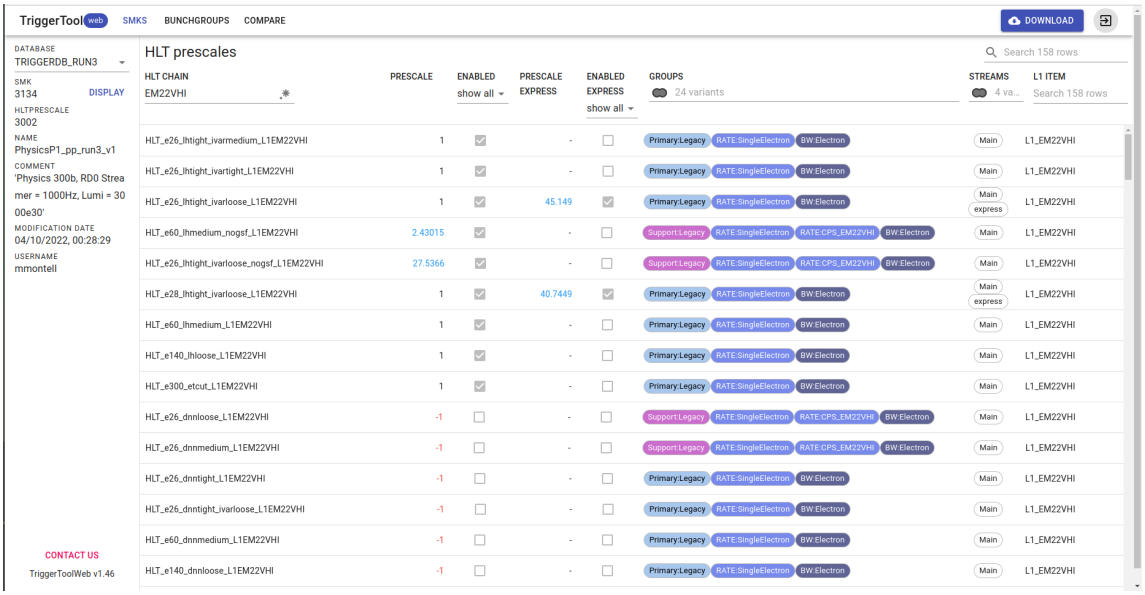


Figure 5.8: HLT Prescale page with additional columns displaying the express stream prescale value and enabled status, the columns are labeled prescale express and enabled express.

Type	Trigger Name	Period	Threshold	Description
$E_{\mathrm{T}}^{\mathrm{miss}}$	HLT_xe70_L1XE50	2015	70 GeV	Seeded using the level L1_XE50 (L1_XE55) LAr and Tile calorimeter triggers, calibrated at the EM scale, with a threshold of 50(55) GeV.
	HLT_xe90_mht_L1XE50	2016 (A-D3)	90 GeV	
	HLT_xe110_mht_L1XE50	2016 ($\geq$ D4)	110 GeV	
	HLT_xe110_pufit_L1XE55	2017	110 GeV	
	HLT_xe110_pufit_xe70_L1XE50	2018	110 GeV	
Electron	HLT_e24_lhmedium_L1EM20VH	2015	24 GeV	Seeded using L1 EM20VH level 1 trigger calibrated at the EM scale with a threshold of 20 GeV, require likelihood medium ID.
	HLT_e60_lhmedium	2015	60 GeV	Require likelihood medium ID.
	HLT_e120_lhloose	2015	120 GeV	Require likelihood loose ID.
	HLT_e26_lhtight_nod0_ivarloose	2016 – 2018	26 GeV	Tight likelihood ID required, alignment-robust likelihood tune with no d0 information used, and variable loose isolation required
	HLT_e60_lhmedium(_nod0)	2016 – 2018	60 GeV	Medium ID likelihood required
	HLT_e140_lhloose(_nod0)	2016 – 2018	140 GeV	Loose ID likelihood required
	HLT_e300_etcut	2018	300 GeV	No ID requirements.
Muon	HLT_mu20_iloose_L1MU15	2015	20 GeV	Seeded using L1MU15 level 1 trigger with a threshold of 15 GeV, and requiring loose isolation requirements.
	HLT_mu50	2015 – 2018	50 GeV	No isolation requirements.
	HLT_mu26_ivarmedium	2016 – 2018	26 GeV	Variable cone medium isolation requirements

Table 5.1: Triggers used during the 2015-2018 data collection period. For the HLT chains that don't mention the associated L1 trigger, the default L1 trigger is used. The default L1 trigger is EM22VHI or EM24VHI for single electron, MU20 or MU21 for single muon, depending on the data taken period. (From internal note)

The leptons in question need to be above a certain p_T threshold: above 20 GeV for electrons and above 15 GeV for muons (this changes throughout the data taking period). If the event passes the electron or muon L1 triggers then it will be passed onto the HLT. The HLT has access to information from the inner detector and any other relevant sub-detectors and runs more complex algorithms in order to judge events suitable for saving to permanent storage.

Single electron and muon triggers are favoured over dilepton triggers due to the complexity of calibrating both single lepton triggers and dilepton triggers, also adding the dilepton triggers only provides $\sim$ 2% increase on the signal acceptance (from a study done in the last round of VHbb analysis). The L1 item name and HLT chain name of a single electron trigger are `L1_EM20VH` and `HLT_e24_lhmedium_L1EM20VH`, respectively. The L1 trigger name contains `EM20` which explains it is looking for a single electron in the EM calorimeter with a p_T of at least 20 GeV. The letters `VHI` displayed in the L1 trigger name after the minimum p_T requirement relate to L1 computations: `V` means the p_T requirement varies in η to account for energy loss; `H` means hadronic core isolation is applied; `I` means electromagnetic isolation is applied. The HLT name displays `e24_lhmedium` with the L1 item appended on the end, meaning it will run an algorithm to check if the electron passes the likelihood medium selection criteria. Electrons can be selected for with loose, medium or tight requirements, also the algorithms can be cut-based or likelihood fits. Similarly, for muons the L1 item is `L1_MU15` with the HLT chain `HLT_MU20_iloose_L1MU15`. The HLT will run the `iloose` algorithm to determine the whether the muon is sufficiently isolated and passes the loose selection criteria. The missing energy triggers are favoured over the single muon triggers for 2L muon events

with $p_T^V > 250\,\text{GeV}$, this is due to the lower efficiency of the muon spectrometer at detecting high energy muons. If two electrons or two muons from the same event pass the HLT triggers then the event can be recorded for use in the VHbb analysis.

Chapter 6

VHbb analysis

Higgs boson decays to a pair of bottom quarks ($H \to b\bar{b}$) have been observed in the resolved topology [42] and in combination with different production modes (VH + ttH + VBF in [43]). The ATLAS collaboration studied the boosted VHbb for the first time in [44], an analysis that established a new reconstruction technique (large-R jet + track-jets). The $H \to c\bar{c}$ decay does not have experimental evidence yet due to its small branching fraction (2.89%) and [45] helped establish a new limit on this process' cross-section.

The analysis is searching for the $H \to b\bar{b}$ decay along with a leptonically decaying vector boson (W/Z). This is split into three separate channels, the 0, 1 and 2-lepton channels, relating to the $Z \to \nu\nu$, $W \to e\nu/W \to \mu\nu$ and $Z \to ee/Z \to \mu\mu$ decays respectively. The search for a leptonically decaying W/Z boson, as opposed to the hadronic decays, reduces the multi-jet background, specifically due to the good performance of the ATLAS calorimeter and muon systems in reconstructing leptons. After event selection the dominant backgrounds are V + jets, $t\bar{t}$, single-top and diboson. The analysis more recently has added a search for $H \to c\bar{c}$ events as well. Figure 6.1 displays the different analysis regimes and the phase space they cover in the transverse momentum plane. It is necessary to look at boosted topologies as they can account for a large fraction of the Higgs events produced at ATLAS; as the LHC energy is ramped up this will only become more important.

6.1 Analysis overview

The analysis is split into $VH(\to b\bar{b})$ boosted and resolved, and a separate $VH(\to c\bar{c})$ analysis split by a dedicated hadronic jet identification algorithm. The $VH(\to b\bar{b})$ resolved regime is characterized by two b-tagged small radius ($R = 0.4$) jets; the $VH(\to b\bar{b})$ boosted analysis

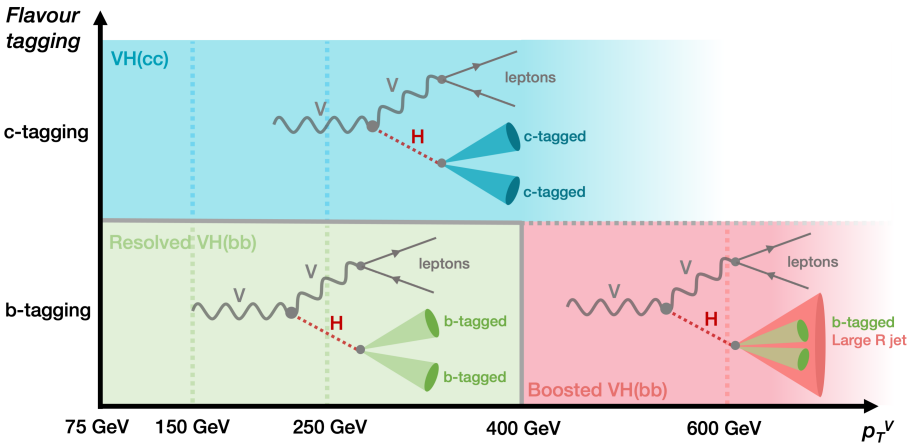


Figure 6.1: Figure showing the separate regimes for the $VH \rightarrow bb(cc)$ analysis, plotted for the different tagging strategies and the p_T^V regions they cover. (From internal note)

is characterized by two b-tagged track jets which must be contained within a large radius (large-R) jet ($R \approx 1$); and the $VH(\rightarrow c\bar{c})$ analysis is characterized by the presence of c-tagged jets and is studied using the resolved topology with small radius jets ($R = 0.4$).

Most of my effort was focused on the boosted analysis, specifically the 2-lepton channel multivariate analysis (discussed in section 6.3). The resolved and boosted regimes are split by a cut at a p_T^V of 400 GeV, this is to remove any overlap between the two analyses (this is discussed in more detail in section 6.2.1). Cross-check analyses were also performed as sanity checks to this analysis: these are the SM Diboson search with a Z boson decaying to b-quark pair (such as ZZ, WZ and WW processes); the cut-based analysis used in previous iterations of the analysis; and also the previous version of the $VH(\rightarrow c\bar{c})$ analysis.

6.1.1 Data and simulated samples

The current analysis uses data collected by the ATLAS detector during the 2015, 2016, 2017 and 2018 proton-proton runs with a centre-of-mass energy of 13 TeV and corresponding to an integrated luminosity of 139 fb^{-1}. All events must be on the good run list which is a list of events that were captured when the detector was in full working order and therefore able to fully capture the event. All signal and background processes are simulated using Monte Carlo techniques excluding the 1L multi-jet background and the 2L $t\bar{t}$ background in the resolved analysis which are estimated using data-driven techniques.

Simulated samples are used to compare with data and also find the largest signal-to-background ratio regions. Monte Carlo (MC) generators are the tool that particle physicists use to simulate both background and signal events. Signal samples are simulated using POWHEG-BOX + GoSAM + MiNLO generators [49] with PYTHIA 8 [50] used for parton

Process	Matrix Element	PDF set (ME)	Parton Shower	σ order	$\sigma\times$ Br [pb]
$qq \to WH \to \ell\nu b\bar{b}$	PowHeg-Box v2 + GoSam + MinLo	NNPDF3.0NLO	Pythia-8.245	NNLO(QCD)+ NLO(EW	2.69×10^1
$qq \to ZH \to \nu\nu b\bar{b}$	PowHeg-Box v2 + GoSam + MinLo	NNPDF3.0NLO	Pythia-8.245	NNLO(QCD)+ NLO(EW	8.91×10^2
$qq \to ZH \to \ell\ell b\bar{b}$	PowHeg-Box v2 + GoSam + MinLo	NNPDF3.0NLO	Pythia-8.245	NNLO(QCD)+ NLO(EW	2.99×10^2
$gg \to ZH \to \nu\nu b\bar{b}$	PowHeg-Box v2	NNPDF3.0NLO	Pythia-8.307	NLO+NLL	1.43×10^2
$gg \to ZH \to \ell\ell b\bar{b}$	PowHeg-Box v2	NNPDF3.0NLO	Pythia-8.307	NLO+NLL	4.80×10^3
$qq \to WH \to \ell\nu c\bar{c}$	PowHeg-Box v2 + GoSam + MinLo	NNPDF3.0NLO	Pythia-8.245	NNLO(QCD)+ NLO(EW	1.34×10^2
$qq \to ZH \to \nu\nu c\bar{c}$	PowHeg-Box v2 + GoSam + MinLo	NNPDF3.0NLO	Pythia-8.245	NNLO(QCD)+ NLO(EW	4.42×10^3
$qq \to ZH \to \ell\ell c\bar{c}$	PowHeg-Box v2 + GoSam + MinLo	NNPDF3.0NLO	Pythia-8.245	NNLO(QCD)+ NLO(EW	1.48×10^3
$gg \to ZH \to \nu\nu c\bar{c}$	PowHeg-Box v2	NNPDF3.0NLO	Pythia-8.307	NLO+NLL	7.10×10^4
$gg \to ZH \to \ell\ell c\bar{c}$	PowHeg-Box v2	NNPDF3.0NLO	Pythia-8.307	NLO+NLL	2.38×10^4
$Z \to \nu\nu+$ jets	Sherpa 2.2.11	NNPDF3.0NNLO	Sherpa 2.2.11	NNLO	1.07×10^4
$W \to \ell\nu+$ jets	Sherpa 2.2.11	NNPDF3.0NNLO	Sherpa 2.2.11	NNLO	6.02×10^4
$Z \to \ell\ell+$ jets	Sherpa 2.2.11	NNPDF3.0NNLO	Sherpa 2.2.11	NNLO	6.30×10^3
$t\bar{t}$	PowHeg-Box v2	NNPDF3.0NLO	Pythia-8.230	NNLO+NNLL	832
single-top (s)	PowHeg-Box v2	NNPDF3.0NLO	Pythia-8.230	NLO	3.35
single-top (t)	PowHeg-Box v2	NNPDF3.0NLO	Pythia-8.230	NLO	70.4
single-top (Wt)	PowHeg-Box v2	NNPDF3.0NLO	Pythia-8.230	Approx. NNLO	71.7
$qq \to WW$	Sherpa 2.2.11	NNPDF3.0NNLO	Sherpa 2.2.11	NLO	45.7
$qq \to WZ$	Sherpa 2.2.11	NNPDF3.0NNLO	Sherpa 2.2.11	NLO	21.7
$qq \to ZZ$	Sherpa 2.2.11	NNPDF3.0NNLO	Sherpa 2.2.11	NLO	6.53
$gg \to VV$	Sherpa 2.2.2	NNPDF3.0NNLO	Sherpa 2.2.2	NLO	3.09

Table 6.1: The nominal Monte Carlo samples for Signal and Background processes in the $VH \to b\bar{b}(c\bar{c})$ analysis and the corresponding cross sections at $\sqrt{s} = 13\,\text{TeV}$. The PDF sets listed are for the Matrix element only. The MC generators used are PowHeg [46], Sherpa [47] and Pythia [48]. (from internal note)

shower and hadronization simulations. The inclusive signal cross-sections (ZH,WH) are calculated at NNLO in QCD [51] and NLO in EW [52]. The V+jets samples are simulated using SHERPA 2.2.1 and the $t\bar{t}$, single-top and Wt are generated using POWHEG-BOX interfaced to Pythia8. All samples and their MC generation software are shown in table 6.1. The diboson samples are simulated using Sherpa 2.2.11; the modelling plots are displayed and discussed later in section 6.4. The cross-section of each simulated process is calculated to at least next-to-leading-order (NLO) precision. The ATLAS detector is simulated using the GEANT 4 software.

The multi-jet background is not simulated but instead it is estimated using data-driven techniques. It is found that the multi-jet contribution to the 0 and 2-lepton channels is negligible, but for the 1-lepton channel the contribution is non-negligible. The contribution is estimated separately for the 2 and 3-jet categories and for the electron and muon sub-channels. In the 2-jet category the multi-jet background is found to contribute 1.9% (2.8%) to the electron (muon) sub-channels and in the 3-jet category it is found to contribute 0.2% (0.4%) to the electron (muon) sub-channels.

Data-driven techniques are also used to estimate the top quark background ($t\bar{t}$, single-top) in the resolved 2-lepton analysis. In this channel, there is a control region (CR) requiring a pair of opposite-flavour opposite-sign leptons ($e\mu$) which is very pure with top backgrounds. The lepton flavour requirement is the only thing that separates this CR from the siganl region (SR). The observed distribution from the CR is used as a direct replacement for MC

generated top backgrounds.

6.1.2 Object and event selection

The search for VH, $H \to b\bar{b}$ focuses on finding two b-jets, arising from b-quark decays, and either 0, 1 or 2 charged leptons. The inner detector picks up tracks which can be used to calculate the interaction vertex from which they came. The vertex with the largest sum of squared transverse momentum is considered the primary vertex.

Leptons

Electrons will deposit energy in the calorimeter which can be matched to one of the tracks found in the inner detector. The calorimeter applies clustering algorithms to extract the lepton signal from background electronic noise and pile-up effects [53]. For an electron to be 'detected' it must have $p_T > 7\,\text{GeV}$ and $|\eta| < 2.47$. Muons have the same $p_T > 7\,\text{GeV}$ restriction but have a slightly wider acceptance angle of $|\eta| < 2.7$ (due to detector geometry). Hadronically decaying τ-leptons are only used in this analysis to cross-check with b-tagged jets for mis-identification, the requirements on the τ are $p_T > 20\,\text{GeV}$ and $|\eta| < 2.5$. In the 0-lepton channel there are no 'hits' in the detector that could be used to reconstruct the Z boson as neutrinos are invisible to the detector. Instead, the missing transverse momentum E_T^{miss} vector is calculated as the negative vector sum of everything in the event originating from the primary vertex and used to reconstruct the Z boson. Overlap removal is carried out to check for any double counting of reconstructed leptons, jets and hadronic τ-leptons.

Jets

Jets are reconstructed using the anti-k_t jet finding algorithm [54]. The anti-k_t algorithm calculates distances between particles, creating clusters of particles that are in close proximity to each other. This clustering process is iterated and if a particle (or cluster of particles) is sufficiently isolated then it becomes a jet. There are three different types of jets used in the analysis; these are the small-R jets, large-R jets and variable-R track jets. Small-R jets are calorimeter jets with radius $R = 0.4$, and are used primarily in the resolved $VH(\to b\bar{b})$ analysis and in the $VH(\to c\bar{c})$ analysis. Large-R jets are calorimeter jets with radius $R = 1$, and are used to reconstruct the Higgs candidate in the boosted $VH(\to b\bar{b})$ analysis. Large-R jets have a large radius to accommodate the high transverse energy Higgs boson which produces boosted decay products (small angle between the pair of b-quarks). The general

rule for choosing the radius parameter of the large-R jet is:

$$\Delta R(b_1, b_2) \sim \frac{m_H}{p_T^H} \frac{1}{\sqrt{z(1-z)}} \sim \frac{2m_H}{p_T^H} \tag{6.1}$$

where b_1 and b_2 are the b-quarks from the Higgs decay, the mass and transverse momentum of the Higgs boson is denoted by m_H and p_T^H, respectively, and z is the momentum fraction carried by b_1. This produces a ΔR value around 1 for the boosted case ($p_T^H > 250\,\text{GeV}$ and $m_H = 125\,\text{GeV}$). Variable-R track jets have a jet p_T dependent distance parameter $R(p_T) = \rho/p_T$ where ρ is the rate at which the jet size decreases with p_T. The maximum and minimum radii cutoffs for these jets are $R_{max} = 0.4$, $R_{min} = 0.02$ and the ρ parameter is set to 30 GeV. These parameters are optimized for double b-tagging efficiency in the boosted $H \to bb$ decays. All jets that didn't originate from the primary vertex are removed from the event. The small-R jets are required to have $p_T > 20\,\text{GeV}$ if they pass through the inner tracker ($\eta < 2.5$), and if they are outside of the range of the tracker ($2.5 < |\eta| < 4.5$) then they are required to have $p_T > 30\,\text{GeV}$. The large-R jet must have $p_T > 250\,\text{GeV}$ and lie in the range $|\eta| < 2$. The variable-R track jets must have $p_T > 10\,\text{GeV}$ and $|\eta| < 2.5$. To determine which fundamental particle our jet originated from, one must use a jet flavour tagging algorithm.

Flavour tagging

Jet flavour tagging is the process of determining the flavour of a jet - for the $VH(\to b\bar{b})$ and $VH(\to c\bar{c})$ analyses this is done using the DL1r algorithm [55]. DL1r is a high-level flavour-tagger which takes the output of low-level taggers as inputs; such as the IP2D, IP3D, SV1, JetFitter and RNNIP algorithms [56]. The low-level tagger outputs are combined along with jet information like p_T and $|\eta|$ using neural network techniques to classify jet flavours. There are two types of low-level taggers: impact-parameter based taggers (such as IP2D, IP3D and RNNIP) which exploit the large impact parameter values associated with B-hadron decays; and reconstructing displaced vertices (SV1 and JetFitter). The analysis focuses on Higgs decaying to b/c-quarks and so it is crucial for the analysis to distinguish between b/c flavour jets and τ/light flavour jets. The DL1r algorithm produces a score for each jet and by setting limits on this score the analysis creates working points (WPs), shown in table 6.2. For the resolved regime and the $VH(\to c\bar{c})$ regime the so-called pseudo-continuous flavour tagging scheme (PCFT) is implemented splitting jets into different regions dependent on DL1r

PCFT bin	bin name	DL1r$_c lim_{low}$	DL1r$_c lim_{up}$	DL1r$_b lim_{low}$	DL1r$_b lim_{up}$
0	untagged	-	0.6735	-	3.245
1	c-loose	0.6735	1.76	-	3.245
2	c-tight	1.76	-	-	3.245
3	b-70	-	-	3.245	4.565
4	b-60	-	-	4.565	-

Table 6.2: Table of PCFT bins and the lower and upper limits imposed on the DL1r scores for c and b tagging denoted by DL1r$_{c/b} lim_{low/up}$(from internal note).

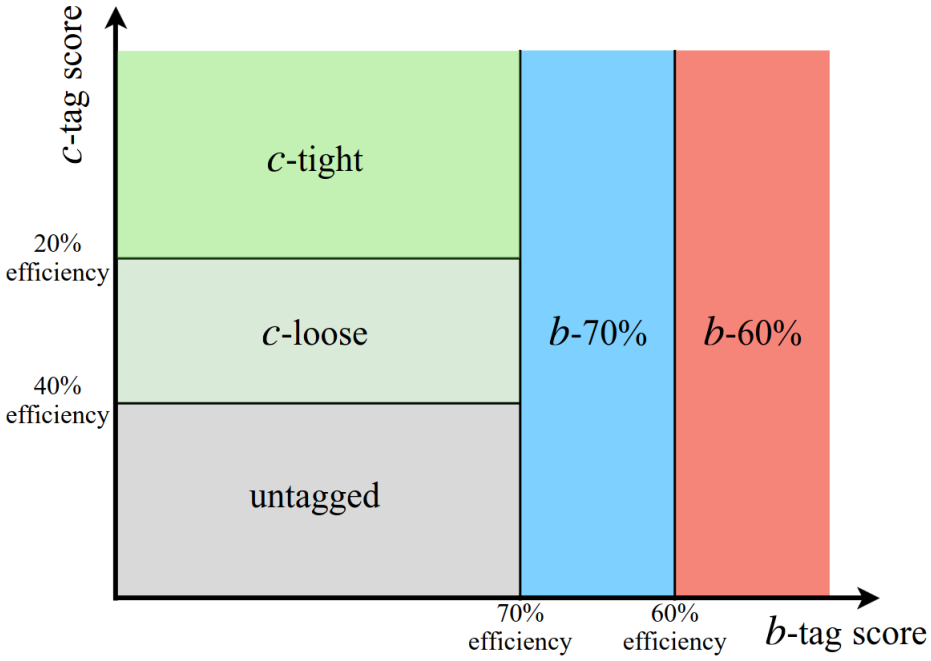


Figure 6.2: Figure displaying the different flavour tagging regions on an axis of the DL1r scores for c and b tagging.

scores/WPs (shown in figure 6.2). This firstly checks if a given jet is b-tagged with two WPs at 60% and 70%. If the jet doesn't pass these requirements it is then checked with the c-tagging WPs which are derived within the context of the analysis. Namely, there are *loose* and *tight* c-tagging working points with an efficiency of 20% with additional rejection mechanisms for b-jets and light-flavour jets in the *tight* classification. In the boosted analysis the jet flavour tagging is done slightly differently using the pseudo-continuous b-tagging method (PCBT) using the DL1r algorithm [57]. The sub-jets in the large-R jet are the objects that are being tagged. The analysis uses the WP at 85% tagging efficiency, b-jets then fall into four bins based on the "quality" of the tag. The PCBT method is split into the following bins $< 60\%$, $60 - 70\%$, $70 - 77\%$, and $77 - 85\%$ which will be some of the key variables used later on in the MVA training.

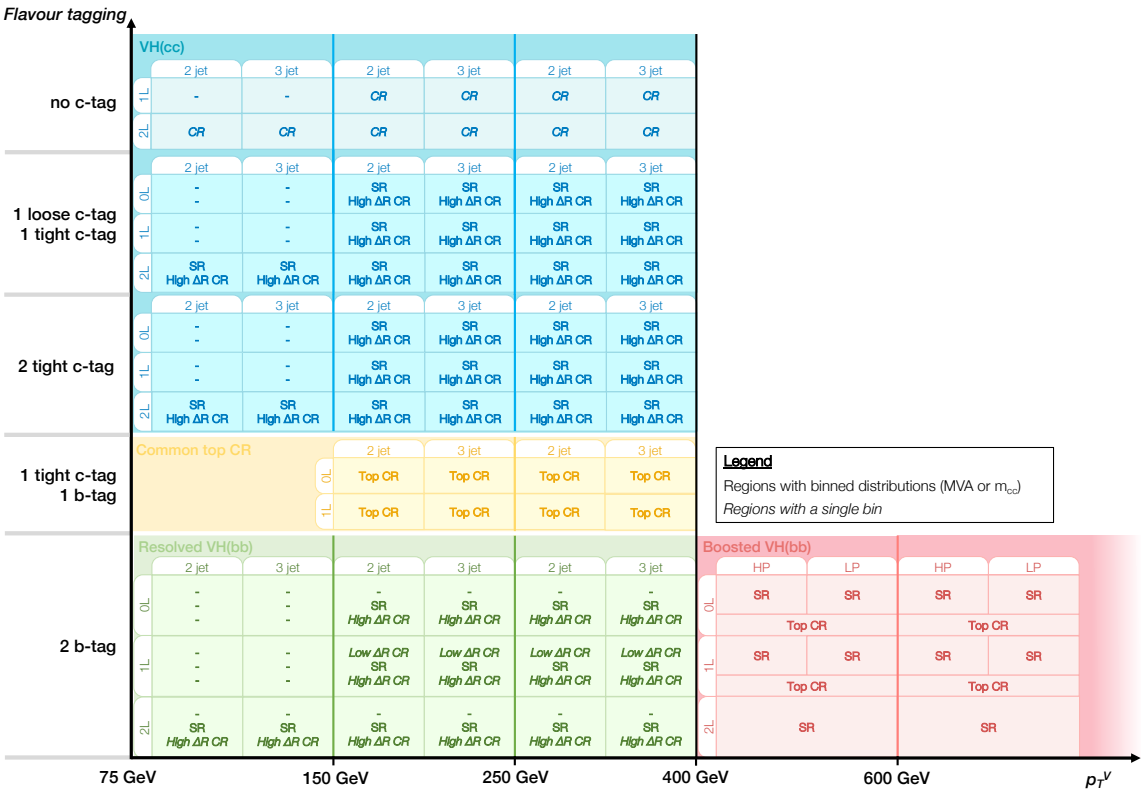


Figure 6.3: Figure showing the analysis tagging strategy regions for the $VH \to bb(cc)$ analysis. (From internal note)

Event selection

Events are selected and put into signal regions (displayed in figure 6.3). There are three lepton channels for events with 0, 1 or 2 reconstructed charged leptons. These channels search for the $ZH \to \nu\nu b\bar{b}$, $WH \to l\nu b\bar{b}$ and $ZH \to llb\bar{b}$ signals, respectively, with each channel requiring exactly two b-tagged jets. In the resolved analysis, events are also split into 2-jet and 3-jet regions. In the 0 and 1-lepton channels the 3-jet region allows only one additional jet (no 4-jet events), whereas, in the 2-lepton channel the 3-jet region allows any event with three or more jets. This is to account for the large $t\bar{t}$ background in events with more than three jets in the 0 and 1-lepton channels. The resolved strategy splits the 2-lepton channel p_T^V regions into: $75 < p_T^V < 150\,\text{GeV}$, $150 < p_T^V < 250\,\text{GeV}$ and $250 < p_T^V < 400\,\text{GeV}$. The 0 and 1-lepton channels do not contain the low-p_T^V region ($75 < p_T^V < 150\,\text{GeV}$), due to the improving signal-to-background ratio with increasing p_T^V.

The boosted strategy has the same three lepton channels as before and looks for the same decay signals but at higher p_T^V. The regions are split into $400 < p_T^V < 600\,\text{GeV}$ and $p_T^V > 600\,\text{GeV}$. There must be at least one large-R jet with 2 b-tagged track jets inside and zero b-tagged jets outside the large-R jet. If there are no other jets outwith the large-R jet

then these events are considered the *high purity signal region* (HP SR); if there are additional jets in the event then it's classified as the *low purity signal region* (LP SR). The HP and LP regions only exist in the 0 and 1-lepton channels. The 0 and 1-lepton channels also add a $t\bar{t}$ CR, this is defined as having at least one additional b-tagged jet outside the large-R jet.

6.1.3 Systematic uncertainties

There are four sources of systematic uncertainty to consider: experimental uncertainties, errors in the modelling of simulated background processes and errors in simulating the VH $H \to b\bar{b}$ signal samples, and those related to estimating the multi-jet background.

Experimental uncertainties are unavoidable and arise from the inability to obtain perfect information from an event. The dominant sources of uncertainty come when reconstructing jets. Jets can be tricky to identify and measure their exact energy. The b-tagging algorithm, as well as c-tagging and light-flavour-tagging algorithms, have an efficiency calculated by testing them on simulated sample jets with known flavours. The b-tagging algorithm has a tuned efficiency of 70%, with a chance to misidentify a light-flavour jet (c-jet) of 0.3% (12.5%). The uncertainties on the tagging efficiencies are 2% for b-jets, 10% for c-jets and 40% for light flavoured jets. There are additional uncertainties when the b-tagging algorithm is used on jets with $p_T > 300\,\text{GeV}$ as the misidentification of hadronically decaying τ-leptons must be considered. Further experimental uncertainties arise from the reconstruction, identification, isolation and trigger efficiencies of muons and electrons, although the contribution of these errors have a smaller impact than the jet tagging uncertainties. Since the missing energy calculation requires the negative vector sum of every reconstructed object in the event, any errors on these objects are propagated through as a total error on the E_T^{miss}.

There are three types of modelling uncertainty: normalisations, acceptance differences across different analysis regions and the shapes of differential curves of important kinematic variables. I studied the diboson modelling for the $VH(\to b\bar{b})$ analysis; this is discussed later in section 6.4. The normalisation uncertainty is calculated at the same time as the nominal value using a MC generator. The uncertainty on regional acceptance differences can be derived from comparisons between data in control regions, or between the regular simulated samples and alternative simulated samples. Uncertainties in the shapes of m_{bb} and p_T^V curves are considered as they effectively control the overall shape variation of the BDT_{VH} discriminant.

Simulated V+jets samples are split into three categories: vector boson with associated

heavy flavour jets (V+HF) which includes $V+bb$, $V+bc$, $V+bl$ and $V+cc$; and the remaining categories are $V+cl$ and $V+ll$. The last two categories account for less than 1% of the total background; the largest contribution comes from the V+HF category. $t\bar{t}$ normalisation and acceptance uncertainties are calculated separately for the 0/1-lepton regions and the 2-lepton region, this is due to the fact that this process has very different signatures when looking at the different lepton channels. In the 0 and 1-lepton channels the normalisation ratio uncertainties are considered for the 2-jet and 3-jet categories and the W+HF control and signal regions. In the 2-lepton channel the normalisations are left as floating points and are later adjusted using a comparison with the $e\mu$ control regions. Single top-quark production takes the t-channel and Wt-channel and derives uncertainties on the normalisations of these processes, the acceptance and the shapes of the m_{bb} and p_T^V distributions. Diboson production is made up of WZ, WW and ZZ events, WW has a very small contribution to the background ($< 0.1\%$). The normalisation uncertainties and relative acceptance errors are derived for WZ and ZZ production events.

Modelling uncertainties in the signal samples come from the calculation of the VH cross-section and the branching ratio uncertainty of the $H \to bb$ decay. These uncertainties are assigned following recommendations from the LHC Higgs Cross Section Working Group.

Uncertainties in the calculation of the multi-jet background can impact the 1-lepton channel in two ways: they can change the shape of the m_T^W distribution which would alter the normalisation values, or they can directly impact the multi-jet BDT distributions affecting the global likelihood fit. The uncertainties discussed in this section are implemented into the analysis during the processes of statistical analysis (described in section 6.1.5) and modelling (described in section 6.4).

6.1.4 Multivariate analysis

Multivariate analysis (MVA) techniques are ubiquitous in the field of particle physics. With a set of discriminating input variables one can train an MVA to be very effective at maximising signal retention while also reducing background leading to better significance values than other analysis techniques. The MVA algorithm is fed a set of input variables and trained on samples, learning the structure of signal events, to create a one dimensional discriminator that classifies events into signal-like or background-like categories. The analysis previously used MVA methods on the resolved analysis only, so one improvement that could be made to the analysis was to introduce this method in the boosted analysis. Boosted decision trees (BDT)

provided by the TMVA package of ROOT is the MVA technique used in this analysis [58].

Boosted decision trees

The BDT method splits the total data sample into two independent folds - one for training and one for testing. These folds are usually created by adding odd numbered events to one fold and even numbered events to the other, but other ways of splitting events also exist i.e. one can have a 70/30 split in the training and testing folds. A number of cuts are made on the variables fed into the training, the samples are split into signal-like and background-like bins. BDTs are iterative, so a misclassification value is noted after each iteration of the training. This value counts the number of events that ended up in the wrong bin i.e. signal events in the background-like bin or background events in the signal-like bin. The iterative process continues until misclassification is minimised and the BDT stabilises.

A BDT typically denotes a 'forest' of trees which classify events into signal and background leaves. The forest usually contains hundreds of trees, the trees make cuts on variables to split events into leaves. An example of a tree is shown in figure 6.4. The root node contains all of the events which then get split into sub nodes by applying a cut on a variable. More cuts are applied until the events are split into leaves which ideally contain only signal or background events. A boosting method is applied to calculate the missclassification of events in the leaves and re-weights events accordingly. The next tree is then made using these new weights and the boosting method is used again up until we reach the total number of trees that the user has defined. This re-weighting of events is cumulative. Finally, an ensemble average of all the trees is used to classify the events using their new weights.

The BDT can be adjusted by altering the hyperparameters. These are settings that control how the BDT functions. The hyperparameters are optimised for each unique data set input to the BDT. Hyperparameter optimisation is discussed later in section 6.3.8 in the context of the 2L boosted $VH(\rightarrow b\bar{b})$ MVA: one method of optimising is by adjusting each hyperparameter to deduce which of them have the most impact on the BDT output and then examining different values of the high impact hyperparameters to maximise the BDT output significance.

MVA training regions

In this analysis, MVA training is split up for each analysis regime (resolved, boosted and $VH(\rightarrow c\bar{c})$) into different lepton channels (0-lepton, 1-lepton, and 2-lepton) and sometimes

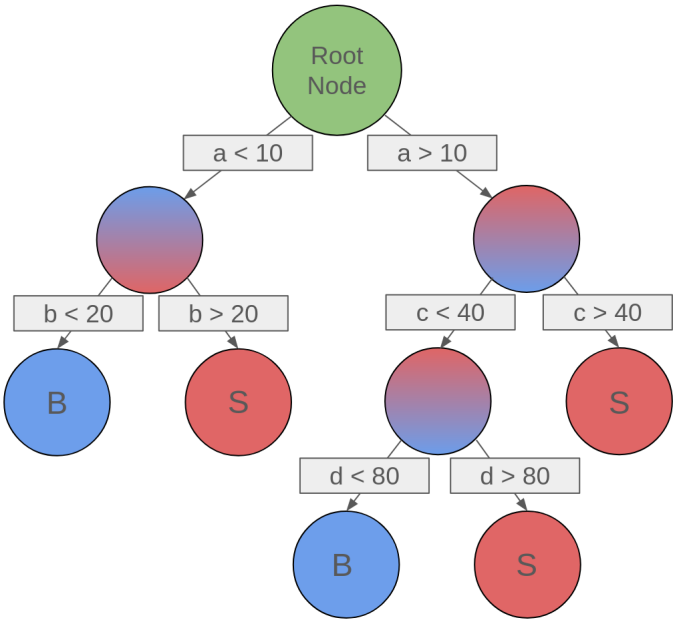


Figure 6.4: Diagram of an example BDT single tree with a depth of 3 making cuts on variables a, b, c and d. The leaves are denoted by the red and blue circles with an S for signal and B for background.

split even further i.e. based on number of jets in a event. The resolved $VH(\rightarrow b\bar{b})$ analysis and $VH(\rightarrow c\bar{c})$ analysis follow the same layout: MVA training regions for all three lepton channels are split at 150 GeV ($75 < p_T^V < 150$ GeV, $p_T^V > 150$ GeV) with separate training regions for 2-jet and 3-jet events, occasionally a $\geq$ 4-jet training as well. All training regions are shown in figure 6.5. The boosted $VH(\rightarrow b\bar{b})$ analysis has a single $p_T^V > 400$ GeV region for each lepton channel with a $\geq$ 2-jets requirement. Although the resolved and boosted topologies are split at 400 GeV in the p_T^V phase space this cut is not applied in the MVA training and is instead introduced in the evaluation stage of the MVA.

6.1.5 Statistical analysis

The likelihood function $\mathcal{L}(\mu, \theta)$ is the basis of the statistical analysis method. It is derived from the product of Poisson probability terms over the bins of the input distributions, shown in equation 6.2 - where N_i, s_i and b_i are the total number of measured events, the predicted number of signal events, and the predicted number of background events from simulations in bin i, respectively. The signal strength, μ, and the systematic uncertainties, θ, are the parameters of the likelihood function. The signal strength is the ratio between measured $\sigma \times BR$ and the theoretical value predicted by simulation.

$$\mathcal{L}(\mu) = \prod_{i \in bins} \text{Pois}(N_i|\mu s_i + b_i) = \prod_{i \in bins} \frac{(\mu s_i + b_i)^{N_i}}{N_i!} e^{-(\mu s_i + b_i)} \tag{6.2}$$

Figure 6.5: Figure showing the MVA training regions for the $VH \to bb(cc)$ analysis in terms of the p_T^V region, lepton channel, jet-multiplicity and analysis topology. For $VH(\to c\bar{c})$ the letters T, L and N refer to the quality of the c-tagged jets where T is tight c-tag, L is loose c-tag and N is not c-tagged. (From internal note)

$$\mathcal{L}(\theta) = \prod_{\theta \in \vec{\theta}} \frac{1}{\sqrt{2\pi}} e^{-(\theta^2/2)} \tag{6.3}$$

The systematic uncertainties are introduced in the form of nuisance parameters (NP), $\vec{\theta}$ which are modelled as Gaussian distributions centered at 0 and with a variance of 1, shown in equation 6.3. NP's are parameters that are not of direct interest but need to be accounted for in the fit. They are incorporated by modifying the expected signal and background yields: $(s_i, b_i) \rightarrow (s_i(\vec{\theta}), b_i(\vec{\theta}))$. These NP's are given by ATLAS groups dedicated to calculating all the systematic uncertainties that arise from the detector/experimental setup. There are other NP's that are unknown pre-analysis but are calculated solely from data, for example using control regions, and are denoted by $\vec{\tau}$. The last type of uncertainty included in the fit originates from the lack of statistics in certain regions and are labeled as $\vec{\gamma}$, these modify the number of expected background events in a bin, i: $b_i(\vec{\theta}) \rightarrow \gamma_i b_i(\vec{\theta})$. The likelihood function for background statistical uncertainties takes the form:

$$\mathcal{L}_{BkgStat}(\gamma) = \prod_{i \in bins} \text{Gauss}(\beta_i | \gamma_i \beta_i, \sqrt{\gamma_i \beta_i}) \tag{6.4}$$

where,

$$\beta_i = \frac{1}{\sigma_{rel}^2} \tag{6.5}$$

and σ_{rel} is the relative statistical uncertainty on the predicted number of background events. So, the complete likelihood function can be written as the product of equations 6.2, 6.3 and 6.4 as shown below:

$$\mathcal{L}(\vec{\mu}, \vec{\theta}, \vec{\gamma}, \vec{\tau}) = \prod_{i \in bins} \text{Pois}(N_i | \vec{\mu} s_i(\vec{\theta}) + b_i(\vec{\theta}, \vec{\gamma}, \vec{\tau})) \times \prod_{\theta \in \vec{\theta}} \frac{1}{\sqrt{2\pi}} e^{-(\theta^2/2)} \times \prod_{i \in bins} \text{Gauss}(\beta_i | \gamma_i \beta_i, \sqrt{\gamma_i \beta_i}) \tag{6.6}$$

The complete likelihood function can be maximised to determine the parameters of interest and their uncertainties. The likelihood function is converted to the log likelihood function as it is easier to calculate. To test a specific hypothesis for a fixed value of $\vec{\mu}$, we can create the profile likelihood ratio, $\lambda(\vec{\mu})$, displayed in equation 6.7, where $(\hat{\hat{\vec{\theta}}}, \hat{\hat{\vec{\gamma}}})$ are the values that maximise the likelihood function for the fixed value of $\vec{\mu}$ and $(\hat{\vec{\mu}}, \hat{\vec{\theta}}, \hat{\vec{\gamma}})$ maximise the likelihood function irrespective of the selected μ-value.

$$\lambda(\vec{\mu}) = \frac{\mathcal{L}(\vec{\mu}, \hat{\hat{\vec{\theta}}}, \hat{\hat{\vec{\gamma}}})}{\mathcal{L}(\hat{\vec{\mu}}, \hat{\hat{\vec{\theta}}}, \hat{\hat{\vec{\gamma}}})} \tag{6.7}$$

A test statistic, q_μ, is defined from the profile likelihood ratio as:

$$q_{\vec{\mu}} = \begin{cases} -2 \ln \lambda(\vec{\mu}) & \hat{\mu} \geq \mu \\ 0 & \hat{\mu} < \mu \end{cases} \tag{6.8}$$

One can define the *p-value*, the probability that the observed data is incompatible with a given hypothesis. If we consider the null hypothesis ($\mu = 0$) i.e. observing signal by pure chance assuming the standard model was the incorrect underlying theory, then the equation for the *p-value* takes the form:

$$p_{\vec{\mu}} = \int_{q_{\vec{\mu},obs}}^{\infty} f(q_{\vec{\mu}}|\mu = 0) dq_{\vec{\mu}} \tag{6.9}$$

where the probability density function of $q_{\vec{\mu}}$ is denoted by $f(q_{\vec{\mu}}|\mu = 0)$ and $q_{\vec{\mu},obs}$ is the observed value of $q_{\vec{\mu}}$ from data. Finally, one can calculate the significance, Z, of the signal i.e. the number of Gaussian standard deviations above the background using the $p - value$:

$$Z = \Phi^{-1}(1 - p) \tag{6.10}$$

where Φ^{-1} is the inverse Gaussian cumulative distribution function. The background samples are calculated using MC templates.

6.2 Overlap and inclusive Higgs studies

Initial studies that were contributed to the VH($H \rightarrow bb$) analysis include quantifying the overlap between the resolved and boosted analysis strategies. Initially, there was no limit set to the maximum p_T^V of an event passing the resolved selection, so many events ended up passing selection for both boosted and resolved strategies and there was a large overlap due to this. Another study was to check the top quark related Higgs production modes ($t\bar{t}H$) to see if they would be a significant background process in the analysis.

6.2.1 Overlap study

There existed an overlap between the two analysis strategies (boosted and resolved). The initial motivation to split analysis regimes was for the resolved strategy to find low-p_T^V events

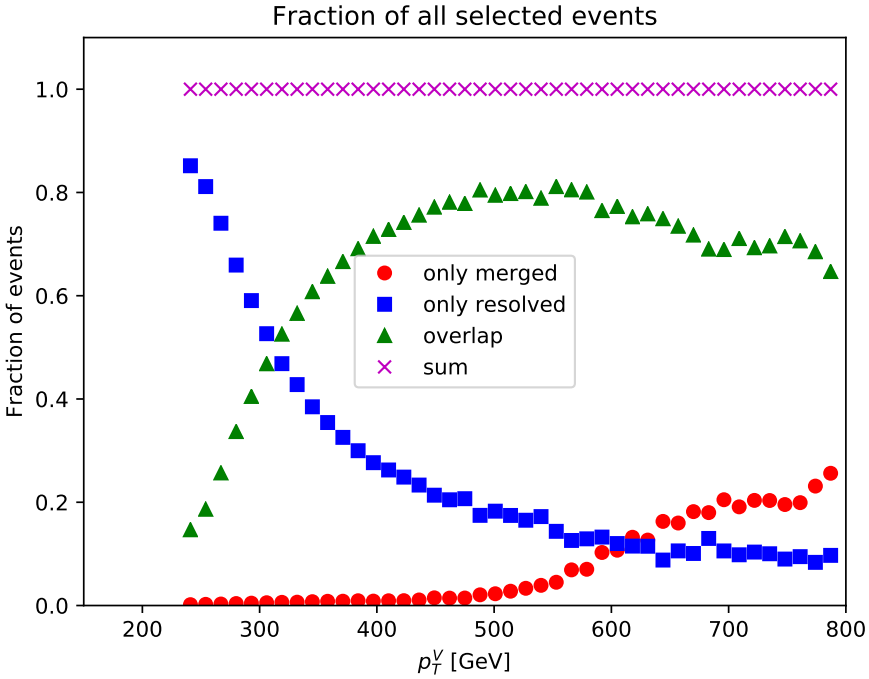


Figure 6.6: The fraction of signal MC events in the 2-lepton channel that are selected by the resolved analysis only (blue squares), merged analysis only (red circles) and both the resolved and merged analyses (green triangles) distributed subject to the event p_T^V. The purple crosses denote the sum of all the selected events which adds up to one across the whole range, showing that every event is accounted for in this plot in the range $p_T^V > 250$ GeV.

with well separated small-R b-jets stemming from a soft Higgs [42], whereas the boosted strategy (also referred to as the merged strategy) was to search for a single large-R jet with two b-tagged track jets inside decaying from a high-p_T Higgs [44]. The issue came when trying to combine the two analyses as there was trouble merging the two strategies due to the way jets were classified in each.

The overlap must be understood for each region of the analysis (all three lepton channels, for 2-jet and 3-jet regions). Figure 6.6 shows the p_T^V distribution of signal events simulated by MC in the 2-lepton channel. As can be seen in the plot there is a large overlap as we increase in p_T^V above 300 GeV, and an increase in the fraction of events picked up by the boosted analysis is only really seen above a p_T^V of around 500 GeV. Figure 6.7, shows the same plot but for data events taken during Run 1 and Run 2 of the LHC, the overall trend is similar but with a smaller total overlap and a much smaller contribution from events selected only by the merged analysis. The tail end of the plots are diffuse due to the lack of statistics at those higher p_T^V ranges. The overlap was smaller for actual LHC data analysis when compared to the MC generated samples but it still posed an issue. So in our "legacy" analysis, we tried different approaches, but in the end we decided to use the resolved (boosted) approach for

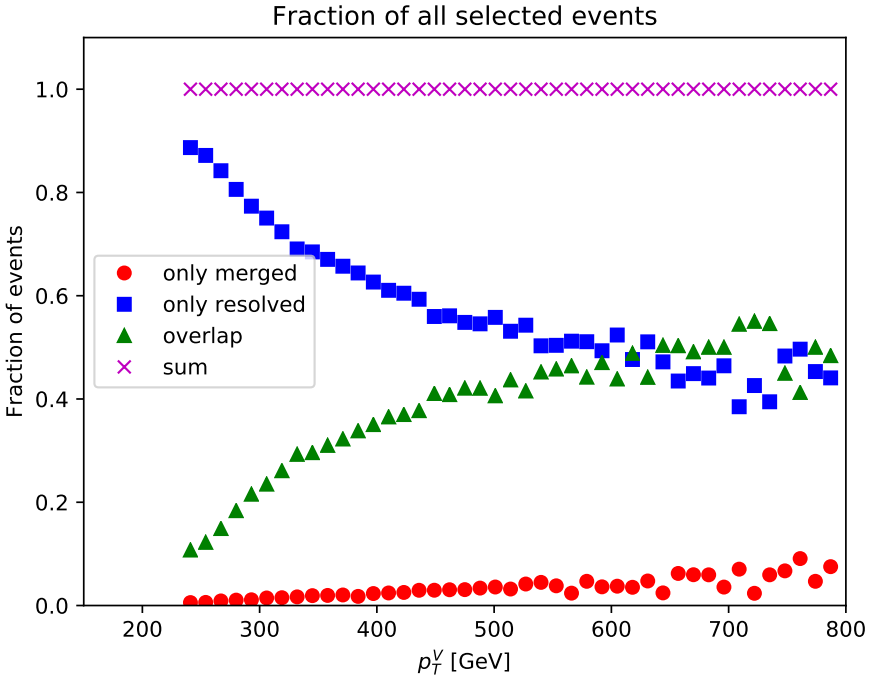


Figure 6.7: The fraction of Run1 + 2 data events in the 2-lepton channel that are selected by the resolved analysis only (blue squares), merged analysis only (red circles) and both the resolved and merged analyses (green triangles) distributed subject to the event p_T^V. The purple crosses denote the sum of all the selected events which adds up to one across the whole range, showing that every event is accounted for in this plot in the range $p_T^V > 250$ GeV.

$p_T^V < (>) 400\,\text{GeV}$.

6.2.2 Inclusive Higgs samples study

Another study was conducted for boosted and resolved analysis strategies in each p_T^V region testing five inclusive Higgs samples, these are ZH, WH (with the Z and W bosons decaying leptonically), $t\bar{t}H$ (with the $t\bar{t}$ pair decaying dileptonically, semi-leptonically and hadronically). The Higgs boson can decay to any of its available decay channels, hence inclusive. The motivation for this study was to achieve a better understanding of the possible contamination from a separate Higgs signal channel ($t\bar{t}$) and study the combination of the various Higgs boson decay channels. The ZH, WH and $t\bar{t}H$ samples were generated using Pythia8 corresponding with Run 2 conditions. The tests were done over all three lepton channels using both the resolved and boosted strategies in the $VHbb$ framework taking plots only from events with 2 b-tagged jets.

Figures 6.8, 6.9 and 6.10 show plots of the large-R jet invariant mass, m_J, distributions for all three lepton channels of the boosted analysis strategy in the $250-400\,\text{GeV}$ and $> 400\,\text{GeV}$ p_T^V regions. These plots are dominated by the ZH and WH samples as these include the

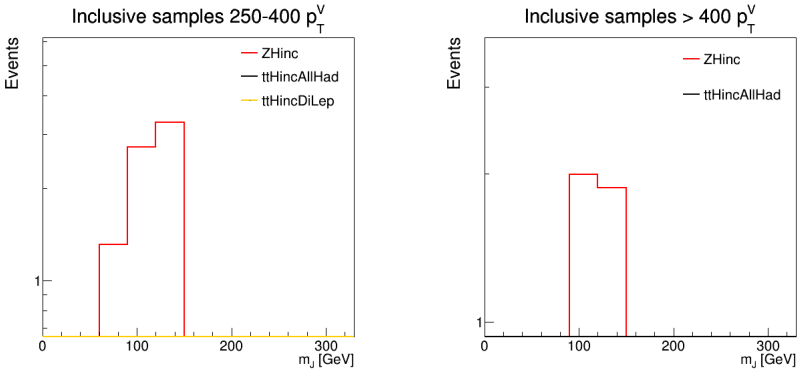


Figure 6.8: Plots show the m_J distribution (mass of the large-R jet) of the inclusive samples in the 2-lepton region of the boosted analysis for the two p_T^V sub-regions.

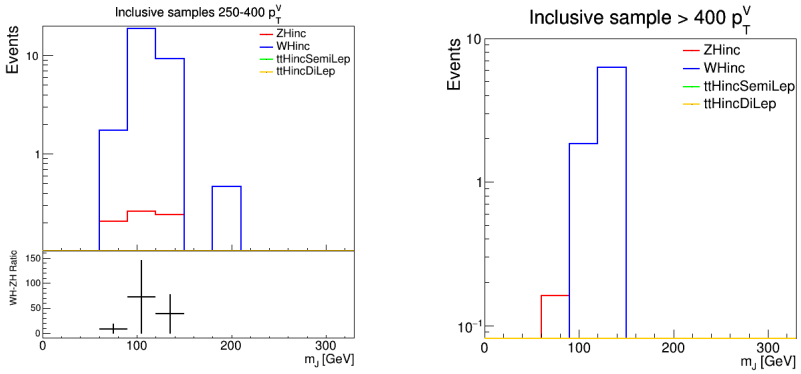


Figure 6.9: Plots show the m_J distribution of the inclusive samples in the 1-lepton region of the boosted analysis for the two p_T^V sub-regions.

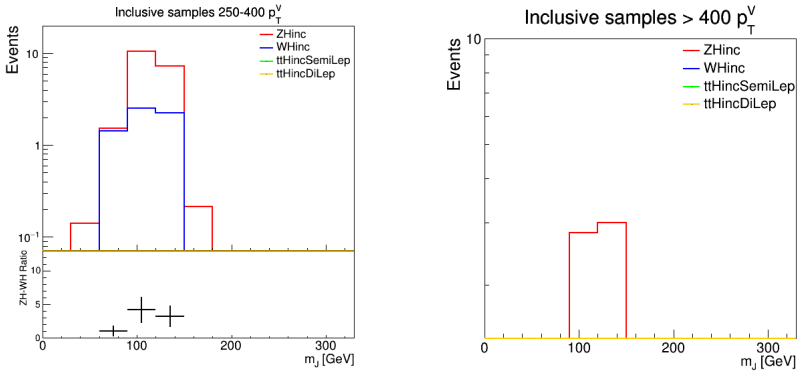


Figure 6.10: Plots show the m_J distribution of the inclusive samples in the 0-lepton region of the boosted analysis for the two p_T^V sub-regions.

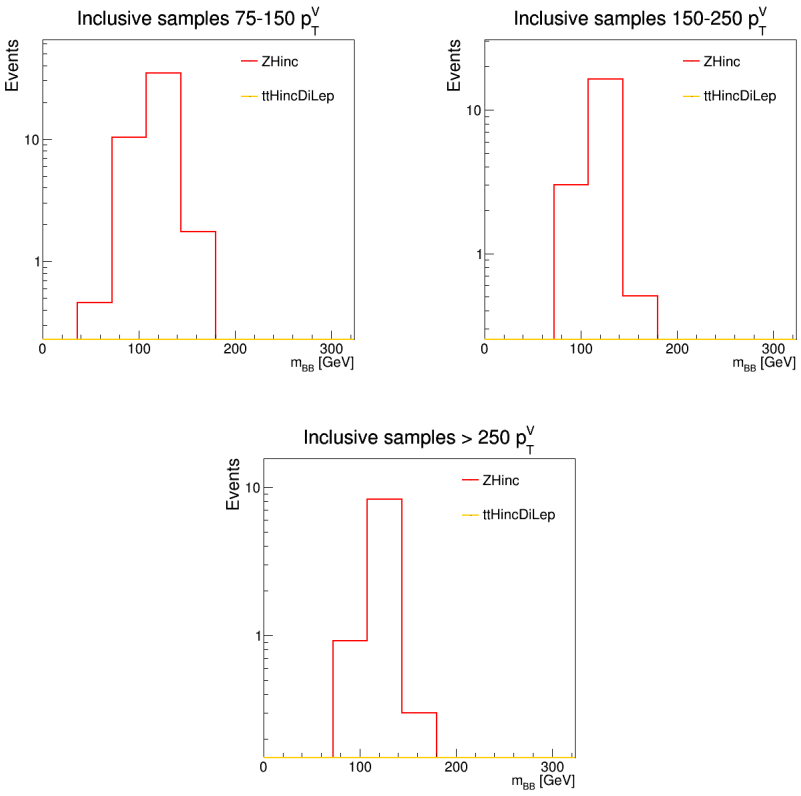


Figure 6.11: Plots show the $m_{(BB)}$ distribution of the inclusive samples in the 2-lepton region of the resolved analysis for the three p_T^V sub-regions.

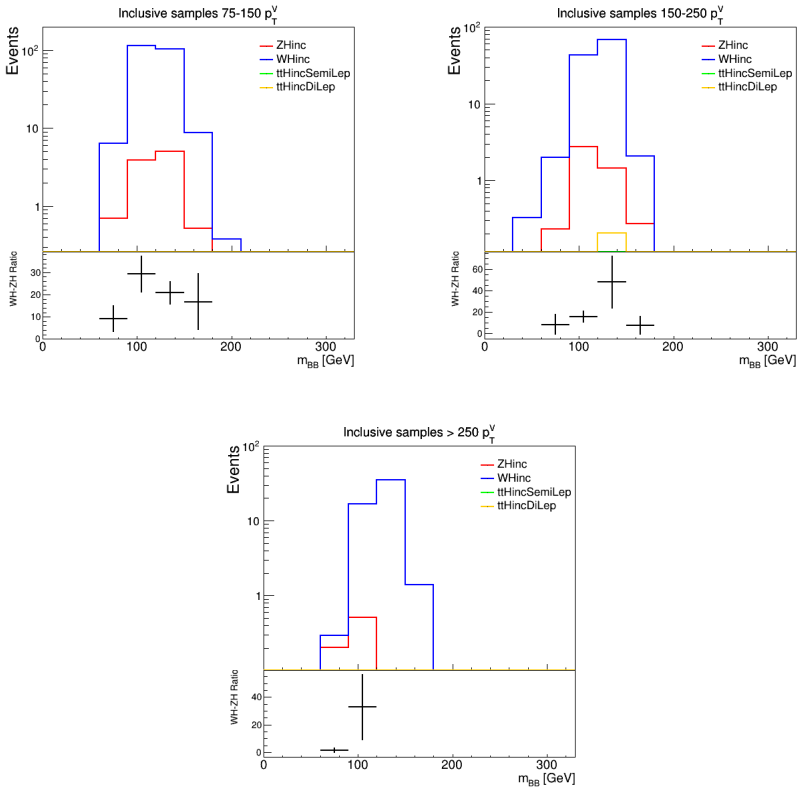


Figure 6.12: Plots show the $m_{(BB)}$ distribution of the inclusive samples in the 1-lepton region of the resolved analysis for the three p_T^V sub-regions.

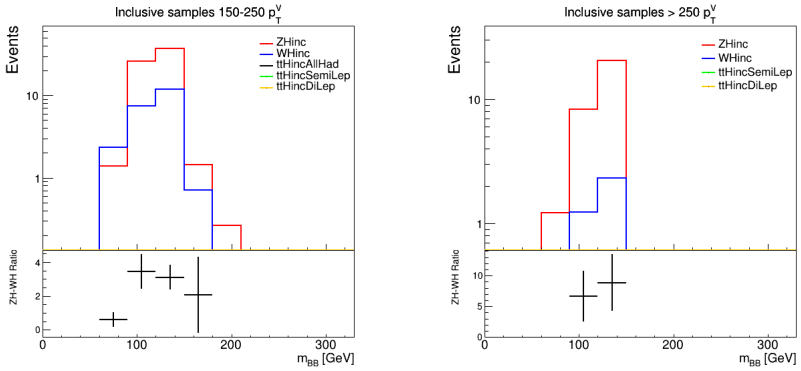


Figure 6.13: Plots show the $m_{(BB)}$ distribution of the inclusive samples in the 0-lepton region of the resolved analysis for the two p_T^V sub-regions.

signal process. The 0 and 2-lepton channels are dominated by the ZH samples and the WH samples dominate the 1-lepton region due to the leptonic decays of the Z and W. These plots show very little contribution from the $t\bar{t}H$ samples and so the $t\bar{t}H$ background is assumed negligible in the $VH(\rightarrow b\bar{b})$ analysis.

Figures 6.11, 6.12 and 6.13 show the distribution of the reconstructed two b-tagged jets mass, m_{BB}, in the resolved analysis regime, for the $75 - 150\,\text{GeV}$, $150 - 250\,\text{GeV}$ and $> 250\,\text{GeV}$ p_T^V regions. The 0-lepton channel doesn't contain the lowest p_T^V region. These plots are dominated by the ZH and WH samples but with considerably more events than the boosted result. This is due to the fact that the boosted regime is searching for higher p_T events with a tighter decay cone of the Higgs. The $t\bar{t}H$ samples are negligible as they don't contribute to the background in either analysis strategy.

6.3 Boosted 2L MVA study

I was tasked with developing the boosted 2L $VH(\rightarrow b\bar{b})$ MVA. This task composed of mostly training the MVA for every new variable added and test variables that did not make the final analysis, as well as tuning the hyperparameters of the BDT.

6.3.1 Significance of MVA output

Significance, Z_{BDT}, values are calculated in each p_T^V region using the following equation:

$$Z_{BDT} = \sqrt{2\sum_{bins}[(S+B)ln(1+S/B) - S]} \tag{6.11}$$

where, S and B are the number of signal events and background events as classified by the

Variables	Low-p_T^V Z	High-p_T^V Z
m_J, p_T^V, $\Delta R(b_1, b_2)$	$1.83\,\sigma$	$1.31\,\sigma$
m_{LL}, lepton imbalance, ΔYVJ	$2.03\,\sigma$	$1.50\,\sigma$
$m_{VHMerged}$, $H_{TBoosted}$	$2.13\,\sigma$	$1.67\,\sigma$
$C2$, $D2$	$2.61\,\sigma$	$1.76\,\sigma$
$nTrackJets$	$2.78\,\sigma$	$2.04\,\sigma$
$p_T AddCaloJets$	$3.26\,\sigma$	$1.75\,\sigma$
$bin\_bTagBTrkJ1$	$3.17\,\sigma$	$1.76\,\sigma$

Table 6.3: Initial variable tests showing increasing significance with each row in the table adding variables on to the previous list; last few rows do not show increasing significance due to overtraining.

BDT in each bin, respectively.

6.3.2 Initial variable tests

I started with simple variables and slowly added more step by step and monitored the output significance at each stage. Here is a list of the variables used at this step:

- m_J, mass of the large-R jet
- p_T^V, transverse momentum of the vector boson
- $\Delta R(b_1, b_2)$, angular separation between the two b-quark track jets
- m_{LL}, combined mass of the lepton pair
- lepton imbalance = $(p_T^{l_1} - p_T^{l_2})/p_T^V$
- ΔYVJ, angular separation between the vector boson and the large-R jet
- $m_{VHMerged}$, combined mass of V and H in the boosted regime
- $H_{TBoosted}$, combined transverse momentum of all jets
- $C2$ and $D2$ provide information on the jet structure [59]
- $nTrackJets$, number of track jets
- $p_T AddCaloJets$, transverse momentum of the additional calorimeter jets
- $bin\_bTagBTrkJ1$, b-tagging information for track jet 1

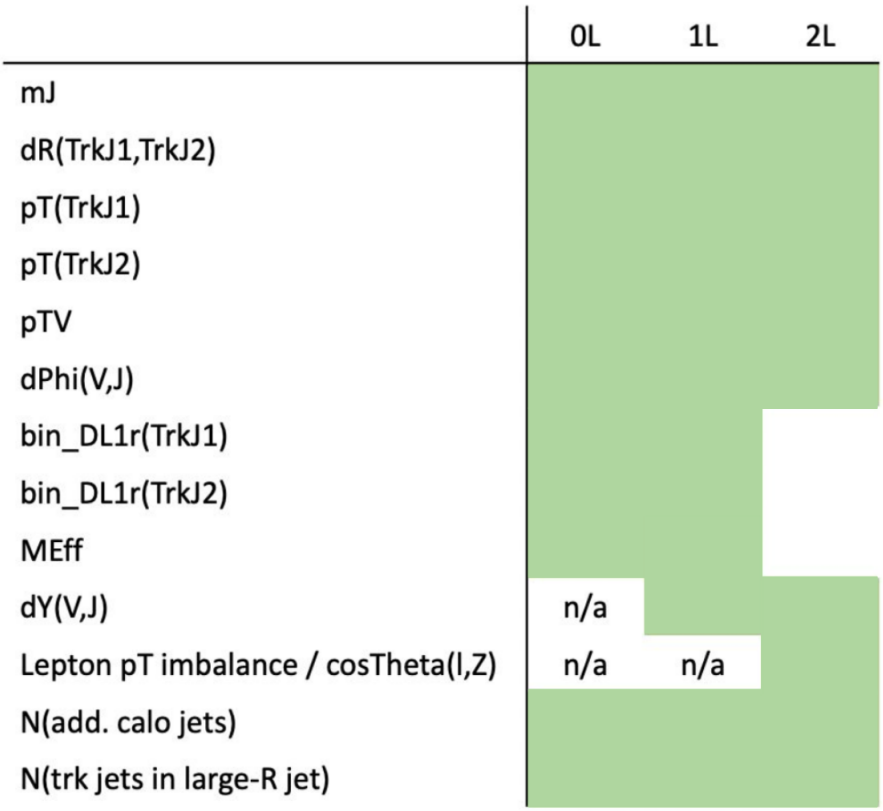


Figure 6.14: Baseline variables for each lepton channel.

As more variables are introduced into the training the significance (calculated from equation 6.11) steadily increases up to $3.24\,\sigma$ and $2.04\,\sigma$ in the low and high-p_T^V regions respectively, shown in figure 6.3. The last two rows see decreases in significance in the high-p_T^V region, this is due to a combination of overtraining and small data samples in the region. As the variables all lead to increased sensitivity, the conclusion is to include them: however, overtraining is an issue that is discussed in more detail in section 6.3.7.

6.3.3 New vs Old MC samples

The analysis was in the process of moving to a new generation of MC generated samples, with improved statistics. This study compares the new (33-05) and old (32-15) generations with four total MVA training sets - two for each sample, one with the baseline variables (figure 6.14) and one with the extended baseline variables (figure 6.15).

List of additional variables introduced in the baseline set:

- $p_T^{J_1}$, transverse momentum of the first b-tagged track jet
- $p_T^{J_2}$, transverse momentum of the second b-tagged track jet
- $\Delta\phi(V, J)$, angular separation between the vector boson and large-R jet

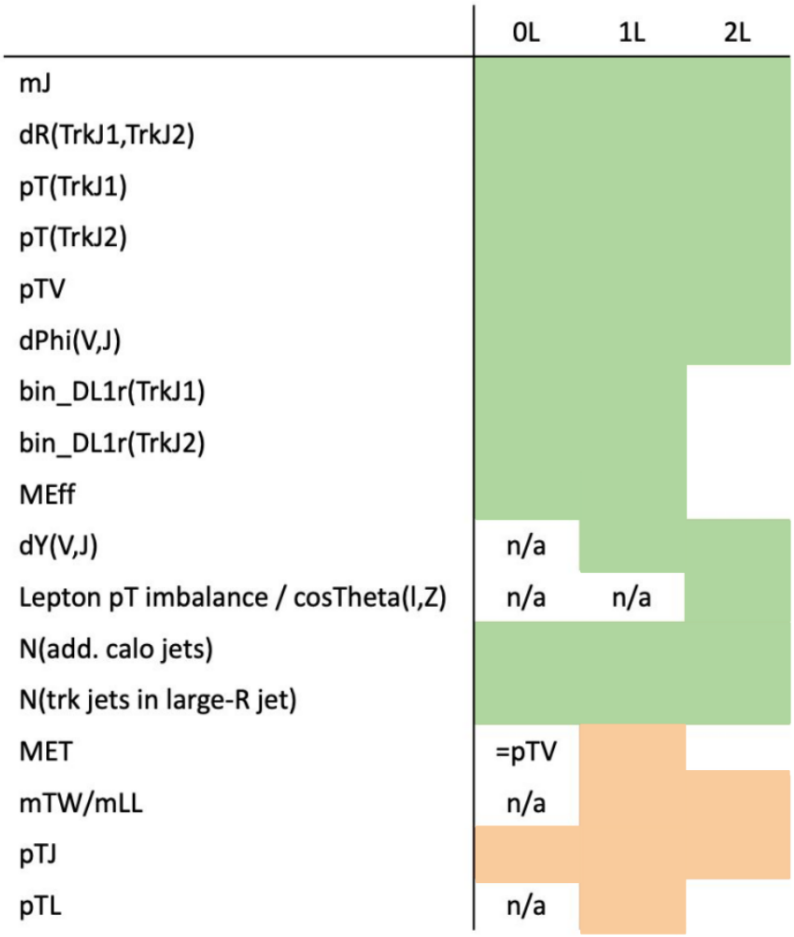


Figure 6.15: Extended baseline variables for each lepton channel.

- $N(add.calojets)$, number of additional calorimeter jets
- $N(trkjetsinlarge-Rjet)$ number of track jets in the large-R jet
- p_T^J transverse momentum of the large-R jet (extended baseline from table 3)

The green boxes represent the variables used in the MVA training and the orange boxes represent newer test variables added in attempts to boost the significance. The white boxes represent variables that were not used (or can't be calculated or are the same as another variable) in that lepton channel.

6.3.4 Variable distribution plots

Variable distribution plots are compared side by side for both the new and old samples and for all extended baseline variables. Most of the variable distributions change very little between the two samples. The plots below show the m_J distributions; this is the most discriminating variable in the analysis with very good separation between signal and background.

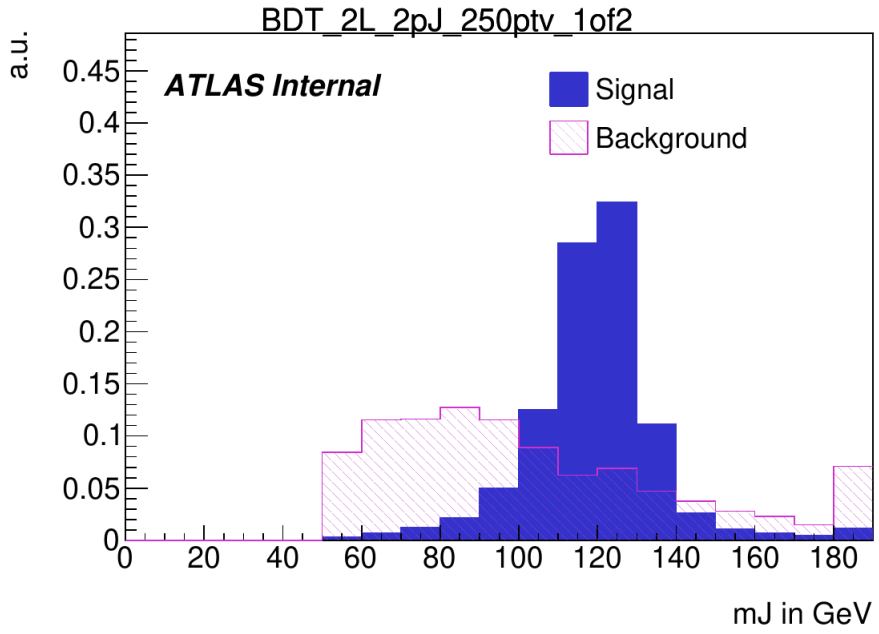


Figure 6.16: m_J distribution for 32-15 sample where the last bin is an overflow.

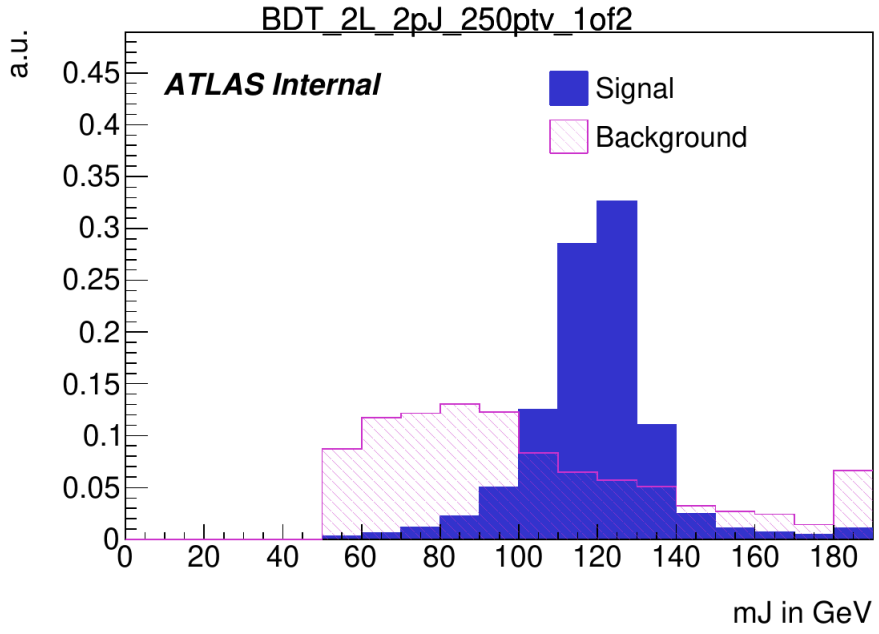


Figure 6.17: m_J distribution for 33-05 sample, compared with previous MC samples for validation.

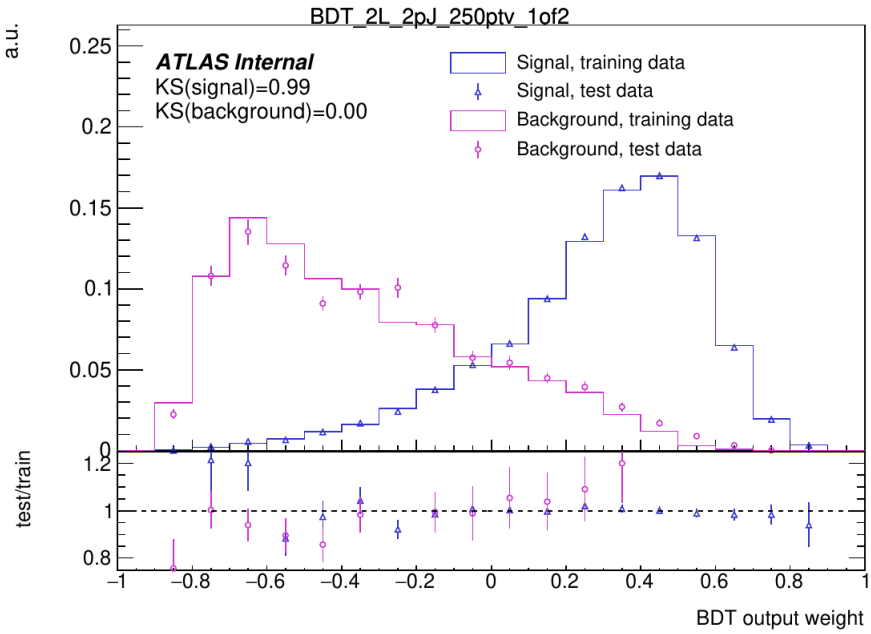


Figure 6.18: Plot of the BDT output to test for overtraining for the 32-15 MC sample, this is compared with the next two MC sample releases.

Figures 6.16 and 6.17 show the m_J distribution for 32-15 and 33-05 samples. The last bin in each plot is an overflow bin. The latter displays a flatter background distribution within the signal zone of around 125 GeV. One would expect this to slightly improve the output significance of the MVA training.

The MVA training is inclusive in p_T^V but the significance is calculated for both high- and low-p_T^V regions. The 32-15 sample significances for the baseline set of variables are $2.76\,\sigma$ in the low-p_T^V region and $1.81\,\sigma$ in the high-p_T^V region. For comparison, the 33-05 sample significances for the baseline variables are $2.86\,\sigma$ in the low-p_T^V region and $2.26\,\sigma$ in the high-p_T^V region. The MVA in both high-p_T^V and low-p_T^V regions sees an increase in significance going from the 32-15 samples to the 33-05 samples.

6.3.5 Overtraining issues

One problem many MVA's face is overtraining which can occur for many reasons. For this analysis the main reasons stem from the small amount of available training data in the high-p_T^V region. Overtraining issues have persisted across every iteration of the MVA training so far. One potential solution is to artificially inflate the statistics in the regions with the least data.

Figures 6.18, 6.19 and 6.20 are the overtraining plots for the 32-15 sample, the 33-05 sample and 33-24 sample. There are two folds in each training - which is split into training

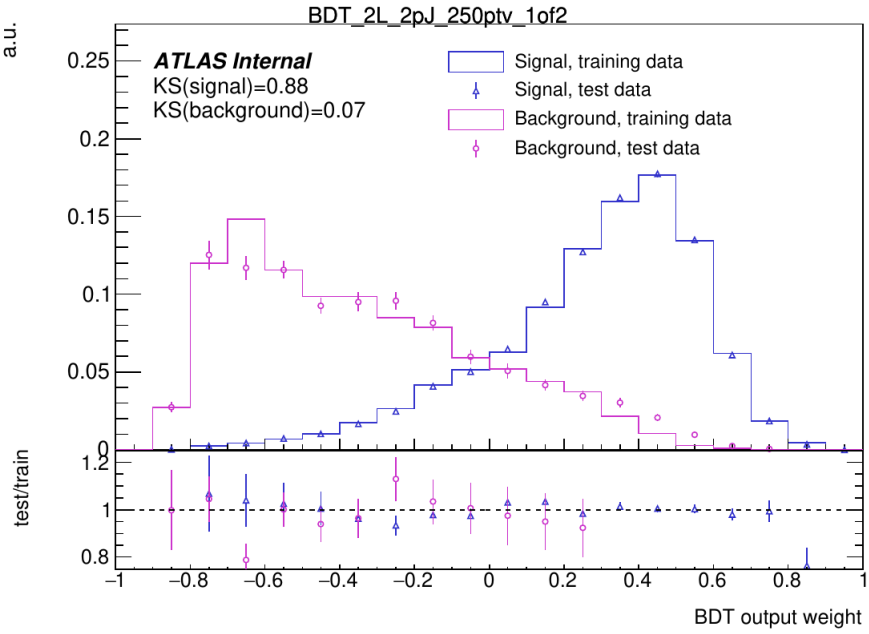


Figure 6.19: Plot of the BDT output to test for overtraining for the 33-05 MC sample, this is compared with the other two MC sample releases.

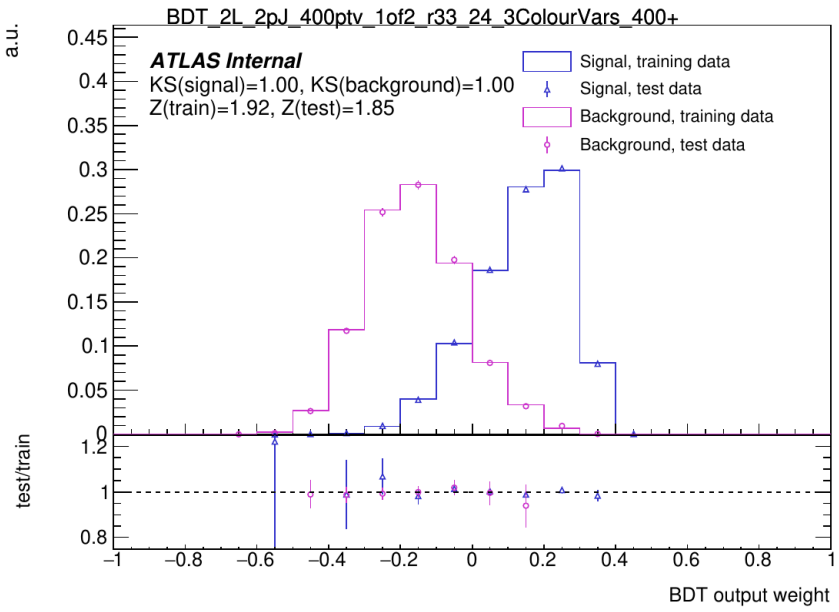


Figure 6.20: Plot of the BDT output to test for overtraining for the 33-24 MC sample, this is compared with the previous two MC sample releases.

data and test data. A perfect MVA with no overtraining would see both folds follow the same distribution exactly for both signal data and background data. The KS score is a test of how much overtraining is present - the KS score can take values between 0 and 1 with a value of 1 means there is no overtraining present. The signal data for all samples has very little overtraining but the background data struggles with this problem, although, a slight improvement can be noticed going from the old samples to the new ones.

6.3.6 Baseline variables

In the list below are the nominal set of variables used in the MVA analysis. New variables were added to the baseline variable set; these are the momentum of each track jet (p_T^J1, p_T^J2, p_T^J3), the number of track jets in the large-R jet, and the b-tagging information on the first two jets:

- m_J, mass of the large-R jet
- $\Delta R(b_1, b_2)$, angular separation between the two b-quark track jets
- p_T^J1, p_T^J2, p_T^J3, transverse momentum of the first track jet, second track jet and third track jet (if one exists)
- p_T^V, transverse momentum of the vector boson
- $abs(\Delta\Phi(V, J))$
- lepton imbalance = $(p_T^{l_1} - p_T^{l_2})/p_T^V$
- ΔYVJ, angular separation between the vector boson and the large-R jet
- $nAddCaloJets$, number of additional calorimeter jets
- $nMatchedTrkJinFJ$, number of track jets in large-R jet
- $bin\_bTagBTrkJ1$, b-tagging information for track jet 1
- $bin\_bTagBTrkJ2$, b-tagging information for track jet 2

6.3.7 Truth flavour tagging vs direct tagging

A huge problem in the two lepton channel has been overtraining in the MVA. This was mainly due to the lack of statistics in this channel which is even more evident in the boosted

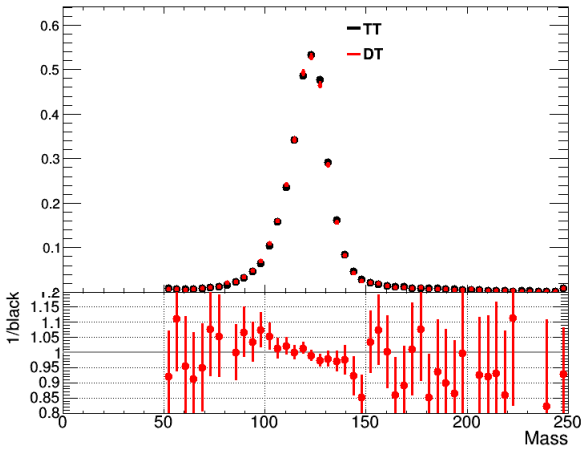


Figure 6.21: Signal distribution plot for m_J comparison between truth-flavour tagged (TT) sample and direct tagged (DT) sample in the $p_T^V > 400$GeV region.

analysis. Initailly, the analysis used a direct tagging (DT) method which only allows events with two tagged b-jets to pass selection. In order to increase the statistics of the training samples truth flavour tagging (TT) was introduced to the analysis. This technique avoids the loss in statistical precision caused by setting a cut value on the tagging scores of b/c-jets in the simulated data; instead by altering the event weight of the event by the jet tagging scores accordingly. Where direct tagging requires two tagged b-jets which drastically limits statistics, truth flavour tagging gets around this by accepting more events and assigning weights based on the probability of being tagged as a b-jet.

The truth flavour tagging algorithm used in this analysis is based on a graph neural network (GNN) created for ATLAS jet tagging. It has been demonstrated in the boosted regime and published in 2022 [60]. It uses a number of jet variables (p_T, η, ϕ and information for the leading p_T hadron in each jet) as the input features to deduce the probability of an event containing b-tagged jets or c-tagged jets in the case of the $VH(\rightarrow c\bar{c})$ analysis. The GNN is trained as inclusively as possible for the boosted analysis but had to be trained separately for each MC generator (Pythia, Sherpa, etc.) due to them having different scale factors.

Initially, the modelling of each variable in the truth flavour tagged sample had to be checked to see if it matches the previous direct tagging (DT) sample. This was done by running a comparison between the first truth flavour tagged results and the direct tagged results. These comparisons led to the discovery of some technical issues with the GNN TT sample production process; after several iterations we achieved a match in the modelling as can be seen in figures 6.21, 6.22 and 6.23.

The distributions in figures 6.21, 6.22 and 6.23 show good correlation between the sam-

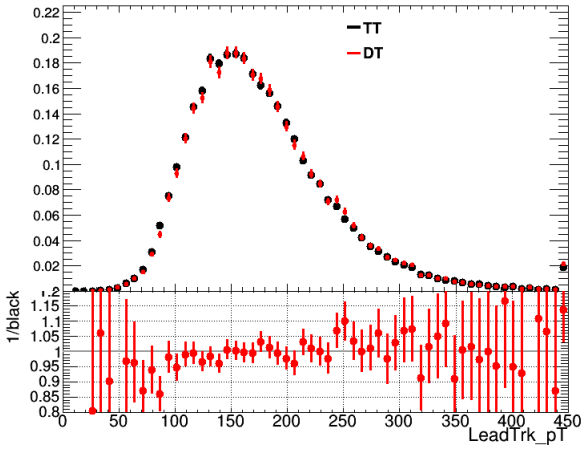


Figure 6.22: Signal distribution plot for p_T of the first track jet comparison between truth-flavour tagged (TT) sample and direct tagged (DT) sample in the $p_T^V > 400$GeV region.

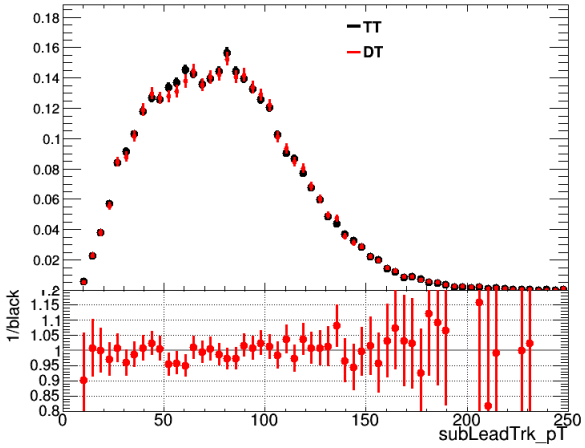


Figure 6.23: Signal distribution plot for p_T of the second track jet comparison between truth-flavour tagged (TT) sample and direct tagged (DT) sample in the $p_T^V > 400$GeV region.

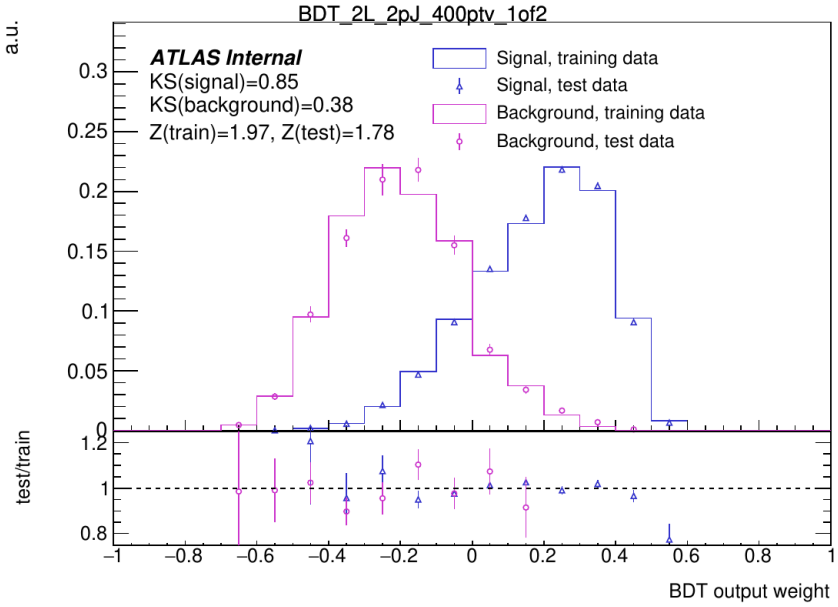


Figure 6.24: Overtraining plot for DT training in the $p_T^V > 400$GeV region.

ples, which is necessary to show that the TT is correctly replicating the DT samples. The distributions around the peak are within 5% of each other and form very similar curves. The variables in these plots are the invariant mass of the large-R jet (m_J), and the transverse momentum of the first and second track jets (p_T^1,p_T^2) which is a subset of the baseline variables used in our analysis.

The reason for switching to truth flavour tagging was to reduce overtraining in the MVA. The overtraining plots in figures 6.24 and 6.25 show a vast improvement going to TT from DT - this can be seen by comparing the KS test scores in the figures. These plots show the signal and background distributions of the BDT score with negative values (along the x-axis) representing background-like and the positive values for signal-like. We are interested in the ratio between the training and test data for both the signal and background distributions, for zero overtraining this ratio would be one. As can be seen below the DT overtraining plot (figure 6.24 has more fluctuations from unity in the ratio plot along the bottom than the TT plot in figure 6.25), therefore the overtraining is much improved by using truth flavour tagging.

The number of events in the TT samples and DT samples are compared in table 6.4. The TT samples greatly increase the available statistics for the analysis, a much needed improvement for the already low number of events in the 2-lepton boosted signal region.

Once it was decided to use truth flavour tagged samples, we had to optimise the hyperparameters in the MVA training to maximise the significance.

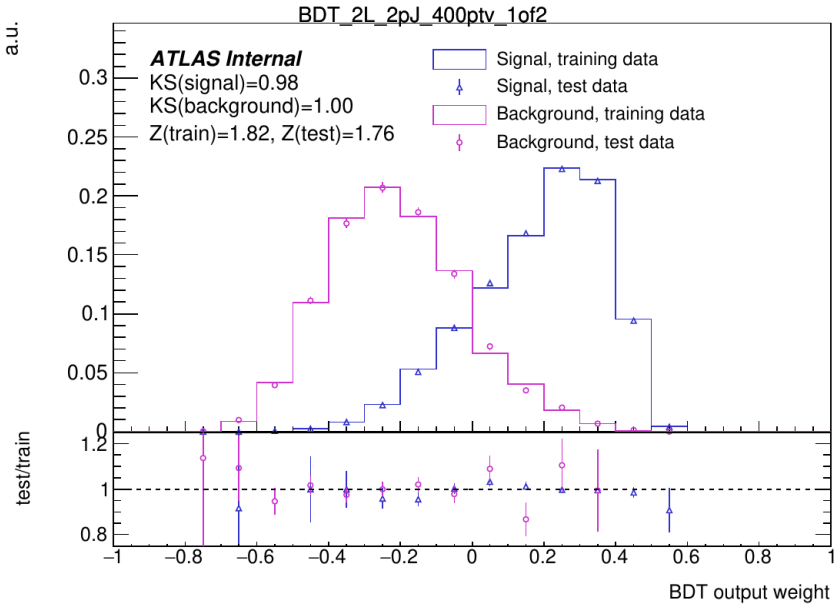


Figure 6.25: Overtraining plot for TT training in the $p_T^V > 400$GeV region.

p_T^V region	Event type	Truth flavour tagged sample	Direct tagged sample
$400 - 600$ GeV	background	1,383,307	12,852
$400 - 600$ GeV	signal	246,073	77,437
600+ GeV	background	761,054	4,536
600+ GeV	signal	111,729	33,178

Table 6.4: The number of events in 2L boosted samples with the truth flavour tagging and direct tagging methods for signal and background in the p_T^V regions $400 < p_T^V < 600$ GeV and $p_T^V > 600$ GeV.

6.3.8 Hyperparameter optimisation

Initially, the training was run with the standard TMVA hyperparameters [58]. The process of optimising these parameters took a large amount of computing power (using condor batch system). The hyperparameters and the range of values that they were optimised over are listed below:

- nTrees: $200, 500, 800$ (number of trees)
- MaxDepth: $2, 4, 6$ (maximum depth of each tree)
- BoostType: AdaBoost (type of learning algorithm)
- AdaBoostBeta: $0.1, 0.3, 0.5$ (rate of learning for the MVA)
- nCuts: $20, 50, 100$ (number of divisions in the range of each variable that we can cut on)
- MinNodeSize: $2, 5, 7$ (minimum percentage of events to make a new node (0.2%, 0.5%, 0.7%))

With the truth tagged samples using the baseline variables we scanned over all the hyperparameter values then ranked them on the output significance in the $p_T^V > 400\,\text{GeV}$ region. The training with the best hyperparameter configuration used the values: nTrees= 200, MaxDepth= 4, BoostType=AdaBoost, AdaBoostBeta= 0.3, nCuts= 20, and MinNodeSize= 5. These hyperparameters where implemented as the nominal hyperparameter settings.

One issue with optimising the hyperparameters is that whenever something new is added to the nominal samples or the MVA training (like a new variable) then the optimisation must be redone.

6.3.9 Colour ring variable testing

New variables were tested in the 2L MVA training in order to improve statistical significance. After defining a baseline set of variables, new variables were introduced such as the colour ring [61] [62]. This is a colour flow sensitive variable that can act as a colour-singlet (like $H \to b\bar{b}$) tagger and reject colour-octet (like gluon $\to b\bar{b}$) background events. The colour ring observable is constructed by taking the ratio of matrix elements for signal and background. The colour ring $\mathcal{O}$ is calculated using the angles between the b-jets (θ_{ab}), and the angle between the decaying gluon/Higgs and the subsequent b-jets (θ_{ak}/θ_{kb}) as shown below:

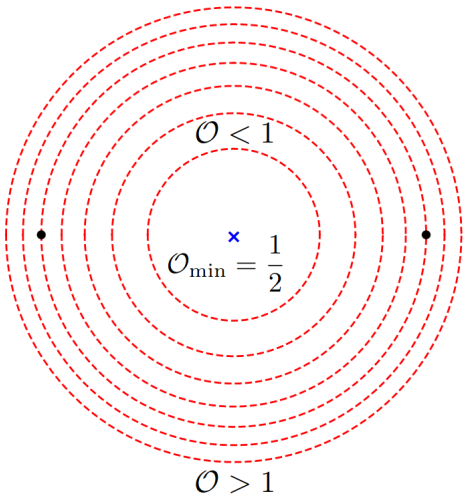


Figure 6.26: Illustration of the geometry selected for by the observable $\mathcal{O}$. The location of the final-state hard partons are denoted by the black dots, and the $\mathcal{O} = 1$ contour passes through the black dots. The region inside the $\mathcal{O} = 1$ circle has observable value less than 1 and the region outside is greater than 1. The point at the center of the circle, directly between the bottom quarks, is where the observable takes its minimum value, as labeled [62].

$$\mathcal{O} \approx \frac{\theta_{ak}^2 + \theta_{bk}^2}{\theta_{ab}^2} \tag{6.12}$$

The colour ring can take values greater than or less than 1; colour-singlet final states predominantly take values of $\mathcal{O} < 1$ and colour-octet states predominantly take values of $\mathcal{O} > 1$. Figure 6.26 displays the geometry selected for by the colour ring observable.

Figure 6.27 displays large-R jet structure distributions using colour ring, dipolarity, pull angle and D_2 observables comparing the acceptance rates for $Z + H(\to b\bar{b})$ vs. $Z + b\bar{b}$. In addition, the ROC curves are shown for each observable. The colour ring performs the best out of all of these jet sub-structure observables being able to distinguish $H(b\bar{b})$ events from $g(b\bar{b})$ with great accuracy. The ROC curves show that the false-positive rate for colour ring stays down for the longest while still producing good signal acceptance.

The colour ring weight variable, $\mathcal{O}_w$, is the colour ring but weighted by particle transverse momentum, and the alternative colour ring variable, $\mathcal{O}_A$, comes from subtracting the angle between the b-jets from the other two angles instead of dividing by it, shown in equations 6.13 and 6.14.

$$\mathcal{O}_w \approx \frac{1}{\sum_i p_T, i} \sum_i p_T, i \cdot \mathcal{O}_i \tag{6.13}$$

$$\mathcal{O}_A \approx \theta_{ak}^2 + \theta_{bk}^2 - \theta_{ab}^2 \tag{6.14}$$

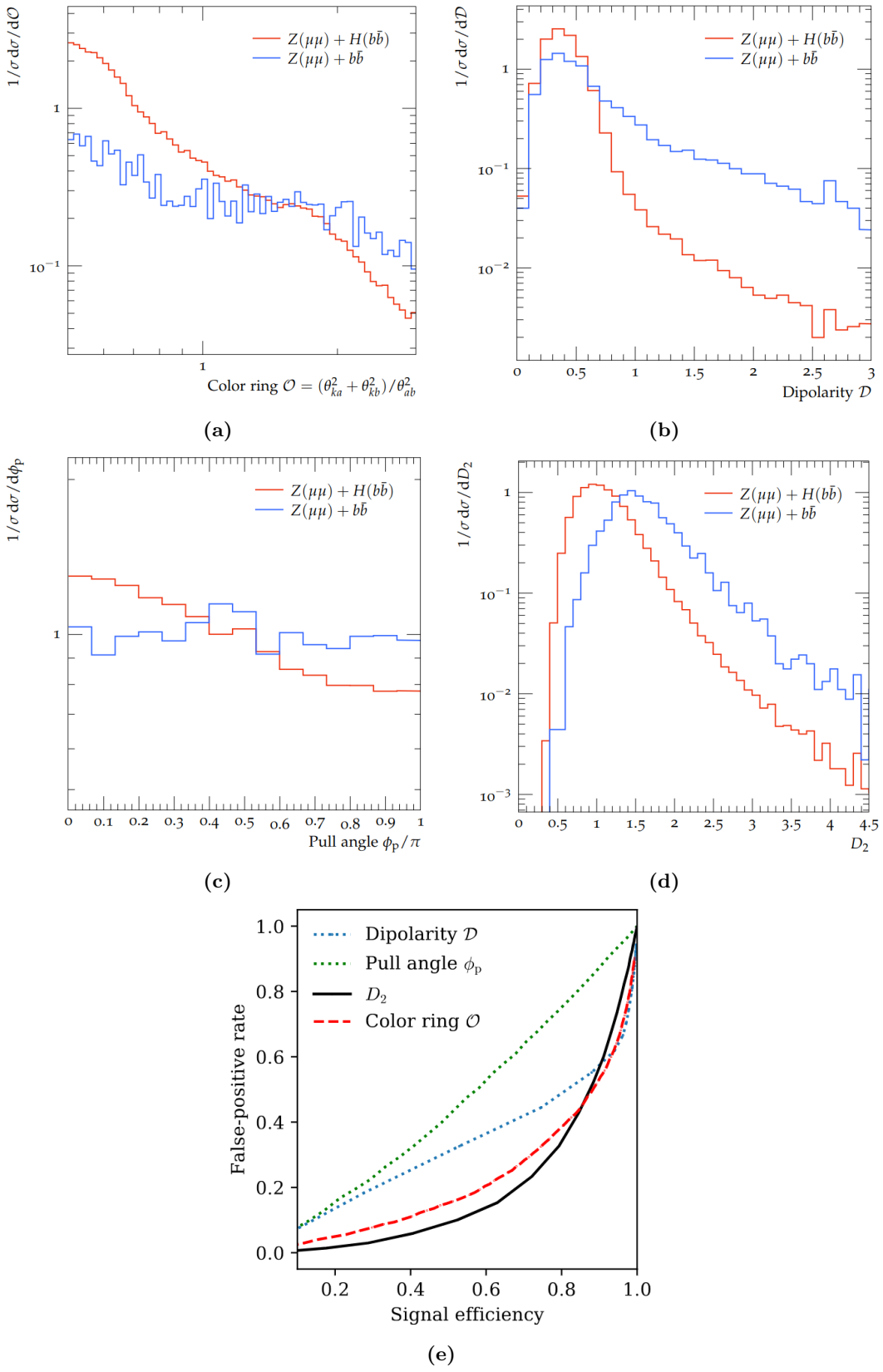


Figure 6.27: Large-R jet structure distributions for $Z(\mu\mu)H(b\bar{b})$ vs. $Z(\mu\mu)b\bar{b}$ using the observables (a) Colour ring, (b) Dipolarity, (c) Pull Angle, and (d) D_2. Plot (e) displays the ROC curves for each observable where $H(b\bar{b})$ is the signal. These curves characterise the trade-off between the (desirable) true-positive event identification rate (aka efficiency), and the (undesirable) false-positive identification rate [62].

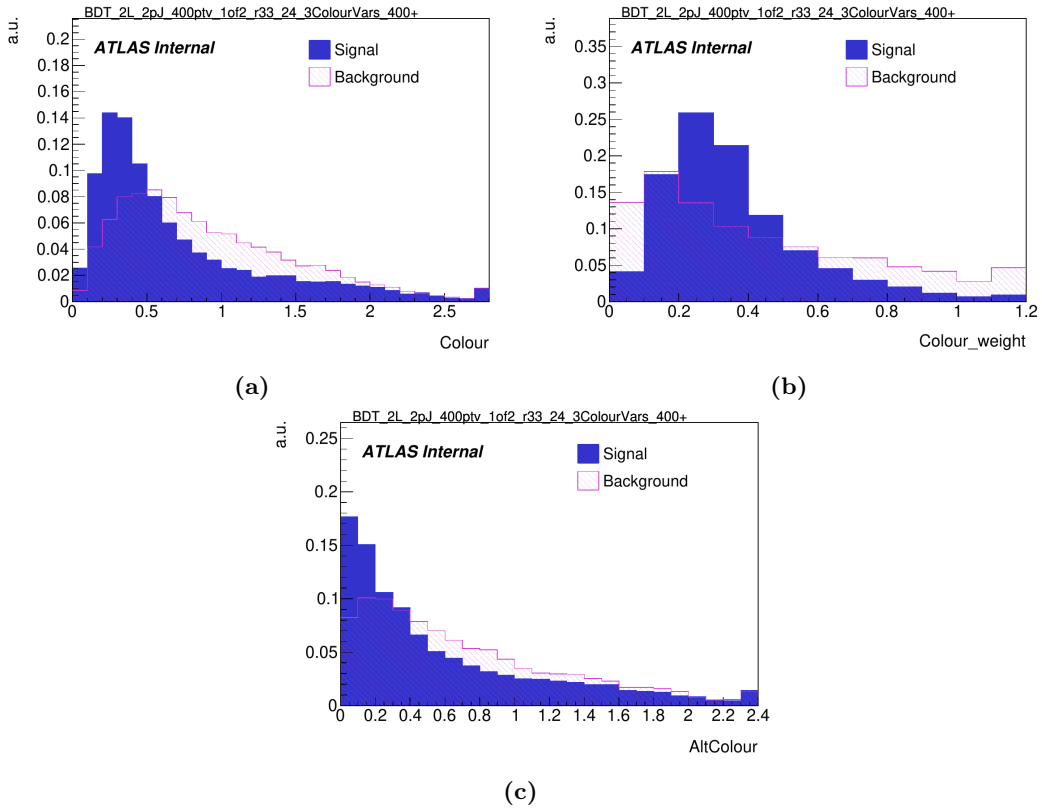


Figure 6.28: Signal and background distributions for (a) Colour ($\mathcal{O}$), (b) Colour weight ($\mathcal{O}_w$), (c) alternative Colour ($\mathcal{O}_A$) in the $p_T^V > 400$ GeV region of the $VH(\to b\bar{b})$ boosted analysis using MC samples.

Figure 6.28 displays the signal and background distributions for all three new colour ring variables. These variables have high discriminating power as can be seen by the large signal spikes in each plot.

Some problems were faced with introducing the colour ring variables, mainly that the alternative colour variable requires three track jets to be calculated. In an event with fewer than three track jets the value for alternative colour would be the negative placeholder value of -99. Instead of splitting the analysis into separate 2 jet and 3+ jet regions (which would lead to reduced statistics), we altered the negative placeholder value in the MVA training to be -1 instead of -99 - which is much friendlier to the MVA.

6.3.10 Colour ring results

The MVA training was conducted with: the three colour variables separately; colour and colour weight; and all three together, for a total of five different configurations. All training was done with the newest MC sample release r33-24 with truth tagged samples and with separately optimised hyperparameters for each set of variables. The hyperparameters are optimised for the $p_T^V > 400\,\text{GeV}$ region as this region suffers most from overtraining and lack

p_T^V region	Colour	Colour weight	AltColour	Colour + Colour weight	All 3 colour variables
250-400 GeV	2.50	2.45	2.51	2.47	2.48
400+ GeV	1.90	1.89	1.92	1.91	1.93

Table 6.5: Table of significances from the colour ring MVA training results.

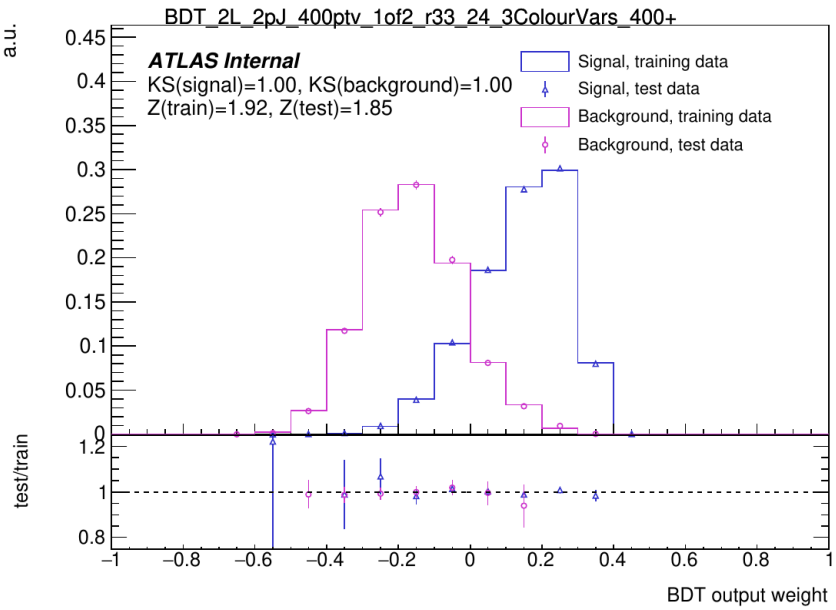


Figure 6.29: Overtraining plot for all three colour variables in the 400+ p_T^V region.

of statistics.

As is seen in table 6.5 the slightly better performing configuration in the $p_T^V > 400\,\text{GeV}$ region is using all three colour ring variables (overtraining plot shown in figure 6.29). There is very little change in significance between the configurations in both regions separately, so all could be used if issues are found with any single variable. This result is an improvement on previous iterations of the MVA that didn't include colour ring variables. The overtraining plot for the MVA training that includes all three colour ring variables shown in figure 6.29 shows great progress in solving the overtraining issues that previously plagued this region of the analysis.

The final colour ring distribution plots are shown in figure 6.30 with data overlaid to demonstrate the variable modelling. The distribution plots for all three colour ring variables display nice signal peaks in the MC simulations which is also shown in the ratio plots at the bottom of each figure. The legend on the right of each plot contains the background and signal processes (signal shown in red). The colour weight plot is blinded due to the signal to background ratio being above the blinding limit (15% for the analysis). The analysis

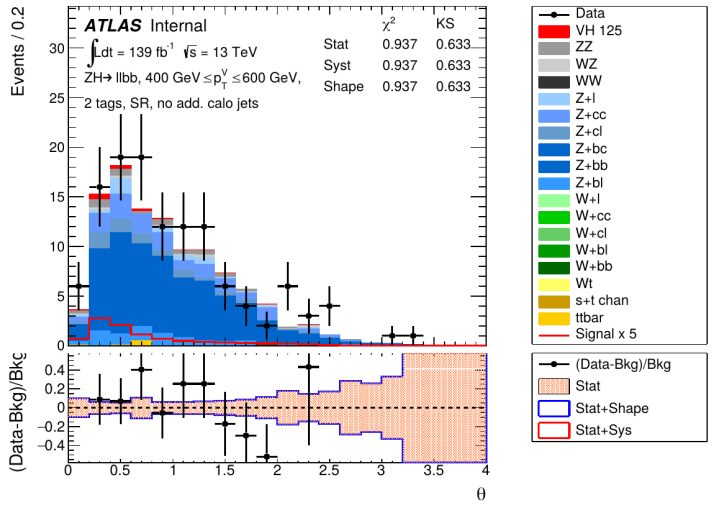
adopted the Colour variable ($\mathcal{O}$) into the baseline set of variables after the improved results had been presented. This section looked at the colour ring results, section 6.5 will look at the overall $VH(\rightarrow b\bar{b})$ analysis results from Run 2.

6.4 Modelling

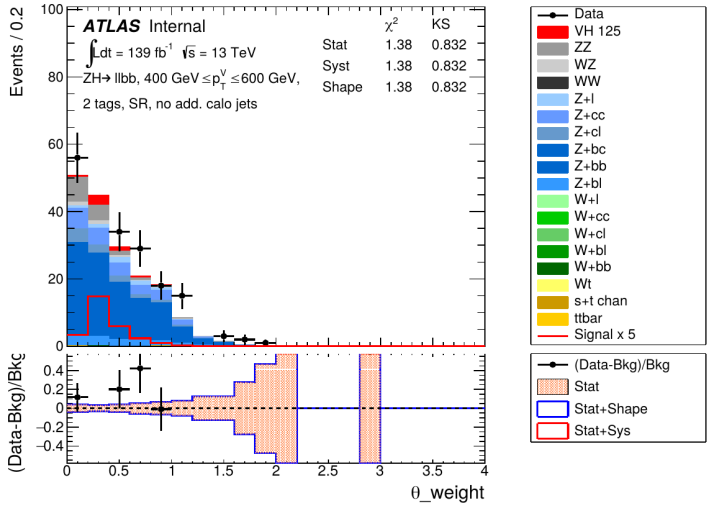
The modelling of the VH signal is done coherently across both boosted and resolved $VH(\rightarrow b\bar{b})$ analysis categories. It follows to a large extent the modelling approach adopted by the individual analyses. This section provides a brief overview of the modelling strategy itself, as well as covering the (small) modifications necessary for the combination analysis. The signal model treats the three main signal processes, $qq \rightarrow WH$, $qq \rightarrow ZH$, and $gg \rightarrow ZH$ separately. In all cases, a number of systematic uncertainties are assigned that either affect primarily the VH production, or primarily the $H \rightarrow b\bar{b}$ decay.

Many systematic uncertainties are taken into consideration for this analysis. Experimental uncertainties are taken for every physics object used in the analysis, such as the jets, leptons and E_T^{miss}. A large source of these experimental uncertainties comes from the flavour tagging algorithms. These flavour tagging uncertainties are derived from specific analyses used solely for calculating these uncertainties, either from the flavour tagging CP group (for the b-tagging in the boosted regime) or within the analysis itself (for resolved and $VH(\rightarrow c\bar{c})$ regimes). As well as these uncertainties there are modelling uncertainties for every signal and background process in the analysis. Usually, these are derived as two-point systematics using the nominal and alternative MC samples. The uncertainties derived from each process can be divided into normalisation/acceptance uncertainties and shape uncertainties.

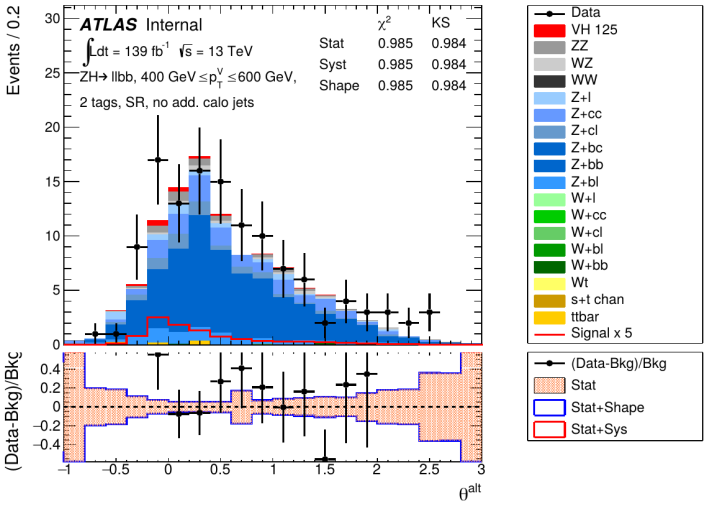
Normalisation uncertainties are calculated in the region of highest purity for the specified process and are taken as an overall value for the total yield. These uncertainties are then extrapolated to all other regions and extra uncertainties are composed for this extrapolation. Acceptance uncertainties look to measure any discrepancies in the distribution of events in the analysis phase space caused by the MC generator. This can affect the ratio of events in different regions or cause some events on the fringe of regions to be put in the wrong region. Shape uncertainties quantify the shape of the final fitting discriminant when comparing between nominal and alternative MC generators. These are calculated using a machine learning n-dimension reweighting algorithm known as CARL. CARL is trained over the MC samples (nominal and alternative) to determine the shape uncertainty on the MVA output and m_{cc}



(a)



(b)



(c)

Figure 6.30: Variable distribution plots for (a) Colour ($\mathcal{O}$), (b) Colour weight ($\mathcal{O}_w$), (c) alternative Colour ($\mathcal{O}_A$) with MC simulation histograms and experiment data in the context of the $VH(\to b\bar{b})$ boosted analysis.

(for the $VH(\rightarrow c\bar{c})$ analysis).

The $VH(\rightarrow b\bar{b})$ boosted analysis previously used Sherpa version 2.2.1 (plotted in figures 6.34 - 6.37) originally to model the background diboson processes (ZZ, WZ, WW). This was to be updated in the most recent iteration of the analysis, so the newer version Sherpa 2.2.11 was tested (figures 6.31 - 6.33). The following plots are for the $2L$ channel displaying the ZZ process; these are plotted for the important MVA variables ($boostedMVA$, m_J and p_T^V). The newer version contains electroweak correction (EWK) plots (figure 6.31) which were not present in the previous Sherpa iteration. The other types of corrections are the quantum chromodynamics corrections (QCD) and particle density function corrections (PDF). The histograms represent the matrix only calculations where as the data points contain the extra correction value. Comparing the new and old Sherpa versions it is clear in the PDF plots that the newer version has smaller error bars leading to better error interpretation. In the QCD scale correction plots it can be seen in the ratio plots that we have less variance across the full range of the plots in the newer version. The older Sherpa version plots are split into signal and background categories. In the 2L channel these plots are for $ZZbb$ and $ZZbkg$ processes. Overall, the newer Sherpa version models these processes better with sharper peaks and smaller uncertainties in all of the variable plots.

Variable distribution plots for the 2L boosted $VH(\rightarrow b\bar{b})$ analysis are shown in figures 6.38-6.42, namely the variables p_T^V, $\Delta R(bb)$, $p_T^{b_1}$, $p_T^{b_2}$ and m_{LL}. The plots display data collected during Run 2 of the LHC, as well as, simulated samples of signal and background processes (specific processes are outlined in the legend on the right).

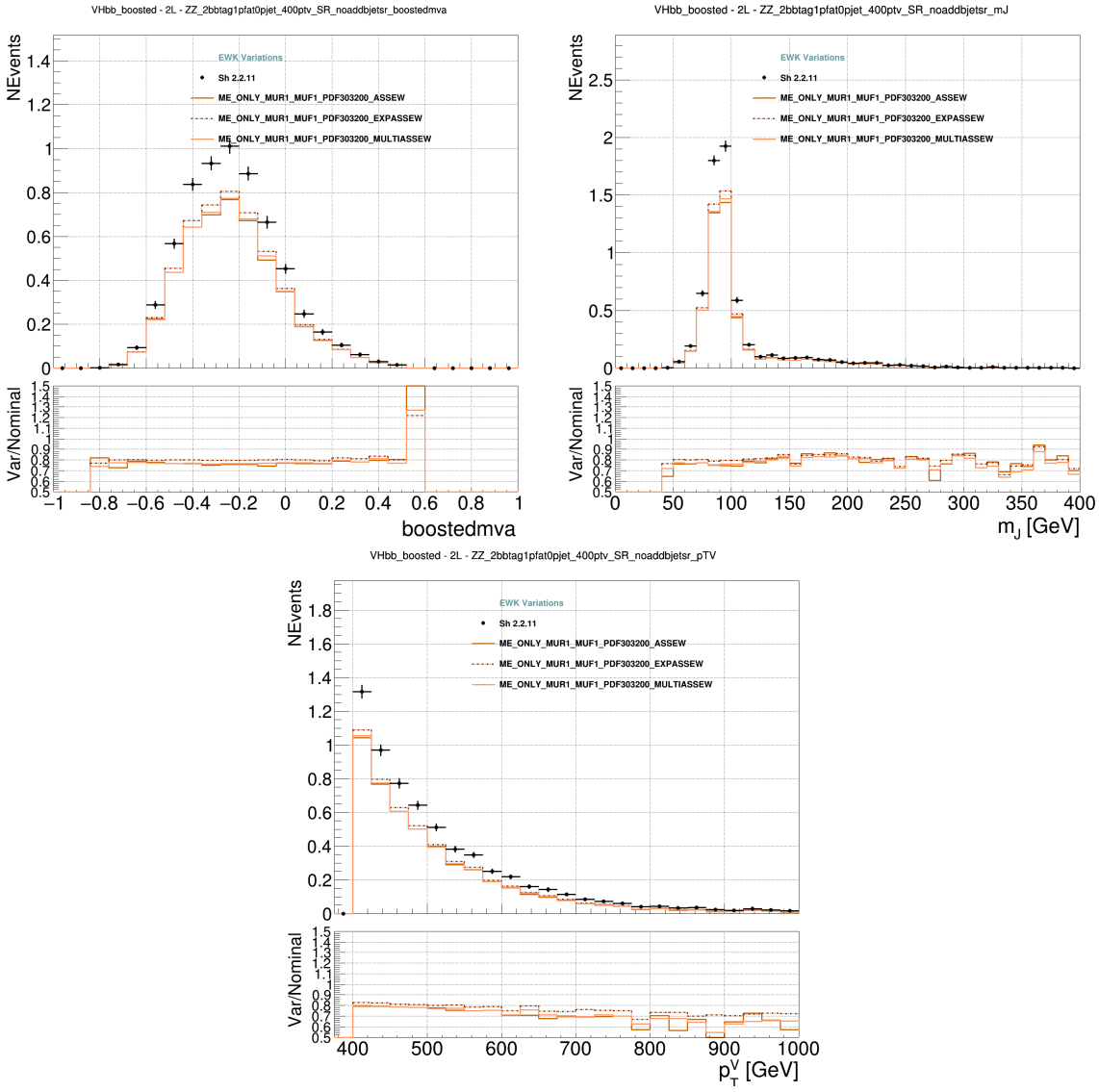


Figure 6.31: Diboson modelling plot displaying the boostedmva, m_J, and p_T^V variables electroweak correction using the Sherpa 2.2.11 sample.

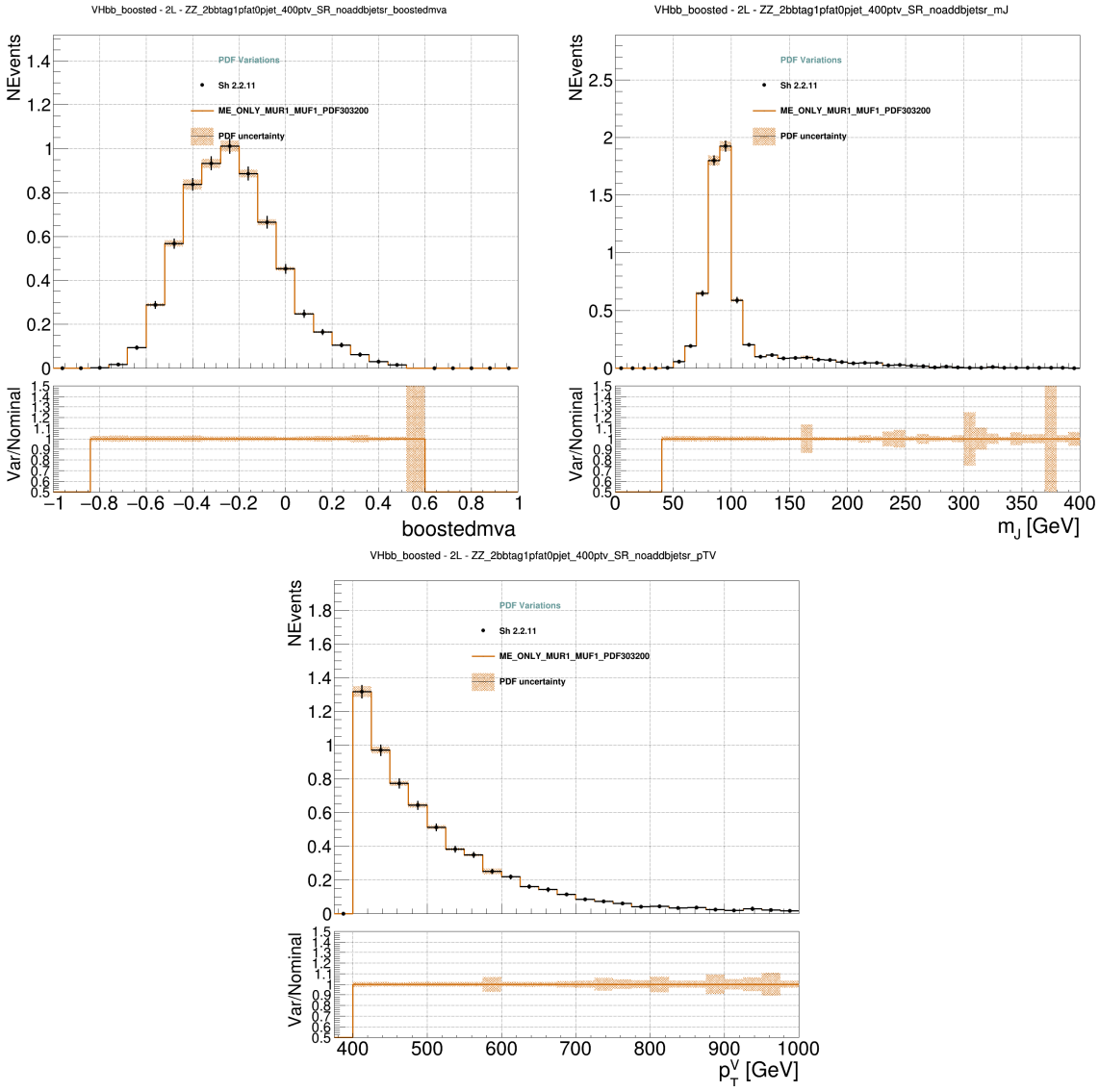


Figure 6.32: Diboson modelling plot displaying boostedmva, m_J, and p_T^V variables particle density function correction using the Sherpa 2.2.11 sample.

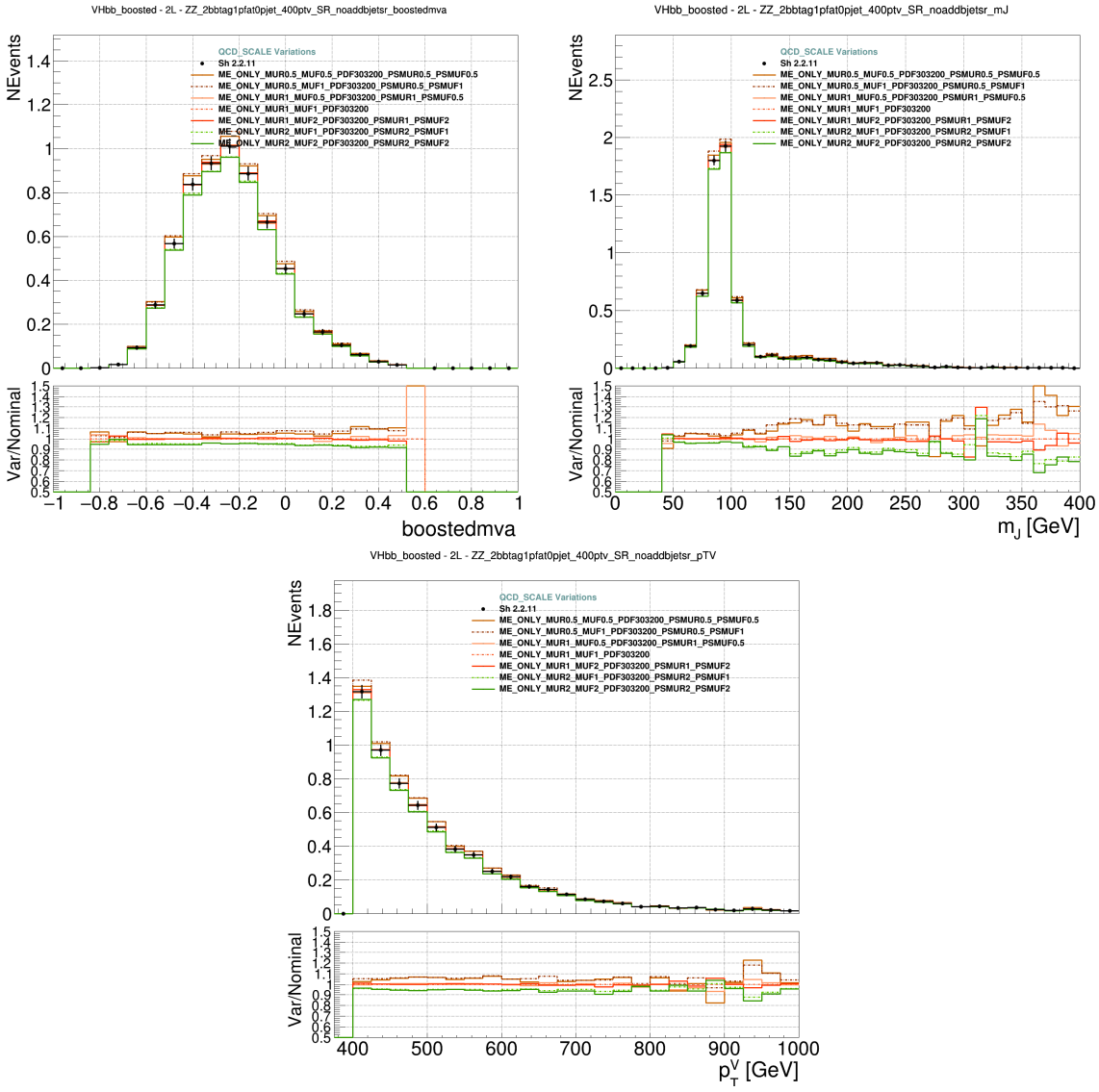


Figure 6.33: Diboson modelling plot displaying boostedmva, m_J, and p_T^V variables QCD scale correction using the Sherpa 2.2.11 sample.

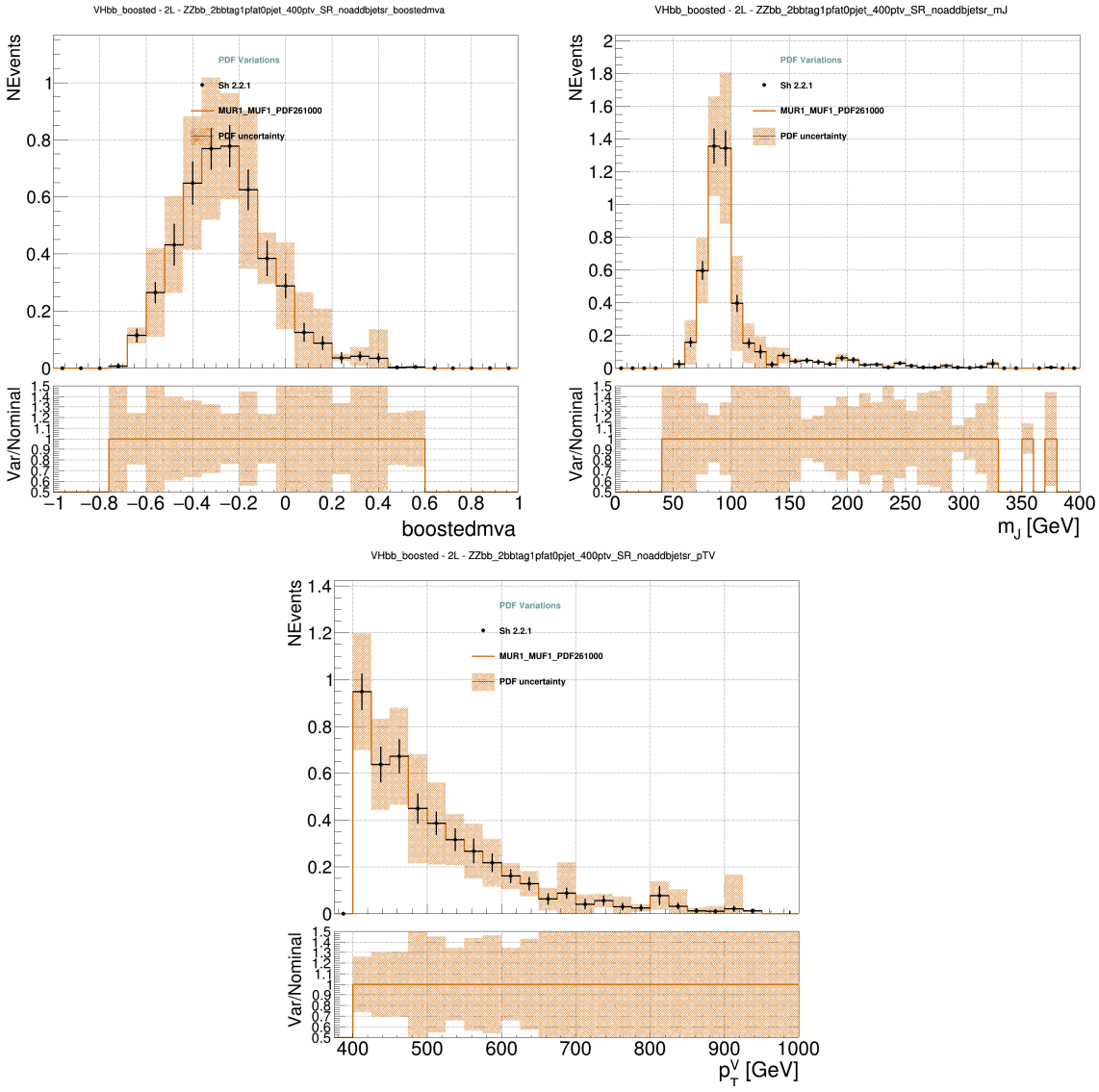


Figure 6.34: Diboson signal modelling plot displaying boostedmva, m_J, and p_T^V variables particle density function correction using the Sherpa 2.2.1 sample.

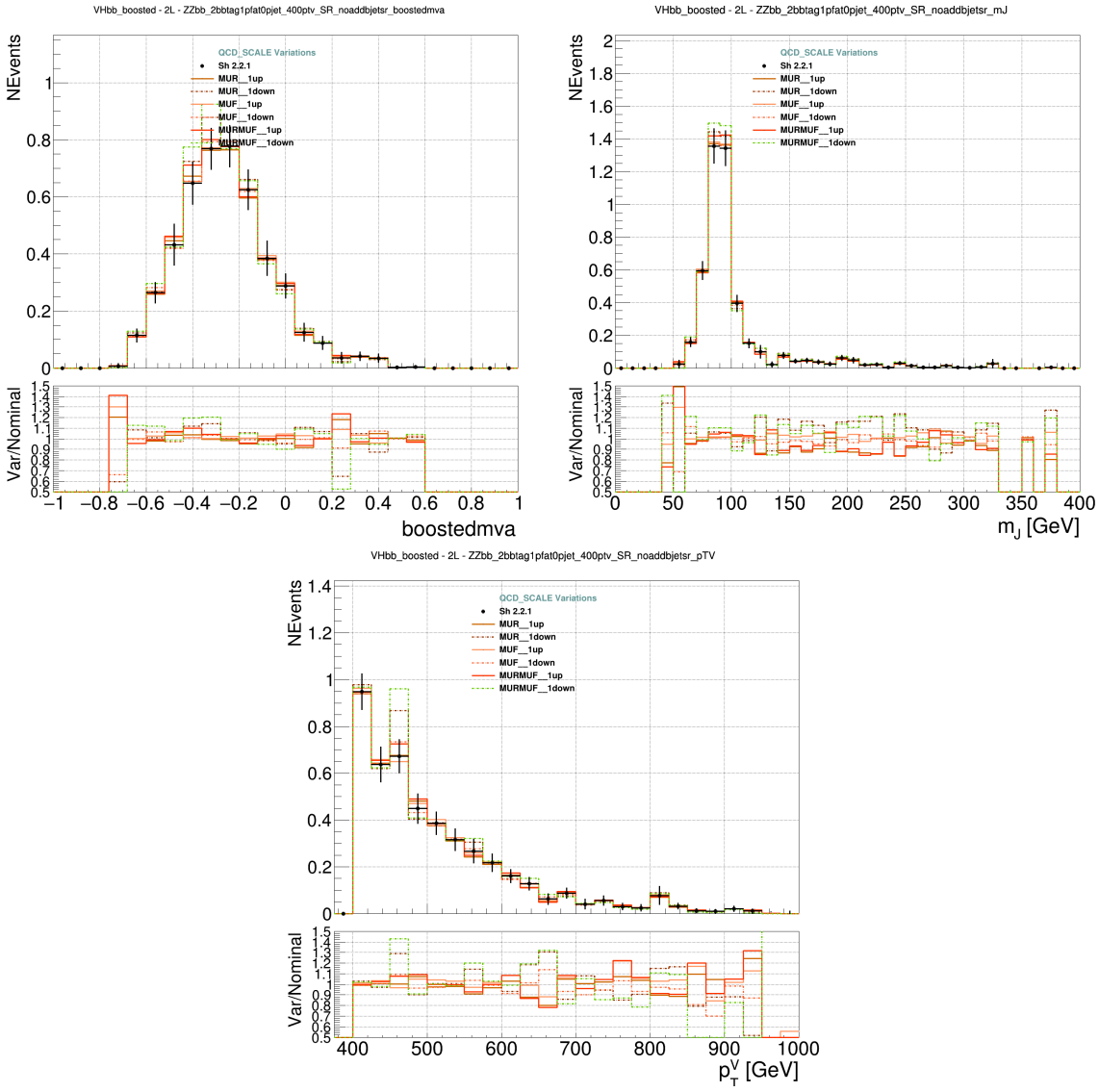


Figure 6.35: Diboson signal modelling plot displaying boostedmva, m_J, and p_T^V variables QCD scale correction using the Sherpa 2.2.1 sample.

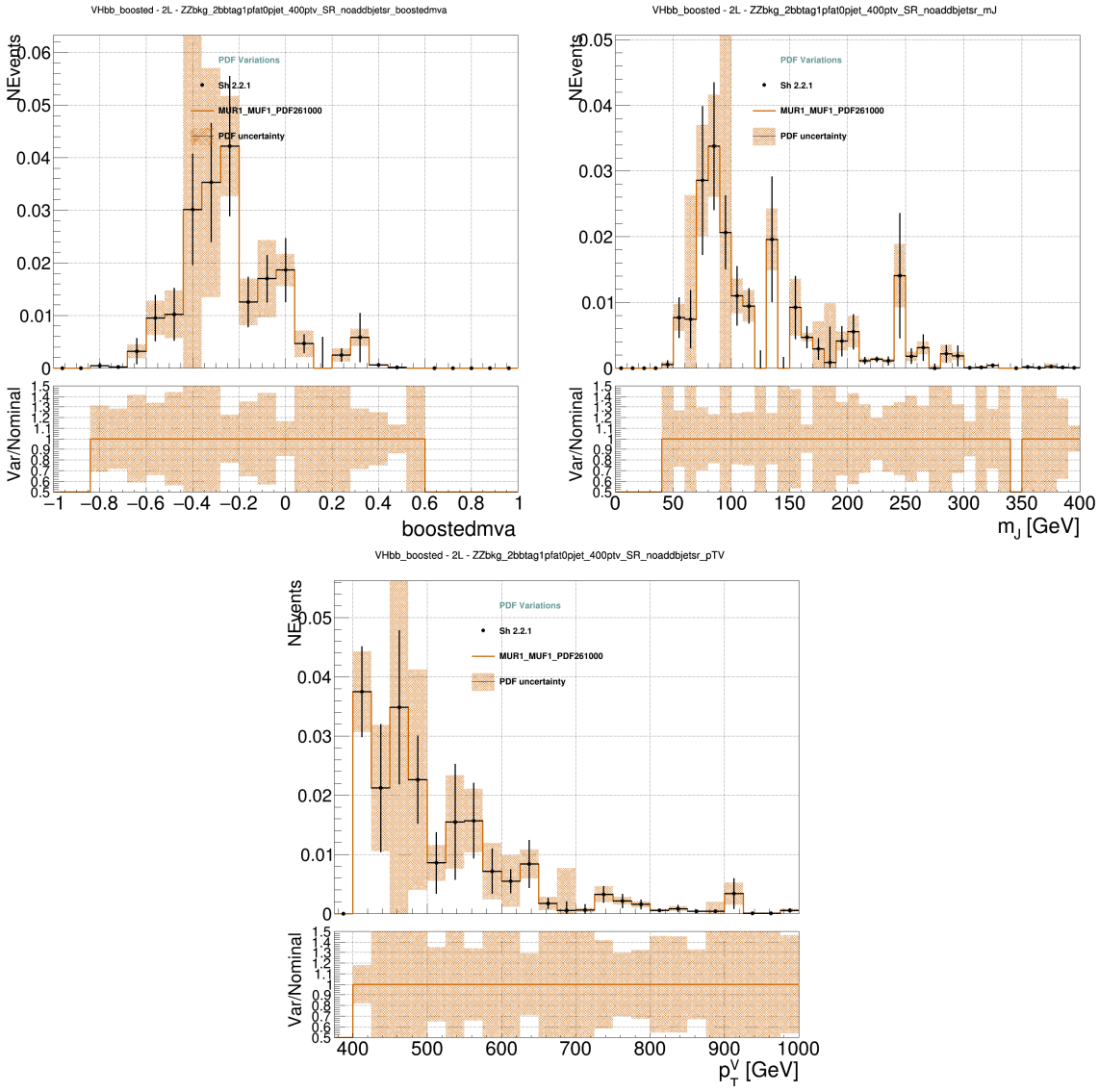


Figure 6.36: Diboson background modelling plot displaying boostedmva, m_J, and p_T^V variables particle density function correction using the Sherpa 2.2.1 sample.

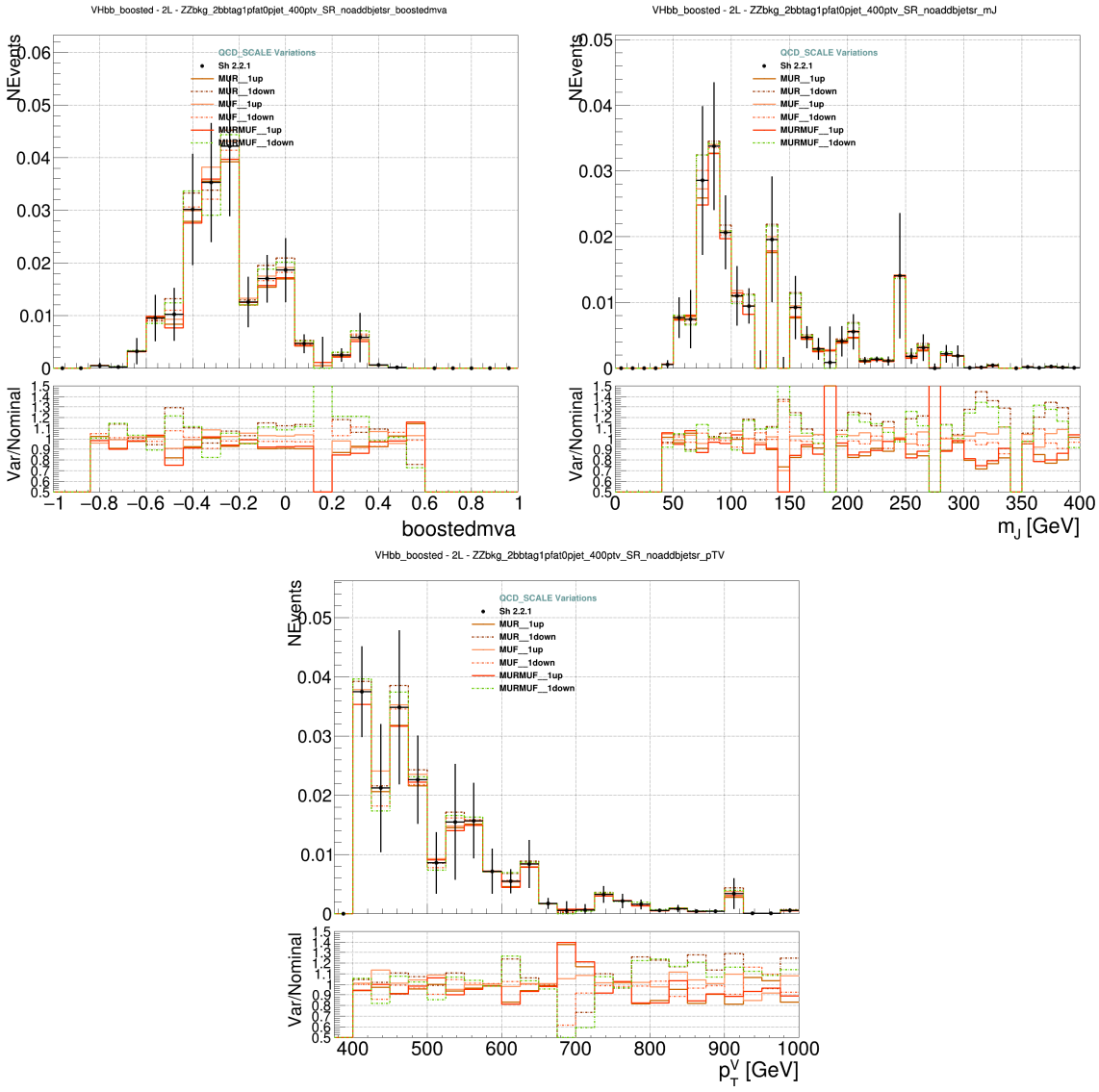


Figure 6.37: Diboson background modelling plot displaying boostedmva, m_J, and p_T^V variables QCD scale correction using the Sherpa 2.2.1 sample.

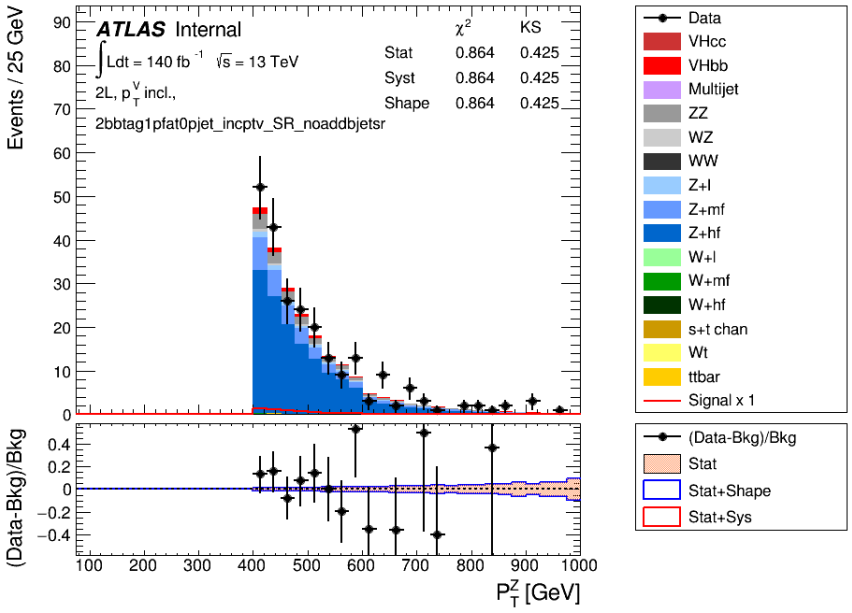


Figure 6.38: p_T^V variable distribution for the 2L boosted $VH(\to b\bar{b})$ analysis with 139 fb^{-1} of data displaying both simulated values and data points.

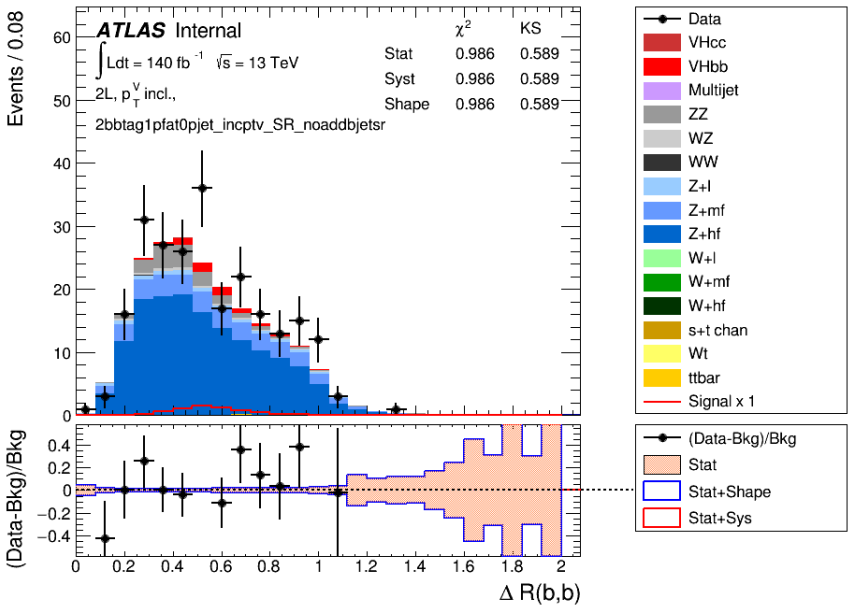


Figure 6.39: $\Delta R(bb)$ variable distribution for the 2L boosted $VH(\to b\bar{b})$ analysis with 139 fb^{-1} of data displaying both simulated values and data points.

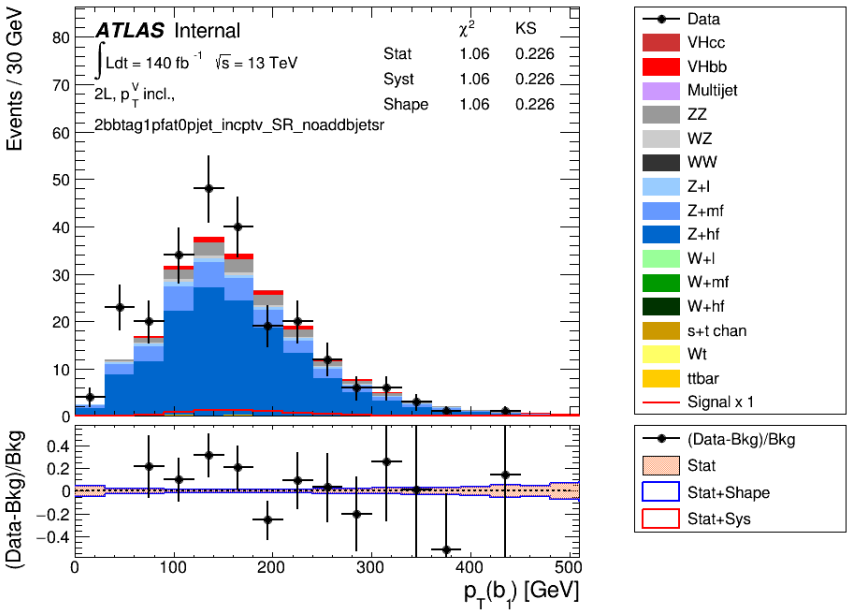


Figure 6.40: $p_T^{b_1}$ variable distribution for the 2L boosted $VH(\to b\bar{b})$ analysis with 139 fb^{-1} of data displaying both simulated values and data points.

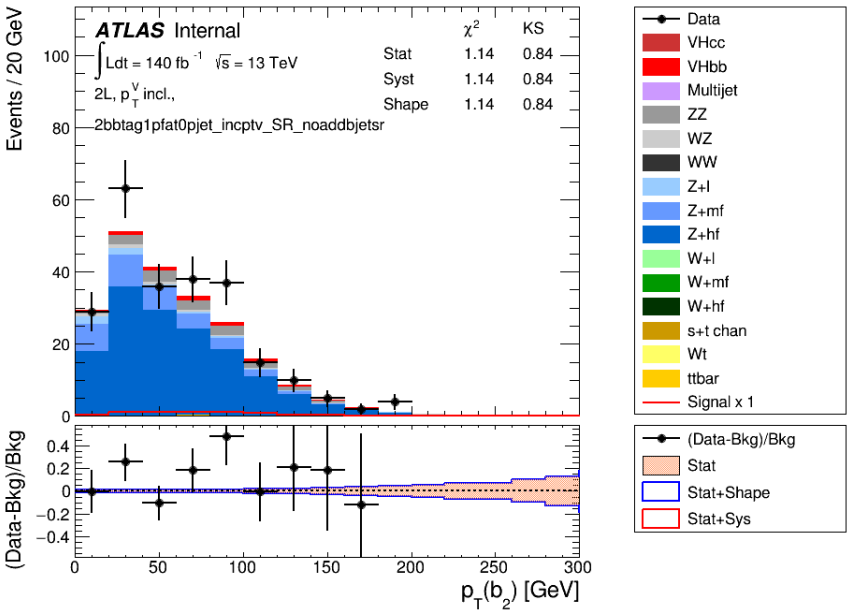


Figure 6.41: $p_T^{b_2}$ variable distribution for the 2L boosted $VH(\to b\bar{b})$ analysis with 139 fb^{-1} of data displaying both simulated values and data points.

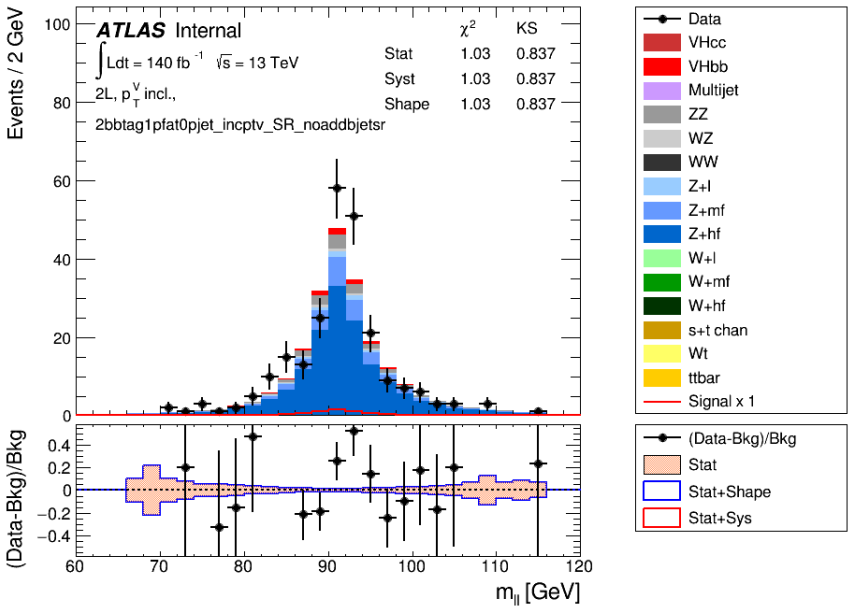


Figure 6.42: m_{LL} variable distribution for the 2L boosted $VH(\to b\bar{b})$ analysis with 139 fb^{-1} of data displaying both simulated values and data points.

6.5 Summary of Run 2 boosted $VH(\to b\bar{b})$ results

6.5.1 Previous Run 2 boosted $VH(\to b\bar{b})$ results

The results presented in this section are from the 2021 boosted $VH(\to b\bar{b})$ paper [44] which used the cut-based analysis technique. The m_J distributions for all three lepton channels are shown in figure 6.43, with all signal strengths, background normalisations and NP's set at the best-fit values. Good agreement is observed in all channels between the data and predictions. These are post-fit plots meaning the backgrounds and NPs have been set to best-fit values and have reduced uncertainties.

When combining all three lepton channels, in the boosted regime, for a Higgs mass of 125 GeV, the observed excess with respect to the background-only hypothesis has a significance of 2.1 standard deviations, which compares to the predicted value of 2.7 standard deviations. The fitted signal strength for the $VH(\to b\bar{b})$ process is:

$$\mu_{VH}^{bb} = 0.72^{+0.39}_{-0.36} = 0.72^{+0.29}_{-0.28}(\text{stat.})^{+0.26}_{-0.22}(\text{syst.}) \tag{6.15}$$

The larger uncertainty in this result is statistical in nature and includes the contribution from floating background normalisations unconstrained in the fit. The systematic uncertainty is dominated by large-R jet calibration, specifically in m_J resolution. The second most dominant systematic uncertainty is the background modelling. There is a non-negligible impact

to the systematic uncertainty associated with the limited size of MC simulation samples.

Figure 6.44(a) displays the signal regions m_J distribution summed over all lepton channels and p_T^V regions ($250 < p_T^V < 400\,\text{GeV}$, $p_T^V > 400\,\text{GeV}$), after subtraction of all background regions excluding the WZ and ZZ diboson processes and is weighted by the ratio of Higgs boson signal to background yields in each region respectively. The fitted values of Higgs boson signal strength, μ_{VH}^{bb}, is shown in figure 6.44(b) for each lepton channel individually, as well as separate values shown for each p_T^V region in each channel, combined values for each p_T^V region and a total combined value. It includes a total uncertainty along with the uncombined values for statistical and systematic uncertainties.

6.5.2 Run 2 boosted legacy analysis results

The preliminary results shown in this section are from the Run 2 boosted legacy analysis. Figure 6.45 shows the m_J distributions for the combined $VH(\to b\bar{b})$ boosted analyses i.e. for all three lepton channels, inclusive for all p_T^V regions. They require a large-R jet containing 2 b-tags. Figure 6.45(b) displays the m_J distributions with all the background processes subtracted except for the WZ and ZZ diboson processes, and figure 6.45(c) is the same as (b) but is also weighted by the Higgs boson signal over background ratio.

6.5.3 Run 2 combined $VHbb(cc)$ analysis results

Figure 6.46 shows the pre-fit BDT output distribution for the 2-lepton channel boosted method, from the combined $VHbb(cc)$ analysis which collectively takes results from the resolved, boosted analyses and the $VH(\to c\bar{c})$ analysis. The distributions are well separated for signal and background events, with good agreement shown for all bins. These plots display the p_T^V regions ($400 < p_T^V < 600\,\text{GeV}$) and ($p_T^V > 600\,\text{GeV}$), due to removing any overlap between resolved and boosted regions the split is placed at $p_T^V = 400\,\text{GeV}$.

This analysis used $139\,\text{fb}^{-1}$ of data collected with the ATLAS detector during Run 2 of the LHC. The signal strength of the $VH(\to b\bar{b})$ process is measured at:

$$\mu_{VH}^{bb} = 0.93^{+0.15}_{-0.14} = 0.93^{+0.10}_{-0.10}(\text{stat.})^{+0.11}_{-0.11}(\text{syst.}) \tag{6.16}$$

The observed excess with respect to the background-only hypothesis has a significance of 7.34 standard deviations, compared with the expected significance of 7.87 standard deviations.

The measured signal strength for the 2-lepton channel only is:

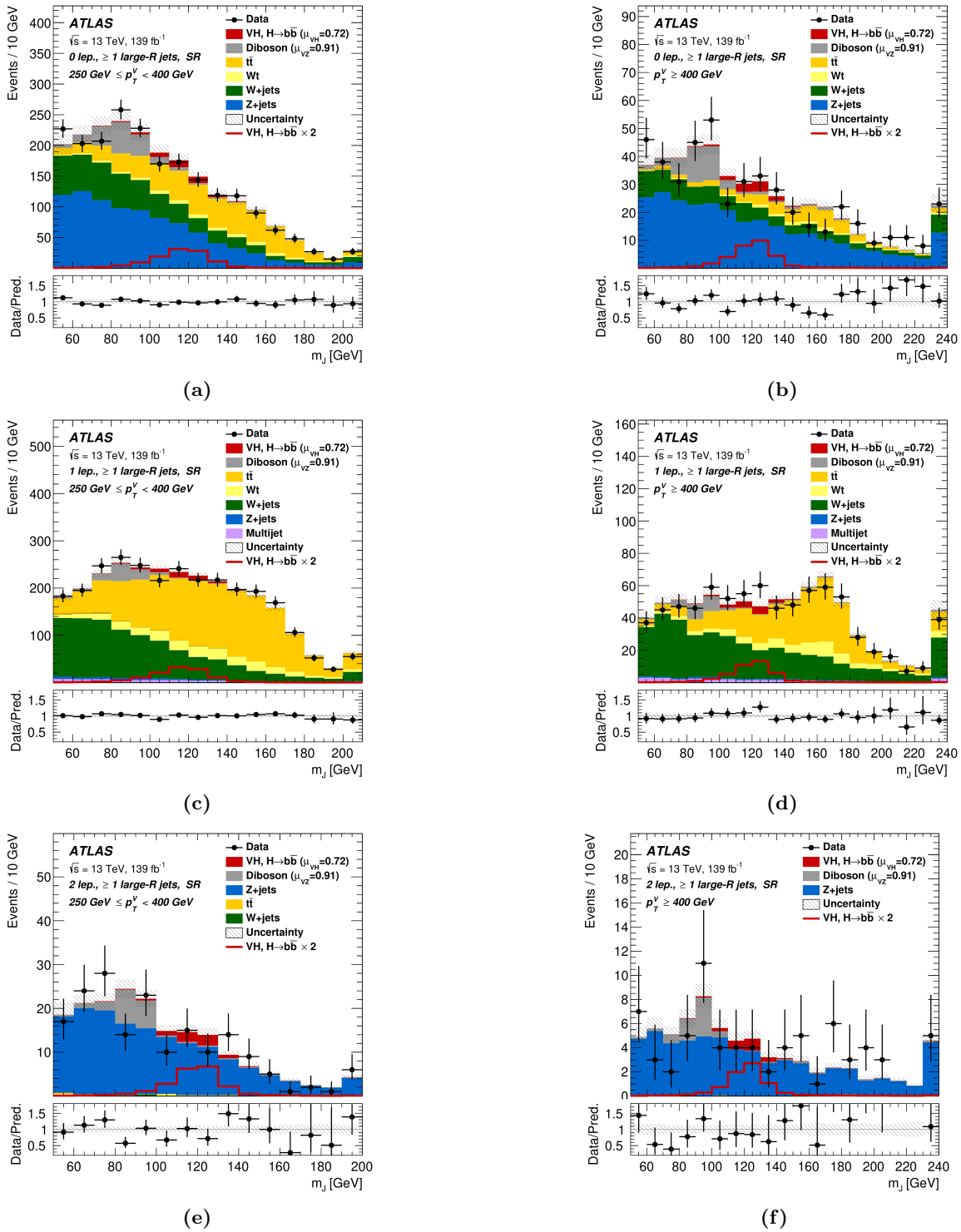
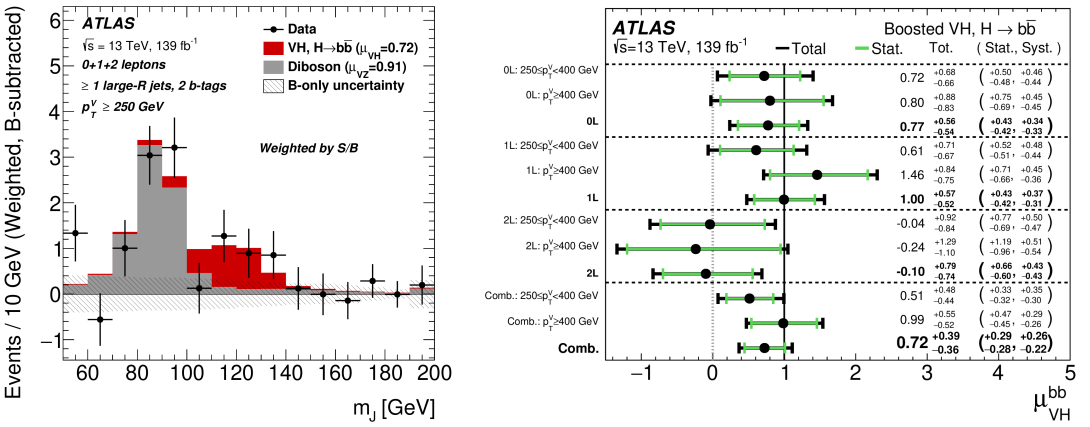


Figure 6.43: The m_J post-fit distributions in (a, b) the 0-lepton, (c, d) 1-lepton and (e, f) 2-lepton signal regions for 2 b-tagged events for (a, c, e) 250GeV$< p_T^V <$ 400 GeV and (b, d, f) $p_T^V \geq 400$ GeV. The low-purity and high-purity categories in the case of the 0-lepton and 1-lepton channels are merged in this figure. The background contributions after the likelihood fit are shown as filled histograms. The Higgs boson signal ($m_H = 125$ GeV) is shown as a filled histogram on top of the fitted backgrounds normalised to the signal yield extracted from data ($\mu_{VH}^{bb} = 0.72$), and unstacked as an unfilled histogram, scaled by the SM prediction times a factor of two. The size of the combined statistical and systematic uncertainty for the sum of the fitted signal and background is indicated by the hatched band. The highest bin in the distributions contains the overflow. The ratio of the data to the sum of the fitted signal and background is shown in the lower panel.



(a) **(b)**

Figure 6.44: (a) m_J distribution in data after subtraction of all backgrounds except for the WZ and ZZ diboson processes. The contributions from all lepton channels and signal regions are summed and weighted by their respective values of the ratio of fitted Higgs boson signal and background yields. The expected contribution of the associated WH and ZH production of a SM Higgs boson with $m_H = 125$ GeV is shown scaled by the measured combined signal strength ($\mu^{bb}_{VH} = 0.72$). The diboson contribution is normalised to its best-fit value of $\mu^{bb}_{VZ} = 0.91$. The size of the combined statistical and systematic uncertainty is indicated by the hatched band. (b) Fitted values of the Higgs boson signal strength parameter, μ^{bb}_{VH}, for $m_H = 125\,\text{GeV}$ for the 0-, 1- and 2-lepton channels in different p^V_T regions separately and for various combinations.

$$\mu^{bb}_{VH} = 0.93^{+0.27}_{-0.25} = 0.93^{+0.21}_{-0.21}(\text{stat.})^{+0.16}_{-0.13}(\text{syst.}) \tag{6.17}$$

This corresponds with a observed (expected) significance of 4.13 (4.46) standard deviations above the background-only hypothesis.

These results can be compared with the 2018 $VH(\to b\bar{b})$ observation paper [63] which uses 79.8fb^{-1} of data collected during Run 2 of the LHC with the ATLAS detector. The measured signal strength was:

$$\mu^{bb}_{VH} = 1.16^{+0.27}_{-0.25} = 1.16^{+0.16}_{-0.16}(\text{stat.})^{+0.21}_{-0.19}(\text{syst.}) \tag{6.18}$$

The observed (expected) significance with respect to the background-only hypothesis was 4.9 (4.3) standard deviations. Additionally, the 2-lepton channel measured signal strength for the $VH(\to b\bar{b})$ process was:

$$\mu^{bb}_{VH} = 1.38^{+0.46}_{-0.42} \tag{6.19}$$

Corresponding with a observed (expected) significance of 3.4 (2.6) standard deviations above the background-only hypothesis. Due to the enhancements made in the analysis (outlined in this thesis) the $VH(\to b\bar{b})$ signal strength measurement and corresponding signal sensitivity

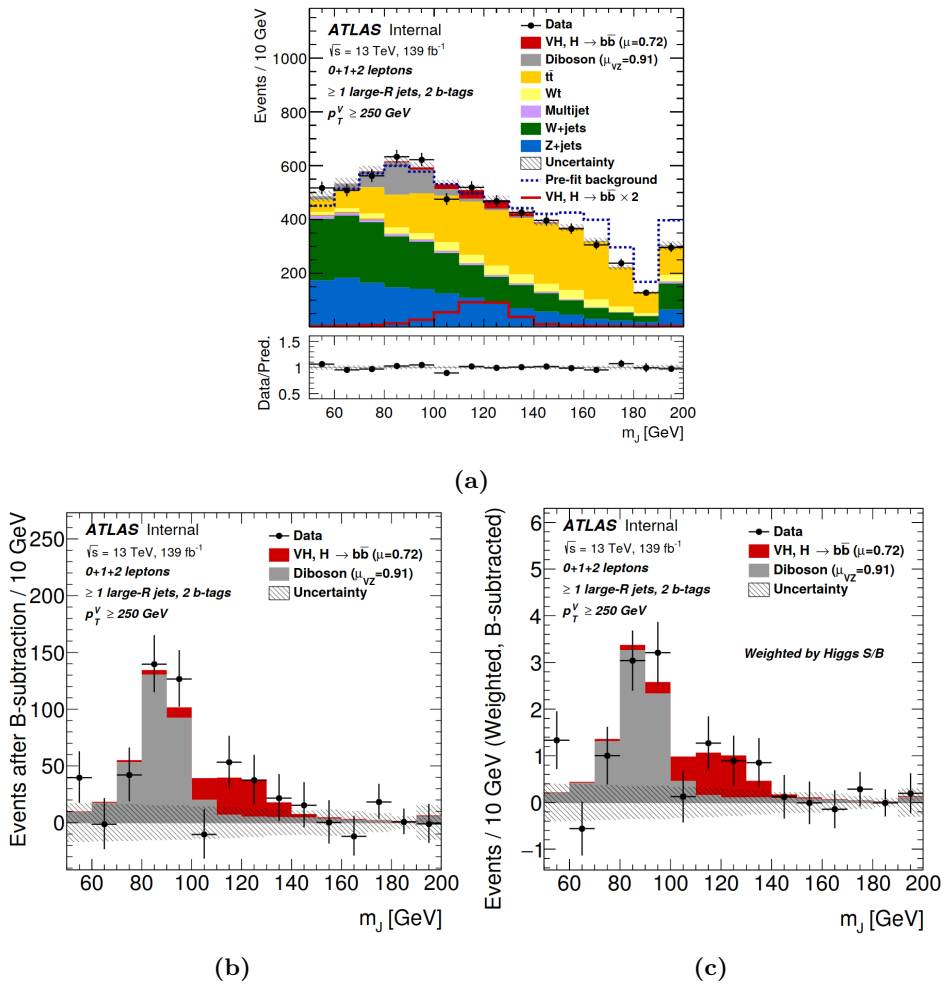
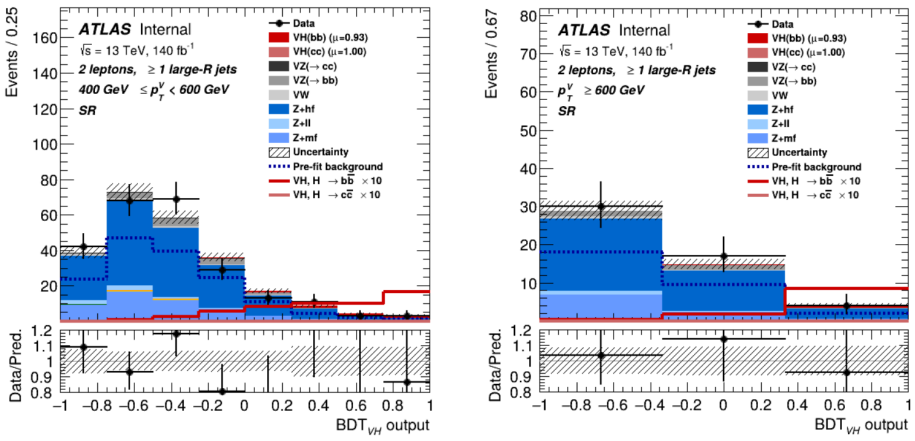


Figure 6.45: m_J distribution for the full Run-2 dataset combining all three lepton channels (a) by stacking them and (b) after background subtraction and (c) weighted by the signal/background ratio after background subtraction. The diboson signal is depicted in light gray, the Higgs signal in red. The last bin is an overflow bin which collects all events with a value of m_J greater than the maximum value on the axis.



(a) **(b)**

Figure 6.46: The pre-fit distributions of BDT output for the p_T^V regions of (a) $400 < p_T^V < 600\,\text{GeV}$ and (b) $p_T^V > 600\,\text{GeV}$ in the 2-lepton boosted analysis with 2 b-tags within the large-R jet.

have greatly improved over the course of the legacy analysis for both the combined and 2-lepton results. The most recent results in this thesis are still internal ATLAS results and should not be made public.

Chapter 7

Conclusion

The search for VH production with an associated $H \to bb$ decay was conducted with the ATLAS detector at the Large Hadron Collider. The analysis collected 139 fb^{-1} of proton-proton collision data during Run 2 at a centre-of-mass energy of 13 TeV. The results section outlines three separate analyses papers: the initial Run 2 boosted $VH(\to b\bar{b})$ cut-based analysis; the Run 2 boosted $VH(\to b\bar{b})$ legacy analysis (yet to be released); and the Run 2 combined $VH(\to b\bar{b})$ legacy analysis (which includes resolved, boosted and $VH(\to c\bar{c})$ analyses).

The combined 2-lepton analysis sees an improvement in observed (expected) significance from 3.4 (2.6) to 4.13 (4.46) standard deviations. The measured signal strength for the 2-lepton $VH(\to b\bar{b})$ process went from $\mu_{VH}^{bb} = 1.38^{+0.46}_{-0.42}$ to $\mu_{VH}^{bb} = 0.93^{+0.27}_{-0.25} = 0.93^{+0.21}_{-0.21}(\text{stat.})^{+0.16}_{-0.13}(\text{syst.})$ in the combined Run 2 legacy analysis. This improvement is a result of using MVA techniques in the boosted regime, adding improved variables to the MVA training and introducing truth flavour tagging to the analysis.

My work on TTWeb was to develop features in preparation for the start of Run 3 of the LHC. I developed the left panel to add key information and help improve navigation on the app. I developed the bulk change feature which allows the user to edit multiple lines in parallel. I added express stream information to the HLT prescales page which provided key information for early configuration runs during Run 3. I created a new view page for displaying both L1 prescale keys and HLT prescale keys on the same page, known as the combined view page.

The overlap study between the boosted and resolved analysis regimes shows that there is $\sim 70\%$ overlap in simulated events and $\sim 50\%$ overlap in the data. This overlap was totally removed by setting a hard cut between the resolved and boosted analyses at $p_T^V = 400\,\text{GeV}$.

The inclusive Higgs samples study shows that the $t\bar{t}H$ contribution is rather small compared with the ZH and WH samples.

The 2-lepton MVA training started with testing out variables until the baseline set was introduced. The analysis went from the old set of samples (32-15) to an updated set (33-05) which led to improved output signifcances: $Z_{250-400GeV} = 2.86\sigma$; $Z_{400+GeV} = 2.26\sigma$. The 2-lepton MVA training continued to improve. Adding truth tagging has countered the overtraining issue that plagued the 2L region. The process of optimising hyperparameters was developed to yield better significances. Colour ring variables produced some exciting improvements and were therefore added to the baseline variable list improving the overall significance in the analysis.

Bibliography

[1] S. Abachi et al. Observation of the top quark. *Phys. Rev. Lett.*, 74:2632–2637, 1995.

[2] F. Abe et al. Observation of top quark production in $\bar{p}p$ collisions. *Phys. Rev. Lett.*, 74:2626–2631, 1995.

[3] K. Kodama et al. Observation of tau neutrino interactions. *Phys. Lett. B*, 504:218–224, 2001.

[4] Georges Aad et al. Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC. *Phys. Lett. B*, 716:1–29, 2012.

[5] Peter W. Higgs. Broken symmetries and the masses of gauge bosons. *Phys. Rev. Lett.*, 13:508–509, Oct 1964.

[6] F. Englert and R. Brout. Broken symmetry and the mass of gauge vector mesons. *Phys. Rev. Lett.*, 13:321–323, Aug 1964.

[7] P.A. Zyla et al. Particle Data Group. Review of Particle Physics. *Progress of Theoretical and Experimental Physics*, 2020(8):083C01, 08 2020.

[8] Wikimedia Commons. Standard Model of Elementary Particles. `https://commons.wikimedia.org/wiki/File:Standard_Model_of_Elementary_Particles.svg`, 2023. (Accessed: January 2023).

[9] Department of Physics Oxford University. Why two higgs are better than one. `https://www.physics.ox.ac.uk/news/why-two-higgs-are-better-one`, 2024. (Accessed: February 2024).

[10] Institute of Physics. Higgs boson and the history of the universe. `https://www.iop.org/higgs-boson-and-history-universe`, 2024. (Accessed: February 2024).

[11] Dwayne Spiteri. *Higgs boson studies: associated production with a vector boson and decay into b-quarks using the ATLAS Run-2 dataset.* PhD thesis, Glasgow U., 2021.

[12] Ana Lopes. CMS homes in on Higgs boson's lifetime. https://home.cern/news/news/physics/cms-homes-higgs-bosons-lifetime, 2021. (Accessed: February 2024).

[13] CERN. CERN Accelerator Complex. https://www.lhc-closer.es/taking_a_closer_look_at_lhc/0.cern_accelerator_complex, 2023. (Accessed: 19th October 2023).

[14] G. Aad et al. The ATLAS Experiment at the CERN Large Hadron Collider. *JINST*, 3:S08003, 2008.

[15] S. Chatrchyan et al. The CMS Experiment at the CERN LHC. *JINST*, 3:S08004, 2008.

[16] A. Augusto Alves, Jr. et al. The LHCb Detector at the LHC. *JINST*, 3:S08005, 2008.

[17] K. Aamodt et al. The ALICE experiment at the CERN LHC. *JINST*, 3:S08002, 2008.

[18] CERN. CERN Aerial view. http://cds.cern.ch/record/841506/files/, 2023. (Accessed: 19th October 2023).

[19] CERN. LHC Magnet Types. https://lhc-machine-outreach.web.cern.ch/components/magnets/types_of_magnets.htm, 2023. (Accessed: 13th October 2023).

[20] CERN. Pulling Together: Superconducting Magnets. https://www.home.cern/science/engineering/pulling-together-superconducting-electromagnets, 2023. (Accessed: 13th October 2023).

[21] CERN. The Proton Synchrotron. https://home.cern/science/accelerators/proton-synchrotron, 2023. (Accessed: 20th October 2023).

[22] CERN. The HL-LHC project. https://hilumilhc.web.cern.ch/content/hl-lhc-project, 2023. (Accessed: 31st October 2023).

[23] S. Dittmaier et al. Handbook of LHC Higgs Cross Sections: 1. Inclusive Observables. 1 2011.

[24] Georges Aad et al. Improved luminosity determination in pp collisions at sqrt(s) = 7 TeV using the ATLAS detector at the LHC. *Eur. Phys. J. C*, 73(8):2518, 2013.

[25] Morad Aaboud et al. Luminosity determination in pp collisions at $\sqrt{s} = 8$ TeV using the ATLAS detector at the LHC. *Eur. Phys. J. C*, 76(12):653, 2016.

[26] Luminosity determination in *pp* collisions at $\sqrt{s} = 13$ TeV using the ATLAS detector at the LHC. 6 2019.

[27] Jeremiah Jet Goodson. *Search for supersymmetry in states with large missing transverse momentum and three leptons including a Z-boson.* PhD thesis, Stony Brook U., 5 2012.

[28] ATLAS inner detector: Technical design report. Vol. 1. 4 1997.

[29] ATLAS inner detector: Technical design report. Vol. 2. 4 1997.

[30] M. S. Alam et al. ATLAS pixel detector: Technical design report. 5 1998.

[31] Georges Aad et al. Operation and performance of the ATLAS semiconductor tracker in LHC Run 2. *JINST*, 17(01):P01013, 2022.

[32] E. Abat et al. The ATLAS Transition Radiation Tracker (TRT) proportional drift tube: Design and performance. *JINST*, 3:P02013, 2008.

[33] A. Airapetian et al. ATLAS calorimeter performance Technical Design Report. 12 1996.

[34] ATLAS liquid argon calorimeter: Technical design report. 12 1996.

[35] Nikiforos Nikiforou. Performance of the ATLAS Liquid Argon Calorimeter after three years of LHC operation and plans for a future upgrade. 2013. Comments: 12 pages, 25 figures, Proceedings of talk presented in: Advancements in Nuclear Instrumentation Measurement Methods and their Applications, Marseille, 2013.

[36] ATLAS tile calorimeter: Technical design report. 12 1996.

[37] Georges Aad et al. Improved luminosity determination in pp collisions at sqrt(s) = 7 TeV using the ATLAS detector at the LHC. *Eur. Phys. J. C*, 73(8):2518, 2013.

[38] Georges Aad et al. Resolution of the ATLAS muon spectrometer monitored drift tubes in LHC Run 2. *JINST*, 14(09):P09011, 2019.

[39] ATLAS magnet system: Technical design report. 4 1997.

[40] Georges Aad et al. Operation of the ATLAS trigger system in Run 2. *JINST*, 15(10):P10004, 2020.

[41] ATLAS. Trigger and Data Aquisition. `https://atlas.cern/Discover/Detector/Trigger-DAQ`, 2024. (Accessed: February 2024).

[42] Georges Aad et al. Measurements of WH and ZH production in the $H \to b\bar{b}$ decay channel in pp collisions at 13 TeV with the ATLAS detector. *Eur. Phys. J. C*, 81(2):178, 2021.

[43] M. Aaboud, G. Aad, B. Abbott, et al. Observation of h→bb¯ decays and vh production with the atlas detector. *Physics Letters B*, 786:59–86, 2018.

[44] Georges Aad et al. Measurement of the associated production of a Higgs boson decaying into b-quarks with a vector boson at high transverse momentum in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector. *Phys. Lett. B*, 816:136204, 2021.

[45] Georges Aad et al. Direct constraint on the Higgs-charm coupling from a search for Higgs boson decays into charm quarks with the ATLAS detector. *Eur. Phys. J. C*, 82:717, 2022.

[46] Stefano Frixione, Paolo Nason, and Carlo Oleari. Matching NLO QCD computations with Parton Shower simulations: the POWHEG method. *JHEP*, 11:070, 2007.

[47] T. Gleisberg, Stefan. Hoeche, F. Krauss, M. Schonherr, S. Schumann, F. Siegert, and J. Winter. Event generation with SHERPA 1.1. *JHEP*, 02:007, 2009.

[48] Torbjorn Sjostrand, Stephen Mrenna, and Peter Z. Skands. A Brief Introduction to PYTHIA 8.1. *Comput. Phys. Commun.*, 178:852–867, 2008.

[49] Gionata Luisoni, Paolo Nason, Carlo Oleari, and Francesco Tramontano. $HW^{\pm}$/HZ + 0 and 1 jet at NLO with the POWHEG BOX interfaced to GoSam and their merging within MiNLO. *JHEP*, 10:083, 2013.

[50] Torbjorn Sjostrand, Stephen Mrenna, and Peter Z. Skands. A Brief Introduction to PYTHIA 8.1. *Comput. Phys. Commun.*, 178:852–867, 2008.

[51] Oliver Brein, Abdelhak Djouadi, and Robert Harlander. NNLO QCD corrections to the Higgs-strahlung processes at hadron colliders. *Phys. Lett. B*, 579:149–156, 2004.

[52] Richard D. Ball et al. Parton Distribution Benchmarking with LHC Data. *JHEP*, 04:125, 2013.

[53] Georges Aad et al. Topological cell clustering in the ATLAS calorimeters and its performance in LHC Run 1. *Eur. Phys. J. C*, 77:490, 2017.

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

[55] Georges Aad et al. ATLAS flavour-tagging algorithms for the LHC Run 2 pp collision dataset. *Eur. Phys. J. C*, 83(7):681, 2023.

[56] Georges Aad et al. ATLAS flavour-tagging algorithms for the LHC Run 2 pp collision dataset. *Eur. Phys. J. C*, 83(7):681, 2023.

[57] Georges Aad et al. ATLAS b-jet identification performance and efficiency measurement with $t\bar{t}$ events in pp collisions at $\sqrt{s} = 13$ TeV. *Eur. Phys. J. C*, 79(11):970, 2019.

[58] J. Therhaag. Tmva toolkit for multivariate data analysis in root. *PoS ICHEP2010*, 510:50, 2010. ed. by B. Pire et al.

[59] Andrew J. Larkoski, Ian Moult, and Duff Neill. Analytic Boosted Boson Discrimination. *JHEP*, 05:117, 2016.

[60] Flavor Tagging Efficiency Parametrisations with Graph Neural Networks. 2022.

[61] Andy Buckley, Giuseppe Callea, Andrew J. Larkowski, and Simone Marzani. Colouring the Higgs boson. *SciPost Phys. Proc.*, 10:001, 2022.

[62] Andy Buckley, Giuseppe Callea, Andrew J. Larkoski, and Simone Marzani. An Optimal Observable for Color Singlet Identification. *SciPost Phys.*, 9:026, 2020.

[63] Morad Aaboud et al. Observation of $H \to b\bar{b}$ decays and VH production with the ATLAS detector. *Phys. Lett. B*, 786:59–86, 2018.