Search for the Higgs Boson Decay into Bottom and Charm Quarks Using Proton-Proton Collisions at $\sqrt{s} = 13$ TeV

Dissertation
zur
Erlangung des Doktorgrades (Dr. rer. nat.)
der
Mathematisch-Naturwissenschaftlichen Fakultät
der
Rheinischen Friedrich-Wilhelms-Universität Bonn

von
Elisabeth Schopf
aus
Berlin

Bonn, April 2018

Dieser Forschungsbericht wurde als Dissertation von der Mathematisch-Naturwissenschaftlichen Fakultät der Universität Bonn angenommen und ist auf dem Hochschulschriftenserver der ULB Bonn http://hss.ulb.uni-bonn.de/diss_online elektronisch publiziert.

1. Gutachter:	Prof. Dr. Norbert Wermes
2. Gutachter:	Prof. Dr. Ian Brock

Tag der Promotion:	17.07.2018
Erscheinungsjahr:	2019

Contents

CHAPTER 1

Introduction

To seek for explanations of the underlying mechanisms of the world we live in is a distinctive desire of humankind. The earliest attempts to do so resulted in various mythologies of different cultures, some of which are still known today. As an example, the Greek mythology tells that a Titan by the name of "Atlas" is holding the sky because it would otherwise crash down on earth. Nowadays, one approach to explain the phenomena around us is to study the fundamental particles and their interactions in the field of particle physics. This is the aim of an experiment called "ATLAS".

The theory of particle physics, the Standard Model, is a gauge theory that generates interactions between particles based on a symmetry principle. Although the Standard Model is very successful in explaining the particle interactions we observe and also made predictions that were later discovered to be true, it had one flaw: it could not explain why particles have a mass without having a broken symmetry as a result. This was the case until a mechanism of spontaneous symmetry breaking was proposed by Francois Englert and Robert Brout. Almost at the same time, Peter Higgs formulated a very similar idea. Since his name is so memorable everything connected to this mechanism was named after him: Higgs mechanism, Higgs field, Higgs boson. This newly introduced theory assumes that the universe was in a meta-stable state and that the underlying symmetry was, at some point, spontaneously broken. Thus giving mass to the mediator particles of the weak interaction and restoring the symmetries of the Standard Model. Luckily, from an experimental particle physicist's point of view, this theory also postulated a new particle: the Higgs boson whose mass is a free parameter of the theory. After many decades of unsuccessful searches for the Higgs boson, it was legitimate to ask: Do we need another theory to explain the masses of the fundamental particles? A last big effort was made with the construction of a proton-proton collider, the Large Hadron Collider, whose energy range covers all possible Higgs boson masses. The ATLAS and CMS experiment were designed with the task of finding the Higgs boson in mind. Finally, in 2012 we got our answer: there is a new boson that looks like the boson we were looking for and it has a mass of 125 GeV. Shortly after this discovery, Peter Higgs and Francois Englert — Robert Brout was already deceased at that time — were awarded the Nobel Prize in physics for their work. However, to gain confidence in the claim that this new particle is indeed the long searched for Standard Model Higgs boson, we try to measure as many of its properties as possible. An important piece of the puzzle are the couplings of the Higgs boson to other Standard Model particles which are predicted to be proportional to the masses of the particles. In addition, there is a conceptual difference between fermions and bosons since the acquisition of mass of the weak bosons is a natural consequence of the Higgs mechanism whereas mass acquisition of fermions via couplings to the Higgs field was added "ad hoc". One way to probe these couplings is via the decay of the unstable Higgs boson into Standard Model particles. So far definitive proof has only been found for decays into bosons although the decay with the largest probability of 58% is the decay into a pair of fermions, namely bottom quarks. The reason why the decay into bottom quarks is not discovered yet, are the challenges that are connected to the hadronic signature of this decay. Due to the nature of the strong force quarks cannot exist as free particles but undergo fragmentation instead and form bundles of hadrons, so-called jets. Jets are a very common signature in hadron collisions. Therefore the challenge of the Higgs boson to bottom quarks decay channel is to distinguish these processes from other processes that form jets, which have an approximately 10^7 larger production cross section than the production of a Higgs boson. But also the decay into lighter fermions is interesting since the current experimental constraints still allow large potential enhancements of these decays, e.g. the coupling to charm quarks might be of a similar order than the couplings to bottom quarks.

This thesis presents a search for the Higgs boson decay into bottom quarks, $H \rightarrow b\bar{b}$, and for the Higgs boson decay into charm quarks, $H \rightarrow c\bar{c}$, in associated production with a Z boson. It uses 36.1 fb^{-1} of proton-proton collision data from the LHC that were recorded with the ATLAS experiment at a collision energy of 13 TeV. The associated production with a Z boson is probed to identify the events of interest via the leptonic Z boson decay products — neutrinos, electrons or muons — and reduce the amount of events from Standard Model processes that produce similar signatures as the signal. To study $Z(H \rightarrow b\bar{b})$ final states an analysis that searches for a new boson A decaying as $A \rightarrow Z(H \rightarrow b\bar{b})$ is performed with the first 3.2 fb^{-1} of proton-proton collision data at 13 TeV. In the following the underlying theoretical framework and the experimental set-up are introduced in chapters 2 and 3, respectively. The phenomena involved in proton-proton collisions and the simulation of such collisions is discussed in chapter 4. Chapter 5 explains the reconstruction and identification of particles that are produced in the proton-proton collision and were recorded with the ATLAS detector. The necessary statistical tools to analyse the large amount of proton-proton collision data are introduced in chapter 6. To enhance the sensitivity of the $H \rightarrow b\bar{b}$ analysis a new correction method, using multivariate techniques, for the jets that are formed by the fragmentation of the bottom quarks is introduced and studied in chapter 7. Afterwards the analysis strategy and results of the $Z(H \rightarrow b\bar{b})$ measurement are presented and the analysis strategy is further validated using WZ and ZZ events, which is all detailed in chapter 8. Next, in chapter 9 a novel search for the $H \rightarrow c\bar{c}$ decay in associated production with a Z boson that decays into a pair of electrons or muons is presented. The search strategy is introduced and an upper limit on the cross section times branching ratio for this production and decay channel is set. Eventually, the analysis and exclusion limits obtained for the cross section times branching ratio for the $A \rightarrow Z(H \rightarrow b\bar{b})$ decay for hypothetical A boson masses between 220 GeV and 2 TeV are presented in chapter 10. The results of those analyses are summarised in chapter 11.

CHAPTER 2

Theory Introduction

Our current description of elementary particles and their interactions is based on the mathematical foundation of the so-called Standard Model (SM) of particle physics. Historically, the fundamental idea reaches back to the beginning of the 20th century, when it was discovered that electrons do not only have particle but also wave properties. From there quantum field theory evolved, which views particles as excitations of a more fundamental entity called a field. It was developed from the combination of the dual wave-particle descriptions introduced in quantum mechanics and the theory of special relativity. The SM is a specific quantum field theory, named a gauge theory. In a gauge theory, the forces that describe the interactions between matter fields are generated by a symmetry principle, similar to the way energy and momentum conservation arise from time and space translational invariance in classical mechanics.

A key ingredient of the SM is the existence of a special field, called the Higgs field. Particles interacting with the Higgs field acquire mass proportionally to the strength of this interaction. The discovery of a boson in 2012 whose properties are consistent with the SM Higgs boson, e.g. the quantum associated with the Higgs field, was a breakthrough and made the SM a consistent, even though still incomplete, theory of nature.

2.1 The Standard Model as a Gauge Theory

In general, a quantum field theory is a theory where matter and forces are described by quantum fields. The excitation of these fields are physical, i.e. observable, particles. One way to excite the field is to provide energy to the vacuum, which is the basic idea of particle colliders.

In the SM the matter fields are fundamental whereas the force fields arise from a symmetry principle. This is achieved by requiring the Lagrangian[1] of the matter field to be symmetric under local transformations. In order to ensure local symmetry extra gauge fields are introduced in the Lagrangian. These extra fields are the force fields and terms that can be interpreted as interactions of the matter field with the force field arise from the modified Lagrangian. The mathematical framework to describe this desired behaviour of the SM is a theory of symmetry such as group theory. Over the years, it evolved that the necessary group that needs to be introduced in the SM to reproduce reality is:

$$U(1)_Y \times SU(2)_L \times SU(3)_c \tag{2.1}$$

the factor groups generate the fundamental forces observed in particle interactions. In order to do

[1] The Lagrangian is expressed as $T - V$ with T the kinetic and V the potential energy and the equation of motions can be derived from it.

calculations, e.g. to predict cross sections of a certain SM interaction which may become infinitely complicated, a so-called perturbative expansion can be used, which offers an approximate solution to the problem. One formalism to do perturbative calculations are Feynman rules. Feynman rules represent the symmetries of the interaction Lagrangian and a graphical representation of those rules are Feynman diagrams [1, 2].

2.2 Elements of the Standard Model

In order to describe the world we live in, three particles need to be introduced in the SM: up- (u) and down (d) quarks that form protons and neutrons and electrons (e) that form atoms together with the neutrons and protons. To explain the phenomenon of radioactive β decay a fourth particle was postulated and subsequently discovered: the neutrino (ν). Those four particles are grouped in two families: one quark and one lepton family. However, over the years two more families have been discovered each, which are copies of the first family, i.e. comprise the same physical properties, but are more massive. All quarks and leptons, which are fermions, are listed in table 2.1. In addition, each of these particles has a corresponding anti particle, which exhibits the same features but has opposite quantum numbers.

The three fundamental forces — electromagnetic, weak and strong force — are included in the SM as well. The interactions generated by these forces are described via the exchange of force carrier particles, which are bosons. The photon is associated to electromagnetic interactions, the $W^{\pm}$ and Z boson to weak interactions and the gluon to strong interactions. The fermions and bosons possesses "charge" properties that determine how they participate in the different interactions. The charges of the fermions are listed in table 2.1 as well. All particles that have an electric charge take part in electromagnetic interactions and can couple to the photon. The weak isospin is the "charge" of the weak interaction and determines the couplings to the $W^{\pm}$ and Z boson. The charge of strong interactions is the so-called colour charge, which is either red, green or blue. Only the quarks and the gluons themselves carry colour charge and participate in strong interactions.

The electromagnetic and weak interactions are incorporated in the SM Lagrangian in a unified way, called electroweak (EW) theory. However, to obtain massive W^{+}, W^{-} and Z bosons and massless photon, which are observed in experiments, the symmetry of the EW gauge group has to be spontaneously broken with the pattern:

$$U(1)_L \times SU(2)_Y \rightarrow U(1)_{\text{em}} \tag{2.2}$$

Spontaneous symmetry breaking means that the Lagrangrian stays symmetric under local transformations but the vacuum, which is one of its solutions is not symmetric any more. The mechanism that introduces this symmetry breaking is the Higgs mechanism and it is explained in the next section. Strong interactions, described by the theory of quantum chromo dynamics (QCD), are included in the SM Lagrangian separately from the EW theory. The gauge group of QCD is $SU(3)_c$. A unique feature of QCD interactions is that the strength of the coupling, given by the coupling constant α_{S}, decreases with energy. As a consequence, at low energies quarks and gluons cannot be isolated, which is why there are no free quarks and gluons in nature but only hadrons, e.g. protons, which are bound colour-neutral states. In contrast, at high energies, like the ones produced in particle colliders, quarks and gluons behave as quasi-free particles, which is called asymptotic freedom. The transition between the quasi-free states to observable colour neutral states produces the typical features observed in hadron colliders, so called jets [1–3]. More details about jets are given in chapters 4, 5.5 and 7.

	electric charge	weak isospin	colour charge	Fermions			Bosons
Quarks:	$+\frac{2}{3}e$ $-\frac{1}{3}e$	$+\frac{1}{2}$ $-\frac{1}{2}$	r, g, b r, g, b	$\begin{pmatrix} u \\ d \end{pmatrix}$	$\begin{pmatrix} c \\ s \end{pmatrix}$	$\begin{pmatrix} t \\ b \end{pmatrix}$	γ g
Leptons:	$-e$ 0	$-\frac{1}{2}$ $+\frac{1}{2}$	– –	$\begin{pmatrix} e \\ \nu_e \end{pmatrix}$	$\begin{pmatrix} \mu \\ \nu_\mu \end{pmatrix}$	$\begin{pmatrix} \tau \\ \nu_\tau \end{pmatrix}$	$W^{\pm}$ Z H

Table 2.1: The fundamental fermions — quarks and leptons — and bosons of the SM. Fermions are ordered by increasing mass from left to right. Bosons are ordered by increasing mass from top to bottom. The electrical charges (in units of electron charge e), the weak isospin and the colour charge (either red, green or blue) of the fermions are given as well. The quarks are: up (u), down (d), charm (c), strange (s), top (t) and bottom (b). The leptons are: electron (e), muon (μ) and τ-lepton (τ) and their corresponding neutrinos ν. The photon (γ) is the exchange particle of the electromagnetic force, the $W^{\pm}$ and Z boson of the weak force, the gluons (g) of the strong force and the Higgs boson (H) is the consequence of the Higgs mechanism.

2.3 The Higgs Mechanism

Spontaneous symmetry breaking s introduced in the SM through the Higgs mechanism. The easiest formulation is to introduce a fundamental scalar. The Lagrangian of a scalar field ϕ close to the phase transition, i.e. from not-broken to broken EW symmetry, is:

$$\mathcal{L}_\phi = \frac{1}{2}\partial_\mu \phi \partial^\mu \phi - \frac{1}{2}\mu^2 \phi^2 - \frac{1}{4}\lambda \phi^4 \tag{2.3}$$

with $\lambda > 0$ to ensure a stable potential and $\mu^2 < 0$ to produce a potential that has a shape which actually breaks the vacuum symmetry. The simplest potential that exhibits this property has the shape of a Mexican hat. The most general form of ϕ for which the Lagrangian is $U(1) \times U(2)$ invariant, is a complex doublet of the form:

$$\phi = \begin{pmatrix} \phi_1 + i\phi_2 \\ \phi_3 + i\phi_4 \end{pmatrix} \tag{2.4}$$

Its vacuum solution has to be not symmetric, e.g.:

$$\phi_0 = \begin{pmatrix} 0 \\ \eta \end{pmatrix} \tag{2.5}$$

The effects of the Lagrangian can be studied by introducing excitations of the field around the vacuum:

$$\phi \sim \begin{pmatrix} 0 \\ \eta + \frac{h(x)}{\sqrt{2}} \end{pmatrix} \tag{2.6}$$

This choice breaks the EW symmetry along the directions, corresponding to 3 out of 4, leaving the generator of electromagnetic interactions untouched. According to the Nambu-Goldstone theorem, this would give rise to 3 massless scalar fields which are not observed in nature. The Higgs mechanism explains how these 3 extra degrees of freedom are "eaten" by the three weak bosons and become their longitudinal degree of freedom. Thus they acquire mass. The extra degree of freedom $h(x)$ that is left over, corresponds to an additional scalar boson, the SM Higgs boson, and its mass is a free parameter of the model. A boson that, so far, exhibits all properties of this boson was discovered in 2012 by the

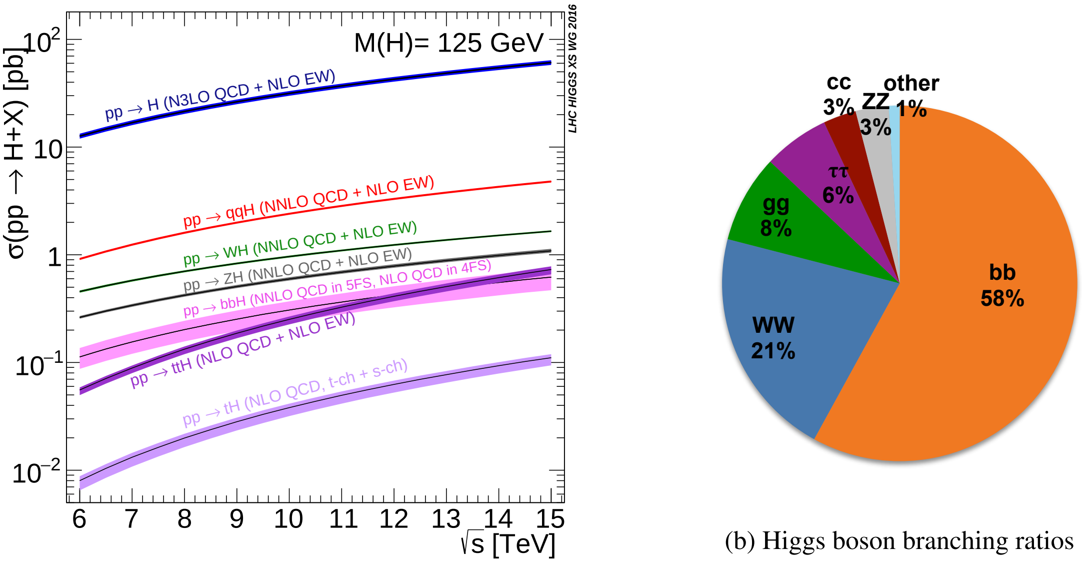
ATLAS and CMS collaboration. The mass of the discovered boson is measured to be at approximately 125 GeV, which has implications on its production cross section and the branching ratio (BR) of its decays as discussed in the next section.

Although the masses of the weak bosons arise from the Higgs mechanism itself, the masses of the fermions have to be added to the Lagrangian in an "ad hoc" way. Since the Higgs mechanism already successfully introduces masses for bosons in the SM it is assumed that the fermions acquire their masses from interactions with the Higgs field as well. The so-called Yukawa coupling, originally designed to describe forces between nucleons via the exchange of pions, offers a formalism to describe interactions between fermion fields and scalar fields. Thus it is used to introduce Higgs fermion coupling terms in the Lagrangian, which does not break the symmetry of the Lagrangian. If the fermion masses are generated in this way the coupling strength between the Higgs and fermion fields has to be proportional to the mass of the fermion. Thus the measurement of fermionic Higgs boson decays, e.g. $H \rightarrow b\bar{b}$, is crucial to confirm the Yukawa nature of the fermion Higgs couplings and therefore the mass generation mechanism for fermions [1, 2].

2.3.1 Higgs Boson Production and Decay

In proton-proton (pp) collisions, as investigated for this thesis, the Higgs boson is produced via different production channels that have different production cross sections. The production cross sections of them are shown in figure 2.1 as a function of the collision energy. The channel with the highest cross section is the production via so-called gluon fusion, denoted as $pp \rightarrow H$, where the gluons produce a loop of heavy particles to which the Higgs boson couples. All other production channels are classified via the additional particles that are produced in association with the Higgs boson. The second largest production cross section, which is approximately 10 times lower than gluon fusion, is so-called (weak) vector boson fusion ($pp \rightarrow qqH$). The third largest production cross section is associated production with a weak boson: $pp \rightarrow WH$ and $pp \rightarrow ZH$. This is the production channel targeted in the analyses of this thesis. The cross section for WH production is approximately 1.5 times higher than for ZH production due to the two possibilities for WH production (W^+H and W^-H), the lower mass of the W boson compared to the Z boson and the parton distribution functions (PDF) content of the protons in the collision (more details on PDFs are given in section 4.1). Associated production of the Higgs boson with quarks of the third quark family have even lower production cross sections. The Higgs boson production cross sections have a dependence on the collision energy, $\sqrt{s}$, and they increase for increased $\sqrt{s}$. The production cross section for Higgs bosons also depends on the mass of the Higgs boson itself with lower cross sections for high Higgs boson masses. This was taken into consideration in the design of the Large Hadron Collider (LHC). Since the mass of the Higgs boson is not given by the SM the goal was to cover a large range of possible Higgs boson masses.

The Higgs boson is not a stable particle, i.e. it decays shortly after its production. Therefore it has to be detected via its decay products. Since the coupling of the Higgs boson to fermions and bosons is proportional to their mass the probability of a certain decay is proportional to the mass of the decay product as well. However, the Higgs boson BR are further constrained by the Higgs boson mass. A pie chart of the Higgs boson BR for a SM Higgs boson with a mass of 125 GeV is displayed in figure 2.1. Decay channels have higher BR if both or at least one of the decay particles can be produced "on shell", i.e. as real particles with their given mass. This is the case if the Higgs boson mass is larger than the masses of the decay products. Therefore the decay into a pair of bottom quarks has the largest BR since they are the heaviest pair of SM particles whose combined mass is smaller than the SM Higgs boson mass. The BR of the $H \rightarrow b\bar{b}$ decay is 58%. The decay into massless particles, e.g. gluons and photons, is possible as well. Similar to the gluon fusion production channel the Higgs produces a triangle of heavy particles to which the gluons and photons couple.



(b) Higgs boson branching ratios

(a) Higgs boson production cross sections

Figure 2.1: The production cross section as a function of the *pp* collision energy (a) and branching ratios (b) for a SM Higgs boson [4]. In both cases the Higgs boson mass is assumed to be 125 GeV. The precision of the cross section is given by the orders of QCD and EW theory up to which the cross sections are calculated.

CHAPTER 3

The LHC and ATLAS Experiment

In the last decades particle accelerators have been proven to be a useful tool to study elementary particle physics. Most accelerators such as the LHC are built as a ring which allows the particles to pass the acceleration chain many times and gain more energy. Once they reached a certain energy they are brought to collision. The LHC is designed to collide protons at a collision energy of 13 TeV. Collision experiments with protons are also called "discovery experiments". Protons can be accelerated to higher energies than e.g. electrons without substantial loss due to synchrotron radiation. This effect describes the radiation of photons from charged particles that are bent which causes the initial particle to lose part of its energy. The energy loss ΔE due to synchrotron radiation per turn in the accelerator is proportional to [5]:

$$\Delta E \propto \frac{E^4}{\rho m^4} \tag{3.1}$$

where E is the accelerated particle's energy, m its mass and ρ the bending radius. Due to the relatively large mass of protons, $m \approx 1\,\text{GeV}$, allows to achieve high energies without sizeable energy loss from synchrotron radiation. If the LHC operates at its design energy the energy loss of the protons per turn is roughly 7 keV which is small compared to their energy of 7 TeV. In addition, a large energy regime can be covered in pp collision, without additional tuning of the collision energy, due the non-fundamental nature of protons. At high energies the constituents, which carry a fraction of the proton momentum, interact [5–8].

3.1 The Large Hadron Collider

The LHC is a superconducting ring accelerator located at the Conseil Européen pour la Recherche Nucléaire (CERN) near the city of Geneva in Switzerland. The accelerator was built in an underground tunnel of 27 km length which was previously used for the Large Electron Positron Collider (LEP). The focus of the physics program at the LHC is pp collisions but it is also operated with heavy ions such as lead. After initial problems the LHC was operated at $\sqrt{s} = 7\,\text{TeV}$ and $\sqrt{s} = 8\,\text{TeV}$ in the years of 2011 and 2012, respectively. This run period is commonly referred to as run 1. Since 2015 the LHC is operated at a collision energy of $\sqrt{s} = 13\,\text{TeV}$. This run period will continue until the end of 2018 and is referred to as run 2. The data set used in this thesis was collected at 13 TeV during LHC operation in 2015 and 2016.

In order to reach an energy of several TeV, the protons, which are extracted from an ionised hydrogen source, are first inserted into a linear accelerator to accelerate them to an energy of 50 MeV. Afterwards they pass through a chain of ring accelerators to increase their energy to 450 GeV before being injected into the LHC to be accelerated to their final energy. The acceleration is achieved with electric fields from superconducting radio frequency cavities. The LHC also contains superconducting dipole magnets which bend the protons to keep them on their circular trajectory in the LHC ring. The field strength of the magnets has to be finely tuned with each turn since the protons gain more and more energy which requires higher field strengths. In order to collide the protons, two beams of protons are circulating in opposing directions in the LHC ring which encompasses two beam pipes. An ultra-high vacuum is achieved in the beam pipes to avoid collisions of the protons with molecules in the beam pipe, which would distort the beams. The beam pipes cross at four points along the LHC ring to collide the protons. At the four interaction points the following experiments are set up: ATLAS, ALICE, CMS and LHCb. This thesis uses the data collected by the ATLAS experiment which is discussed in more detail in section 3.2.

The proton beams consist of proton packets, so-called bunches, which contain $O(10^{11})$ protons each. Up to roughly 2800 bunches per beam can circulate in the LHC at the same time. The nominal distance between proton bunches, which travel almost at the speed of light, results in the beams crossing at the interaction points each 25 ns. At the beginning of the 2015 data taking period this so-called bunch spacing was increased to 50 ns. All beam properties are combined into a property to quantify an accelerator's performance: the luminosity L. It is defined as [8]:

$$L = \frac{N_\mathrm{b}^2 n_\mathrm{b} f_\mathrm{rev} \gamma_r}{4\pi \epsilon_n \beta^*} F \tag{3.2}$$

with N_b the number of protons per bunch, n_b the number of bunches, f_rev the revolution frequency, γ_r the relativistic gamma factor, ϵ_n the normalised transverse beam emittance, β^* the beta function of the beam at the collision point and F the geometric luminosity reduction factor since the beams cross under a certain angle at the interaction points. The normalised transverse beam emittance and beta function are both measures for the dimensions of the beams, i.e. how focussed the beams are. The LHC is designed to deliver a luminosity of up to $10^{34}\,\mathrm{s}^{-1}\,\mathrm{cm}^{-2}$. The luminosity is directly proportional to the number of collisions. The integrated luminosity $\mathcal{L}$ is the luminosity integrated over time $\mathcal{L} = \int L dt$ and represents the amount of delivered data during a specific time period, t. The integrated luminosity of good quality physics data collected by the ATLAS detector during the 2015+2016 LHC run period is $\mathcal{L} = 36.1 \times 10^{39}\,\mathrm{cm}^{-2} = 36.1\,\mathrm{fb}^{-1}$ compared to an integrated luminosity of $42.7\,\mathrm{fb}^{-1}$ delivered by the LHC. With the cross section, σ, for a certain physics process at a given collision energy this can be translated into an amount of recorded events of this physics process [8]:

$$N = \sigma \mathcal{L} \tag{3.3}$$

The total amount of recorded pp collisions ($\sigma(pp \to \text{anything}) \approx 100\,\mathrm{mb}$) by the ATLAS experiment in 2015+2016 is $O(10^{15})$. This large amount of collisions are necessary to study SM processes, e.g. Z boson production, with a high statistical precision as well as measure and discover rarely occurring processes, e.g. production of a Higgs boson. Therefore, in addition to the high collision energy, a high luminosity was a main design criterion for the LHC. Figure 3.1 shows the cross sections of several SM processes as a function of the pp collision energy. In the energy regime of the LHC the production of b-quarks, high energy jets and W/Z bosons is abundant. In comparison, the production cross section of a SM Higgs boson with a mass of 125 GeV is several orders of magnitude lower. The figure additionally shows that the production of heavy particles, like the top quark, has a strong dependence on the collision energy [8–12].

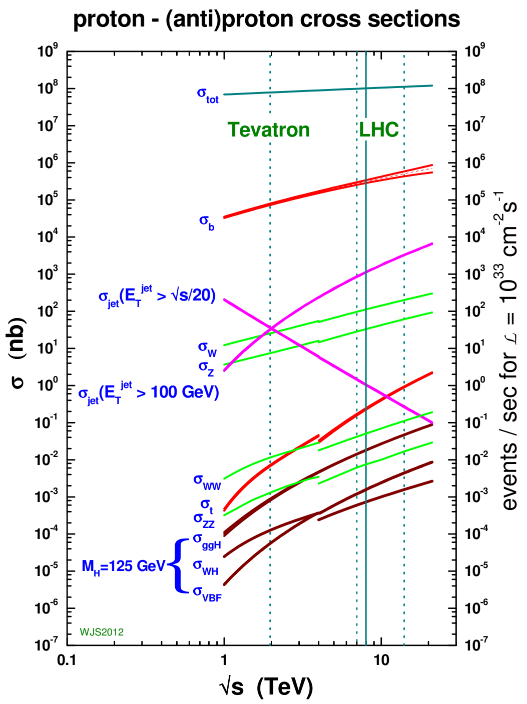


Figure 3.1: The cross section of several SM processes as a function of the collision energy $\sqrt{s}$ of proton-(anti)proton colliders. At $\sqrt{s} = 4\,\text{TeV}$ a switch-over is made from proton-antiproton colliders, like the Tevatron, to proton-proton (pp) colliders, like the LHC. This causes discontinuities in some of the cross sections. The solid vertical line indicates the LHC at $\sqrt{s} = 8\,\text{TeV}$. The dashed line next to it on the right side indicates the LHC at its design energy of $\sqrt{s} = 14\,\text{TeV}$ [13].

3.2 The ATLAS Experiment

The ATLAS experiment is an experimental set-up to study pp collisions at the LHC. It has a very broad physics program and therefore uses a general purpose detector that will be described in section 3.2.1. One of the great achievements of the ATLAS experiment was the discovery of a new particle in 2012 whose properties are consistent with the SM Higgs boson [14]. Since then many measurements have been made to investigate the nature of the discovered boson and its properties. Besides the studies on the Higgs boson, the ATLAS experiment analyses other SM processes, including top quark physics, and searches for new physics phenomena, e.g. dark matter and supersymmetric particles, using the data obtained with the ATLAS detector.

3.2.1 The ATLAS Detector

The ATLAS detector is a multi-layer detector which detects the pp collisions of the LHC. It has a cylindrical layout around the beam pipe and is symmetric around the beam pipe as well as in forward and backward direction with respect to the beam pipe. The ATLAS detector contains several sub-detectors which make use of a wide range of detection techniques. The combination of the information gathered in the various sub-detectors is used to reconstruct the properties of the particles in the recorded event as well as identify those particles. The coordinate system, which is used to reconstruct particles recorded in the ATLAS detector, has its origin in the primary vertex (PV). The z-axis points along the beam pipe, the y-axis points upwards and the x-axis is perpendicular to $\vec{z}$ and $\vec{y}$ and points towards the centre of the LHC ring. The azimuthal angle ϕ describes angles in the xy-plane, i.e. around the beam axis, whereas the polar angle θ describes angles in the yz-plane, i.e. with respect to the beam axis. Instead of θ more commonly the pseudorapidity η is used since differences in η are Lorentz invariant under boost in z-direction:

$$\eta = -\ln\left(\tan\frac{\theta}{2}\right) \tag{3.4}$$

According to this definition $\eta = 0$ is perpendicular to the z direction and $|\eta| = \infty$ is pointing in z direction. Angular distances of particles recorded in the ATLAS detector are expressed in the $\eta\phi$-plane by the distance measure ΔR:

$$\Delta R = \sqrt{(\Delta\phi)^2 + (\Delta\eta)^2} \tag{3.5}$$

Often properties are given in the plane transverse to the beam, i.e. in the xy-plane. For example the transverse momentum p_T of a particle is defined as: $p_\mathrm{T} = \sqrt{p_x^2 + p_y^2} = |\vec{p}|\sin\theta$, with the momentum $\vec{p} = (p_x, p_y, p_z)$ [15].

The detector can be divided in three main components from inside to outside and a schematic picture of the ATLAS detector with its components is shown in figure 3.2:

Inner Tracker: The inner tracker is designed to measure trajectories, called tracks, of charged particles and covers a range of $-2.5 < \eta < 2.5$. The whole inner tracker is penetrated by a magnetic field which is oriented parallel to the beam. Therefore the momentum of a charged particle traversing through the detector is deduced from its bending radius. In addition, the PV, i.e. the coordinates where the hard scattering event happened, is reconstructed from the tracks. Furthermore, a secondary vertex (SV) (or more than one), i.e. the coordinates of vertices displaced from the PV may be reconstructed in the inner tracker as well. Secondary vertices usually originate from the decay of particles into secondary particles after a certain flight length. The inner most layers of the inner tracker are three layers of silicon pixels (pixel detector) for high precision spatial measurement close to the PV. For run 2 a fourth layer of silicon pixels was added in between the first layer and the beam pipe. The pixel detector is followed by layers of silicon strips (semiconductor tracker (SCT)) for additional high precision tracking. The SCT is followed by the transition radiation tracker (TRT) which consists of gas filled straw tubes. The TRT has a lower spatial resolution compared to the pixel detector and SCT but provides more measurement points along a charged particle's trajectory. In addition, the gaps between the tubes are filled with fibres which are chosen such that transition radiation is only invoked for traversing electrons. This information is used to aid the identification of electrons [15–18].

Calorimeter System: The calorimeter system consists of an inner electro-magnetic calorimeter (ECAL) and an outer hadron calorimeter (HCAL) and covers up to $-4.9 < \eta < 4.9$. The calorimeters measure the energy of electrons and photons and hadrons respectively. Both of them are so-called sampling calorimeters. This means they use alternating layers of an absorber material and an active material. The absorber material is lead in the ECAL and iron and copper in the HCAL. Through interactions with the absorber material the traversing particles lose energy and create secondary particles. The secondary particles then create tertiary particles and so on. The whole cascade of these particles is called a shower. Electrons and photons interact via Bremsstrahlung and electron-positron pair production with the absorber material and the resulting shower only consists of electrons, positrons and photons. Hadrons interact with the absorber material via a wide range of nuclear interactions, e.g. ionisation or spallation, resulting in a mixtures of electrons, photons and hadrons in the shower. The average length a particle can travel trough material before losing a characteristic amount of its energy depends on the material but it is in general much larger for hadrons than for electrons and photons. Therefore the absorber material and thickness of the two calorimeters are chosen such that electrons and photons deposit all their energy in the ECAL and hadrons all their energy in the HCAL if they carry typical energies as expected at the LHC ($O(10^1$ to $10^3)$ GeV). In both calorimeters the shower particles create a signal in the layers of the active material. The active material is liquid argon in the barrel ($|\eta| < 1.475$) and forward region ($1.375 < |\eta| < 3.2$) of the ECAL. To avoid "cracks" in between calorimeter modules around the z-axis the absorber and active material are layered in an accordion geometry. The HCAL active layers consist of scintillators in the barrel region ($|\eta| < 1.7$) and with liquid argon in the forward region ($1.5 < |\eta| < 4.9$). The signal induced in the active layers is proportional to the energy of the incoming particle. Due to the sampling structure and non-uniformities especially in hadron induced showers the energy measurement has a significant uncertainty. The energy resolution σ_E/E in the barrel region of the ECAL is $\propto 10\%/\sqrt{E(\text{GeV})}$ and the one of the HCAL is $\propto 50\%/\sqrt{E(\text{GeV})}$. Since the calorimeters consist of many separate modules a direction information is recorded as well. The resolution in $\Delta\phi$ and $\Delta\eta$ is 0.025 for the ECAL and 0.1 for the HCAL [7, 15–17].

Muon Spectrometer: The muon spectrometer measures the trajectory of muons and covers $-2.7 < \eta < 2.7$. It is built within a toroidal magnetic field which allows to measure the momentum of the muons complementary to the measurement in the inner tracker. Since muons are the only known (and detectable) particles that pass through the inner tracker and calorimeter system the muon spectrometer allows to identify muons. The muons only lose a small amount of their energy when they pass through the calorimeter system because of their high mass compared to electrons. The muon spectrometer consists of so called precision chambers and trigger chambers which both utilise several detection techniques. The precision chambers, e.g. drift tubes, provide a very good spatial resolution. The trigger chambers, e.g. resistive plate chambers, have a high efficiency and provide fast signals with a low dead time. The information of the trigger chambers are not used to reconstruct muons properties but to provide information for the trigger system (see also the following section) [16, 17].

3.2.2 The Data Acquisition System

The ATLAS trigger system has to select events of interest for the ATLAS physics program to be recorded since the total amount of pp collisions is too large to be recorded. The trigger system reduces the pp collision rate of up to 40 MHz down to a trigger rate of approximately 1 kHz in two steps. The first step, level 1 trigger (L1), is a hardware based trigger which uses information from the calorimeter system and the muon spectrometer. Therefore it is able to trigger on electrons/photons, jets, hadronically decaying τ-leptons and muons which pass a certain p_T threshold that is defined. These are identified with the help of the signal they induce in several close by modules in the corresponding sub-detectors. The L1 is also able to combine the information of the whole calorimeter system to select events with overall high transverse energy or high missing transverse energy[1]. The events that pass the L1 are then processed by the high level trigger (HLT) which does partial reconstruction of the events to further select the events to be recorded. The HLT uses the information from the L1, i.e. the modules that invoked the trigger

[1] Transverse and missing transverse energy is defined in section 5.4. Missing transverse energy is an indicator for neutrinos or new unknown weakly interacting particles.

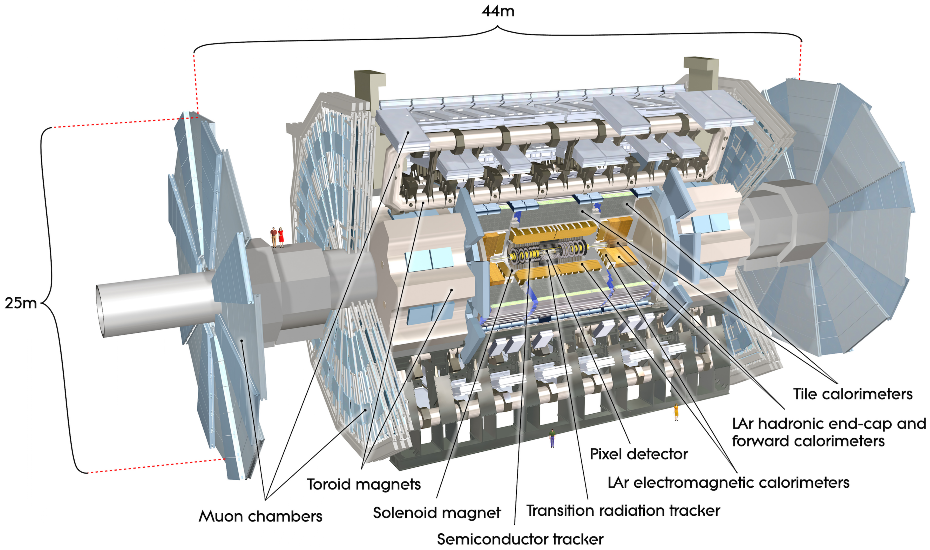

44m
25m
Tile calorimeters
LAr hadronic end-cap and forward calorimeters
Pixel detector
LAr electromagnetic calorimeters
Toroid magnets
Muon chambers
Solenoid magnet
Transition radiation tracker
Semiconductor tracker

Figure 3.2: Schematic picture of the ATLAS detector and its components [19].

decision. Therefore not the full event in the full detector volume is reconstructed. The reconstruction of tracks with information from the inner tracker and muon spectrometer allows to identify electrons and muons but also more complex signatures like hadronically decaying τ-leptons and b-jets. The calorimeter information is used to reconstruct photons, jets and missing transverse energy. The final trigger decision is made based on a combination of several of the reconstructed properties, e.g. high p_T, high isolation which means not much activity around the reconstructed object, possible τ-lepton or b-jet, possible decay of a heavy object which results in high transverse energies, etc. Many possible combinations are defined which make an event eligible to be recorded for physics analysis. It is a trade-off between manageable trigger rates and high trigger efficiency for events of interest. In general, particles with high transverse momenta are a good indicator that a hard scattering event took place. On the other hand high p_T jets are a very common signature due to the structure of protons and the resulting QCD nature of pp collisions. Therefore it is difficult to use jets as trigger objects without having an unmanageable trigger rate. This imposes a challenge for the Higgs decays that are investigated in this thesis: $H \to b\bar{b}$ and $H \to c\bar{c}$. Chapter 8 details how these problems are partially circumvented [15, 20].

CHAPTER 4

Proton-Proton Collisions and Their Simulation

The signature of proton-proton collisions is highly complex due to the inner structure of the protons and the involvement of QCD processes. As described in chapter 2 QCD exhibits certain special features such as confinement. The following section explains the phenomena involved in pp collisions. Furthermore, in order to develop analysis strategies for pp collision data and test predictions of theoretical models against data, simulated events are used. The generation of a simulated pp collision event is closely related to the physics processes present in pp collisions and is explained in the next section. The simulation models that are relevant for the analyses of this thesis are given in section 4.2.

Protons are composite objects and their constituents are referred to as partons. The partons are further distinguished into: valence quarks, gluons and sea quarks. There are three valence quarks in the proton, two up quarks and one down quark, which define the fundamental properties, i.e. quantum numbers, of the proton. The valence quarks in the proton are bound by the strong force and interact via gluons. The gluons can spontaneously form a virtual quark anti-quark pair which annihilates back into a gluon. These quarks are called sea quarks. There are three different scenarios that may happen when two protons collide (in order of most likely to least likely): elastic scattering, soft inelastic scattering or hard scattering. In an elastic scattering event only small momenta are transferred between the protons and the protons interact as a whole. The inner structures of the protons do not become visible in these reactions, the protons stay intact and no new particles are produced. Due to the low momentum transfer these reactions have a low transverse momentum with respect to the proton beam. In soft inelastic scattering events one or both of the protons are destroyed, e.g. the protons are excited and subsequently decay. Nevertheless the momentum transfer is still relatively low and mostly pions and other light hadrons with low transverse momenta are produced. In order to produce new and high mass particles, e.g. Z bosons or Higgs bosons, a hard scattering event has to take place. The momentum transfer in these reactions is very high and instead of the protons (or parts of them) interacting as a whole, single partons from each proton interact. Although being bound inside the protons the partons can be described as "free" particles in these reactions due to asymptotic freedom, as described in chapter 2. The two interacting partons in a hard scattering event may be valence quarks, gluons, sea quarks or a combination of two of them. Each parton carries a fraction x_i of the proton's momentum. Hence the actual centre of mass energy $\sqrt{\hat{s}}$ is smaller than the collision energy $\sqrt{s}$:

$$\hat{s} = x_1 x_2 s \tag{4.1}$$

The protons in hard scattering events are destroyed and the remnants are no longer colour neutral and hadronise (hadronisation is explained in more detail in the following section). During a hard scattering event new particles are produced and they carry high transverse momenta [3, 21, 22].

4.1 Physics Beyond the Proton-Proton Collisions

Several phenomena occur in pp collision events. The so-called factorisation theorem allows to split this event in several parts, which are described separately in the simulation of an event. A schematic picture of these parts is shown in figure 4.1. The main parts are: 1. the incoming protons and their parton contents, which are described by parton distribution functions (PDF), 2. the hard scattering event producing new particles, which is described by the matrix element (ME), 3. the transition of intermediate partons to observable colour neutral hadrons, which involves a parton shower (PS) and subsequent hadronisation. This can also be expressed as a differential cross section $d\sigma_{pp\to X}/dO$, with O, denoting a specific set of observables (transverse momenta, jet multiplicities, etc.), which has to be calculated:

$$\frac{d\sigma_{pp\to X}}{dO} = \sum_{i,j} \int dx_1 dx_2 \int d\Omega\, f_i(x_1,\mu) f_j(x_2,\mu)\, \frac{d\hat{\sigma}_{ij\to x}(x_1,x_2,\Omega,\alpha_s(\mu),\mu)}{d\hat{O}}\, D_{x\to X}(\Omega,\mu) \tag{4.2}$$

The previously mentioned parts can also be identified in equation 4.2: PDF f_i, f_j for the colliding partons i, j, the matrix element (ME) or partonic cross section $d\hat{\sigma}/d\hat{O}$ and the transition function D. In addition, there are several energy scales involved: the fraction of the protons' momenta that is carried by the colliding partons x_1, x_2, the renormalisation and factorisation scale which are usually chosen at the same value μ, the coupling strength at this given scale $\alpha_s(\mu)$ and the available phase space Ω. Some of these parts can be precisely calculated whereas others have to be modelled by introducing approximations or "cookbook recipes". These pieces and how they are simulated are explained in the following [5, 22].

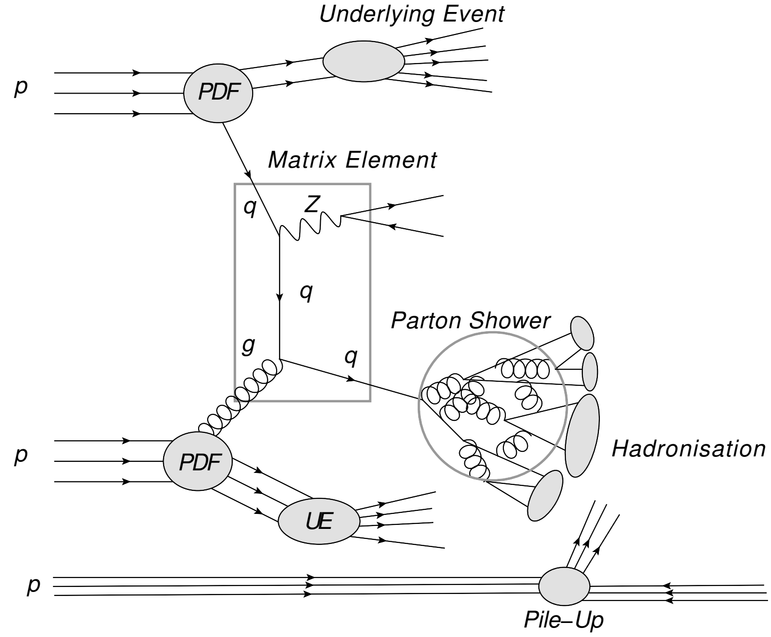


Figure 4.1: Schematic picture of a proton-proton collision. Shown are the incoming protons and their parton content described by the PDFs, the partonic interaction given by the ME, the PS and hadronisation of the quarks and gluons in the event and additional activity in the event originating from the underlying event and pile-up.

Parton distribution functions (PDF): To calculate a partonic cross section it is essential to know which parton takes part in the collision and the momentum it carries. The PDF describe the probability distribution to "find" a certain parton with a certain fraction of the proton's momentum

x inside the proton. These distributions depend on the energy scale because higher energies correspond to smaller wavelengths and therefore smaller structures that can be resolved. Since it is not possible to calculate QCD effects at low energies such as present in the proton the PDF have to be measured. Measurements that are sensitive to the PDF are: probing the structure of protons in electron proton collisions, Drell-Yan production and production of jets in proton-proton collisions. The results of these measurements are combined in a fit to extract the full distributions for all partons and values of x and to extrapolate them to a desired energy scale μ. Several collaborations use different strategies concerning which measurements are included and how the included results are weighted and provide their PDF sets to be used in Monte Carlo (MC) simulations. Therefore the choice of the PDF set influences the predictions of the simulation. The PDF collaborations also provide the uncertainties of the PDF fit in so-called error sets which can be used to assess the systematic uncertainty of the simulation that originates from the chosen PDF set [22].

Matrix element (ME): The ME describes the hard scattering cross section on parton level. It can be calculated from QED and QCD theory with the help of Feynman rules. In principle the calculation of the partonic cross section has to take into account all possible diagrams for a given process including all virtual and real corrections. In leading order (LO), i.e. only taking the diagrams with the largest contribution to the process into account, there are often only a few diagrams to be calculated. In next-to-leading order (NLO), i.e. the introduction of either a loop or radiative correction, the amount of possible diagrams already multiplies. Therefore the amount of computation time and power to calculate all possible diagrams becomes quickly too large. As a result most simulation programs that are utilised for the calculation of the ME do not go beyond NLO precision[1]. In general it is desirable to use the simulation model that provides the highest possible order. In all simulations that are used for ATLAS analyses the ME does not only include the production of particles but also the decay of these particles in case they are short lived standard model particles, e.g. $Z \rightarrow \ell^- \ell^+$ or $H \rightarrow b\bar{b}$. All diagrams in the ME calculation are calculated at a fixed energy scale μ. Therefore ME simulations that have the same precision may still differ due to the μ they chose [22, 23].

Factorisation and renormalisation scale μ: The factorisation scale μ_F and renormalisation scale μ_R are two cut-off scales on the virtuality of intermediate particles and they separate short distance from long distance effects. Substructures of loops and real emissions below these scales are only approximated instead of calculated. In the case of loops missing higher order virtual emissions are approximated with an additional term given by the so-called renormalisation group equation. Factorisation is used in two places in equation 4.2: in the separate description of the PDFs and partonic cross section and in the PS, which is part of the transition function D. Most of the simulation models choose the same value for both scales $\mu = \mu_\mathrm{F} = \mu_\mathrm{R}$. In general, μ is an arbitrary scale but from experience it is useful to set μ in the order of the hard scattering scale, e.g. the centre of mass energy, the mass of the heavy particle that is produced, the momentum of the produced particles, etc. The choice of μ changes the predicted outcome of the simulations. This dependence is reduced if higher orders of α_S are included in the ME calculation. In addition, many simulated data sets offer the possibility to vary μ from its nominal value to assess the uncertainty that originates from its choice [5, 21, 22].

[1] The "counting" of orders in simulations is in most cases not equivalent to counting the orders in α_s and α. Instead LO corresponds to the minimal order possible in QCD and QED to create a certain process, e.g. $gg \rightarrow ZH$ already includes a triangle in LO. NLO corresponds to the addition of one loop (l) or one radiative correction (r) in QCD, i.e. $r + l = 1$. Next-to-next-to-leading order (NNLO) corresponds to $r + l = 2$. If also higher orders in QED are taken into account this is mentioned separately.

Parton shower (PS): The PS is the connection between the ME on parton level and the observable hadrons. Partons that are produced at ME level will never form hadrons since their energy is too high to form a bound state. Instead these partons undergo several stages of radiation until a certain scale is reached ($\mathcal{O}(1\text{ GeV})$) and the partons build hadrons. The general idea of the PS is: instead of calculating one parton producing n other partons the process of one parton producing two partons, e.g. $u \rightarrow \bar{u}g$, is calculated n times[2]. The evolution of the parton shower is ordered by an evolution parameter t and each additional branching of a parton is lower in t than the previous branching. The evolution parameter is roughly proportional to the energy scale. It also describes the "resolvability" of the produced parton, e.g. will the parton produce its own jet (usually included in the ME) or will it just contribute to the substructure of the jet (usually included in the PS). Depending on the simulation model different evolution parameters are chosen, which influences the predicted outcome of the simulation. In addition, in some cases the ME is calculated using a different simulation interface than for the PS. In these cases the two have to be "matched" to resolve overlap between the ME and PS, e.g. in NLO generators the radiation of one extra parton is already included and therefore the contribution of this diagram has to be removed from the PS, otherwise the probability for a certain process could exceed unity. There are several dedicated matching schemes developed. PS algorithms are often fine tuned using measurements. The general idea of this tuning is to partially compensate for neglected higher order effects[3] [22, 23].

Hadronisation: The PS is only evolved up to a certain scale where the creation of new partons cannot be described by perturbative QCD any more. At this scale the formation of hadrons — hadronisation — has to take place. All common simulation models used in ATLAS analysis use either the string fragmentation model or cluster fragmentation model. The string fragmentation model connects a quark and an anti-quark that are produced in the PS via a "colour string" as a representation of the field of the strong force between their colour charges. In this picture gluons from the PS represent kinks in the strings with two string pieces attached. If the two quarks move apart the energy stored in the field grows until the energy is large enough to create a new quark anti-quark pair. At this point the string breaks. This process continues until all energy is absorbed such that no further string breaks occur and neighbouring quarks are grouped together into hadrons. Cluster fragmentation first breaks each gluon from the PS into a quark anti-quark pair. It then pairs each quark with an anti-quark which is neighbouring in momentum space. These clusters then decay, via the creation of quark anti-quark pairs in the cluster, into hadrons absorbing the energy stored in them. Both models also allow baryons to form if diquark (loosely bound states of two quarks) pairs are produced instead of quark anti-quark pairs. In addition, there are several parameters that tune these models to match the observed fractions of heavy to light hadrons, mesons to baryons and observed spin states [5, 22].

Hadron decays: The hadrons created in the hadronisation may not be stable. "Not stable" typically means they decay inside the beam pipe or detector volume. These decays are modelled inside the simulations using the hadron life time, the branching ratios of hadron decays and the hadron decay width measured in data. Therefore they strongly depend on the available measurements and their uncertainties. Some simulation models also include additional factors such as helicity, spin correlation, CP violation or B meson oscillation [22].

Underlying event (UE): The underlying event describes additional interactions beyond the hard scat-

[2] Some MC generators that simulate the PS also include QED radiation, e.g. $u \rightarrow \bar{u}\gamma$

[3] The art of tuning is to conserve the universality of the model and its predictions instead of tuning a MC simulation to statistical features of a specific process.

tering event and its PS and hadronisation described above. The energy scale of these additional interactions is only a few GeV, i.e. small compared to the energy scale of the hard scattering event. Therefore the underlying event usually does not produce additional jets but rather a uniformly distributed underlying activity in the form of hadrons[4]. In simulations, similar to the cut-off scale for hadronisation, a minimal scale is introduced for additional parton interactions. Interactions above this scale will be simulated as additional activity with PS and hadronisation. A suitable choice for the cut-off scale is estimated from the proton radius, the impact parameter of the protons (how "head on" a collision is) and the collision energy. It also depends on the utilised PDF set. The underlying event (UE) simulation is often additionally fine tuned with results obtained from data [22, 23].

Pile-up: Pile-up events are additional pp reactions during the crossing of two proton bunches. Technically pile-up events are separate collisions. Therefore they are simulated separately from the hard scattering event. However, in the detector they cannot be separated and therefore particles from pile-up events are part of the event signature. By construction the pile-up events do not involve a hard scattering event and are modelled as the exchange of gluons at small centre of mass energies. The resulting final state is then simulated using hadronisation models [22, 23].

4.2 Monte Carlo Simulations Used in the Analyses

The simulation models of pp collisions use random number Monte Carlo methods. So called MC generators already provide a framework to generate these events. A variety of MC generator programs exist to simulate ATLAS data. Thus the ATLAS collaboration provides recommendations on the MC generator choice for a particular process, which describes the process to best possible precision. In addition, alternative choices are provided and used to evaluate systematic uncertainties on the recommended simulation models. Some of these generators such as Pythia [24], Herwig [25, 26] and Sherpa [27] are so-called multi-purpose generators and generate a full event. There are other more specialised generators, such as Powheg [28] and MadGraph5_aMC@NLO [29, 30], which only simulate the hard scattering event. These generators have to be interfaced to one of the other generators to describe the PS thus introducing the necessity to "match" the ME and PS simulation to remove overlap between them [31]. These five generators are the ones that are used for the analyses in this thesis and their main features are:

- **Pythia:** Pythia is a multi-purpose generator that provides LO ME calculations. The PS and UE event model implemented in Pythia uses string fragmentation. The Pythia PS and UE event model is further tuned using measurements. Special sets of tuned parameters are provided. The most common ones used for Pythia are AZNLO [32] and A14 [33]. These tuning sets also provide variations that are used to study the effects on the simulation given by the tuning. Since Pythia only provides LO ME precision it is not used as a stand-alone generator any more. However, the PS model exhibits good agreement with data therefore Pythia is often used to provide the PS description for a simulation model.

- **Herwig:** The Herwig multi-purpose generator is not used as a stand-alone generator any more. However it is still used to provide the PS description for a simulation model. The Herwig hadronisation uses the cluster fragmentation model and special tunes for the Herwig PS exist. Therefore many simulations that use Pythia to describe the PS are also generated with Herwig instead, which is used to assess the systematic effect of the PS model on the simulation.

[4] For example at the LHC at a collision energy of 13 TeV this underlying activity is about 3.3 GeV per unit of ΔR with large fluctuations in the order of 2 GeV.

- **Sherpa:** Sherpa is a stand-alone multi-purpose MC generator. It provides NLO precision for a variety of processes. Therefore it is often the preferred generator for processes with additional jets, which Sherpa can include in the ME calculation. In contrast to Pythia and Herwig there are no "hybrid" versions of Sherpa, which use Sherpa only to generate a certain part of the event [27].

- **Powheg:** The Powheg MC generator only generates the ME and provides NLO precision for a variety of processes. In the simulations used in this thesis the default is to interface Powheg to Pythia for the description of the PS. However, often an alternative version of Powheg +Herwig is provided as well to study systematic effects between the two.

- **MadGraph5_aMC@NLO:** MadGraph5_aMC@NLO provides, similar to Powheg, only ME calculations. In most cases MadGraph5_aMC@NLO is not used as the default generator and only used to study systematic effects. For the PS description it is either interfaced to Pythia or to Herwig.

All generators used in the analyses are detailed in table 8.2 in chapter 8. Independent of the generators that were used to generate the ME and PS all ATLAS simulated data sets model the pile-up events with Pythia. This means, the simulation of the hard scattering event is overlayed with these simulated pile-up events. In addition, the simulated events are reweighted such that the distribution of the number of pile-up events agrees with the actual average number of reconstructed pile-up events in data. All simulations, except for those generated with Sherpa, use the EvtGen [34] program to describe the decay of hadrons containing bottom or charm quarks.

The next step in the simulation of an ATLAS event is the simulation of the interaction of the final state particles with the detector. This is explicitly tailored towards the experiment. The ATLAS detector is modelled with the Geant 4 package [31, 35]. Every single piece of the detector is implemented as a separate piece (detector modules but also material from support structures) with a precise location in space. The response of these pieces to particles traversing through them is implemented in the simulation as well based on test measurements. The detector simulation also includes a description of the magnetic fields inside the detector. This simulation is constantly updated if changes appear (defect modules, etc.). To get the final and full simulation of an event in the ATLAS detector, the simulated event from the MC generator is inserted into the detector simulation. The last step is to digitise the detector's response to a particle interacting with it, i.e. translate the interaction into electrical signals. This allows to, eventually, pass the simulated event through the same trigger and reconstruction algorithms, which are described in chapter 5, as for data.

CHAPTER 5

Object Reconstruction in ATLAS Data

Particles produced in pp collisions, as described in the previous chapter, pass through the ATLAS detector and produce signals in form of electrical signals. The event reconstruction starts with the signals in the detector and reconstructs physical objects out of it. The aim is to identify these physics objects and reconstruct their properties, e.g. 4-momentum, charge. This chapter summarises the event reconstruction relevant for this thesis.

5.1 Tracks and Vertices

Tracks are reconstructed from the signals in the inner tracker using specialised algorithms. In general, the reconstruction of a track consists of two steps: pattern recognition and track fitting with more details given in [15, 36]. The track reconstruction efficiency has a strong dependence on η since the amount of material, i.e. possible losses due to interactions with it, is larger for larger η. The efficiency also depends on the p_T of the track and reaches a maximum efficiency of $(85-90)\%$ for $p_\mathrm{T} > 5\,\mathrm{GeV}$.

Vertices identify points in space which have a high density of tracks. Similar to track reconstruction, vertex reconstruction uses pattern recognition to identify potential vertex candidates first and then a fit is performed to determine its properties such as its coordinates. The primary vertex (PV) is defined as the reconstructed vertex with the highest $\sum p_\mathrm{T}^2$ of the tracks associated to it [15]. SV are displaced vertices and there are several algorithms available to identify SV, which are mostly used for the purpose of b-tagging, see section 5.5.1.

5.2 Leptons

In contrast to the physics definition of a lepton, i.e. electrons, muons, τ-leptons and neutrinos, in ATLAS analyses the term often refers to long-lived charged leptons. Therefore the definition of the term lepton, as it will be used throughout this thesis, includes only electrons and muons since τ-leptons decay before they can be detected directly and neutrinos do not leave any signal in the detector at all.

5.2.1 Electrons

Electrons passing through the ATLAS detector leave a track in the inner tracker and then deposit all their energy in the ECAL. Therefore they are reconstructed from signals in the ECAL and inner tracker. The reconstruction of electrons starts with clusters in the ECAL. A cluster is the combination of neighbouring calorimeter modules. The size of an ECAL cluster, i.e. the size of the grid of ECAL modules to be combined, is fixed and based on the typical size for electron induced showers. The energy of the cluster needs to exceed 2.5 GeV to qualify as an electron candidate. A reconstructed track has to match the ECAL cluster in η and ϕ direction. If no such track is found the ECAL cluster is considered a photon candidate. The 4-momentum of the electron candidate is given by the energy of the ECAL cluster and the η and ϕ of the reconstructed track. To account for energy losses in the inner tracker and the intrinsic calorimeter resolution the energies of the electron candidates are calibrated using simulated events and $Z \to e^+e^-$ events. In order to identify an electron candidate several properties that distinguish them from other particle, e.g. a track+ECAL cluster inside a jet distinguished them from photons, are combined in a multivariate algorithm. Based on this multivariate algorithm different quality levels are defined for the electron candidate based on the rejection mis-identified electrons. For the analyses presented here the *loose* quality was chosen which identifies electrons with an efficiency of $\approx 95\%$ and the mis-identification of other objects as electrons is below 1%. To further suppress the mis-identification an isolation criterion is defined for the electron candidate. It is defined as the sum of the p_T of tracks in the vicinity of the electron candidate with respect to the p_T of the electron candidate. In this thesis a *loose* track isolation is chosen which has an efficiency of 99% [15, 37]. Additional more analysis specific requirements may be defined to minimise the amount of mis-identified electrons, as described in section 8.2.

5.2.2 Muons

Muons are reconstructed by combining inner tracker and muon spectrometer information. The default is to reconstruct the track of a muon candidate from its signals in the muon chambers. The next step is to extrapolate the reconstructed muon spectrometer track to the inner tracker to find a reconstructed inner tracker track that matches its direction and momentum[1]. If no track in the inner detector can be found, but the track reconstructed in the muon spectrometer can be extrapolated back to the PV taking into account the energy loss, this muon candidate is used in the analysis as well. Three identification quality levels are defined. The analyses described here utilise the *loose* muon identification quality. The reconstruction efficiency for *loose* muons is $\approx 97\%$ and the mis-identification of other signatures as muons is smaller than 1%. The *loose* muon definition does not require a fully reconstructed track in the muon spectrometer for $|\eta| < 0.1$, where the muon spectrometer is only partially instrumented. The 4-momentum of the muon is determined from the combined track in the inner tracker and muon spectrometer or from only one of them if they are not both present. A *loose* isolation, based on tracks reconstructed in the inner tracker, to further select muons is defined in the same way as for electrons [15, 38]. Analysis specific muon requirements are defined in section 8.2.

5.3 Hadronic τ-leptons

Hadronically decaying τ-leptons are used to discard the events or single jets as described in section 8.2. Jets reconstructed with the anti-k_T algorithm with a radius parameter of $R = 0.4$, as described in section 5.5, are inputs to the τ-lepton identification algorithm. Various properties of the τ-lepton candidate jet are combined in a multivariate algorithm to distinguish it from quark/gluon induced jets. One distinctive feature is the presence of an odd number of charged hadrons, i.e. tracks. The τ-lepton identification algorithm defines different quality levels for the τ-lepton candidate jet. *Medium* quality is used in this thesis, which has an identification efficiency of $\approx 52\%$ for hadronic τ-leptons and mis-identifies $\approx 1\%$ of jets as τ-leptons [39]. This means the rejection of *medium* hadronic τ-leptons removes

[1] "Match" is defined very generously since the muon loses parts of its energy, $O(3\,\mathrm{GeV})$, when it passes through the calorimeter system

more than half of the hadronic τ-leptons with only a small loss of potential signal jets. All analysis specific hadronic τ-lepton requirements are defined in section 8.2.

5.4 Missing Transverse Energy

The sum of the transverse momenta of all objects in an event should add up to 0 since the proton beams only have momentum in the longitudinal direction before they collide. This assumption does not hold if one or more neutrinos are produced. Since neutrinos only interact weakly the probability of them interacting within the ATLAS detector is effectively zero. Therefore the momentum they carry away will be detectable as an energy imbalance $E_{\mathrm{T}}^{\mathrm{miss}}$. It is calculated from its components $E_{x(y)}^{\mathrm{miss}}$ in $x(y)$ direction:

$$E_{\mathrm{T}}^{\mathrm{miss}} = \sqrt{(E_x^{\mathrm{miss}})^2 + (E_y^{\mathrm{miss}})^2} \tag{5.1}$$

Each of the components is calculated as the negative vectorial sum of the calibrated momenta[2] of all reconstructed electrons, photons, muons, hadronic τ-leptons and jets. An additional so-called soft term is added which includes the momentum sum of all tracks that are matched to the PV but are not associated to any of the aforementioned reconstructed objects. Mismeasurements of the visible objects in an event and detector acceptance effects lead to $E_{\mathrm{T}}^{\mathrm{miss}} > 0$ even in events without any neutrinos [40].

5.5 Jets

A jet is a collimated bundle of particles, mostly hadrons. They are created since quarks/gluons cannot exist as free particles due to confinement (see chapter 2). Jets in the ATLAS detector are reconstructed from close-by energy depositions in the calorimeter system.

The first step in the reconstruction of jets is to form clusters of calorimeter signals. If a calorimeter module registers an energy deposit above a certain energy threshold it is clustered together with the energy depositions in its neighbouring modules. If one (or more) of these neighbouring modules is again over a certain energy threshold also its neighbouring modules are added to the cluster. This continues until no more modules above a certain energy threshold are found. If the energy distribution of the modules of the cluster has local maxima it is likely that the cluster corresponds to more than one particle. In that case the initial cluster is split such that each resulting cluster encompasses one of the local maxima. The general idea behind this procedure is that one cluster corresponds to the energy depositions of one particle inside the jet. This relation does not always hold and depends on the distance between particles and the size of the calorimeter modules which is not uniform. Each reconstructed calorimeter cluster is treated as a pseudoparticle with no mass, i.e. $E_{\mathrm{cluster}} = |\vec{p}|_{\mathrm{cluster}}$. Therefore its 4-momentum vector is given by the total energy of all calorimeter signals belonging to the cluster and the average energy weighted η and ϕ of these signals.

In the next step the reconstructed calorimeter clusters have to be combined to form a reconstructed jet. The challenge is to efficiently pick up as many of the clusters that were initiated by the fragmentation of the initial quark/gluon. At the same time it is desirable to reduce the amount of unrelated clusters, e.g. from pile-up, in the reconstructed jet. Therefore jets are reconstructed with flexible jet finding algorithms. There are a wide variety of these algorithms and the so-called anti-k_{T} algorithm [41] is

[2] The name "missing transverse energy" originates from the fact that it is largely based on the measurements of the calorimeters and the calorimeters measure energy. Nevertheless with $E \approx |\vec{p}|$ and the angular information a momentum vector can be constructed which is used in the calculation of $E_{\mathrm{T}}^{\mathrm{miss}}$.

chosen to reconstruct jets in the ATLAS detector. Its sensitivity to pile-up and the underlying event (UE) is particularily small compared to other algorithms. Its advantage is that soft, i.e. low p_{T}, particles do not modify the jet while hard, i.e. high p_{T}, particles do. The anti-k_{T} jet finding algorithm defines two distances d_{ij} and d_{iB} :

$$d_{ij} = \min\left(\frac{1}{p_{\mathrm{T}i}^2}, \frac{1}{p_{\mathrm{T}j}^2}\right)\frac{\Delta R(i,j)^2}{R^2} \tag{5.2}$$

$$d_{iB} = \frac{1}{p_{\mathrm{T}i}^2} \tag{5.3}$$

The jet finding starts with a reconstructed cluster i with transverse momentum $p_{\mathrm{T}i}$ and reconstructed cluster j with transverse momentum $p_{\mathrm{T}j}$ which have the distance $\Delta R(i, j)$ in the $\eta\phi$-plane. If the resulting d_{ij} is smaller than d_{iB} they are combined. The combined object then represents object i for the next iteration and another reconstructed cluster is object j. The combining stops as soon as $d_{ij} \geq d_{iB}$ and the resulting object is the reconstructed jet. This stopping criteria depends largely on the radius parameter R in formula 5.2. Jets used in this thesis are reconstructed with a radius parameter of $R = 0.4$. The anti-k_{T} algorithm is also used to reconstruct jets from tracks instead of calorimeter clusters. The radius parameter for these track jets is set to $R = 0.2$. Track jets do not contain information about the neutral particles inside a jet. In this thesis track jets are used in chapter 7 for the definition of SV properties.

The 4-momentum of the reconstructed jet is given by the sum of the energies of all its clusters, the energy weighted average η and ϕ of its clusters and the mass which is the invariant mass of all massless clusters. The axis of the jet is given by the energy weighted (η, ϕ) coordinate pair. The jet axis is an important quantity because distances between a jet and any other object is defined as the distance of that object to the jet axis. The 4-momentum is further modified in the calibration step [15, 41]. The calibration and why it is necessary is explained in chapter 7.

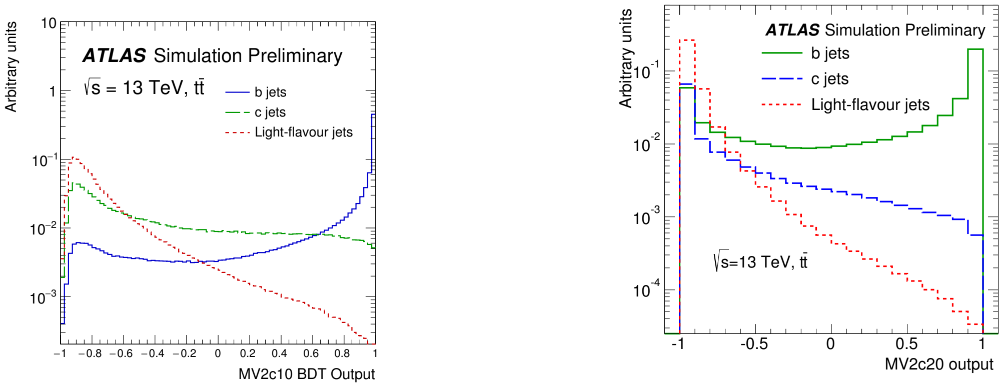
All gluons and quarks hadronise and form jets which is why their experimental signatures are very similar. Nevertheless jets that are initiated by b-quarks and c-quarks have certain features that distinguish them from jets originating from other quarks and gluons and to a certain extent from each other. These features are caused by their relatively high mass compared to other quarks. Since the identification of the flavour of the originating parton is a crucial part of this thesis the following definitions are used to identify the "flavour" of a jet in simulated events: if a b-hadron is found inside the jet ($\Delta R(\text{jet}, \text{hadron}) < 0.3$) the jet is labelled a b-jet, if no b-hadron but a c-hadron is found inside the jet the jet is labelled a c-jet and if neither a b- nor a c-hadron was found inside the jet the jet is labelled a light jet. The entirety of b-jets and c-jets is also referred to as heavy flavour jets.

5.5.1 *b*-tagging

The technique to distinguish b-jets from c-jets and light jets is called b-tagging. It is solely based on information from tracks. Therefore tracks are associated to the jet if they are within a certain distance ΔR to the jet. The specific ΔR depends on the p_{T} of the jet, e.g. it is 0.45 for $p_{\mathrm{T}}^{\mathrm{jet}} = 20\,\mathrm{GeV}$ and only 0.26 for $p_{\mathrm{T}}^{\mathrm{jet}} = 150\,\mathrm{GeV}$. All b-tagging techniques make use of the relatively long life time ($\tau \approx 1.5\,\mathrm{ps}$) of b-hadrons, i.e. hadrons containing a b-quark. A b-hadron forms during the hadronisation. If it has a typical transverse momentum of $O(10^1\,\mathrm{GeV})$ its flight length is, according to $l = \beta\gamma c\tau$[3], a couple of mm before it decays. The charged decay products are measured as tracks in the inner tracker. Due to the long

[3] With $\beta = v/c$ and $\gamma = (\sqrt{1 - (v/c)^2})^{-1}$ and c the speed of light and v the velocity of the particle.

flight length two important quantities can be resolved in the inner tracker: the impact parameter of the tracks and the decay vertex. The impact parameter is defined as the point of closest approach to the PV in the transverse plane d_0 or longitudinal plane z_0. The significance of this displacement for various types of jets is combined in a basic tagging algorithm. The decay vertex formed by the tracks of the decay products is reconstructed as a SV from the tracks associated to the jet. The properties of this secondary vertex and its associated tracks provide more information to distinguish b-jets. Another approach is a dedicated fitting algorithm that reconstructs the full decay chain of the b-hadron, including tertial vertices. It provides additional vertex and track information. The standard ATLAS b-tagging algorithm combines the jet's 4-momentum information, information from the basic impact parameter based algorithm and information about the vertices and the tracks into a Boosted Decision Trees (BDTs) based multivariate algorithm. This b-tagging algorithm (short: "tagger") is called MV2c. The official recommendation to analyse ATLAS data from 2015 was to use the MV2c20 tagger and for the data from 2015+2016 to use the MV2c10 tagger. In the training b-jets were considered as "signal" and an admixture of 7%(20%) c-jets and 93%(80%) light jets as "background" for the MV2c10(MV2c20) tagger. The background admixture stirs the separation power with respect to c-jets relative to light jets. The response of the BDTs for b-jets, c-jets and light jets is shown in figure 5.1 for the MV2c10 and MV2c20 tagger. The b-jets accumulate at high BDTs output scores whereas the light jets accumulate at low output scores, c-jets also have a maximum at low output scores but a much longer tail towards higher output scores. In analyses, which aim to select b-jets, jets have to pass a minimal requirement on the tagger's BDTs score. There are recommended requirements which provide a fixed b-jet identification efficiency in a reference process, in this case top quark pair production. The analyses presented here use the recommendation to achieve 70% b-jet identification efficiency for both taggers. This requirement falsely identifies 8%(12%) of c-jets as b-jets and 0.3%(4%) of light jets as b-jets for the MV2c10(MV2c20) tagger which was used. The identification and mis-identification efficiencies depend on the p_{T} of the jet but are relatively stable for all jets between 20 GeV and 200 GeV. These efficiencies were found to differ in data. Therefore an additional calibration in the form of weights for simulated events are given for each tagger and the selected b-jet identification efficiency. These so-called scale factors were derived separately for b-jets, c-jets and light jets. They are dependent on the jet p_{T} (and for light jets also on the η as well). They were derived from the comparison of data and simulated events. [15, 42–45].



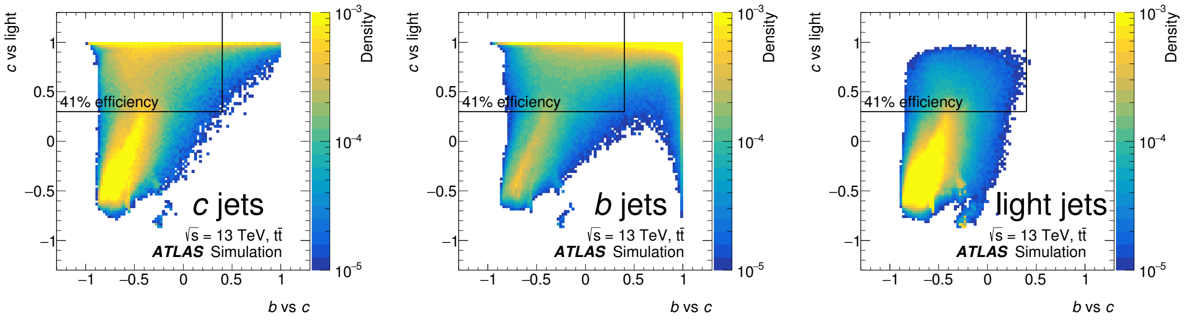
(a) MV2c10

(b) MV2c20

Figure 5.1: The response of the BDTs for b-jets, c-jets and light jets for the a) MV2c10 [42] and b) MV2c20 [43] tagger. A b-jet identification efficiency of 70% corresponds to output(MV2c10) > 0.8244 or output(MV2c20) > −0.0436

5.5.2 *c*-tagging

To identify c-jets very similar techniques to those of b-tagging are exploited. Also c-hadrons have a relatively long lifetime $\tau \approx (0.5 - 1.0)$ ps but it is on average not as long as for b-hadrons. This makes it possible to distinguish them from b-jets and light jets. The challenge is that the features of c-jets have tails towards light jet like features on one side and towards b-jet like features on the other side. For c-tagging a combination of two different taggers are used: MV2c100 and MV2cl100. Both of them utilise the same algorithm and input variables as for b-tagging but different data sets for the training. MV2c100 is trained with b-jets as "signal" and c-jets as "background". MV2cl100 is trained with c-jets as "signal" and light jets as "background". These dedicated trainings increase the respective separation power. The distribution of the BDTs output score is shown in figure 5.2 as a two dimensional plot of both taggers. It shows that b-jets accumulate at high BDTs scores for both trainings with a long tail towards low scores, i.e. c-jet like, in the MV2c100 tagger. Light jets accumulate at low BDTs scores for both trainings. The c-jets accumulate in two regions: at low scores in both trainings and at low scores for MV2c100 but high scores for MV2cl100. The accumulation at low scores in both trainings stems mostly from jets where the SV or track displacement could not be resolved. These c-jets cannot be distinguished from light jets. Nevertheless the c-jets that accumulate at low MV2c100 scores but high MV2cl100 scores can be selected with requirements for the respective scores. The requirements are chosen such that the c-jet identification efficiency is 41%, the b-jet mis-identification efficiency is 25% and the light jet mis-identification efficiency is 5% in $t\bar{t}$ events. To eliminate differences in the efficiencies in data and simulated events scale factors are derived for simulated events. This is done in the same way as for b-tagging [7, 46]. The usage of these c-tagging techniques in ATLAS was first studied for the $ZH \rightarrow \ell^{-}\ell^{+}c\bar{c}$ analysis which is described in section 9.



(a) c-jets

(b) b-jets

(c) light jets

Figure 5.2: The output distribution of the c-tagging algorithm for a) c-jets, b) b-jets and c) light jets. The response of the BDTs for the discrimination of c-jets and light jets (MV2cl100 tagger) is shown on the y-axis and the response of the BDTs for the discrimination of b-jets and c-jets (MV2c100 tagger) is shown on the x-axis. A c-jet identification efficiency of 41% corresponds to output(MV2c100) < 0.4 and output(MV2cl100) > 0.3 which is also indicated in the distributions.

5.5.3 Truth tagging

Truth tagging is used to reduce the statistical error of simulated data sets due to the usage of b- and c-tagging techniques. If b-tagging or c-tagging requirements are imposed in analyses, such as the $H \rightarrow b\bar{b}$ and $H \rightarrow c\bar{c}$ analyses, the amount of selected events is significantly reduced. Truth tagging replaces the application of tagging requirements: instead of rejecting simulated events all simulated events are used and reweighted such that the effect of tagging requirements is "simulated". The truth tagging event weight per event correspond to the efficiency that this event would pass the tagging requirements. It depends on the flavour — determined by the hadrons inside the jets, as previously explained — the p_{T} and η of each jet in the event. This probability reweighting reproduces the kinematic distributions in the simulated data set expected from the direct usage of tagging techniques, but preserved the statistical power of the full data set. In the analyses truth tagging is used for simulated data sets that have a low acceptance in the analyses phase spaces due to the tagging identification efficiencies.

CHAPTER 6

Statistical Tools

Statistical data analysis methods are crucial to perform analyses of the large data sets collected by the ATLAS experiment. The size of the data set imposes challenges when it is analysed but also opportunities to profit from the large amount of information contained in it to improve the analysis. This chapter introduces two statistical tools, which are used in this thesis. Multivariate techniques are used to improve the analysis with the help of machine learning techniques and are discussed in section 6.1. The techniques used to extract the final result of the analysis and interpret its statistical implications are explained in section 6.2.

6.1 Multivariate Techniques

The general idea of any multivariate method is to predict the outcome y of an experiment based on a set of n features x of the experiment:

$$y = f(\mathbf{x}) \qquad \text{with: } \mathbf{x} = (x_1, x_2, ..., x_n) \tag{6.1}$$

The simplest example is linear regression where the outcome y is a linear function of x and by identifying $f(x)$ a prediction for y can be made given x. However, in particle physics, or the "real world" in general, the identification of $f(\mathbf{x})$ could be highly complex and in any case can only be approximated as $\hat{f}(\mathbf{x}) = \hat{y} \approx y$ from a given set of observations $(\mathbf{x}_i, y_i)$. In addition, in particle physics, the parameter space given by $\mathbf{x}$ has a high dimensionality. There are many techniques to approximate $f(\mathbf{x})$ for high dimensional problems. The methods that are used in this thesis employ machine learning algorithms, more specifically Boosted Decision Trees (BDTs). The approximation $\hat{f}(\mathbf{x})$ built by the BDTs is not analytical and more details on the design of BDTs are given in the next section. These non-analytical methods have the advantage that they are more flexible in determining $\hat{f}(\mathbf{x})$ since they are not restricted to a certain assumption of its functional form. Therefore they can make use of features that have only an indirect correlation with y. On the other hand these methods need in general a substantial amount of observations to provide a reliable estimate, $\hat{y}$. Otherwise, due to its flexibility, it finds a $\hat{f}(\mathbf{x})$ such that is matches all given observations, which is called *over training*. In high energy physics the available simulated MC data sets provide this large set of observations, which are used in the machine learning algorithm to build the prediction model $\hat{f}(\mathbf{x})$ that estimates $\hat{y}$. A distinction is made between prediction models that are quantitative, called regression, or categorical, called classification. The question asked in the regression is: Given features $\mathbf{x}$ what is the most likely value for y? For classification it is: Given features $\mathbf{x}$ how likely is it that the observed event belongs to category S or B? Both functionalities are employed in this thesis — to estimate the transverse momentum of b-jets based on the features of the b-jet (regression) and to estimate the probability that an event is a signal event or a background event based on the features of the event (classification) [47, 48].

6.1.1 Training and Application

The prediction model of a machine learning algorithm is built in the so-called training. In the cases used in this thesis, this means the algorithm is presented with simulated events which exhibit features $\mathbf{x}$ and the outcome of the measurement y is known, e.g. the b-jet energy or if the event is a signal or background event. In the following all descriptions focus on the practical examples of this thesis. Since it is impossible to feed all available information to the learning algorithm the features that are assumed to characterise y have to be selected manually. These so-called input variables have to be selected according to the given problem and are usually kinematic features of the object (b-jets for regression) or the full event (classification). In case of regression the outcome y is also called the target of the training. The aim of the training is to minimise the difference between y and $\hat{y}$ in case of the regression or minimise the amount of events that were classified wrong in the case of classification. For an optimal minimisation enough information, e.g. in the form of relevant input variables, has to be provided in the training and the algorithm has to be flexible enough to build an accurate model of the input variables' phase space. Once the algorithm has built its prediction model it is applied to a given data set with unknown y, e.g. recorded ATLAS data. Based on the observations $\mathbf{x}$ given by the measured input variables' values the prediction model is evaluated and provides the estimate $\hat{y}$, e.g. for the b-jet p_T or the "signal-likeness" of the event. The application is only successful if the relation between the input variables and the outcome is similar between the training data set and the data set the prediction model is applied to [49, 50].

In the case of a classification problem the distribution of the evaluated outputs for the events in a given data set can be directly used as the discriminant of the analysis. The signal events are assigned a high "signal-likeness" probability whereas it is opposite for background events. Thus separating the signal and background events given by the distinctive shapes of the respective classifier output distributions.

6.1.2 Boosted Decision Trees

Boosted decision trees (BDTs) are the algorithms that are used for the regression and classification problems in this thesis. The concept of a single decision tree is depicted in figure 6.1. A single tree is "grown" by dividing the training data set into two sub sets (nodes) based on a selection requirement c_1 for an input variable x_i. Those two sub sets are divided again based on a c_2 and c_3 for input variable x_j. The same variable may be used several times in one tree. Splitting the training data set into smaller and smaller sub sets continues until a certain stop criterion is fulfilled, e.g. a minimum number of observations left in the sub sets or a maximum number of nodes or a maximum depth of the tree, which are the most common stopping criteria. The last nodes in the tree represent estimated values $\hat{y}_k$, e.g. the estimated b-jet p_T or "signal-likeness" of the event, given the conditions $(x_i < c_1, x_j > c_2, ...)$ for the input variables. For classification the "signal-likeness" is directly given by the relative amount of signal and background events that end up in a certain output node. The split at each node is determined by finding the input variable and corresponding requirement, which gives the best estimate for the outcome y or provides the largest separation between signal and background events. For the BDTs used in this thesis this is parametrised as:

$$\text{Mean squared error (regression): } \frac{1}{N}\sum_{i=1}^{N}(y - \langle y \rangle)^2 \tag{6.2}$$

$$\text{Gini index (classification): } p \cdot (1 - p) \tag{6.3}$$

For regression trees it is given by the average difference between the target value y and the average of all target values $\langle y \rangle$ in the resulting nodes. The requirement is therefore chosen such that the average difference in the nodes is minimised. Classification trees use the so-called Gini index, which uses the purity p of a node given by the fraction of signal events in a node compared to the total amount of events in the node[1]. The optimal requirement for splitting nodes in classification trees is therefore defined as the requirement that yields the largest difference between the Gini index of the daughter nodes compared to the mother node. However, the prediction power of a single tree is limited which is why so-called boosting is introduced. The general idea of boosting is to build new trees using weighted events according to the output of the previous tree. The weight depends on how "wrong" the outcome is predicted in the previous tree thus assigning a higher importance to this event in the next tree. Two boosting algorithms are used in this thesis, which use different definitions of "wrong". The regression BDTs utilise the gradient boost algorithm and the weight is given by the so-called Huber loss [51] which uses a modification of the simple difference between y and $\hat{y}$ to be robust against the small number of very large differences whose weights would otherwise dominate the performance of the next trees. The classification BDTs use adaptive boosting which uses a common boost weight given by the relative amount of misclassified events. The boosting step, i.e. growing of a new tree, is usually repeated many hundred times. The final estimate $\hat{y}$ of an event with observed features $\mathbf{x}$ of the prediction model is given by the weighted average of the estimate of the single trees [48–50].

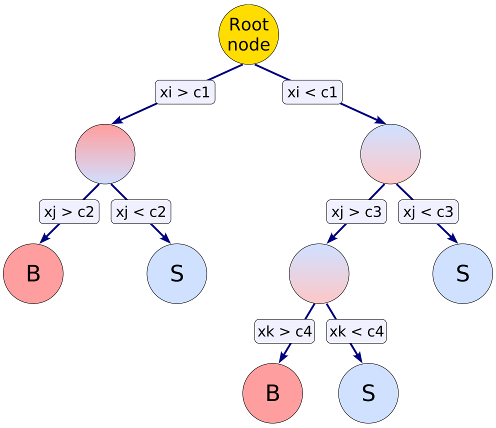


Figure 6.1: Schematic picture of a single decision tree. Each node represents a sub set of the training data set with different signal purities. The data set is divided using requirements c for the input variables x. The last nodes of the tree correspond to the output nodes and express the "signal-likeness" of the events in the sub set of the output node ("B" = background like, "S" = signal like). In a regression tree the output nodes are replaced by estimates of the target value y [50].

[1] Since high purity in background events is as favourable as a high purity in signal events the Gini index has its maximum value if the node contains 50% signal events and 50% background events.

6.2 Hypothesis Tests and Profile Likelihood Fit

Since particle physics processes are probabilistic all results make statistical statements of the forms: How likely is it that the data matches a background-only or background-plus-signal hypotheses? The formalism to answer this question is based on Poissonian probability likelihoods L:

$$L(\mu) = \text{Pois}(\text{data} \mid \mu \cdot s + b) \quad \overset{\text{bins}}{\longrightarrow} \quad \prod_{i=1}^{N_{\text{bins}}} \text{Pois}(n_i \mid \mu \cdot s_i + b_i) \tag{6.4}$$

which expresses: Given the data what is the probability for the signal(s)-plus-background(b) model? Usually the amount of signal events is not fixed in these models and allowed to contain an additional scaling factor μ, which is referred to as the signal strength. The signal strength is 0 if there is only background, 1 if there are signal events as expected and any other (positive value) if there are signal events but the amount is lower or higher than expected. The simplest experiment is a counting experiment that expects to observe certain amounts of events for the background-only and background-plus-signal hypothesis respectively and the probability that the data count matches either expectation is given by the respective Poisson probabilities. In many physics analyses a so-called discriminant is used instead of the total count of events. The discriminant is a binned distribution of events, i.e. each bin is a counting experiment by itself. Usually the discriminant is a distribution that has different shapes for background and signal events. For binned distributions the likelihood function is therefore given by the product of the Poisson probabilities of the single bins, as defined in equation 6.4 as well. However, in physics analyses the exact expected distribution, i.e. content per bin, of signal and background events is not known due to systematic uncertainties θ. This means, the amount of signal (or background) events in a given bin are functions of θ. The uncertainties are considered as additional parameters in the statistical model, which have to be determined from data and are referred to as nuisance parameters. The nuisance parameters automatically decrease the statistical power of the result since there are more possible ways to interpret the signal as a fluctuation of the background. Usually nuisance parameters are not introduced as free parameters but as parameters with outer constraints, which are given by e.g. complementary measurements. They are in most cases implemented as a Gaussian multiplicator in equation 6.4 called a Gaussian prior. This means the probability for the amount of events in a given bin to change towards other values is given by a Gaussian whose mean is the nominal amount of events and its width is the size of the uncertainty. Hence, variations "far away" from the nominal value are disfavoured and are penalised in the fitting procedure. Due to the physical meaning of the nuisance parameters it is important to parametrise their correlations correctly, e.g. if the background model is the sum of two background models that are affected by the same uncertainties or if the size of the uncertainties are different for the two background models but the underlying physical source is the same. In those cases nuisance parameters should not be introduced twice since their values should be determined in a consistent way, i.e. the nominal should not be allowed to be shifted in one direction for one background component and in the other direction for the other background component. This is usually meant by "correlated" or "un-correlated" uncertainties or nuisance parameters [52, 53].

The final hypothesis test uses a so-called test statistic t based on the ratio of two likelihood functions, which have to be maximised:

$$t = -2 \ln \lambda(\mu) = \frac{L(\,\mu, \hat{\hat{\theta}}\,)}{L(\,\hat{\mu}, \hat{\theta}\,)} \tag{6.5}$$

with $\hat{\hat{\theta}}$ denoting the values of θ that maximise L for a given μ (conditional likelihood) and $L(\hat{\mu}, \hat{\theta})$ the maximised unconditional likelihood with $\hat{\mu}$ and $\hat{\theta}$ as its estimators. The maximisation is what happens if "a fit to data" is performed, which estimates the values of $\hat{\theta}$ and $\hat{\mu}$ in the unconditional case. Thus if the data disfavours a given hypothesis for μ, $L(\mu, \hat{\hat{\theta}})$ and $L(\hat{\mu}, \hat{\theta})$ do not yield the same maximised result, i.e. the test statistic corresponds to the incompatibility between the data and a model [52, 53].

6.2.1 Measurements, Significances and Limits

To make a quantitative statement about the result of an analysis the fit to data is performed under different hypotheses. In the conventions of particle physics a measurement "discovered" a new phenomenon if the background-only hypothesis is rejected by 5σ. To set an upper limit on an observable, e.g. the signal strength, the signal-plus-background hypothesis for a given signal strength has to be excluded with a 95% probability. The following definitions are used in this thesis:

- **Measurements** To measure a signal the question to answer is: How likely is it to reject the background-only hypothesis? Equation 6.5 is evaluated with $\mu = 0$. If the background-only hypothesis can be rejected, the maximum for $L(\hat{\mu}, \hat{\theta})$ is achieved for $\hat{\mu} > 0$ thus leading to a minimum for the test statistic due to the discrepancies between the conditional and unconditional likelihood. The $\hat{\mu}$ for which equation 6.5 is minimal corresponds to the measured signal strength in data [52].

- **Significances** If the background-only hypothesis is rejected the question is: How large is the probability that the observed signal is caused by a fluctuation in the background? The significance quantifies the disagreement between background-only and background-plus-signal hypothesis given $\hat{\mu}$ and $\hat{\theta}$ obtained from the fit. It is assumed that the measured signal variable, e.g. the signal strength, is a Gaussian distributed variable. Thus its significance with respect to the expected value, e.g. $\mu = 0$, is given by: $(\mu_{\text{obs}} - \mu_{\text{exp}})/\sigma_{\text{obs}}$, with the measurement error σ_{obs}. The significance is also utilised to determine the sensitivity of an analysis based on its expected median value to reject the predicted background model if the data would reproduce exactly the expected background-plus-signal distributions. To calculate this a representative data set, e.g. using simulated events, is constructed by setting all parameters to their nominal value. This data set is commonly referred to as Asimov data set[2]. In the case of signal measurement the Asimov data set would be constructed by replacing $\hat{\mu}$: $\hat{mu} \quad \rightarrow \quad \mu' = 0$ [52, 53].

- **Limits** The question that is asked in case a limit is determined is: Which values of $\hat{\mu}$ are excluded by the data and at which confidence level? The conditional likelihood in this case is constructed from the signal-plus-background hypothesis with a hypothetical signal strength μ. The incompatibility between data and the model is maximal if $\hat{\mu} < \mu$ and the value for which this happens is the maximum strength of a potential signal that may be "hidden" in the data. Thus a limit measures up to which value of μ the data is still compatible with a background only hypothesis. To make this statement a confidence level CL for this compatibility has to be defined. Similar to the significance it is defined assuming that the measured variable $\hat{\mu}$ is Gaussian distributed and the probability to obtain μ instead is smaller than $1 - CL$. Thus the signal-plus-background hypothesis for a given μ is excluded with a certain probability that corresponds to the confidence level. Usually the signal confidence level CL_s is defined as the final measure, which is the ratio of the signal plus background confidence level CL_{s+b} and the background confidence level CL_b. An expected limit may be calculated as well with the help of an Asimov data set [52–54].

[2] Based on the short story "Franchise" by Isaac Asimov where a vote is conducted such that group of voters is replaced by a single representative voter.

CHAPTER 7

Jet Energy Regression

This chapter will introduce a multivariate energy correction for b-jets (in the following referred to as "jet energy regression"). Many physics processes at the LHC contain b-jets, for example the $H \rightarrow b\bar{b}$ decay. Therefore it is important to measure their properties as precise as possible. Several effects lead to a substantial deterioration of the energy measurement of b-jets in particular, but also all jets. The main contributing effects are the intrinsic calorimeter resolution, particles of the jets that are not contained in the reconstructed jet cone (so-called out-of-cone leakage) and semi-leptonic decays. As detailed in section 5.5 all jets that were reconstructed in the ATLAS detector are calibrated in multiple steps. First, the direction of the reconstructed jets is determined such that the jet axis points to the reconstructed hard scattering vertex. Afterwards the jet is corrected for pile-up contributions as a function of several pile-up observables, e.g. the number of pile-up vertices. Based on the simulated true jet energy as predicted by MC simulations first a global calibration factor is applied to the four-momentum of the jet. Additional MC based calibration factors are applied sequentially as functions of calorimeter, track and muon-segment variables to reduce flavour dependence and out-of-cone leakage effects. Jets in data events undergo an additional step where they are calibrated using in situ measurements. This chain of calibrations, referred to as standard ATLAS jet calibration (SJC) in the following, is described in detail in [55]. All described calibration steps use simulated or ATLAS data events inclusive in jets flavours. Due to a larger production cross section those data sets are dominated by light jets. There are substantial differences between the structure of b-jets and light jets that create different responses in the HCAL. Undetected neutrinos and possibly muons that are only partially detected in the HCAL, which carry away portions of the jet's energy are the largest source of b-jet energy mismeasurements. The BR for these semi-leptonic decays is in the order of 10% per lepton flavour for the b-hadron contained in the b-jet[1] [7]. The jet energy regression introduced in this chapter is designed to account for these differences and possible other sources of energy mismeasurement that are not included in the standard ATLAS calibration and meant to be applied only to b-jets on top of the SJC. The following sections will motivate the choice of input variables and target variable, present training cross checks and a validation in Zb events as well as discuss systematic uncertainties of this method. Corrections for b-jets utilising multivariate regression was initially studied and utilised in the CDF [56] and CMS [57] collaborations.

[1] The exact BR depends on the quark content of the b-hadron

7.1 Regression Set-Up

Due to the mentioned sources of jet energy mismeasurements the resolution of the b-jet energy is degraded and the total energy of b-jets is on average underestimated. The aim of the jet energy regression is to correct the energy with a multivariate algorithm, in this case BDTs. For details on multivariate regression with BDTs refer to sec. 6.1.2. In order to achieve this goal input variables for the BDTs, which are correlated with the jet energy and describe the sources of mismeasurements, have to be identified. To train the BDTs a simulated $t\bar{t}$ sample is utilised. The sample was generated at NLO with the Powheg generator and interfaced to Pythia 6 for the parton shower, underlying event and multiple parton interactions description. The details are listed in tab. 8.2 and described in sec. 4.2. The $t\bar{t}$ simulated data set is chosen since it provides a large training data set of over 2 million b-jets from the $t \rightarrow bW$ decay. It is a well established MC data set for b-jet related algorithms. In addition, the $t\bar{t}$ data set covers the b-jet p_{T} range that is relevant for many SM analyses. To define the target for the jet energy regression training a representation for the b-jet's original energy has to be identified. The simplest choice would be the b-quark from the top-quark decay before PS and hadronisation take place. On the other hand this initial b-quark is not very well defined in simulations since it is not a physical, i.e. measurable, object but only an intermediate particle and its 4-momentum is just a mathematical construct [58]. Therefore for the jet energy regression truth jets are used which are clustered from stable truth particles, i.e. physical particles in the simulation before they are affected by detector effects. They are reconstructed with the same anti-k_{T} algorithm with a radius parameter of $R = 0.4$ as the calorimeter jets. A truth particle qualifies as stable if its lifetime $c\tau$ is larger than 10 mm. Neutrinos, electrons, muons and associated photons are not used in the definition of a truth jet, except if they originate from hadron decays [58]. The p_{T} distribution of the reconstructed b-jets and the truth b-jets is displayed in fig. 7.1. The reconstructed b-jet p_{T} spectrum predicts more low p_{T} jets in comparison to the truth b-jet p_{T} spectrum which suggests that the p_{T} of the reconstructed jets is on average underestimated.

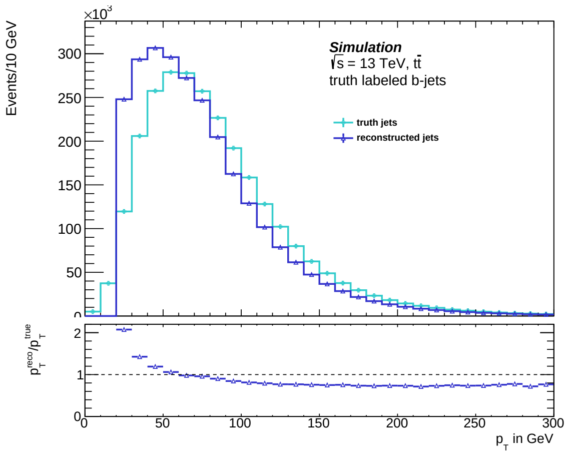


Figure 7.1: The p_{T} spectrum of reconstructed b-jets and truth b-jets in a simulated $t\bar{t}$ sample.

To avoid that the target for the jet energy regression is extended over a couple of magnitudes, the momentum of the truth jet $p_{\mathrm{T}}^{\mathrm{true}}$ is not directly used as the target. Instead the fraction of the transverse momentum of the reconstructed b-jet $p_{\mathrm{T}}^{\mathrm{reco}}$ and the transverse momentum of the truth jet $p_{\mathrm{T}}^{\mathrm{true}}$ is used[2]. This target represents the correction factor C for the full b-jet 4-momentum which consistently corrects

[2] For historic reasons the method presented here is called jet **energy** regression but it is using the transverse momentum following ATLAS conventions to assess the jet p_{T} resolution as a quality measure.

the momentum, energy and mass of the jet 4-momentum without changing the direction in η and ϕ:

$$C \equiv \frac{p_{\mathrm{T}}^{\mathrm{true}}}{p_{\mathrm{T}}^{\mathrm{reco}}} \tag{7.1}$$

In general, a large range of the target variable limits the precision of the algorithm since the output values only have a certain granularity given by the number of output nodes and trees. A b-jet only enters the training data set if exactly one truth jet was matched to it ($\Delta R < 0.4$) and if it fulfils the *signal* jet criteria as defined in sec. 5. In addition, there must be an associated b-quark with a transverse momentum of at least 4 GeV which has to be a descendant of a top quark. This requirement simply ensures that the jet that enters the training is a b-jet. These selection criteria are summarised in table 7.1.

The set of input variables is optimised by identifying input variable categories, such as b-jet kinematics, associated tracks, etc., and testing BDTs trainings of all possible combinations of variables of a given category. The training that yields the best b-jet p_{T} resolution and contains the smallest amount of variables is selected. Thus category by category variables are added. The full optimisation is described in detail in appendix A.2. The chosen variable set contains the following 9 (out of initially 24) input variables:

- **jet mass:** The reconstructed jet mass m is calculated as the invariant mass of the momenta of the massless HCAL clusters of the jet. It is directly proportional to the jet energy E and absolute momentum p. This information is linked to two effects: the intrinsic calorimeter resolution is proportional to $1/\sqrt{E}$ and the out-of-cone leakage is more significant for jets with low momenta since on average the jet's constituents carry less momentum. Therefore the jet is less collimated and in addition the charged constituents of the jet experience a small bending radius in the magnetic field of the inner detector. These effects lead to larger losses due to out-of-cone leakage.

- **jet width:** The jet width is defined as the average p_{T} weighted distance ΔR of the jet constituents, i.e. clusters the jet was reconstructed from, from the jet axis [38]:

$$\mathrm{width} = \frac{\sum\limits_{\mathrm{constituent}_i} \Delta R(\mathrm{jet{,}constituent}_i) \cdot p_{\mathrm{T}}^{\mathrm{constituent}_i}}{\sum\limits_{\mathrm{constituent}_i} p_{\mathrm{T}}^{\mathrm{constituent}_i}} \tag{7.2}$$

  It is a measure for how collimated the jet is. If a jet is less collimated typically the losses due to out-of-cone leakage are larger due to aforementioned effects.

- **p_{T} fraction carried tracks of the jet:** The p_{T} fraction carried by the tracks associated to the jet is defined as the scalar sum of the transverse momenta of the associated tracks divided by the reconstructed transverse momentum of the jet:

$$\frac{\sum\limits_{\mathrm{track}_i} p_{\mathrm{T}}^{\mathrm{track}_i}}{p_{\mathrm{T}}^{\mathrm{reco}}} \tag{7.3}$$

  The transverse momentum of the tracks is directly proportional to the b-jet p_{T} since on average 2/3 of the b-jet's energy is carried by charged particles. The p_{T} of the tracks is measured in the tracking system which has a better p_{T} resolution for low momenta than the HCAL. Tracks are associated to the jet via ghost association [59] which sets the p_{T} of the tracks to an infinitesimal value during jet clustering to keep the jet axis unaffected. Tracks are only considered if they pass basic tracking quality criteria and have a p_T of larger than 0.5 GeV to ensure reliable track reconstruction [60]. If no tracks are associated to the jet this variable is set to a default value of -1.

- **p_T of muons in jet:** The scalar sum of the transverse momenta of all muons that are found within a cone of $\Delta R < 0.4$ around the b-jet axis is a direct measure for the amount of energy that escaped with muons from semileptonic b-hadron decays. Several muons might be present in one b-jet if there are secondary semi-leptonic decays. In addition, this variable is an indicator for the amount of energy that escaped with the corresponding neutrinos of the decay, which leave the detector undetected. For the muons to be considered they have to fulfill at least the *loose* muon quality criterium and their transverse momentum has to be larger than 6 GeV since the muon mis-identification becomes sizeable for looser criteria [38]. If no muons are present in the jet this variable is set to a default value of -1.

- **p_T of muons and electrons in jet:** The scalar sum of the transverse momenta of all muons and electrons found within a cone of $\Delta R < 0.4$ around the b-jet axis is also a measure for energy from semi-leptonic b-hadron decay products that is not included in the reconstructed jet. Since an electron is only reconstructed and identified in 5% of all b-jets the electron information is summed together with the muon information such that the variable is not ignored by the multivariate algorithm. To define this variable the same muons as in the definition for the p_T of muons in jet variable are used. The additional electrons have to fulfil at least the *loose likelihood* quality criteria and their transverse momentum has to be larger than 6 GeV to ensure reliable electron reconstruction and identification [37]. If no muons and no electrons are present in the jet this variable is set to a default value of -1.

- **secondary vertex mass:** The SV mass is the invariant mass of all tracks associated to the SV [61]. It is proportional to the mass of the decaying b-hadron. The SV has to fulfull quality criteria listed in Ref. [42] to ensure that the reconstructed vertex corresponds to a b-hadron decay vertex. Due to the size reduction policy for simulated data sets the SV information is only available for jets that were clustered using tracks instead of HCAL clusters. The track jets are reconstructed with the anti-k_T algorithm with a radius parameter of $R = 0.3$. Their momentum has to larger than 10 GeV and the distance ΔR between the track jet and the reconstructed calorimeter jet has to be smaller than 0.4. If no SV is reconstructed this variable is set to a default value of -1.

- **secondary vertex decay length significance:** The decay length significance is defined as the distance of the SV from the primary vertex in three dimensions L_3D divided by the measurement error on this distance [61]. The decay length is directly proportional to the transverse momentum of the b-hadron and increases with the increasing b-hadron momentum. SV reconstruction is performed as mentioned above. and track jets are used to define this variable as well. If no SV is reconstructed this variable is set to a default value of -1.

- **jet vertex tagger weight:** The jet vertex tagger weight is a measure for the amount of energy contained in the jet that originates from pile-up and not from the fragmentation of the initial b-quark [62]. The amount of pile-up activity and its distribution in the detector volume is independent from the hard scattering. Since pile-up jets have a displaced SV as well they may be identified as b-jets. The jet vertex tagger information helps to distinguish those jets. Otherwise the jet energy regression would treat pile-up jets as b-jets and correct them as such.

- **p_T of close-by jets:** The scalar sum of the transverse momenta of jets found within $0.4 < \Delta R < 1.0$ around the b-jet axis is a measure for the amount of energy carried away by hard gluons or quarks emitted during the jet fragmentation process. These partons can be reconstructed as separate jets and their energy will be missing in the b-jet reconstruction. If no close-by jets are found this variable is set to a default value of -1.

The selection criteria for the objects used to define the jet energy regression input variables are summarised in table 7.2. For information on the reconstruction of these objects refer to section 5. No further event selection is applied for the training because the jet energy regression is applied per b-jet and not per event. The description of the input variables in the simulated $t\bar{t}$ data set in comparison to data was checked in a dedicated validation region which selects events with exactly one muon, one electron and two b-jets and therefore is very pure in $t\bar{t}$ events. The details of the event selection and all histograms are included in app. A.3. Overall a good agreement between data and simulated events is observed for most of the variables. Larger deviations are present in the jet mass and jet width distributions. However, it is important that the correlation between the input variables and the b-jet p_T, which is the feature that is corrected, is correctly described, which is the case. This is reflected in the $p_\mathrm{T}^\mathrm{jet}$ distribution in data and simulated events: the level of agreement before and after the jet energy regression is applied is similar, see fig. 7.2. If an input variable that is not correctly described in simulated events would introduce a bias this would cause an additional level of disagreement between data and simulated events.

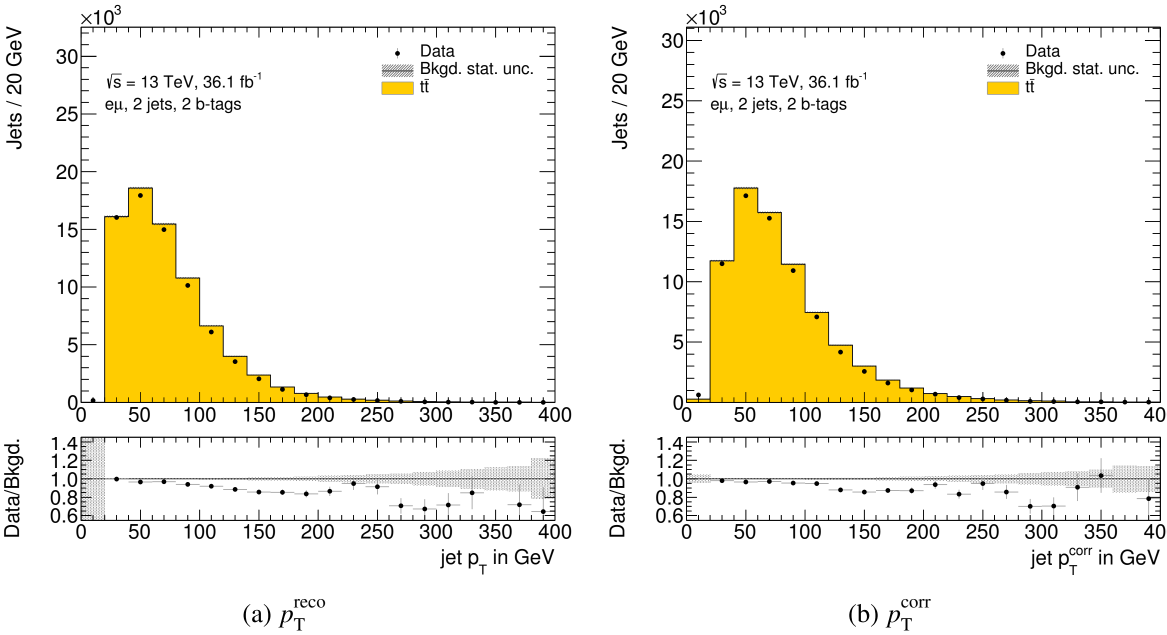
Truth Jets j_truth	b-quark b	Reconstructed Jets j
$\Delta R(j, j_\mathrm{truth}) < 0.4$	$\Delta R(j, b) < 0.4$ descendant of truth top quark p_T >4 GeV	*signal jet* N(associated j_truth) = 1 N(associated b) = 1

Table 7.1: Selection criteria for the reconstructed and truth jets that are used in the jet energy regression. The additional selection of the b-quark ensures that the jet is actually a b-jet.

Tracks	Muons μ	Electrons e	Close by Jets j_near	Track Jets j_track
tracking quality	*Loose* quality	*Loose LH* quality		
$p_\mathrm{T} > 0.5$ GeV	$p_\mathrm{T} > 6$ GeV	$p_\mathrm{T} > 6$ GeV	$p_\mathrm{T} > 15$ GeV	$p_\mathrm{T} > 10$ GeV
ghost associated to j	$\Delta R(\mu, j) < 0.4$	$\Delta R(e, j) < 0.4$	$0.4 < \Delta R(j_\mathrm{near}, j) < 1.0$	$\Delta R(j_\mathrm{track}, j) < 0.4$

Table 7.2: Selection criteria for the objects that are associated to the jet j utilised in the jet energy regression. The same selection is applied for the jet energy regression training and the application of the jet energy regression.

The training of the BDTs is done with TMVA which is a toolkit for multivariate algorithms in ROOT [50]. The boosting is done with the gradient boost algorithm. Except for the number of segments that are allowed on the distributions of the input variables all training parameters were kept at their default value for regression trees. Reducing the number of trees and increasing the depth of the individual trees was briefly tested and no improvement in performance was found, see appendix A.1. The number of segments of the input variables was increased to 1000 since some of the variables, e.g. p_T of the muons inside b-jets, have a large range and more segments allows finer cuts on them. The training data set contains 2.6 million b-jets which permits such a high number of segments. All training parameters are summarised in the appendix A.1. Two separate trainings are run, each with 2.6 million b-jets: one trained with b-jets from events with an even event number (referred to as "even training") and one with b-jets from events with an odd event number (referred to as "odd training"). The even (odd) training is applied to b-jets in events with an odd (even) event number. This way the data sets used for the training and the application are fully orthogonal. To test that the trainings did not become sensitive to statistical

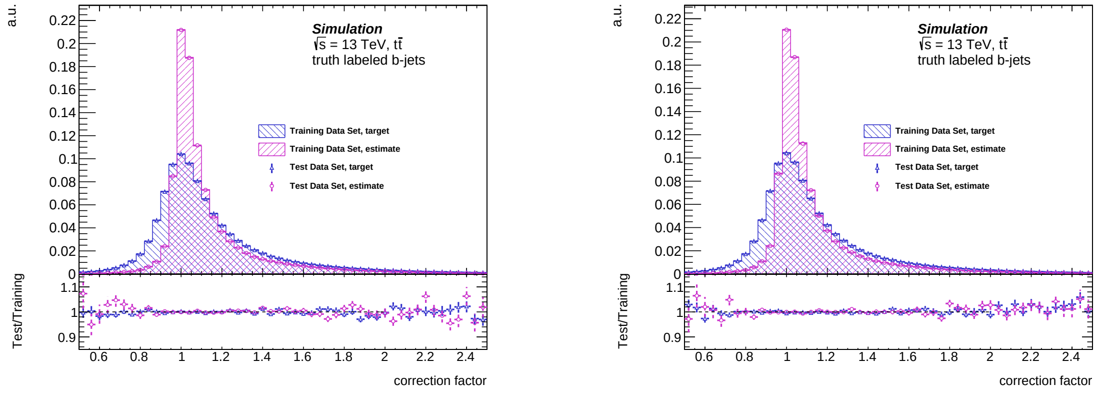


(a) $p_{\mathrm{T}}^{\mathrm{reco}}$

(b) $p_{\mathrm{T}}^{\mathrm{corr}}$

Figure 7.2: Distributions of the transverse momentum of the *b*-jets a) before and b) after the jet energy regression is applied. Shown is a comparison between data and MC in the $t\bar{t}$ validation region.

features in the training data sets, so-called *over training*, the jet energy regression is applied to the very same events it was trained on and to a statistically independent simulated data set. Figure 7.3 shows the distribution of the target (correction factor C) and the estimate of the jet energy regression of the correction factor $C_{\mathrm{est.}}$ for the training data sets and statistically independent test data sets separately for the even and odd training. The distributions of the correction factors is the same, within statistical uncertainties, in the test and training data sets. Therefore it is concluded that *over training* is not present.



(a) Even Training

(b) Odd Training

Figure 7.3: Overtraining check for the even (left) and odd (right) jet energy regression trainings. Shown is the distribution of the correction factor (target) and the estimated correction factor from the training (estimate) for the training data set and an independent test data set. A simulated $t\bar{t}$ data set was utilised in both cases.

7.2 Performance

The performance of the jet energy regression is evaluated comparing the reconstructed transverse momentum of the b-jets $p_{\mathrm{T}}^{\mathrm{jet}}$ after the application of the jet energy regression correction factor to the truth jet's transverse momentum $p_{\mathrm{T}}^{\mathrm{true}}$. Before the jet energy regression is applied, $p_{\mathrm{T}}^{\mathrm{jet}}$ corresponds to the reconstructed transverse momentum after the SJC $p_{\mathrm{T}}^{\mathrm{reco}}$. The jet energy regression corrects $p_{\mathrm{T}}^{\mathrm{reco}}$ by the estimated correction factor $C_{\mathrm{est.}}$ based on the input variables of the b-jet. This results in a corrected transverse momentum $p_{\mathrm{T}}^{\mathrm{corr}}$:

$$p_{\mathrm{T}}^{\mathrm{corr}} = C_{\mathrm{est.}} \cdot p_{\mathrm{T}}^{\mathrm{reco}} \quad \approx \quad C \cdot p_{\mathrm{T}}^{\mathrm{reco}} = \frac{p_{\mathrm{T}}^{\mathrm{true}}}{p_{\mathrm{T}}^{\mathrm{reco}}} \cdot p_{\mathrm{T}}^{\mathrm{reco}} = p_{\mathrm{T}}^{\mathrm{true}} \tag{7.4}$$

For the assessment of the performance the regression is applied to b-jets in a subset of the simulated $t\bar{t}$ data set, which is statistically independent from the training data set. The selection for b-jets and associated objects is the same as in the training and is detailed in table 7.1 and 7.2, respectively[3]. Figure 7.4 compares the distribution $p_{\mathrm{T}}^{\mathrm{jet}}/p_{\mathrm{T}}^{\mathrm{true}}$ after the SJC and after the jet energy regression. It is clearly visible that the distribution is more symmetric and more narrow after tjet energy regression is applied. To quantify the improvement a Bukin function [63] is fitted to the distribution. The Bukin function assumes a Gaussian core and models the left and right tail separately with exponential functions[4]. The application of the jet energy regression improves the position of the peak from 0.9821 ± 0.0002 to 1.0087 ± 0.0009 and the resolution, given by the standard deviation of the Gaussian core, improves by 21.4%. Figure 7.5b) demonstrates that the jet energy regression improves $p_{\mathrm{T}}^{\mathrm{jet}}/p_{\mathrm{T}}^{\mathrm{true}}$ over the whole range of the $t\bar{t}$ p_{T} spectrum. The distribution is symmetrically centered around 1 and the underestimation of the jet p_{T}— by up to 10% in the low $p_{\mathrm{T}}^{\mathrm{jet}}$ bins as shown in figure 7.5a) — is corrected.

In addition, the influence of the input variables on the performance is investigated by adding the input variables one by one. The BDTs are re-trained in each case. For each training the mean and the standard deviation of the $p_{\mathrm{T}}^{\mathrm{corr}}/p_{\mathrm{T}}^{\mathrm{true}}$ distribution is extracted. Figure 7.6 shows that the jet mass is the most important variable to correct the jet p_{T} scale and the sum of the p_{T} of the muons inside the jet is the most important variable to correct the jet p_{T} resolution. Variables added after the sum of the p_{T} of the muons inside the jet yield an additional 5% improvement in the jet p_{T} resolution.

7.3 Zb Validation

The regression is validated in events where a Z boson is produced in association with one b-jet. In those events the transverse momentum of the Z boson and the transverse momentum of the b-jet should be balanced. The p_{T} of the Z boson can be measured very precisely for the $Z \to \mu^+\mu^-$ decay due to the good muon p_{T} resolution. Figure 7.7 shows examples of Zb production processes. To select the desired Zb process, the events are required to pass the standard single muon trigger and must contain exactly two muons and one jet that pass the standard quality requirements as detailed in chapter 4. All objects also have to pass the standard overlap removal procedure as described in section 8.2. In addition, the muons have to be of opposite charge and fulfil the *VH loose* lepton requirement with at least one of them also passing the *VH tight* lepton requirement, as defined in section 8.3. The invariant mass of the muon pair $m_{\mu\mu}$ has to be in agreement with the Z-boson mass ($81\,\mathrm{GeV} < m_{\mu\mu} < 101\,\mathrm{GeV}$). The jet in the event

[3] To access the performance on equal footing as in the training, truth labelling of b-jets is used in this section. In a physics analysis, the truth labelling is replaced with a b-tagging requirement.

[4] All given errors of the fitted parameters are of statistical nature. They were determined by sampling 1000 histograms from varying the bin contents by a random Poissonian term.

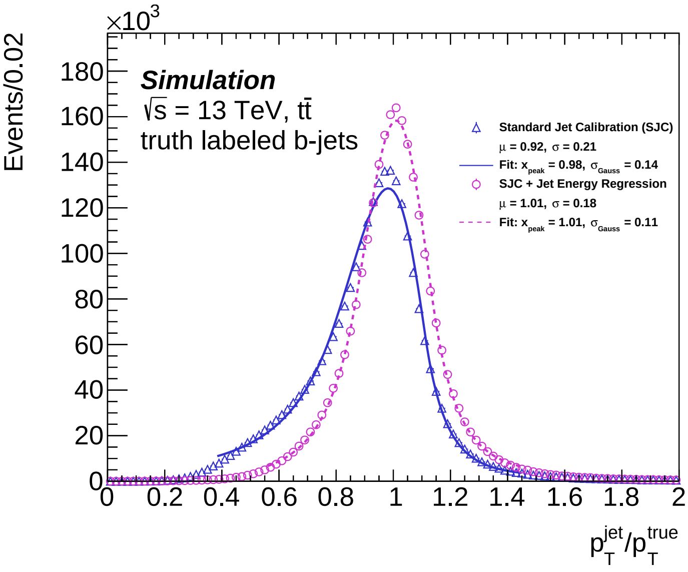
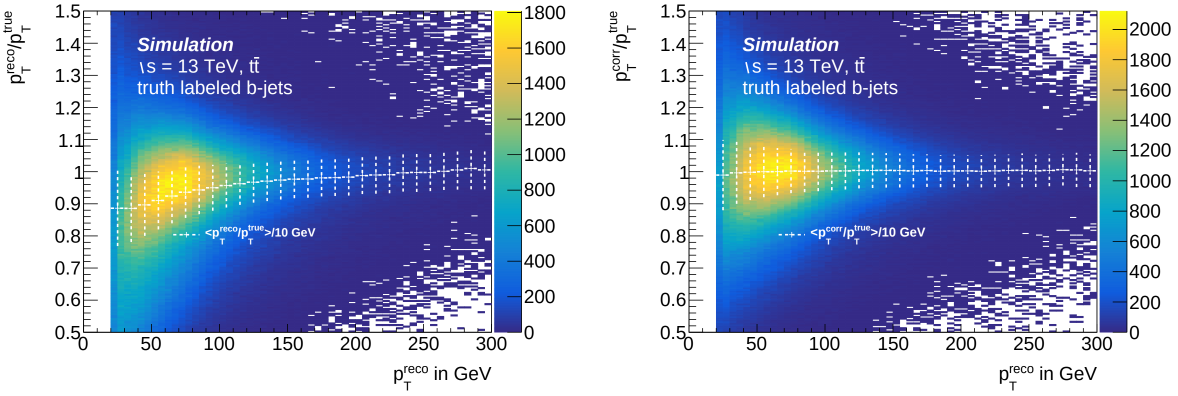


Figure 7.4: The distribution of $p_{\mathrm{T}}^{\mathrm{jet}}/p_{\mathrm{T}}^{\mathrm{true}}$ for b-jets after the SJC (blue) and after the jet energy regression (pink) was applied on top of it in a simulated $t\bar{t}$ data set. The mean μ and standard deviation σ of the distributions (points) are given as well as the peak position x_{peak} and width of the Gaussian core σ_{Gauss} determined by fitting a Bukin function (lines).



(a) SJC

(b) SJC+jet energy regression

Figure 7.5: The distribution of $p_{\mathrm{T}}^{\mathrm{jet}}/p_{\mathrm{T}}^{\mathrm{true}}$ as a function of $p_{\mathrm{T}}^{\mathrm{reco}}$ for b-jets in a simulated $t\bar{t}$ data set: a) after the SJC and b) after the application of the jet energy regression. The white points indicate the average $p_{\mathrm{T}}^{\mathrm{jet}}/p_{\mathrm{T}}^{\mathrm{true}}$ per 10 GeV bin of $p_{\mathrm{T}}^{\mathrm{reco}}$. The error bars correspond to the standard deviation of $p_{\mathrm{T}}^{\mathrm{jet}}/p_{\mathrm{T}}^{\mathrm{true}}$ in each individual bin.

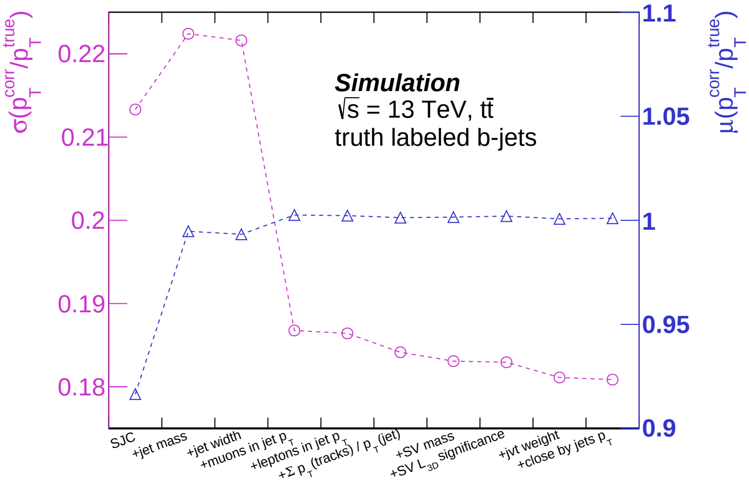


Figure 7.6: Evolution of the mean and standard deviation of the $p_T^{\mathrm{corr}}/p_T^{\mathrm{true}}$ distribution by gradually adding input variables to the jet energy regression training. A simulated $t\bar{t}$ data set was used.

has to fulfil the *signal* jet criteria, as defined in section 8.3, and pass the 60% b-tagging requirement. This tight b-tagging cut was chosen to efficiently suppress contributions from events with a light jet ($Z + l$) and events with a charm jet ($Z + c$). The jet energy regression is applied to all jets that pass the selection. There are no other central or forwards jets allowed in the event since additional jets would disturb the balance of the Zb-system. For this validation the 2015 and 2016 ATLAS data set was used which corresponds to 36.1 fb^{-1}. For a comparison simulated Z+jets events produced with the SHERPA event generator are used. This is the recommended simulated sample for this process for ATLAS analyses and the same as used for the analysis presented here. Contributions from $t\bar{t}$ are taken into account as well and the simulated $t\bar{t}$ events are produced with the POWHEG MC generator interfaced to PYTHIA for the description of the PS and UE. More details on the Z+jets and $t\bar{t}$ MC generators are listed in tab. 8.2. All other processes have negligible contributions in this kinematic phase space. The SHERPA Z+jets simulated data sets have problems with discontinuities in the p_T^Z distribution that are introduced due to the "slicing" that is used to enhance the amount of simulated events in higher p_T phase spaces. To correct this effect a dedicated weighting of simulated events in bins of p_T^Z was introduced to match the data. In addition, it is known that the normalisation of the heavy flavour jets component of the Z+jets samples is not correctly normalised which causes a discrepancy between data and simulated Z+jets events. Therefore a global scaling factor of 1.35, based on the observed normalisation difference between data and simulated events, is introduced. The reweighting and scaling has no influence on the final measure of the Zb validation as documented in appendix A.4. All details on the reweighting and scaling are given there as well.

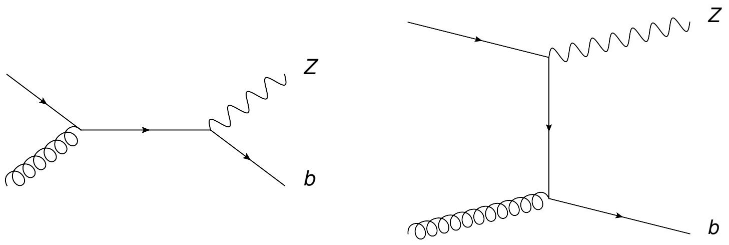


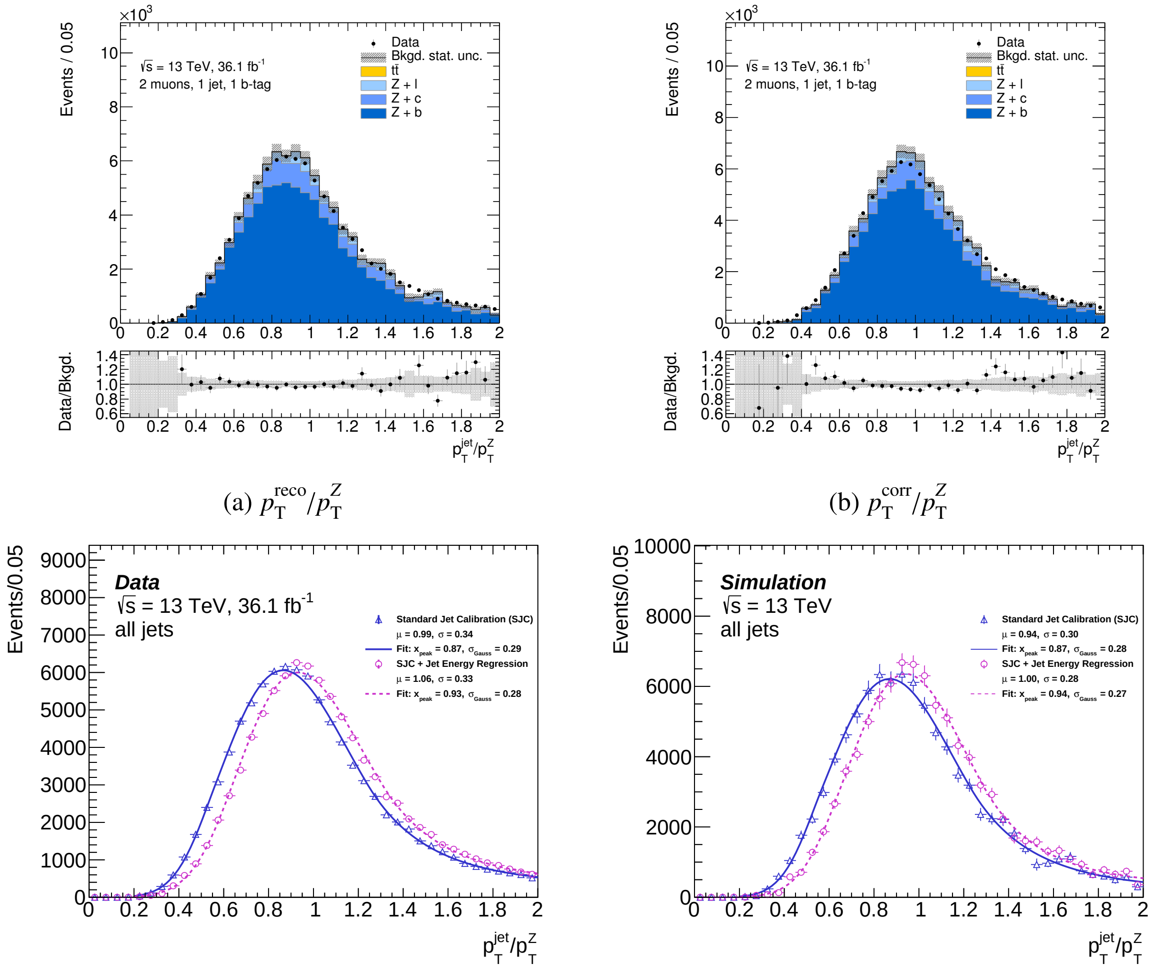
Figure 7.7: LO diagrams that contribute to Zb production.

As motivated beforehand, the p_{T} balance of the Zb-system allows to measure the b-jet transverse momentum resolution which is measured as the b-jet p_{T} relative to the Z-boson p_{T} ($p_{\mathrm{T}}^{\mathrm{jet}}/p_{\mathrm{T}}^{Z}$). The Z-boson p_{T} is reconstructed from the momenta of the two selected muons in the event. For a quantitative assessment a Bukin function is fit to the $p_{\mathrm{T}}^{\mathrm{jet}}/p_{\mathrm{T}}^{Z}$ distribution. The fit results are compared for the b-jets that are calibrated using the standard ATLAS jet calibration (SJC) and after the b-jets were additionally corrected using the jet energy regression. To ensure that the jet energy regression does not introduce artificial discrepancies between data and MC prediction the fit is performed separately for data and MC. Figure 7.8 shows a similar level of agreement between data and MC prediction before and after the application of the jet energy regression. In addition, the jet energy regression improves the peak position from 0.87(0.87) to 0.93(0.94) in data(MC) thus shifting it closer to the expected value of $p_{\mathrm{T}}^{\mathrm{jet}}/p_{\mathrm{T}}^{Z} = 1$. Furthermore, the p_{T} balance is measured in several regions of the Zb phase space and the results are summarised in tab. 7.3. The corresponding histograms and fits are included in appendix A.5. The performance of the jet energy regression for b-jets that include a muon (i.e. a semi-leptonic b-hadron decay took place) and b-jets that do not, is similar. An improvement of the peak position towards the expected value of 1 is observed in both cases. Therefore, the jet energy regression is able to correct for the special case of semi-leptonic b-hadron decays but also improves the p_{T} of the b-jets beyond that as demonstrated with b-jets containing no muons. The p_{T} balance was also measured for different regimes of p_{T}^{Z}: *low* (p_{T}^{Z} <75 GeV), *medium* (75 GeV< p_{T}^{Z} <150 GeV) and *high* (p_{T}^{Z} >150 GeV). These specific values were chosen based on the p_{T}^{Z} categories in the $V(H \to b\bar{b})$ analysis. The improvement in the peak position gets smaller in the high p_{T}^{Z} regime due to intrinsically better $p_{\mathrm{T}}^{\mathrm{jet}}$ resolution for higher transverse b-jet momenta (see also fig. 7.5). In all cases no significant differences are found between the values extracted from the fit to data and the fit to simulated events. As a last check the p_{T} balance of c-jets is studied in simulated events which allow an identification of the jet flavour from MC truth information. Since the only experimental way to identify b-jets is with the help of b-tagging algorithms the jet energy regression is also applied to light jets and c-jets that were mis-identified. It is important to make sure that these mis-identified jets are treated in a sensible way by the jet energy regression. For c-jets an improvement of the peak position towards the expectation of 1 is observable. No measurement is done for light jets since the amount of light jets that pass the b-tagging requirements is statistically limited and no reliable quantitative statement can be made. Due to the good suppression of mis-identified light jets they will not have a significant impact on any analysis that may use the jet energy regression. In conclusion, the performance gain of the jet energy regression was validated using the p_{T} balance of Zb events. It treats jets that pass b-tagging requirements consistently in data and MC.

7.4 Systematic Uncertainties

Three possible sources of systematic uncertainties of the jet energy regression are investigated. Most of the ATLAS analyses that involve b-jets are impacted by these uncertainties and take them into account.

1. **Differences in treatment of data and MC events:** In the previous sections the jet energy regression was tested in a $t\bar{t}$ validation region as well as in a region enriched in Zb events. In both cases, compare fig. 7.2 and 7.8, the data to MC agreement is as good for nominal b-jets, that just underwent the SJC, as for b-jets that were additionally corrected with the jet energy regression. In conclusion, the level of agreement did not become worse due to the application of the jet energy regression and the jet energy regression treats simulated events and data events consistently. Therefore no jet energy regression specific systematic uncertainty has to be assigned to account for effects due to differences in treatment of data and MC events.



(a) $p_{\mathrm{T}}^{\mathrm{reco}}/p_{\mathrm{T}}^{Z}$

(b) $p_{\mathrm{T}}^{\mathrm{corr}}/p_{\mathrm{T}}^{Z}$

(c) Bukin fits to data

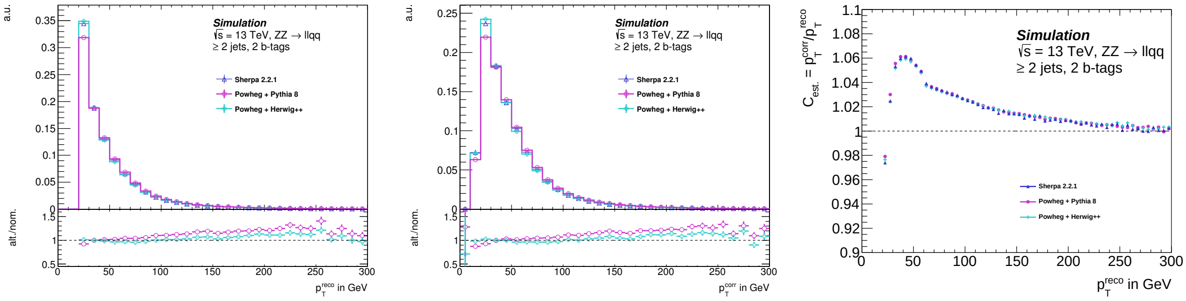
(d) Bukin fits to MC

Figure 7.8: Comparison between data and MC simulation of the p_{T} balance of the Zb system a) before and b) after the jet energy regression is applied. A Bukin function is fit separately to the c) data and d) MC events.

	Data		MC	
	SJC	SJC + regression	SJC	SJC+regression
all jets	0.865 ± 0.002	0.931 ± 0.003	0.869 ± 0.002	0.941 ± 0.002
jets with no muon	0.878 ± 0.003	0.927 ± 0.003	0.884 ± 0.002	0.941 ± 0.002
jets with ≥ 1 muon	0.854 ± 0.007	0.968 ± 0.008	0.839 ± 0.007	0.946 ± 0.007
low p_{T}^{Z}	0.866 ± 0.003	0.951 ± 0.005	0.865 ± 0.003	0.957 ± 0.003
medium p_{T}^{Z}	0.903 ± 0.005	0.931 ± 0.005	0.914 ± 0.002	0.949 ± 0.004
high p_{T}^{Z}	0.944 ± 0.012	0.936 ± 0.011	0.965 ± 0.010	0.966 ± 0.010
c-jets	–	–	0.885 ± 0.005	0.964 ± 0.005

Table 7.3: The peak positions in several phase spaces that were determined by a fit of a Bukin function to the $p_{\mathrm{T}}^{\mathrm{jet}}/p_{\mathrm{T}}^{Z}$ distribution. The results are given for b-jets before and after the jet energy regression was applied and separately for data and MC events.

2. **Differences in the treatment of MC models:** The training of the jet energy regression relies on the Powheg +Pythia description of the input variables and the truth jet p_{T}, which is used for the definition of the target. In particular the latter depends on the parton shower (PS) model that is used in the simulation. Therefore it is important to make sure that the jet energy regression is applicable to simulated events that were generated with other MC models and if additional systematic uncertainties are necessary. The simulation of the $ZZ \to \ell\ell qq$ decay is available with PS modelled by Pythia, Herwig and Sherpa, the three main generators used in ATLAS to model the PS. A comparison of the modelling of the jet p_{T} distribution before and after the application of the jet energy regression is shown in fig. 7.9. The jets shown there have to pass b-tagging requirements and fulfil the *signal* jet criteria. The variation between the three PS models is as large before as after the jet energy regression is applied. In addition, it is demonstrated that the correction factor assigned by the jet energy regression to a jet with a given p_{T} does not depend on the PS model Therefore no additional jet energy regression specific uncertainty due to the particular choice of the PS model in the training is necessary.



(a) $p_{\mathrm{T}}^{\mathrm{reco}}$

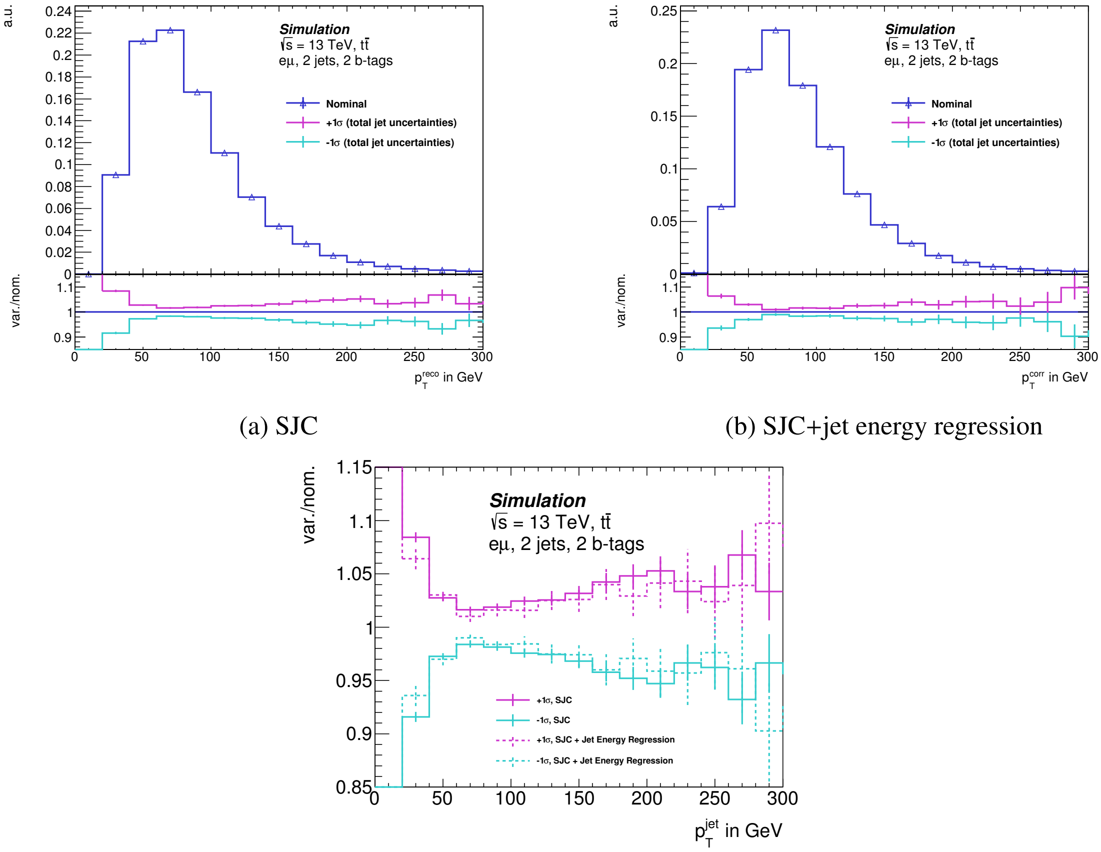
(b) $p_{\mathrm{T}}^{\mathrm{corr}}$

(c) $p_{\mathrm{T}}^{\mathrm{corr}}$

Figure 7.9: Comparison of the $p_{\mathrm{T}}^{\mathrm{jet}}$ modelling a) before and b) after the jet energy regression was applied and c) the applied correction factor as a function of $p_{\mathrm{T}}^{\mathrm{reco}}$ for three different simulation models: Sherpa, Powheg +Pythia and Powheg +Herwig. In a) and b) the lower panel shows the ratio of the alternative (alt.) Pythia and Herwig models to the nominal (nom.) Sherpa simulation model.

3. **Effects of the intrinsic jet energy scale and resolution uncertainties:** Each jet in a MC event that was calibrated with the SJC has a set of systematic uncertainties assigned to it. These uncertainties are connected to the systematic effects of the calibration procedure such as the flavour response, pile-up, etc. They have to be taken into account in every analysis that utilises jets. Therefore it is important to study how jets which are systematically shifted are treated by the jet energy regression. This study was carried out in the phase space of the $t\bar{t}$ validation region (see sec. 7.1). All jet energy regression input variables were varied by $+1\sigma$ and -1σ from their nominal value as given by the systematic uncertainty. Afterwards the jet energy regression is applied using the varied events as inputs. This is done separately for each systematic uncertainty. Finally, the $+1\sigma$ and -1σ variation was averaged and symmetrised for each systematic uncertainty and the total uncertainty was determined as the sum in quadrature of the individual uncertainties. Figure 7.10 shows the impact of the total uncertainty on the $p_{\mathrm{T}}^{\mathrm{jet}}$ distribution with and without the application of the jet energy regression. Within the statistical precision a similar effect on the $p_{\mathrm{T}}^{\mathrm{jet}}$ distribution is observed. Therefore it may be concluded that it is valid to use the intrinsic uncertainties connected to reconstructed jets also for jets that were corrected by the jet energy regression.

As demonstrated the studied intrinsic uncertainties are consistently propagated through the jet energy



(a) SJC

(b) SJC+jet energy regression

(c) Comparison of total uncertainty

Figure 7.10: Impact of the systematic uncertainties connected to the jet reconstruction and calibration on jets a) after the application of the standard ATLAS jet calibration (SJC) and b) after the application of the jet energy regression on top of it. A direct comparison of both cases is displayed in c). The total symmetrised uncertainty is shown.

regression. Although, as all experimental techniques, the jet energy regression certainly has systematic uncertainties, these uncertainties, based on the extent of the studies presented here, are concluded to be negligible compared to the intrinsic systematic uncertainties all analyses are impacted by. Therefore no jet energy regression specific systematic uncertainties have been assigned since the p_{T} of b-jets that were corrected with the jet energy regression react to systematic variations in the same way as uncorrected b-jets.

CHAPTER 8

Search for the Standard Model $H \to b\bar{b}$ Decay

The discovery of a Higgs boson [14, 64] (H) in 2012 raised the question if this particle is the Higgs boson as predicted by the Standard Model (SM). Many of the Higgs boson's properties have been measured in the decay into a pair of bosons, i.e. photons, Z bosons and W bosons. In those channels, the mass has been measured to be approximately 125 GeV. In contrast, the decay into fermions is not discovered, yet. However, those decays are an important piece to establish the Yukawa coupling between the Higgs boson and fermions, which is "ad hoc" introduced in the SM Lagrangian, as described in chapter 2. Since the coupling strength is proportional to the mass of the fermion, decays to a pair of heavy fermions are expected to have a larger branching ratio (BR). However, the mass of the heaviest fermion, the top quark ($m_t \approx 175$ GeV), is much larger than the Higgs boson mass. Therefore this decay is phase space suppressed. The second and third heaviest fermions are the bottom quark ($m_b \approx 4$ GeV) and τ-lepton ($m_\tau \approx 2$ GeV). The Higgs boson (H) decay into a pair of bottom quarks ($b\bar{b}$) is expected to have a BR of 58% and the decay into a pair of τ-leptons has a BR of 6%. Until recently, the only evidence for Higgs boson to fermion couplings was the decay into τ-leptons [65] obtained with run 1 data. The search for the $H \to b\bar{b}$ decay with run 1 data did not yield evidence for this decay. The observed significance of this search with ATLAS data is 1.4 σ [66] and a similar analysis performed with CMS data obtained 2.1 σ [67]. Due to the challenges of analyses targeting the $H \to b\bar{b}$ decay a discovery only recently comes into reach using the large amount of run 2 data.

The analysis presented here sheds more light on the $H \to b\bar{b}$ decay which is — despite its large BR — a truly challenging Higgs boson decay channel. The production cross section for a multitude of jets, multi-jet, is approximately 10^7 times higher than the cross section of the $H \to b\bar{b}$ signal. The first challenge are the available signatures the ATLAS trigger system is capable to record since the rate of produced multi-jet events is simply not manageable. Thus, the ATLAS trigger system only records events with jets that have a transverse momentum of the order of several 100 GeV. Due to the mass of the Higgs boson the p_{T} of the b-jets is not expected to be this high. Special trigger options that target b-jets are also not sufficient for the $H \to b\bar{b}$ channel since the p_{T} threshold is still high ($\approx$ 200 GeV or 150 GeV and 50 GeV for two b-jets) in order to keep the trigger rate manageable and the efficiency is 80% or less depending on the exact trigger [20]. Therefore the search for the $H \to b\bar{b}$ decay targets a specific production channel of the Higgs boson: associated production with a weak vector boson $pp \to VH$, see figure 8.1. The decay of the vector boson into leptons offers a good trigger signature with much lower p_{T} trigger thresholds (starting at $\approx$ 20 GeV). The caveat is the 50 times lower production cross section, compared to the most dominant Higgs boson production channel, which is further reduced by the low BR for leptonic V boson decays. Nevertheless, VH production is so far the most sensitive channel to search for the $H \to b\bar{b}$ decay[1]. The advantage of the VH channel is a largely reduced amount of expected background events due to the low cross sections — compared to multi-jet production — of SM processes that produce leptons and b-jets.

The analysis targets the decay of the Z boson to a pair of neutrinos ($Z \to \nu\bar{\nu}$), electrons or muons (summarised as $Z \to \ell^- \ell^+$). The decay into τ-leptons is excluded from the $Z \to \ell^- \ell^+$ search channel due to the additional complications that are involved in τ-lepton reconstruction and estimation of contributing background processes. The $W^\pm H \to \ell^\pm \overset{(-)}{\nu} b\bar{b}$ channel is also not specifically targeted although it has a much higher amount of expected signal events due to the larger WH production cross section and the larger BR of the $W^\pm \to \ell^\pm \overset{(-)}{\nu}$ decay. However, this advantage is outweighed by the higher amount of expected background events, especially from W+jets and $t\bar{t}$ production, and non-negligible contributions from multi-jet production whose kinematic distributions have a large uncertainty. Hence the choice to investigate $Z(H \to b\bar{b})$ signatures, which have a better expected ratio of signal to background events. Moreover the prediction of the contributing background processes has a higher precision. This search uses good quality data recorded by the ATLAS detector during the 2015 and 2016 data taking periods. During those periods the LHC collision energy was $\sqrt{s} = 13\,\text{TeV}$. The investigated data set corresponds to $36.1\,\text{fb}^{-1}$ of pp collisions.

8.1 Signal and Background Processes

In order to develop an analysis strategy it is crucial to understand the signatures of the $Z(H \to b\bar{b})$ signal processes as well as all SM processes that produce similar signatures and therefore populate the same phase space as the signal processes. These processes are studied using MC simulations. More details on simulations and MC generators for ATLAS analyses are given in sections 4.1 and 4.2. In case the cross section of a certain process is known at a higher order than the MC generator uses for the simulation[2], the events in that simulation are normalised to the known cross section. In the following their experimental signatures are discussed. The cross sections and branching ratio (BR) for all relevant processes are also listed in table 8.1 and the details on the simulation models of these processes are given in table 8.2.

SM VH: The analysis presented here focuses on the $ZH \to \nu\bar{\nu}b\bar{b}$ and $ZH \to \ell^- \ell^+ b\bar{b}$ decays as signal processes. Nevertheless WH events enter the analyses as well since they mimic $(Z \to \nu\bar{\nu})H$ events in case the W boson decays into a hadronically decaying τ-lepton or the lepton from the W boson is not reconstructed[3]. Two Feynman diagrams for ZH production are considered: quark induced production and gluon induced production as shown in figure 8.1. WH processes are only produced via quark induced processes. The calculation of the VH cross section considers up to NNLO QCD and NLO EW effects. The predicted values are: $\sigma(q\bar{q} \to ZH) = 0.76\,\text{pb}$, $\sigma(gg \to ZH) = 0.12\,\text{pb}$ and $\sigma(pp \to WH) = 1.37\,\text{pb}$. The BR for the $Z \to \nu\bar{\nu}$ decay is 20%, it is 6.7% for the $Z \to \ell^- \ell^+$ decay and 21.3% for the $W^\pm \to \ell^\pm \overset{(-)}{\nu}$ decay. The SM BR for the Higgs boson decay into a pair of bottom quarks is 58.2%. The common experimental signature are two jets from the Higgs boson decay. In addition, a significant amount of missing transverse energy or two reconstructed leptons are expected from the Z boson decay. The simulated SM VH data set used in this analysis is

[1] Results targeting $pp \to qqH$ and $pp \to ttH$ production are not sensitive yet and set limits on μ of approximately 4 and 2 times the SM expectation [68, 69].

[2] There are dedicated programs available to calculate cross sections. Those might not necessarily be able to simulate an event.

[3] This could either be the case because the lepton is out of the ATLAS detector's acceptance or because it simply does not meet all of the selection criteria for *VH tight* or *VH loose*. The latter often is the case if the lepton is inside a jet thus not meeting the isolation requirements. For simplicity, all these cases will be further referred to as "not reconstructed".

Process	σ in pb	cross section order
$q\bar{q} \to ZH$	0.76	NNLO(QCD)+NLO(EW)
$gg \to ZH$	0.12	NLO(QCD)
WH	1.37	NNLO(QCD)+NLO(EW)
Z	1 906	NNLO
$(Z \to \ell^-\ell^+)+\geq$ 2jets	54	Measurement
W	20 080	NNLO
$t\bar{t}$	831.76	NNLO
ZZ	22.89	NLO
WZ	77.33	NLO
WW	117.5	NLO
tq (t-channel)	216.97	NNLO
Wt (Wt channel)	71.7	NNLO
tb (s-channel)	10.32	NNLO

Decay	BR
$\sum_{\nu_e,\nu_\mu,\nu_\tau} Z \to \nu\bar{\nu}$	20.0%
$\sum_{e,\mu} Z \to \ell^-\ell^+$	6.7%
$H \to b\bar{b}$	58.2%
$\sum_{e,\mu} W^\pm \to \ell^\pm \overset{(-)}{\nu}$	21.3%
$Z \to q\bar{q}$	69.9% (all flavours)
$Z \to b\bar{b}$	15.1%
$W \to q\bar{q}'$	67.4%

Table 8.1: The production cross section of the signal and background processes as well as the BRs of Z, Higgs and W boson decays that are relevant for the analysis. The values for the Z and W boson decays stem from the combination of several measurements [7]. The values for the Higgs boson decays refer to the expected BR for a SM Higgs boson with a mass of 125 GeV [4].

simulated with the POWHEG generator [28] which has LO precision. There are separate simulations for the $q\bar{q} \to VH$ and $gg \to ZH$ production channel. The $q\bar{q} \to VH$ simulations make additional use of the GOSAM program package [70] and MINLO prescription [71] to include higher order virtual corrections and achieve NLO QCD precision [72]. The PS and UE are simulated with PYTHIA. These simulations provide additional event weights to emulate alternative μ_{R}, μ_{F} and variations of the PDF set and its assumed α_S. The simulations assume a SM Higgs boson with a mass of $m_H = 125\,\mathrm{GeV}$. The simulations explicitly include the V decay into leptons and the $H \to b\bar{b}$ decay. Alternative MC data sets are available that were simulated with a different ME generator or a different PS generator. Since the simulation does not include NLO EW effects which have a sizeable impact on the p_{T}^Z distribution a dedicated correction is applied to all simulated $V(H \to b\bar{b})$ events as a function of p_{T}^V [4].

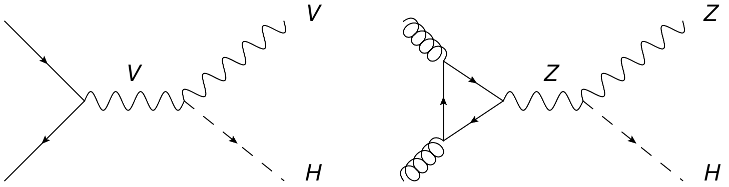


Figure 8.1: Feynman diagrams for the SM VH production; quark induced VH production on the left and gluon induced ZH production on the right.

V+jets: The production of a single vector boson with additional jets, as shown in figure 8.2 is one of the major components contributing to the background for the analyses presented here. These events enter the analysis phase space if the V boson decays into leptons. If it is either a $Z \to \nu\bar{\nu}$ or $Z \to \ell^-\ell^+$ decay with two additional b-quarks, the final state is exactly the same as for the signal processes. In this case, the only distinctions may be made based on the kinematic properties of the objects and events. Z+jets events also enter the analysis phase space even if the jets are not b-jets due to the non-negligible mis-identification probabilities of the utilised b-jet identification techniques. In addition, a fraction of W+jets events may fall into the analysis phase space if the W decays leptonically. In that case these events mimic a $(Z \to \nu\bar{\nu})H$ event if either the lepton is not reconstructed or the lepton is a hadronically decaying τ-lepton. The cross section of single V boson production is known at NNLO precision: $\sigma(pp \to Z) = 1\,906\,\text{pb}$ and $\sigma(pp \to W^{\pm}) = 20\,080\,\text{pb}$ [73]. Nevertheless most of these V bosons are produced without additional jets and therefore these events do not enter the phase space of the $Z(H \to b\bar{b})$ analysis. The cross section of events with a leptonically decaying Z boson $(Z \to \ell^{-}\ell^{+})$ with 2 or more additional jets was measured to be 54 pb [74]. All V+jets simulated events are produced with the Sherpa MC generator. The MEs for V+jets are generated at NLO precision for up to two additional partons and LO precision for up to four partons. In order to avoid overlap between different parton multiplicities a dedicated procedure for the merging of them is used, which introduces a merging and resummation scale. Sherpa provides event weights that allow to vary the factorisation, renormalisation, resummation and merging scale as well as variations for the PDF set. To ensure a sufficient amount of MC events in the kinematically suppressed phase space of high V boson p_{T} and jet p_{T} the simulated events are produced in so-called slices based on the V boson p_{T} and total scalar sum of the p_{T} of all objects in the event. Furthermore, the V+jets simulations are produced separately for V plus light jets, c-jets and b-jets to provide a sufficient amount of events with c-jets and b-jets which would be otherwise suppressed. Only leptonic V boson decays are simulated. Alternatively the V+jets simulated events are produced with the MadGraph5_aMC@NLO generator which achieves LO precision for this process.

Top quark pairs: Another large contributor to the background, is the production of top quark pairs ($t\bar{t}$), as shown in figure 8.2. The production cross section for $t\bar{t}$ events is known at NNLO precision: $\sigma(pp \to t\bar{t}) = 831.8\,\text{pb}$. A top quark decays into a bottom quark and a W boson with a BR of larger than 99%. If the two W bosons in the $t\bar{t}$ event subsequently decay as $W^{\pm} \to \ell^{\pm}\overset{(-)}{\nu}$ the events mimic the $(Z \to \ell^{-}\ell^{+})H$ events since the neutrinos cannot be reconstructed. Top quark pair events may also enter the $(Z \to \nu\bar{\nu})H$ phase space if one of the W bosons decays leptonically and the other one decays into two quarks. If the lepton is not reconstructed or it is a hadronically decaying τ-lepton it mimics a $(Z \to \nu\bar{\nu})H$ event with additional jets. The simulated $t\bar{t}$ events are generated at NLO precision with the Powheg MC generator. It is interfaced to Pythia for the description of the PS and UE. This simulation is also provided with alternative choices for various parameters that lead to more or less QCD radiation. Alternative MC data sets are available with a different ME generator as well as with a different PS model. The mass of the top quark is set to 172.5 GeV in all simulations.

***VV*:** The production of two weak bosons, i.e. ZZ, WZ or WW, is a sub-dominant background process. One example of a Feynman diagram for VV production is shown in figure 8.2. The cross sections are relatively low and are known at NLO precision: $\sigma(ZZ) = 22.9\,\text{pb}$, $\sigma(WZ) = 77.3\,\text{pb}$ and $\sigma(WW) = 117.5\,\text{pb}$. Nevertheless the ZZ events can produce the exact same final state as the signal events if one Z boson decays as $Z \to \ell^{-}\ell^{+}$ or $Z \to \nu\bar{\nu}$ and the other Z boson decays into a pair of bottom quarks (BR=15.1%). Even the invariant mass of the quark pair is close to the one expected from a Higgs boson decay ($m_Z = 91.2\,\text{GeV}$). The WZ events enter the signal phase space if the Z boson decays leptonically and the W boson decays into two quarks since the b-jet identification techniques have a non-negligible probability for mis-identifications. In the $(Z \to q\bar{q})(W^{\pm} \to \ell^{\pm}\overset{(-)}{\nu})$ decay it might happen that the lepton is not reconstructed and therefore these events enter the $(Z \to \nu\bar{\nu})H$ phase space. For the same reason, a small amount of $(W \to q\bar{q}')W^{\pm} \to \ell^{\pm}\overset{(-)}{\nu}$ events enter the $(Z \to \nu\bar{\nu})H$ phase space. The diboson process is simulated with the Sherpa generator. This simulation achieves NLO precision for VV with up to 1 additional parton. The production of 2 or 3 additional partons is included with LO precision. This simulation provides event weights to emulate alternative renormalisation and factorisation scale. Alternative simulations that use a different ME and PS generator are available as well.

Single top quarks: The production of single top quarks is a sub-dominant background process. It is useful to distinguish three different production channels, t-, Wt and s-channel, which are all shown in figure 8.2. These production channels lead to different final states. The production cross sections for these processes are calculated up to NNLO precision: $\sigma(t\text{-channel}) = 217.0\,\text{pb}$, $\sigma(Wt\text{-channel}) = 71.7\,\text{pb}$ and $\sigma(s\text{-channel}) = 10.3\,\text{pb}$. The t- and s-channel mostly contribute to the $(Z \to \nu\bar{\nu})H$ phase space in case the W boson decays leptonically and the lepton is not reconstructed. The Wt channel mostly enters the $(Z \to \ell^-\ell^+)H$ phase space in case both W bosons decay leptonically and an additional jet it produced. The production of single top quarks is simulated at NLO precision with the Powheg generator. The PS is simulated with Pythia. The three production channels are simulated separately. Similar to the simulation of $t\bar{t}$ events the simulations are provided with alternative choices for various parameters that lead to more or less QCD radiation. Simulations with an alternative ME or PS are available for the t-channel and Wt. In addition, the Wt simulation has to resolve interferences with $t\bar{t}$ production. The baseline procedure to do this is called diagram removal (DR). An alternative simulation that uses so-called diagram subtraction (DS) is provided. The top quark mass is set to 172.5 GeV in all simulations.

Multijet: The production of a multitude of jets is a very likely process in a hadron collider and it includes any process that produces a final state with n partons without the presence of any additional particle, e.g. Z boson, top quark, outside of the jets. Nevertheless the analyses presented here suppress the amount of multijet events to a negligible level which was confirmed by studies with simulated events and data-driven methods [75]. The suppression factor in the $ZH \to \ell^-\ell^+b\bar{b}$ decay channel is the requirement of two isolated leptons. An additional suppression factor in both channels is the requirement of heavy flavour jets (b-jets or c-jets) since the production of heavy quarks is suppressed with respect to light quarks. In multijet events the leptons usually originate from hadron decays inside the jet which typically produces non-isolated low p_T leptons. The visible signature of the $(Z \to \nu\bar{\nu})H$ channel is a multitude of jets and therefore an additional event selection has to be made to suppress multijet events. The selection criteria are detailed in section 8.3.

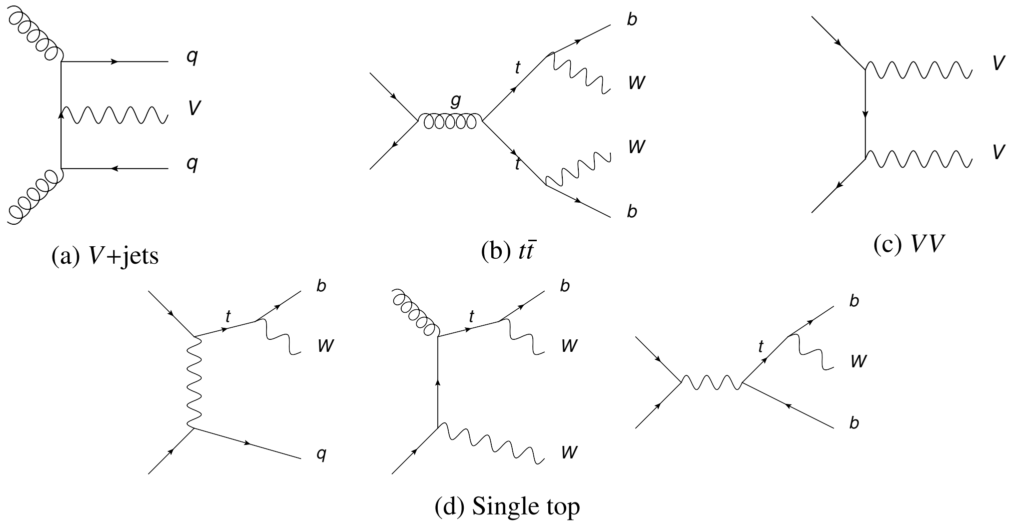
8.2 Object Selection

In order to reconstruct the desired ZH events identification of the respective decay products is crucial. Therefore several sets of requirements are defined for muons, electrons, hadronically decaying tau leptons and jets. All object selections are summarised in tab. 8.3.

Electrons: Electrons are reconstructed from clusters in the ECAL and identified as described in 5. Two sets of electron selection criteria are defined for the analysis described here: *VH loose* and *VH tight*. In order to fulfil the *VH loose* requirement electrons need to pass the *loose* quality criteria. To reject tracks from pile-up, the origin of the electron's track has to be consistent with the primary vertex by imposing cuts on the impact parameter significance $d_0^\text{sig.}$ and the distance of the electron's track to the primary vertex. To avoid that charged particles inside a jet are mis-identified as electrons

Process	PDF	ME	PS+UE	tune	variations
$q\bar{q} \to VH$	**PDF4LHC15NLO** [76]	**Powheg +GoSaM +MiNLO**	**Pythia**	**AZNLO**	μ_{F}, μ_{R}, PDF+α_S
	PDF4LHC15NLO	Powheg +GoSaM +MiNLO	Herwig	H7UE [26]	
	PDF4LHC15NLO	MadGraph5_aMC@NLO	Pythia	A14	μ_{F}, μ_{R}, PDF+α_S, A14
$gg \to ZH$	**PDF4LHC15NLO**	**Powheg**	**Pythia**	**AZNLO**	
V+jets	**NNPDF3.0NNLO** [77]	**Sherpa**	**Sherpa**	**Sherpa**	μ_{F}, μ_{R}, PDF
	CT10	Sherpa	Sherpa	Sherpa	μ_{F}, μ_{R}, resummation scale, merging scale
	NNPDF2.3LO	MadGraph5_aMC@NLO	Pythia	A14	
$t\bar{t}$	**NNPDF3.0NLO**	**Powheg**	**Pythia**	**A14**	high radiation, low radiation
	NNPDF3.0NNLO	Powheg	Herwig	H7UE	
	NNPDF3.0NNLO	MadGraph5_aMC@NLO	Pythia	A14	
singletop	**CT10**	**Powheg**	**Pythia**	**Perugia 2012**	high or low radiation, DS (Wt channel)
	CT10	Powheg	Herwig	UE-EE-5 [78]	
	CT10	MadGraph5_aMC@NLO	Herwig	UE-EE-5	
VV	**NNPDF3.0NNLO**	**Sherpa**	**Sherpa**	**Sherpa**	μ_{F}, μ_{R}, PDF
	CT10	Powheg	Pythia	AZNLO	PS tune
	CT10	Powheg	Herwig	UE-EE	

Table 8.2: Monte Carlo generators and their parameters utilised in the SM ZH analyses in this thesis. The nominal MC generator recommendations for ATLAS analyses of 2015+2016 data are given in bold letters for each physics process. All other MC generators provide alternative models and are used to assess systematic uncertainties. Details about available reweighting schemes for certain generators to assess the effect of alternative set-ups are given in the column "variations".



(a) V+jets

(b) $t\bar{t}$

(c) VV

(d) Single top

Figure 8.2: Feynman diagrams for the production of SM processes that contribute as background events to the ZH analysis phase space: a) weak boson production with two additional quarks, b) production of top quark pairs, c) production of weak boson pairs, d) the production of single top quarks in the t-, Wt and s-channel (from left to right).

a *loose* isolation of the electron's track is required (details are given in sec. 5). In addition, the electron has to have a p_{T} of at least 7 GeV and a pseudorapitidity of $|\eta| < 2.47$. In order to pass the *VH tight* selection the electron need to pass the *VH loose* requirements and its transverse momentum needs to be at least 27 GeV.

Muons: Muons are reconstructed from tracks in the tracking system and/or in the muon spectrometer and deposited energy in the calorimeter and identified as described in chapter 5. In accordance with the electron selection criteria, *VH loose* and *VH tight* muon selection criteria are defined. *VH loose* muons have to pass the *loose* muon quality requirement. Similar to the electron selection, selection criteria are imposed on the impact parameter significance of the muon's track and the distance of the muon's track to the primary vertex to reject pile-up and cosmic muons. The muon also needs to pass a *loose* track isolation requirement. The transverse momentum of the muon has to be larger than 7 GeV and its pseudorapidity has to fulfill $|\eta| < 2.7$. For the *VH tight* selection the muon needs to pass the *VH loose* selection criteria and additionally $|\eta| < 2.5$ and $p_{\mathrm{T}} < 27$ GeV.

Jets: Jets are reconstructed from clusters in the calorimeter system using the anti-k_{T} algorithm with a radius parameter of $R = 0.4$ as described in 5. Two sets of selection criteria are defined: *signal* jets and b-jets. Independently these selection criteria all jets have to pass the jet cleaning procedure which removes calorimeter noise or non-collision background that was mis-identified as a jet [60, 79]. A jet is defined as a *signal* jet if its p_{T} is larger than 20 GeV and $|\eta| < 2.5$. If the p_{T} of the jet is in between 20 GeV and 60 GeV and $|\eta| < 2.4$ it has to pass an additional Jet Vertex Tagger (JVT) requirement in order to be classified as a *signal* jet. The Jet Vertex Tagger combines track and vertex information in a likelihood discriminant to suppress jets that originate from pile-up events [62]. In order to define a Higgs candidate from a $H \to b\bar{b}$ decay the resulting jets have to be identified as b-jets. To classify a jet as a b-jet, multivariate tagging algorithms are utilised which are described in section 5.5.1. A *signal* jet is considered to be a b-jet if it passes the threshold of the MV2c10 algorithm that corresponds to an identification efficiency of 70%.

τ-leptons: In the decay channels considered for the $Z(H \rightarrow b\bar{b})$ analysis τ-leptons are not included in the final state. However if a τ-leptons decays into leptons and the resulting electron/muon passes the aforementioned selection criteria it may be included in the analysis. The decay of τ-leptons into hadrons, which have a much larger BR, have similar signatures as jets and are reconstructed as described in chapter 5. To remove jets from the analysis that are actually hadronic τ-leptons the following τ-identification criteria are defined: fulfil the *medium* hadronic τ-lepton identification criteria, have exactly 1 or 3 associated tracks, their p_{T} is larger than 20 GeV and $|\eta| < 2.5$. The latter excludes the transition region between the barrel and end-cap calorimeters $(1.37 < |\eta| < 1.52)$.

Since object reconstruction and identification is not always unambiguous, e.g. an electron may also be reconstructed as a jet, a dedicated — so-called overlap removal — procedure is applied to remove ambiguous objects from the event. It involves several sequential stages to remove overlap between muons, electrons, jets and hadronic τ-leptons based on criteria involving the distance ΔR between them in combination with their p_{T} and quality. The SM $Z(H \rightarrow b\bar{b})$ analysis involves additional steps to remove overlap between hadronically decaying tau leptons and muons, electrons and jets.

Lepton category	p_{T}	$\lvert\eta\rvert$	quality	$d_0^{\text{sig.}}$	$\lvert\Delta z_0 \sin\theta\rvert$	track isolation
VH loose electron	> 7 GeV	< 2.47	*loose*	< 5	< 0.5 mm	*loose*
VH tight electron	> 27 GeV		same as for *VH loose* electron			
VH loose muon	> 7 GeV	< 2.7	*loose*	< 3	< 0.5 mm	*loose*
VH tight muon	> 27 GeV	< 2.5	same as for *VH loose* muon			

Jet category	quality	p_{T}	$\lvert\eta\rvert$	JVT requirement
forward	jet cleaning	> 30 GeV	$2.5 < \lvert\eta\rvert < 4.5$	-
signal	jet cleaning	> 20 GeV	< 2.5	if $p_{\mathrm{T}} < 60$ GeV and $\lvert\eta\rvert < 2.4$
b-jets	same as *signal* + *b*-tagging at 70% b-jet eff. (MV2c10)			

Tau lepton category	quality	p_{T}	$\lvert\eta\rvert$	number of tracks
hadronic tau	*medium*	> 20 GeV	< 2.5 (excluding $1.37 \leq \lvert\eta\rvert < 1.52$)	1 or 3

Table 8.3: Selection criteria for the lepton, jet and tau categories defined for the $Z(H \rightarrow b\bar{b})$ analysis.

8.3 Event Selection

In order to suppress the amount of expected background events, i.e. increase the relative amount of signal events, several requirements are made on the kinematics of the final states. The selection of events is the same as for the standard ATLAS $V(H \rightarrow b\bar{b})$ analysis and the details are explained in the following [75]. The desired events should contain a $H \rightarrow b\bar{b}$ decay. Therefore the presence of at least 2 *signal* jets is required. Exactly two of these *signal* jets have to be identified as *b*-jets in accordance with the requirements defined in tab. 8.3. *Signal* jets in each event are ordered in the following way (and may be referred to as leading (b_1), sub-leading (b_2) and third jet (j_3)): *b*-jet with higher p_{T} amongst the two

b-jets, b-jet with lower p_{T}, *signal* jet with the highest p_{T} amongst the non-b-jet *signal* jets. At least one of the b-jets has to have a p_{T} of more than 45 GeV. This requirement is motivated by the mass of the Higgs boson of 125 GeV and suppresses background processes that typically produce low p_{T} jets, e.g. multijet and V+jets production. The Higgs candidate in each event is reconstructed from these two b-jets. Due to different background compositions in the $ZH \rightarrow \nu\bar{\nu}b\bar{b}$ and $ZH \rightarrow \ell^{-}\ell^{+}b\bar{b}$ final states, the analysis is performed separately in these two channels. They are referred as 0 lepton ($ZH \rightarrow \nu\bar{\nu}b\bar{b}$ final state) and 2 lepton ($ZH \rightarrow \ell^{-}\ell^{+}b\bar{b}$ final state) channels in the following based on the number of visible leptons. The 0 and 2 lepton specific event selection criteria are described below and are summarised in tab. 8.4.

Events that enter the 0 lepton channel have to pass a $E_{\mathrm{T}}^{\mathrm{miss}}$ trigger. The trigger threshold is $E_{\mathrm{T}}^{\mathrm{miss}} >$ 70 GeV for 2015 data and it was raised to $E_{\mathrm{T}}^{\mathrm{miss}} > 90$ GeV and eventually $E_{\mathrm{T}}^{\mathrm{miss}} > 110$ GeV for 2016 data. Due to this threshold a lower limit of $E_{\mathrm{T}}^{\mathrm{miss}}$ >150 GeV is required to ensure a reasonable trigger efficiency of at least 85%. The $E_{\mathrm{T}}^{\mathrm{miss}}$ trigger becomes fully efficient at 180 GeV. In the transition region differences in the trigger efficiency are observable between data and simulation which are corrected scale factors as a function of $E_{\mathrm{T}}^{\mathrm{miss}}$. Additional systematic uncertainties are assigned for these scale factors. The $E_{\mathrm{T}}^{\mathrm{miss}}$ of the desired signal events originates from the $Z \rightarrow \nu\bar{\nu}$ decay. Therefore the $E_{\mathrm{T}}^{\mathrm{miss}}$ measurement represents the Z boson candidate. Furthermore, the 0 lepton selection requires exactly 0 *VH loose* and *VH tight* leptons. Only events are considered that contain either 2 or 3 *signal* jets. This requirement also allows for signal events with an additional jet from initial or final state radiation to pass the selection. The third jet is not allowed to be a b-jet. The selection of events with more than 3 jets would introduce a large amount of additional background events, mainly $t\bar{t}$, with only a small gain in signal acceptance. A requirement is imposed on the scalar sum of the 2(3) *signal* jets in the event of at least 120 GeV(150 GeV) motivated by the $E_{\mathrm{T}}^{\mathrm{miss}}$ threshold. The expected spatial distribution of the objects in the 0 lepton final state is further exploited to suppress multijet background to a negligible level. In multijet events the $E_{\mathrm{T}}^{\mathrm{miss}}$ usually originates from leptonic hadron decays inside the jets or inaccuracies in the jet energy measurement. In contrast, in $ZH \rightarrow \nu\bar{\nu}b\bar{b}$ events the $E_{\mathrm{T}}^{\mathrm{miss}}$ originates from the Z boson decay and the jets from the Higgs boson decay. A set of "multijet suppression cuts", are defined to suppress the multijet events to a negligible level [75]:

- $\Delta\phi(E_{\mathrm{T}}^{\mathrm{miss}}, \text{Higgs candidate}) > 120^{\circ}$
- $\Delta\phi(\text{Higgs candidate jets}) < 140^{\circ}$
- $\Delta\phi(E_{\mathrm{T}}^{\mathrm{miss}}$, nearest *signal* jet) $> 20^{\circ}(30^{\circ})$ for events with 2(3) *signal* jets
- $\Delta\phi(E_{\mathrm{T}}^{\mathrm{miss}}, p_{\mathrm{T}}^{\mathrm{miss}}) < 90^{\circ}$ with the missing transverse momentum based on tracks p_{T} instead of calorimeter information

The 2 lepton channel utilises single lepton triggers. The minimum p_{T} threshold for the single electron triggers is 24 GeV for 2015 data and 26 GeV for 2016 data. The minimum p_{T} threshold for the single muon triggers is 20 GeV for 2015 data and it was gradually increased from 24 GeV to 26 GeV for 2016 data. The single lepton triggers imply minimum quality and isolation criteria on the electron and muon respectively. The p_{T} thresholds of the *VH tight* lepton definitions are given by the single lepton trigger threshold. The 2 lepton selection requires exactly two *VH loose* leptons of the same flavour — $e^{-}e^{+}$ or $\mu^{-}\mu^{+}$ — from the $Z \rightarrow \ell^{-}\ell^{+}$ decay. At least one of the leptons has to pass the *VH tight* lepton requirements and has to coincide with the object that activated the single lepton trigger. If the lepton pair is a muon pair the two muons are required to be of opposite charge. This requirement is not imposed on electrons since they have a non-negligible charge mis-identification probability [80]. The invariant mass of the two leptons $m_{\ell^{-}\ell^{+}}$ has to be within 81 GeV and 101 GeV to be in agreement with the Z boson mass.

These requirements suppress background processes that contain two leptons that do not originate from a Z boson decay, such as $t\bar{t}$ and multi-jet background. The Z boson candidate is reconstructed from the two leptons in the event and its transverse momentum p_T^Z has to be larger than 75 GeV. This is motivated by the overwhelming amount of background events at low p_T^Z and observed difficulties to model this region of the phase space in the MC simulation. The 2 lepton channel also takes into account events with 3 or more *signal* jets as long as none of the additional jets is identified as a b-jet. In contrast to the 0 lepton channel allowing more than 3 *signal* jets leads to a substantial gain in the sensitivity.

Common selection
≥ 2 *signal* jets 2 b-jets ≥ 1 b-jet with $p_\mathrm{T} > 45\,\mathrm{GeV}$

0 lepton selection	2 lepton selection
0 *VH loose* leptons	2 *VH loose* leptons
0 *VH tight* leptons	≥ 1 *VH tight* lepton
$E_\mathrm{T}^\mathrm{miss} > 150\,\mathrm{GeV}$	$p_\mathrm{T}^Z > 75\,\mathrm{GeV}$
2 or 3 *signal* jets	$81\,\mathrm{GeV} < m_{\ell^-\ell^+} < 101\,\mathrm{GeV}$
$\sum_{i=1}^{N_\mathrm{jet}=2(3)} p_\mathrm{T}^i > 120\,\mathrm{GeV}(150\,\mathrm{GeV})$	
multijet suppression cuts	

Table 8.4: Event selection for the $Z(H \to b\bar{b})$ analysis. Numbers in brackets for the 0 lepton selection indicate the requirements if there are 3 instead 2 *signal* jets present in the event.

The amount of expected and accepted signal events are given in tab. 8.5 which also lists the cross section × BR for the relevant signal processes. In total 2.1% of all expected $ZH \to \nu\bar{\nu}b\bar{b}$ and 8.2% of all expected $ZH \to \ell^-\ell^+b\bar{b}$ events pass the event selection criteria of the 0 lepton and 2 lepton analysis phase space respectively[4]. More than half of all expected signal events are not reconstructed due to limited coverage of the detector as well as trigger and object reconstruction efficiencies. The amount of signal events is further reduced due to the p_T^Z thresholds of 150 GeV(75 GeV) in the 0(2) lepton channel. This affects especially the amount of accepted $ZH \to \nu\bar{\nu}b\bar{b}$ events where the p_T^Z threshold is given by the $E_\mathrm{T}^\mathrm{miss}$ trigger threshold. Since the Z bosons have on average a higher p_T in the $gg \to ZH$ production channel compared to the $q\bar{q} \to ZH$ production channel the acceptance of these events is approximately 2 times higher, in the 0 and 2 lepton channel. Nevertheless the production cross section is 6 times lower for the $gg \to ZH$ production. Another limiting factor for the acceptance of signal events is the identification efficiency of the b-tagging algorithms. The 0 and 2 lepton selection both require two b-jets which corresponds to an efficiency in the order of 50%. In addition to the $Z(H \to b\bar{b})$ events, 0.2% of all expected $W^\pm H \to \ell^\pm \overset{(-)}{\nu} b\bar{b}$ events are accepted in the 0 lepton phase space. These are mostly events where the W boson decays into a tau lepton (and corresponding neutrino) which further decays into hadrons. In the following the WH events will be considered as signal events as well. Although the analysis targets the $Z(H \to b\bar{b})$ processes the ultimate goal of the analysis is to search for the $H \to b\bar{b}$ decay and it is not possible to disentangle the ZH and WH events. The amount of $ZH \to \ell^-\ell^+b\bar{b}$ events accepted in the 0 lepton phase space and of $ZH \to \nu\bar{\nu}b\bar{b}$ and $W^\pm H \to \ell^\pm \overset{(-)}{\nu} b\bar{b}$ events in the 2 lepton phase space is much smaller than 1 event each and therefore negligible.

[4] Assuming SM cross sections and branching ratios.

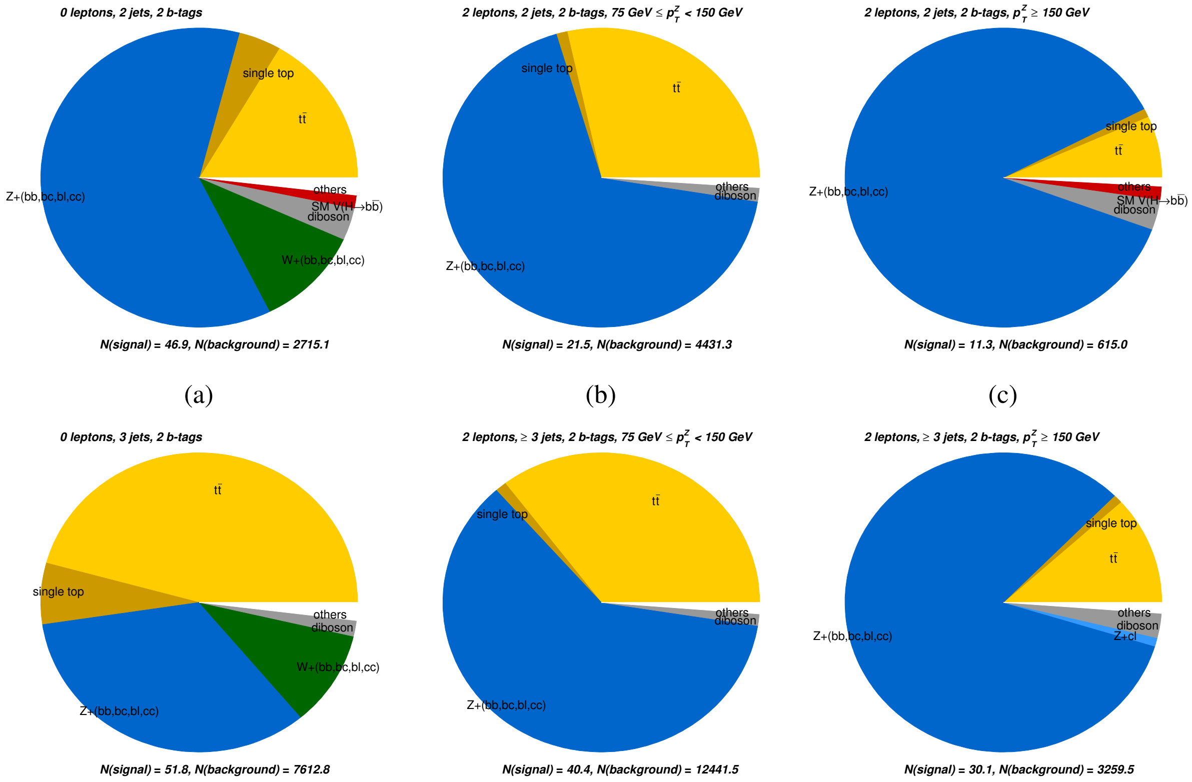
Process	Cross section × BR	Expected N(events)	Selected N(events)	Acceptance
		0 lepton selection		
$ZH \rightarrow \nu\bar{\nu}b\bar{b}$	103.4 fb	3732.7	79.2	2.1%
$W^{\pm}H \rightarrow \ell^{\pm}\overset{(-)}{\nu}b\bar{b}$	269.0 fb	9710.9	18.8	0.2%
		2 lepton selection		
$ZH \rightarrow \ell^{-}\ell^{+}b\bar{b}$	34.7 fb	1252.7	103.0	8.2%

Table 8.5: The cross section × BR — as predicted by the SM — for the considered signal processes, as well as the amount of expected and selected number of events (N(events)) and the overall acceptance. The expected amount of events are calculated using an integrated luminosity of 36.1 fb^{-1} which corresponds to the 2015+2016 ATLAS data set. The selected amount of events were determined from the amount of simulated signal events that pass the selection criteria. The acceptance is calculated as the fraction of the number of selected to expected events.

The events that pass the outlined selection criteria are split into 6 categories. The categories are defined to maximize the sensitivity to different background components which will be exploited by a multivariate analysis (MVA), see section 8.5. The following categories are defined:

- **0 lepton, 2 jets**: events in the 0 lepton channel with exactly 2 *signal* jets which have to be b-jets
- **0 lepton, 3 jets**: events in the 0 lepton channel with exactly 3 *signal* jets; exactly two of the *signal* jets have to be b-jets
- **2 lepton, 2 jets, medium p_{T}^{Z}**: events in the 2 lepton channel with exactly 2 *signal* jets which have to be b-jets; p_{T}^{Z} has to be in between 75 GeV and 150 GeV
- **2 lepton, 3+ jets, medium p_{T}^{Z}**: events in the 2 lepton channel with 3 or more *signal* jets; exactly two of the *signal* jets have to be b-jets; p_{T}^{Z} has to be in between 75 GeV and 150 GeV
- **2 lepton, 2 jets, high p_{T}^{Z}**: events in the 2 lepton channel with exactly 2 *signal* jets which have to be b-jets; p_{T}^{Z} has to be larger than 150 GeV
- **2 lepton, 3+ jets, high p_{T}^{Z}**: events in the 2 lepton channel with 3 or more *signal* jets; exactly two of the *signal* jets have to be b-jets; p_{T}^{Z} has to be larger than 150 GeV

The composition of background and signal processes in these categories is shown in figure 8.3. The categories with the highest relative amount of signal events are the 0 lepton 2 jets category and the 2 lepton 2 jets high p_{T}^{Z} category. In those two categories approximately 1.5% of all events are $Z(H \rightarrow b\bar{b})$ events. The main contributors in all regions are Z+jets and $t\bar{t}$ production with $t\bar{t}$ being enhanced in the 3 (3+) jet categories. The 2 jet high p_{T}^{Z} categories are dominated by Z+jets events and $t\bar{t}$ events only contribute of the order of 10%. Single top production is a sub-dominant contributor to the 2 lepton channel with less than 2%. The 0 lepton channel compositions are more diverse with W+jets contributing approximately 13% and a larger relative amount of single top events. Diboson events are a sub-dominant contributor in all categories varying between 1.5% and 4.5%. Nevertheless these events offer a good opportunity to validate the $H \rightarrow b\bar{b}$ analysis strategy since they also contain a $b\bar{b}$ resonance with a similar mass to the Higgs boson mass. The details of this validation are discussed in section 8.8.2.



(a) (b) (c)

(d) (e) (f)

Figure 8.3: The signal and background processes that contribute to the a) 0 lepton 2 jets, b) 2 lepton 2 jets medium p_T^Z, c) 2 lepton 2 jets high p_T^Z, d) 0 lepton 3 jets, e) 2 lepton 3+ jets medium p_T^Z and f) 2 lepton 3+ jets high p_T^Z category. Shown are the relative contributions for each analysis category. Components that are smaller than 1% are grouped together in "others" (white).

8.4 Improvement of $m_{b\bar{b}}$ Resolution

The invariant mass distribution of the di-b-jet system is the most powerful variable of the $Z(H \to b\bar{b})$ analysis. It has the largest separation power between the signal and the background processes since a peak around the Higgs boson mass is expected for signal events. For all background processes, except for diboson production, a continuously falling $m_{b\bar{b}}$ spectrum is expected since the two b-jet do not originate from the decay of a resonance. As discussed in chapter 7, the $m_{b\bar{b}}$ resolution[5] is deteriorated by the inaccuracies in the b-jet energy measurements. To improve the $m_{b\bar{b}}$ resolution several b-jet energy corrections are explored for this analysis. To avoid differences in acceptance of signal and background events and corresponding uncertainties introduced by the chosen jet energy correction all corrections are applied after the object and event selections are performed. The jet energy corrections are only applied to the two b-jets in the event.

Muon-in-jet correction: The muon-in-jet correction corrects for semi-muonic b-hadron decays. This is only partially accounted for in the jet energy measurement since the muon only deposits a part of its energy in the calorimeter as described in chapter 7. If a muon that passes the *medium* quality

[5] The $m_{b\bar{b}}$ resolution is measured as the width divided by the peak position of the $m_{b\bar{b}}$ distribution.

criteria and has a p_{T} larger than 5 GeV is found within a jet (ΔR(jet,muon) < 0.4), the momentum of the muon is added to the momentum of the b-jet. Therefore the muon in jet correction corrects the energy and direction of the b-jet. No correction is applied in case no muon is found within the jet.

Resolution correction (*PtReco*): The *PtReco* correction accounts for effects beyond the semi-leptonic decays by applying $p_{\mathrm{T}}^{\mathrm{jet}}$ dependent scale factors. Scale factors are derived separately for jets with a muon inside, jets with an electron inside and jets without any muon or electron inside. The definition of a muon inside a jet is the same as for the muon in jet correction. Electrons inside a jet are defined as electrons that pass the *loose* electron quality criteria, have a p_{T} of larger than 5 GeV and are found within ΔR(jet,electron) < 0.4. The scale factors are defined as $p_{\mathrm{T}}^{\mathrm{true}}/p_{\mathrm{T}}^{\mathrm{jet}}$ and calculated using SM $ZH \to \ell^{-}\ell^{+}b\bar{b}$ events. The true jet momentum, $p_{\mathrm{T}}^{\mathrm{true}}$, is defined as the p_{T} of the truth jet[6]. The application of the muon-in-jet and *PtReco* in a sequence yields a better $m_{b\bar{b}}$ resolution than the usage of *PtReco* as a stand-alone-correction. Therefore $p_{\mathrm{T}}^{\mathrm{jet}}$ is defined as the p_{T} of the reconstructed jet after the muon in jet correction is applied.

Kinematic fit (2 lepton only): The kinematic fit is developed for events with two leptons and exactly two b-jets assuming that the di-lepton and the di-b-jet systems are balanced in p_{T}. The momenta of the leptons and b-jets are varied such that their varied values maximise a likelihood function. The likelihood function implements certain constraints given by the expected event kinematics within which the momenta are allowed to vary:

- $m_{\ell^{-}\ell^{+}}$ may vary within the Breit-Wigner distribution given by the Z boson mass as its mean and the Z decay width as its width
- p_x and p_y of the $\ell^{-}\ell^{+}b\bar{b}$ system may vary within a Gaussian distribution with its mean at 0 GeV and a width of 9 GeV. The mean value is motivated by the assumption that the event should be balanced. The width is determined by the width observed in simulated ZH events based on MC truth information.
- The momenta and angles of the b-jets and leptons may vary within a Gaussian distribution with its width given by the measurement uncertainties of these variables.
- An additional prior is set for the energy of the b-jets given by $p_{\mathrm{T}}^{\mathrm{true}}/p_{\mathrm{T}}^{\mathrm{jet}}$ to account for the fact that the reconstructed b-jet energy is on average underestimated. This ratio is defined in the same way as for the *PtReco* correction.

Since the lepton energies are measured with a very good precision the kinematic fit mainly corrects the b-jets' momenta. The kinematic fit uses b-jets after they were corrected by the muon in jet correction. Therefore it always has to be applied in combination with the muon-in-jet correction. The kinematic fit does not yield a significant improvement in the 0 lepton channel since the information of the single leptons cannot be disentangled from the $E_{\mathrm{T}}^{\mathrm{miss}}$ measurement and the information of the leptons' momenta in z-direction is missing. This explains why the kinematic fit is only used in the 2 lepton channel.

Jet energy regression: The jet energy regression is a multivariate jet energy correction that was developed in the context of this thesis and is described in chapter 7. It corrects several effects including semi-leptonic b-hadron decays. It is a stand-alone correction and is applied to calibrated b-jets. It is an alternative to the aforementioned corrections[7].

[6] Using the same truth jets and following the association procedure to the reconstructed jet as described in sec. 7.1.

[7] A combination of the jet energy regression and the kinematic fit might yield an additional improvement of the $m_{b\bar{b}}$ resolution.

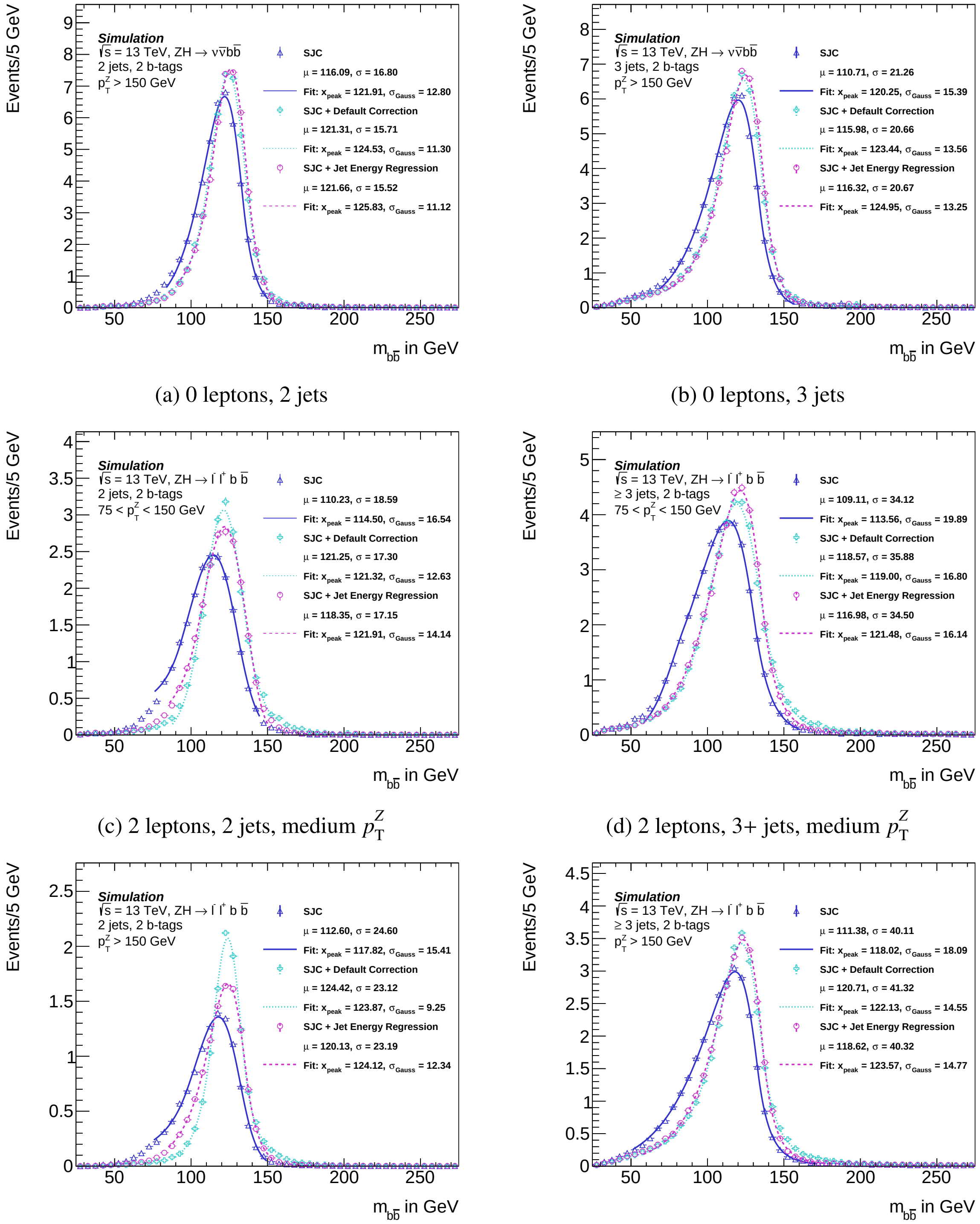
The baseline correction for this analysis is the jet energy regression, which has been developed in the context of this thesis. It is compared to the corrections used in the standard ATLAS $V(H \rightarrow b\bar{b})$ analysis [75] that was performed with LHC data taken in 2015 and 2016. The corrections utilised in that analysis vary based on the lepton channel and the number of *signal* jets. This procedure of corrections is summarised in table 8.6 and will be referred to as *default* correction in the following. As stated in this table the muon-in-jet correction is used for all events in the analysis phase space. In addition, the 0 lepton channel uses the *PtReco* correction. The 2 lepton channel uses the the kinematic fit for events with 2 and 3 *signal* jets. For events with 3 *signal* jets still only the information of the two leptons and the two b-jets is used. Although the third jet disturbs the balance of the di-lepton and di-bjet systems a gain in the $m_{b\bar{b}}$ resolution is still observable. This is not the case for events with 4 or more jets. Therefore in those events the kinematic fit is replaced by the *PtReco* correction.

Phase space	muon-in-jet	*PtReco*	kinematic fit	jet energy regression
Standard $V(H \rightarrow b\bar{b})$ Analysis				
0 leptons, 2 or 3 jets	✓	✓	-	-
2 leptons, 2 jets	✓	-	✓	-
2 leptons, 3 jets	✓	-	✓	-
2 leptons, $\geq$ 4 jets	✓	✓	-	-
Baseline for this thesis				
full phase space	-	-	-	✓

Table 8.6: The corrections used in the standard ATLAS $V(H \rightarrow b\bar{b})$ analyis [75]. In all phase spaces, exactly two b-jets are required and any additional jet has to be classified as *signal* jets as described in section 8.2. In the case of 2 lepton events with 3 jets the kinematic fit does not use the information of the third jet.

Figure 8.4 shows a comparison of the $m_{b\bar{b}}$ distributions without any additional corrections, the *default* correction and the jet energy regression for the $Z(H \rightarrow b\bar{b})$ signal processes. They are shown for all analysis categories, i.e. different jet multiplicities and additionally in the 2 lepton channel for different p_{T}^{Z} regimes. The $m_{b\bar{b}}$ resolution is improved by the application of a jet energy correction in all categories. In the 0 lepton channel the *default* correction and jet energy regression have a similar performance. In the 2 lepton channel the *default* correction yields a better $m_{b\bar{b}}$ resolution than the jet energy regression in the region with exactly 2 b-jets and no additional jets. The difference is especially large in events with a high p_{T}^{Z} (larger than 150 GeV). This is the result of the kinematic fit which makes use of the full event information in contrast to the jet energy regression which only uses information about single jets. In the 2 lepton channel in events with additional *signal* jets (3+ jets) the performance of the *default* correction and jet energy regression is similar with the jet energy regression yielding a better $m_{b\bar{b}}$ resolution in events with $75\,\mathrm{GeV} < p_{\mathrm{T}}^{Z} < 150\,\mathrm{GeV}$. In general, the jet energy regression has advantages with respect to the *default* correction in regions of the phase space that intrinsically have a worse $m_{b\bar{b}}$ resolution, i.e. higher jet multiplicities and lower p_{T}^{Z}. Overall, compared to the case with no jet energy correction, the jet energy regression improves the $m_{b\bar{b}}$ resolution by 15% to 25% depending on the analysis category.

It was not tested in the context of this thesis since it would require to re-derive the constraints for the kinematic fit after the application of the jet energy regression and to re-process all simulated and data events.



(a) 0 leptons, 2 jets

(b) 0 leptons, 3 jets

(c) 2 leptons, 2 jets, medium p_{T}^{Z}

(d) 2 leptons, 3+ jets, medium p_{T}^{Z}

(e) 2 leptons, 2 jets, high p_{T}^{Z}

(f) 2 leptons, 3+ jets, high p_{T}^{Z}

Figure 8.4: Invariant di-b-jet mass, $m_{b\bar{b}}$, distributions after the standard ATLAS jet calibration (SJC) (blue), after the b-jets were corrected with the *default* correction (cyan) and after the b-jets were corrected with the jet energy regression (pink) for different analysis phase spaces. For all distributions the mean μ and standard deviation σ is given as well as the peak position x_{peak} and width of the Gaussian core σ_{Gauss} as determined by fitting a Bukin function to the distributions.

8.5 Multivariate Analysis

To enhance the sensitivity of the analysis a multivariate analysis (MVA) is performed. It utilises BDTs to classify events as background or signal, as described in 6.1.2. Only events that passed the event selection, which is described in section 8.3, are considered in the multivariate analysis (MVA). There is a dedicated training of the BDTs for each of the previously defined 6 analysis categories. The input variables for the MVA were chosen based on their separation power between signal and background. They were optimised for the run 1 $V(H \to b\bar{b})$ analysis. The same variables are used for this analysis as well since the expected kinematic features, such as transverse momenta, change consistently in the signal and background processes. Therefore the features that provide separation power are conserved. The following input variables are used in all analysis categories:

- $m_{b\bar{b}}$: invariant mass of the two b-jets
- $\Delta R(b_1, b_2)$: distance in η and ϕ between the two b-jets
- $p_{\mathrm{T}}^{b_1}$: transverse momentum of the b-jet with the higher p_{T}
- $p_{\mathrm{T}}^{b_2}$: transverse momentum of the b-jet with the lower p_{T}
- p_{T}^{Z}: given by $E_{\mathrm{T}}^{\mathrm{miss}}$ in the 0 lepton channel; vectorial sum of the transverse momenta of the two leptons in the 2 lepton channel
- $\Delta\phi(Z, H)$: distance in ϕ between the Z boson candidate, i.e. $E_{\mathrm{T}}^{\mathrm{miss}}$ in the 0 lepton channel and the di-lepton system in the 2 lepton channel, and the Higgs boson candidate, i.e. the di-b-jet system

The 0 lepton channel utilises two additional variables:

- $|\Delta\eta(b_1, b_2)|$: distance in η between the two b-jets
- H_{T}: scalar sum of $E_{\mathrm{T}}^{\mathrm{miss}}$ and the p_{T} of all *signal* jets present in the event

The 2 lepton channels adds three more variables to the common variables:

- $E_{\mathrm{T}}^{\mathrm{miss}}$: missing transverse energy
- $|\Delta\eta(Z, H)|$: distance in η between the dilepton and di-b-jet system
- $m_{\ell^-\ell^+}$: invariant mass of the dilepton system

The trainings for the 0 lepton 3 jet category and the 2 lepton 3+ jets categories use two more variables that provide information about the third jet:

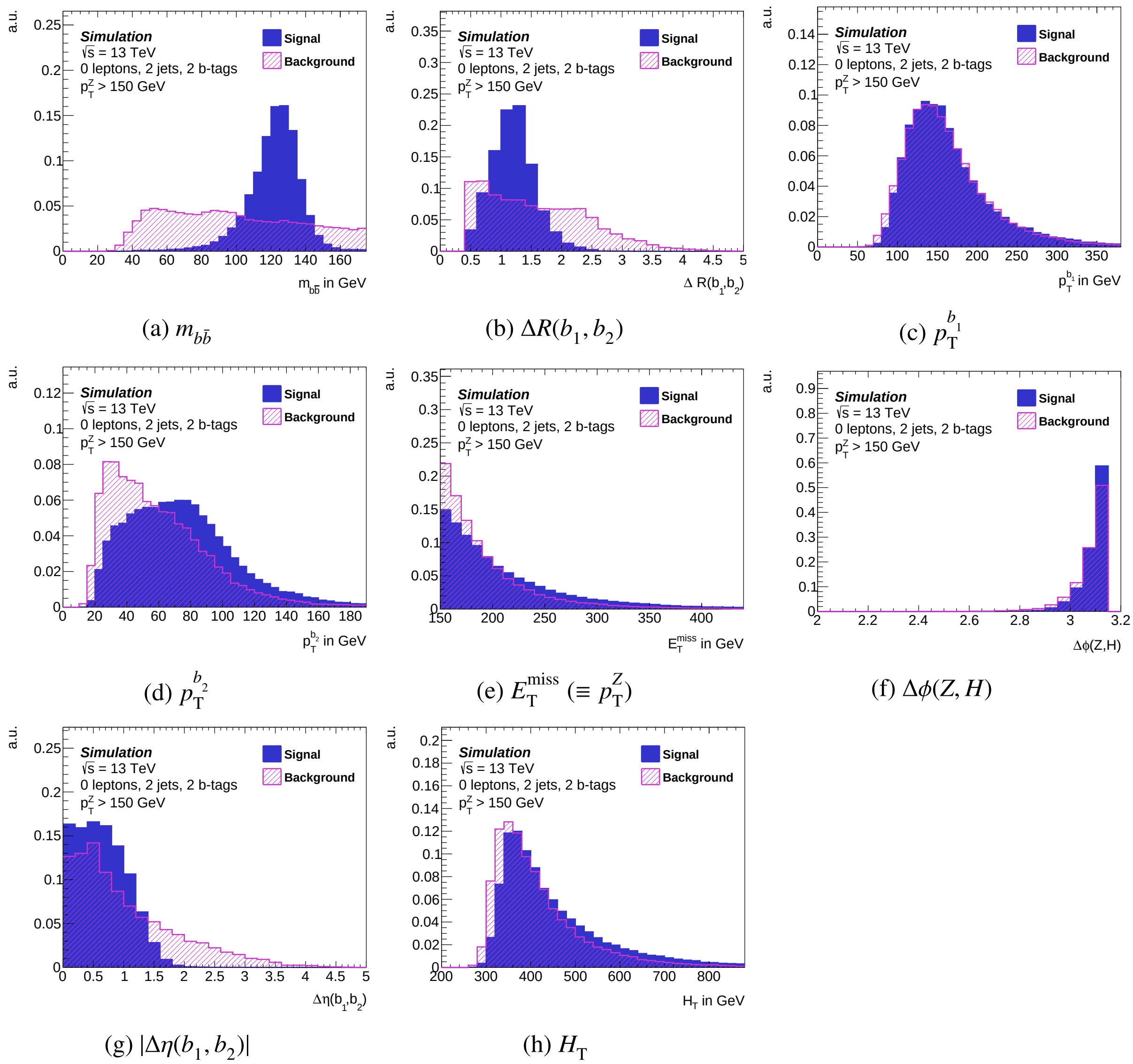
- $p_{\mathrm{T}}^{j_3}$: transverse momentum of the third *signal* jet in the event
- $m_{b\bar{b}j}$: invariant mass of the two b-jets and the third *signal* jet in the event

The input variables that use the information of the two b-jets exploit the correlation of the two b-jets and the constraints on their kinematics given by the Higgs boson decay. These correlations do not exist in the background processes since the b-jets do not originate from the decay of a resonance. The only exception is the diboson process which is a sub-dominant contributor to the overall background. As visible in figures 8.5 and 8.6 the invariant mass of the di-b-jet system, $m_{b\bar{b}}$, and the distance in η and ϕ between the two b-jets, $\Delta R(b_1, b_2)$, have the largest separation power between signal and background. Therefore a good $m_{b\bar{b}}$ resolution is crucial to increase the separation power of this variable. Additional separation power is provided by p_{T}^{Z} since it allows the BDTs to make use of the correlation between the Z boson and Higgs boson kinematics. Especially $\Delta R(b_1, b_2)$ has a strong correlation with p_{T}^{Z} for the signal process: $\Delta R(b_1, b_2)$ becomes smaller with increasing p_{T}^{Z}. An additional gain in sensitivity is achieved by including information about the whole ZH system such as $\Delta\phi(Z, H)$, which is large for the signal since the ZH system is balanced. The 2 lepton channels includes variables to specifically increase the separation power with respect to $t\bar{t}$ events. This is achieved by including $m_{\ell^-\ell^+}$ and $E_{\mathrm{T}}^{\mathrm{miss}}$ since, in contrast to the signal process, there is no Z boson but neutrinos present in the $t\bar{t}$ events that are included in the 2 lepton phase space. The signal and background distributions for all analysis categories and comparisons between data events and simulated MC events are given in appendix B.1.3.

Variables that are b-jet properties or are derived from b-jet properties are used in the MVA after they were corrected as described in sec. 8.4. The baseline and all distributions shown are corrected using the jet energy regression.

8.5.1 MVA Training Set-Up

The BDTs are trained using simulated MC events that passed the analysis' event selection. The sum of all relevant background processes, as listed in sec. 8.1, is used as the background template for the training. Each background process is weighted to its expected amount of events in the final phase space of the analysis. This ensures that the background template for the training represents the accurate background composition. The training parameters of the BDTs, which are listed in appendix B.1, were optimised for the $V(H \to b\bar{b})$ analysis that was performed with ATLAS run 1 data [66]. Tests confirmed that this set-up is still optimal. Since most of the kinematic variables have tails towards very high values the range of the input variables is limited to a range that includes 99% of all signal events. A table of the limits for the input variables is included in appendix B.1. All events above those limits will be artificially set to the maximum value. This procedure is introduced to avoid that the BDTs waste degrees of freedom to categorise the small number of events that accumulate in the tails of these distributions. To ensure that the training data set is fully orthogonal to the data set, same strategy is used as for the jet energy regression, described in section 7.1: the training data sets are split into two sub-sets based on their event number — even or odd — and then evaluated crosswise. The split into even and odd data sets is only done for the training and evaluation. They are not treated as separate categories in the analysis since no difference in terms of physics is expected between them. Finally, to enhance the amount of training events truth tagging, as described in section 5.5.3, is used for all simulated background events. This is crucial to preserve the statistical power of the simulated background data sets which would be depleted if b-tagging techniques were used directly. Truth tagging increases the performance of the BDTs and helps to avoid *over training*. Figure 8.7 includes the BDTs output distributions in the 2 lepton 2 jets high p_{T}^{Z} category and compares the odd training data set and the statistically independent even test data set. This category is the analysis category with the least amount of available training events. Differences between the training and test data set are neither observed for the distribution of the signal events nor the background events. In conclusion, no *over training* is present. To assess the performance of the jet energy regression within this analysis the analysis' BDTs are retrained: 1. using variables that were not corrected with dedicated b-jet corrections and 2. using variables that were corrected using the default corrections of the standard ATLAS $V(H \to b\bar{b})$ analysis. Comparisons of the distributions of the input variables and BDTs output for all three methods are given in appendix B.1.2.



(a) $m_{b\bar{b}}$ (b) $\Delta R(b_1, b_2)$ (c) $p_{\mathrm{T}}^{b_1}$

(d) $p_{\mathrm{T}}^{b_2}$ (e) $E_{\mathrm{T}}^{\mathrm{miss}}$ ($\equiv p_{\mathrm{T}}^{Z}$) (f) $\Delta\phi(Z, H)$

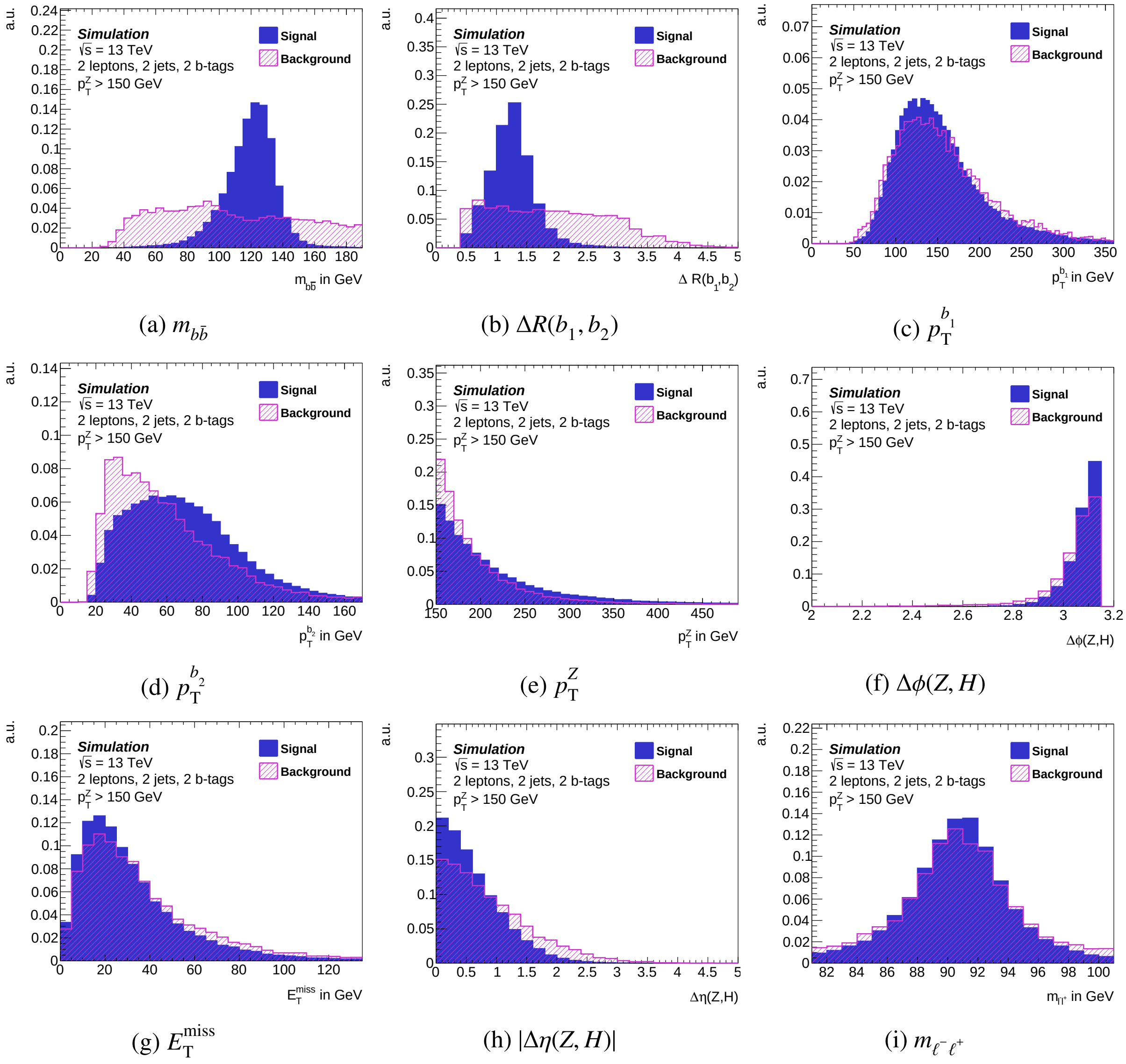
(g) $|\Delta\eta(b_1, b_2)|$ (h) H_{T}

Figure 8.5: The distributions of all input variables for the 0 lepton 2 jet MVA. Shown are the distributions in simulated signal events (blue) and the sum of all simulated background (pink) events. All distributions are normalised to the same area.

8.5.2 MVA Performance

Figure 8.7 shows the BDTs output distributions for the 0 lepton and 2 lepton category with the highest relative amount of signal events: 0 leptons 2 jets and 2 leptons 2 jets high p_{T}^{Z} category. The signal events accumulate at high BDTs output scores whereas the background events accumulate at low BDTs output scores, which is the desired behaviour. This characteristic shape is used as the final discriminant of the analysis.

To assess the statistical power of this discriminant the expected statistical significances are calculated from the simulated events. The full distributions of the BDTs outputs are used and the number of bins and their widths is the same as used to obtain the final result of the analysis (the binning strategy is explained in section 8.7.1). Figure 8.8 shows the statistical significances for each analysis category comparing



(a) $m_{b\bar{b}}$

(b) $\Delta R(b_1, b_2)$

(c) $p_{\mathrm{T}}^{b_1}$

(d) $p_{\mathrm{T}}^{b_2}$

(e) p_{T}^{Z}

(f) $\Delta\phi(Z, H)$

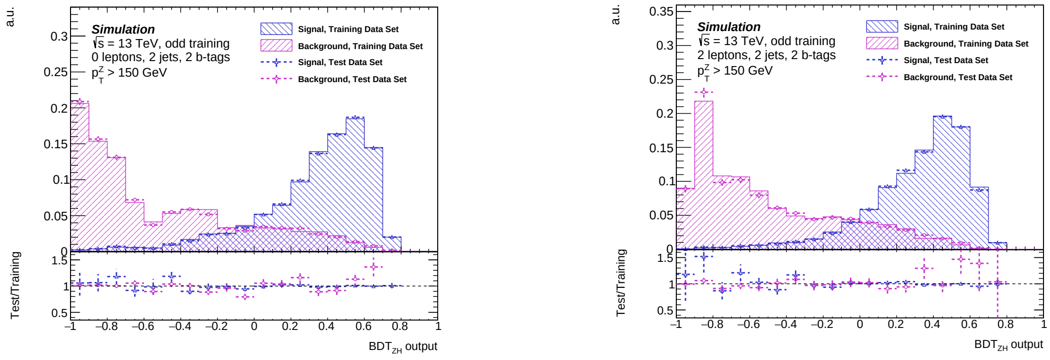
(g) $E_{\mathrm{T}}^{\mathrm{miss}}$

(h) $|\Delta\eta(Z, H)|$

(i) $m_{\ell^-\ell^+}$

Figure 8.6: The distributions of all input variables for the 2 lepton 2 jet high p_{T}^{Z} MVA. Shown are the distributions in simulated signal events (blue) and the sum of all simulated background (pink) events. All distributions are normalised to the same area.

results obtained with no dedicated b-jet correction (nominal), the *default* corrections and the jet energy regression. The performance with the jet energy regression yields an improvement of the significance in all analysis regions compared to the results with no b-jet correction applied. Differences between the *default* correction and the jet energy regression shows similar trends as in the $m_{b\bar{b}}$ resolution: advantages of the jet energy regression in the medium p_{T}^{Z} and 3 (3+) jets categories and an advantage of the *default* correction in the 2 lepton 2 jets high p_{T}^{Z} category. The total significance of all regions combined yields: 3.82σ for the nominal case, 3.95σ for the *default* correction and 3.99σ for the jet energy regression. Figure 8.8 also reveals that the 0 lepton 2 jets category has a much higher significance than the other categories. Although the 2 lepton categories yield lower significances compared to the 0 lepton categories the 2 lepton channel adds gain by including the medium p_{T}^{Z} categories. Thus the effect of the statistical



(a) 0 leptons, 2 jets

(b) 2 leptons, 2 jets, high p_T^Z

Figure 8.7: The distributions of the BDTs output for a) the 0 leptons 2 jets category and b) the 2 leptons 2 jets high p_T^Z category separate for the signal process (blue) and sum of all background processes (pink). For each distribution the BDTs output for the training data set (filled histogram) and a statistically independent test data set (points) is compared. All distributions are normalised to the same area and show simulated events.

advantage of the 0 lepton channel — due to the higher BR of the $ZH \to \nu\bar{\nu}b\bar{b}$ decay — balances with the larger phase space of the 2 lepton channel.

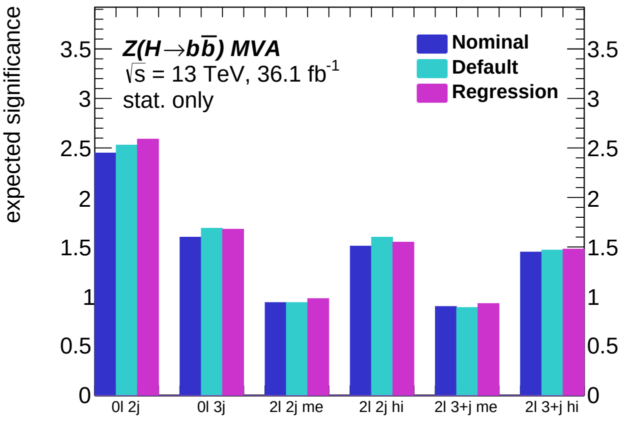


Figure 8.8: The statistical significance of the analysis regions of the $Z(H \to b\bar{b})$ MVA based on the binned BDTs outputs. Different b-jet correction methods are shown: nominal (blue), *default* corrections (cyan) and jet energy regression (pink). The analysis categories are abbreviated as: $0l$ = 0 lepton channel, $2l$ = 2 lepton channel, $2j$ = 2 jets category, $3j$ = 3 jets category, $3{+}j$ = 3+ jets category, me = medium p_T^Z, hi = high p_T^Z.

8.6 Systematic Uncertainties

The analysis techniques that are used in the $Z(H \to b\bar{b})$ analysis introduce systematic uncertainties, which have an impact on the obtained results. In general, two different sources of uncertainties are distinguished: experimental uncertainties, discussed in section 8.6.1, and modelling uncertainties, discussed in section 8.6.2.

8.6.1 Experimental Uncertainties

Experimental uncertainties are connected to the way data is recorded and objects are reconstructed. They follow the official ATLAS recommendations and are independent of the analysis. The impact of the experimental uncertainties on the final result of the analysis depends on the analysis strategy.

Luminosity and pile-up: The knowledge of the recorded luminosity is used to predict the expected amount of events from background and signal processes. Its uncertainty is between 2% and 5% depending on the data set. The luminosity and its uncertainty are derived from measurements of the number of interactions while displacing the proton beams w.r.t. each other in the $x - y$ plane [81, 82]. Since a difference is observed in the average number of pile-up interactions in data and simulated events, the simulated events are rescaled to the distribution of the average number of pile-up interactions in recorded data events. A systematic uncertainty that is as large as the rescaling is assigned to this procedure.

Leptons: Two sources of systematic uncertainties connected to leptons enter the analysis. The first set is connected to the single lepton triggers that are used to identify events of interest. These uncertainties take into account differences in the trigger efficiency in data and simulated events. The second set enters the analysis since the reconstructed leptons are used to define the final analysis phase space and their properties are further used in the analysis, e.g. their p_{T}. Systematic uncertainties are determined for the reconstruction, identification and isolation efficiencies as well as the energy scale and resolution of electrons [37] and muons [38]. Due to fundamental differences in the reconstruction of electrons and muons their systematic uncertainties are derived and treated separately.

Jets: For the ZH analyses several kinematic selections are applied to the reconstructed jets. In addition, reconstructed jets are used to reconstruct the Higgs boson candidate. Thus systematic uncertainties affecting the jet energy affect the analysis. The calibration of the jet energy involves multiple steps, which all introduce additional systematic uncertainties. Therefore a large set of systematic uncertainties is assigned to the jet energy scale based on studies in simulations and measurements in data [55]. This set is reduced to 19 uncorrelated systematic uncertainties. Effects connected to the selection and simulation of Z+jets, γ+jets and multi-jet events for the calibration, as well as pile-up effects, the effect of the η-calibration and differences in the response of gluon-, light- and b-jets are taken into account. The total uncertainty on the jet energy scale is 4.5% at $p_{\mathrm{T}} = 20\,\mathrm{GeV}$ and decreases to 1% at 200 GeV. A separate systematic uncertainty is determined for the jet energy resolution taking into account differences between data and simulated events and experimental uncertainties connected to the measurement of the jet energy resolution [83].

$E_{\mathrm{T}}^{\mathrm{miss}}$: Two sources of uncertainties related to the measured $E_{\mathrm{T}}^{\mathrm{miss}}$ enter the analysis: uncertainties related to the $E_{\mathrm{T}}^{\mathrm{miss}}$ trigger and uncertainties connected to the usage of $E_{\mathrm{T}}^{\mathrm{miss}}$ properties in the analysis. The systematic uncertainties of the objects that are used to calculate $E_{\mathrm{T}}^{\mathrm{miss}}$, e.g. leptons, jets, are propagated to the calculation of $E_{\mathrm{T}}^{\mathrm{miss}}$. Additional systematic uncertainties connected to the $E_{\mathrm{T}}^{\mathrm{miss}}$ scale and resolution, the efficiency of the track reconstruction that enter the soft term as well as the model that describes the underlying event are determined. They take into account differences in data and MC events [40, 84].

***b*-tagging:** The identification of b-jets is a crucial part for the analysis described here in order to reconstruct the Higgs boson candidate and suppress background processes containing light jets. They are identified by placing a requirement on the output distribution of the multivariate taggers.

To eliminate differences in the efficiency between data and simulation imposed by the requirement, scale factors are applied (see section 5). The systematic uncertainties originate from various sources connected to the derivation of these scale factors such as description of kinematic distributions and variations of flavour fractions in the simulations. All uncertainties are then decomposed into sets of uncorrelated uncertainties for b-jets, c-jets and light jets. The reduction set chosen depends on how sensitive the analysis is to the effects that are accounted for in the systematic uncertainties. The $Z(H \to b\bar{b})$ analysis utilises the set which provides 3 uncertainties for b-jets and c-jets each and 5 uncertainties for light jets. The size of the uncertainties depend on the p_{T} of the jets and are in the order of 2% for b-jets, 10% for c-jets and 30% for light jets [85].

8.6.2 Modelling Uncertainties

Modelling uncertainties are systematic uncertainties that are connected to the description of the signal and background processes in the MC simulations. These uncertainties originate from the assumptions that enter the simulations. They are determined by comparing a model that uses the baseline assumptions of the analysis, e.g. given by the choice of the MC generator, to an alternative model that uses alternative assumptions. In the following the baseline models of the analysis are referred to as nominal models. Modelling uncertainties largely depend on the analysis phase space and have to be derived specifically for the $Z(H \to b\bar{b})$ analysis. Based on the effect on the analysis conceptually three different types of uncertainties are distinguished: normalisation uncertainties, acceptance uncertainties, due to the analysis selection, and shape uncertainties. Technically, normalisation uncertainties and acceptance uncertainties have the same effect, they express the uncertainties on the absolute number of the events. Nevertheless the distinction is made since they often have different sources. Normalisation uncertainties are connected to an underlying source that affects all events of a certain process alike, e.g. an uncertainty on the total cross section of a process. A normalisation uncertainty $\Delta_{\text{norm.}}$ is derived from the comparison of the amount of expected events in the nominal model N_n and in the alternative model N_a:

$$\Delta_{\text{norm.}} = 1 - \frac{N_a}{N_n} \tag{8.1}$$

Acceptance uncertainties only affect the number of events in certain analysis categories and are therefore more analysis specific. They account for migration effects between analysis categories due to the selection of the analysis phase space and the split of events into analysis categories. Those migration effects are connected to uncertainties on the properties that are used in the event selection and categorisation, e.g. the uncertainty on the number of jets in an event could cause more events in the 2 jets category with respect to the 3(3+) jets category. An acceptance uncertainty $\Delta_{\text{acc.}}$ is derived from the comparison of the relative amount of events predicted by the nominal model in one analysis category N_n^i with respect to another category N_n^j and the same fraction in the alternative model. It is parametrised as an uncertainty on N_n^i:

$$\Delta_{\text{acc.}} = 1 - \frac{N_a^i/N_a^j}{N_n^i/N_n^j} \tag{8.2}$$

Shape effects are treated separately, i.e. a variation in the shape does not change the amount of events. This is a technical choice but has the advantage that it allows to "diagnose" the dominating effects of modelling uncertainties, i.e. if a variation in the shape or the normalisation has a larger impact on the result. This is useful information for future analyses and production of simulated data sets. Shape uncertainties in the $Z(H \to b\bar{b})$ analysis are assigned as a variation of the BDTs output distribution. Since the shape of the BDTs output distribution depends on the shapes of the distributions of the input variables a change in the BDTs output encodes many effects. To partially disentangle these effects, in this analysis shape uncertainties are parametrised as variations of the $m_{b\bar{b}}$ and p_{T}^{Z} distribution and one uncertainty is derived for each. This parametrisation approximately separates the underlying sources of effects that impact the analysis. Shape variations in the p_{T}^{Z} distribution are more impacted by uncertainties in the ME calculation, e.g. missing higher order effects, or uncertainties in the PDF set whereas the $m_{b\bar{b}}$ distribution is impacted by PS and related effects for the signal and most background processes. The shape variations of a variable x is derived from the comparison of the distribution in the nominal model $f(x_n)$ to the distribution in the alternative model $f(x_a)$:

$$\Delta_{\text{shape}} = 1 - f\left(\frac{x_a}{x_n}\right) \approx 1 - \frac{f(x_a)}{f(x_n)} \tag{8.3}$$

The shape uncertainty is approximated by fitting an analytical function f to the binned distribution of x_a/x_n to dim the effect of statistical fluctuations in the distribution of x_a/x_n. The final uncertainty for the BDTs output is determined by reweighting the events in dependence of $m_{b\bar{b}}$ or p_{T}^{Z} according to their shape variations, i.e. $\Delta_{\text{shape}}(m_{b\bar{b}}) \equiv$ event weight. The BDTs are then re-evaluated with those reweighted events and the resulting shape variation is the shape uncertainty. This strategy also takes into account the change in shape of variables that are correlated to $m_{b\bar{b}}$ or p_{T}^{Z}. Since $m_{b\bar{b}}$ and p_{T}^{Z} are only loosely correlated a double counting of effects is avoided as well. In the following these prescriptions are referred to as $m_{b\bar{b}}$ and p_{T}^{Z} shape uncertainties for simplicity.

A set of normalisation, acceptance and $m_{b\bar{b}}$ and p_{T}^{Z} shape uncertainties is derived separately for the signal processes and each of the background processes. The uncertainties are derived from the comparison of the nominal MC model used in the analysis to alternative models, which are all listed in table 8.2. For the derivation of systematic uncertainties conceptually two different alternative models are distinguished: 1. those that change a specific parameter of the nominal simulation, e.g. just a different normalisation scale, 2. those that use a different generator to simulate the event or parts of it (e.g. only the PS generator is exchanged). Since the latter often encodes a multitude of changes — renormalisation scale, factorisation scale, PS tune — the final systematic uncertainty is derived from either (1) or (2) to avoid double counting of effects. Both options are considered and the one that shows the larger variation is assigned as an uncertainty. If normalisation and acceptance uncertainties are derived from option (1) the sum in quadrature of all available variations is assigned as an uncertainty. If shape uncertainties are derived from option (1) the envelope of all variations is assigned as an uncertainty. If shape uncertainties are derived from option (2) in most cases one of the alternative MC generators predicts a much larger variation of the shape compared to the other alternative generators. Thus the largest variation is assigned as an uncertainty. An exception to this strategy are the Z+jets shape uncertainties which are derived from comparisons of data and simulated events since the 2 lepton channels offers a high purity of Z+jets events. In general, if an uncertainty is expected and observed to be similar in different analysis categories a common uncertainty is assigned. In the 2 lepton channel all uncertainties are derived by combining the medium and high p_{T}^{Z} category. Systematic uncertainties due to the p_{T}^{Z} categorisation are accounted for by the p_{T}^{Z} shape uncertainty. The size and source of all modelling uncertainties are summarised in table 8.7 for the signal processes and table 8.8 for the background processes. If the source is given as "various" the uncertainty is derived from the comparison to a generator that uses different models for various parts of the simulation and the exact changes cannot be disentangled. Uncertainties to account for systematic uncertainties due to the categorisation based on the jet multiplicity and p_{T}^{Z} and $m_{b\bar{b}}$ shape uncertainties are derived for all processes. In the following process specific systematic uncertainties are explained and all details are given in appendix B.3.

$\boldsymbol{V(H \rightarrow b\bar{b})}$**:** In addition to the analysis specific uncertainties, theory uncertainties on the production cross section and $H \rightarrow b\bar{b}$ BR are considered [4, 86, 87]. The acceptance uncertainties for the jet multiplicty categorisation that originate from renormalisation scale μ_R and factorisation scale uncertainties μ_F are calculated with the recommended Stewart-Tackman (ST) method [88]. An additional shape uncertainty for the NLO EW reweighting is considered as well.

Uncertainty	Source	$ZH \rightarrow \nu\bar{\nu}b\bar{b}$		$W^{\pm}H \rightarrow \ell^{\pm}\overset{(-)}{\nu}b\bar{b}$		$ZH \rightarrow \ell^{-}\ell^{+}b\bar{b}$	
		$0\ell\,2j$	$0\ell\,3j$	$0\ell\,2j$	$0\ell\,3j$	$2\ell\,2j$	$2\ell\,3{+}j$
$\Delta\sigma(q\bar{q} \rightarrow VH)$	QCD	0.7%		0.7%		0.7%	
$\Delta\sigma(gg \rightarrow ZH)$	QCD	27%		–		27%	
$\Delta\sigma(q\bar{q} \rightarrow VH)$	PDF+α_S	1.6%		1.9%		1.6%	
$\Delta\sigma(gg \rightarrow ZH)$	PDF+α_S	5%		–		5%	
$\Delta\mathrm{BR}(H \rightarrow b\bar{b})$	QCD, EW, m_b,α_S	1.7%		1.7%		1.7%	
$\Delta_\mathrm{norm.}$	PS, UE	10.0%		12.1%		13.9%	
$\Delta_\mathrm{acc.}(3j$ w.r.t. $2j)$	PS, UE	-	13.0%	-	12.9%	-	13.4%
$\Delta_\mathrm{norm.}(2j)$	$\mu_\mathrm{R}, \mu_\mathrm{F}$	6.9%	-	8.8%	-	3.3%	-
$\Delta_\mathrm{norm.}(3j)$	$\mu_\mathrm{R}, \mu_\mathrm{F}$	-7.0%	+5.0%	-8.6%	+6.8%	-3.2%	+3.9%
$\Delta_\mathrm{norm.}(\geq 4$ veto$)$	$\mu_\mathrm{R}, \mu_\mathrm{F}$	-	-2.5%	-	3.8%	-	-
$\Delta_\mathrm{norm.}$	PDF+α_S	1.1%		1.3%		0.5%	
$\Delta_\mathrm{shape}(p_\mathrm{T}^Z)$	missing higher orders (EW)	S		S		S	
$\Delta_\mathrm{shape}(p_\mathrm{T}^Z)$	$\mu_\mathrm{R}, \mu_\mathrm{F}$	S	S	S	S	S	S
$\Delta_\mathrm{shape}(m_{b\bar{b}})$	$\mu_\mathrm{R}, \mu_\mathrm{F}$	S	S	S	S	S	S
$\Delta_\mathrm{shape}(p_\mathrm{T}^Z)$	PDF+α_S	S		S		S	

Table 8.7: Modelling uncertainties assigned to the signal processes. Given is the type of the systematic uncertainty, the source and the size for each analysis category. Shape uncertainties are labelled as "S" and correspond to a functional form. Therefore no precise value can be given. The analysis categories are abbreviated as: 0ℓ = 0 lepton channel, 2ℓ = 2 lepton channel, $2j$ = 2 jets category, $3j$ = 3 jets category, $3{+}j$ = 3+ jets category. In the 2 lepton channel the same uncertainties are assigned to the medium and high p_T^Z category. If no sign for the uncertainty is given it is symmetric around the nominal value.

V+jets: The V+jets uncertainties contain a large set of uncertainties to account for differences in the flavour composition of the jets. Due to the usage of b-tagging techniques the acceptance of V+jets events depends on the jets' flavour. A change in the flavour composition changes the accepted amount of V+jets events and influences the BDTs output shape as well.

$t\bar{t}$: The systematic uncertainties of the $t\bar{t}$ model in the 0 lepton and 2 lepton channel are derived separately due to different $t\bar{t}$ final states that are probed in these channels.

Diboson: The diboson samples are normalised to the MC generator cross section and an additional uncertainty is assigned to it. To calculate the μ_R and μ_F acceptance uncertainties the ST method, as for the signal, is used. Since WW events only contribute less than 1% to the diboson background events only a normalisation uncertainty is assigned.

Single top: Analysis specific uncertainties are derived for t-channel and Wt production. Only a normalisation uncertainty is assigned to the s-channel events since it contributes less than 1% to the amount of single top events.

8.7 Statistical Analysis

To extract the final result from data a profile likelihood fit, as described in section 6.2, is utilised which is implemented in the RooStats framework [89, 90]. The various steps and parts that have to be taken into

Uncertainty	Source	Z+jets				W+jets	
		$0\ell\,2j$	$0\ell\,3j$	$2\ell\,2j$	$2\ell\,3{+}j$	$0\ell\,2j$	$0\ell\,3j$
$\Delta_{\text{norm.}}(V + (bb, bc, bl, cc))$	-	float				33%	33%
$\Delta_{\text{norm.}}(V + cl)$	μ_{R}, μ_{F}, merging & resummation scale	23%				37%	
$\Delta_{\text{norm.}}(V\text{+light})$	μ_{R}, μ_{F}, merging & resummation scale	18%				32%	
$\Delta_{\text{acc.}}(V + (bb, bc, bl, cc),\ 0\ell$ w.r.t $2\ell)$	μ_{R}, μ_{F}, merging & resummation scale	7%		-	-	-	-
$\Delta_{\text{acc.}}(V + bc$ w.r.t $V + bb)$	various	40%		40%	30%	15%	
$\Delta_{\text{acc.}}(V + bl$ w.r.t $V + bb)$	various	25%		28%	20%	26%	
$\Delta_{\text{acc.}}(V + cc$ w.r.t $V + bb)$	various	15%		16%	13%	10%	
$\Delta_{\text{shape}}(m_{b\bar{b}})$	various	S				S	
$\Delta_{\text{shape}}(p_{\text{T}}^{Z})$	various	S				S	

Uncertainty	Source	$t\bar{t}$			
		$0\ell\,2j$	$0\ell\,3j$	$2\ell\,2j$	$2\ell\,3{+}j$
$\Delta_{\text{norm.}}$	-	float		float	float
$\Delta_{\text{acc.}}(2j$ w.r.t $3j)$	ME, PS, μ_{R}, μ_{F}	9%	-	-	-
$\Delta_{\text{shape}}(m_{b\bar{b}})$	ME	S		S	S
$\Delta_{\text{shape}}(p_{\text{T}}^{Z})$	ME	S		S	S

uncertainty	source	ZZ				WZ		WW
		$0\ell\,2j$	$0\ell\,3j$	$2\ell\,2j$	$2\ell\,3{+}j$	$0\ell\,2j$ or $2\ell\,2j$	$0\ell\,3j$ or $2\ell\,3{+}j$	all
$\Delta_{\text{norm.}}$	various	20%				26%		25%
$\Delta_{\text{norm.}}$(categories)	PS, UE	5.6%		5.8%		3.9%		-
$\Delta_{\text{acc.}}(3j$ w.r.t $2j)$	PS, UE	-	7.3%	-	3.1%	-	10.8%	-
$\Delta_{\text{acc.}}(0\ell$ residual)	PS, UE	6%		-	-	11% (only 0ℓ)		-
$\Delta_{\text{norm.}}(2j)$	$\mu_{\text{R}}, \mu_{\text{F}}$	10.3%	-	11.9%	-	12.7%	-	
$\Delta_{\text{norm.}}(3j)$	$\mu_{\text{R}}, \mu_{\text{F}}$	-15.2%	+17.4%	-16.4%	10.1%	-17.7%	21.2%	-
$\Delta_{\text{norm.}}(\geq 4$ veto)	$\mu_{\text{R}}, \mu_{\text{F}}$	-	+18.2%	-	-	-	+19.0% (only 0ℓ)	-
$\Delta_{\text{shape}}(p_{\text{T}}^{Z})$	PS, UE	S		S		S		-
$\Delta_{\text{shape}}(p_{\text{T}}^{Z})$	$\mu_{\text{R}}, \mu_{\text{F}}$	S		S		S		-
$\Delta_{\text{shape}}(m_{b\bar{b}})$	PS, UE	S						-
$\Delta_{\text{shape}}(m_{b\bar{b}})$	$\mu_{\text{R}}, \mu_{\text{F}}$	S						-

uncertainty	source	t-channel		Wt		s-channel
		$0\ell\,2j$ or $2\ell\,2j$	$0\ell\,3j$ or $2\ell\,3{+}j$	$0\ell\,2j$ or $2\ell\,2j$	$0\ell\,3j$ or $2\ell\,3{+}j$	all
$\Delta_{\text{norm.}}$	μ_{R}, μ_{F}, PDF, α_{S}	4.4%		6.2%		4.6%
$\Delta_{\text{norm.}}$(analysis)	μ_{R}, μ_{F}, PDF, α_{S}	17%	20%	35%	41%	-
$\Delta_{\text{shape}}(m_{b\bar{b}})$	μ_{R}, μ_{F} (t-channel); DS (Wt)	S		S		-
$\Delta_{\text{shape}}(p_{\text{T}}^{Z})$	ME (t-channel), DS (Wt)	S		S		-

Table 8.8: Modelling uncertainties assigned to the background processes. Given is the type of the systematic uncertainty, the source and the size for each analysis category. Shape uncertainties are labelled as "S" and correspond to a functional form. Therefore no precise value can be given. The analysis categories are abbreviated as: 0ℓ = 0 lepton channel, 2ℓ = 2 lepton channel, $2j$ = 2 jets category, $3j$ = 3 jets category, $3{+}j$ = 3+ jets category. In the 2 lepton channel the same uncertainties are assigned to the medium and high p_{T}^{Z} category. Absolute normalisations of the main backgrounds, Z+jets and $t\bar{t}$, are determined from the fit to data and are listed here as "float". If the source is given as "various" the uncertainty was derived from the comparison to a simulation from a different MC generator which incorporates different PDF sets, tunes, PS models, scales, merging schemes etc. or from a comparison to data. If the source is given as "ME" the uncertainty was derived from a comparison to a model of a MC generator that uses the same PS model as the nominal model but has a different ME generator which also includes PDF differences, merging between ME and PS differences etc.

account in the set-up of the fit model are explained in the following sections.

8.7.1 Fit Inputs

The input to the fit are the binned BDTs output distributions of all analysis regions, in the following also referred to as signal regions (SRs). The following separate signal and background components are considered in the fit:

- **Signal**: $gg \to ZH \to \nu\bar{\nu}b\bar{b}$, $qq \to ZH \to \nu\bar{\nu}b\bar{b}$, $W^{\pm}H \to \ell^{\pm}\overset{(-)}{\nu}b\bar{b}$, $gg \to ZH \to \ell^{-}\ell^{+}b\bar{b}$, $qq \to ZH \to \ell^{-}\ell^{+}b\bar{b}$
- ***V*+jets**: $Z+bb$, $Z+bc$, $Z+bl$, $Z+cc$, $Z+cl$, $Z+ll$, $W+bb$, $W+bc$, $W+bl$, $W+cc$, $W+cl$, $W+ll$
- **$t\bar{t}$**
- **single top**: t-channel, Wt, s-channel
- **diboson**: ZZ, WZ, WW

The template for each of these components is given by their corresponding simulated data. To enhance the statistical power of the simulated $V + cl$, $V + ll$ and WW components truth tagging is used. In the 2 lepton channel additional $t\bar{t}$ dominated analysis categories are introduced in the fit. The selection of the events is the same as for the already introduced analysis categories, with the exception that the leptons have to be of opposite flavour, i.e. one muon and one electron per event. This excludes all events that contain a Z boson. Events that fulfil this requirement are split into 2 jets and 3+ jets events as well as medium and high p_{T}^{Z} events in accordance with the 2 lepton SRs. These additional categories, referred to as $e\mu$ CRs, contain from 88% up to 97% $t\bar{t}$ events. The remaining parts of these categories are composed of single top events. Other physics processes contribute less than 0.5% to the $e\mu$ CRs. These CRs are introduced as coarsely (50 GeV bin width) binned distributions of $m_{b\bar{b}}$ into the fit. The 2 jets high p_{T}^{Z} category is an exception since the amount of events is very low. Therefore this category is introduced as a single bin, i.e. only the total number of events and no shape information is used. The same systematic uncertainties as for the 2 lepton SRs are assigned to the $e\mu$ CRs since the same kinematic phase space is selected — only the selected lepton flavours are different. This is also the reason why no additional acceptance uncertainties are applied to account for differences between the SRs and CRs. The distributions of the $e\mu$ CRs are shown in appendix B.2. A summary of all categories and corresponding distributions in the fit is given in table 8.9. A simultaneous fit is performed in all these analysis categories. The binning of the BDTs output distributions is optimised to enhance the sensitivity encoded in the shape

		0 leptons		2 leptons	
		2 jets	3 jets	2 jets	3+ jets
SRs	medium p_{T}^{Z}			BDT	BDT
	high p_{T}^{Z}	BDT	BDT	BDT	BDT
$e\mu$ CRs	medium p_{T}^{Z}			$m_{b\bar{b}}$	$m_{b\bar{b}}$
	high p_{T}^{Z}			N(events)	$m_{b\bar{b}}$

Table 8.9: The analysis categories, including CRs, and the corresponding distribution of each cetagory. "BDT" refers to the BDTs output distribution.

of these distributions. The following transformation prescription, which was developed for the ATLAS $V(H \to b\bar{b})$ analysis with run 1 data [66], is used to merge bins:

$$Z = z_s \frac{n_s}{N_s} + z_b \frac{n_b}{N_b} \tag{8.4}$$

with N_s and N_b the total number of signal and background events, n_s and n_b the current number of signal and background events, and z_s and z_b the signal and background transformation parameters. The transformation starts with a very finely binned histogram of the BDTs output and starts to merge bins beginning from low BDTs output scores. With each added bin n_s and n_b increases until $Z > 1$. The bins that were merged up to that point represent a new bin and the merging continues with the next bins. The bin widths are controlled by the parameters z_s and z_b which have been optimised to $z_s = 10$ and $z_b = 5$. These values also ensure that the statistical uncertainty of the simulated events in each bin is not larger than 20%. The transformation parameters are correlated to the resulting amount of background and signal dominated bins. The general idea is to have a finer binning in regions with higher amounts of signal to optimally use the characteristic shape of the BDTs discriminant. A coarser binning is determined for the background dominated bins as they do not contain signal information. Nevertheless these bins are crucial in the fit to obtain information about the background and help to constrain the systematic uncertainties of the background model.

8.7.2 Normalisations and Systematic Uncertainties

The normalisation of 6 of the signal and background components are free parameters (*floating*) in the fit and they are determined from data:

- μ: The normalisation of the signal, referred to as signal strength, is the primary target of the analysis. The signal strength is defined as the ratio of the observed to expected number of signal events, i.e. for a SM Higgs boson μ is expected to be 1.
- $Z + (bb, bc, bl, cc)$ normalisation in the 2 jets categories: Comparisons of data and simulated events suggest that the normalisation of Z boson production in association with heavy flavour jets is underestimated in simulated events. Thus it is a free parameter in the fit. Categories with a high purity of these events and the bins of the low BDTs output scores offer the possibility to extract this normalisation. There is one common floating normalisation for the 0 and 2 lepton channel.
- $Z + (bb, bc, bl, cc)$ normalisation in the 3 and 3+ jets categories: Due to the different final state this normalisation in the 3 (3+) jets categories is free floating separately from the 2 jets categories.
- $t\bar{t}$ normalisation in the 0 lepton channel: The 0 lepton channel probes an "unusual" $t\bar{t}$ phase space since objects have to be mis-identified or not reconstructed in order to fulfil the 0 lepton selection criteria. To not rely on simulations to properly describe the normalisation of these events the $t\bar{t}$ normalisation is free floating in the 0 lepton channel.
- $t\bar{t}$ normalisation in the 2 lepton 2 jets categories: Since the $e\mu$ CRs offer an excellent opportunity to extract the $t\bar{t}$ normalisation in the 2 lepton channel it is extracted from the fit rather than to rely on the simulations.
- $t\bar{t}$ normalisation in the 2 letpon 3+ jets categories: The $t\bar{t}$ normalisation in the 3+ jets categories is floating separately to the 2 lepton channel since a different final state is probed there including higher order $t\bar{t}$ production. Again the $e\mu$ CRs offer enough information to introduce this free parameter in the fit.

The systematic uncertainties as described in section 8.6 are introduced as additional nuisance parameters in the fit. In contrast to the free floating parameters their size have external constraints. They are implemented as Gaussian priors. The bins of the distribution in a given category and the nuisance parameters acting on them are treated as correlated. Normalisation and acceptance uncertainties are fully correlated across bins. For uncertainties that encode a change in the shape of the distribution the bin-by-bin correlation is given by the shape variation. Additional correlations are introduced across categories in case the underlying source of an uncertainty is expected to be independent of the categories' definitions. Each experimental uncertainty is fully correlated across all categories. Each modelling uncertainty listed in tables 8.7 and 8.8 is correlated across categories but not correlated across processes (components). Exception are the $t\bar{t}$ and V+jets component. The $t\bar{t}$ modelling uncertainties are treated as not correlated between the 0 and 2 lepton channel due to the very different $t\bar{t}$ signatures entering these channels. The Z+jets shape uncertainties are treated as fully correlated across all Z+jets components. The Z+jets uncertainties that affect the $Z + (bb, bc, bl, cc)$ component are treated as fully correlated across these components. The same procedure is used for W+jets events. Additional nuisance parameters are introduced in the fit to account for the limited amount of available simulated events. For each bin the statistical error of the simulated data in that bin is introduced as Poissonian priors following the technique described in [91].

8.7.3 Smoothing and Pruning

To prevent artefacts of statistical fluctuations influencing the performance of the fit, a so-called smoothing procedure is applied. Certain experimental uncertainties such as the jet energy scale uncertainties lead to bin-by-bin acceptance effects in the distributions of the fit. This migration of events causes large statistical fluctuations which are non-physical. To avoid these effects the smoothing procedure from the standard ATLAS $V(H \to b\bar{b})$ analysis, which was used in the run 1 and run 2 analysis, is adopted. This smoothing procedure assumes that there is only one maximum in the varied distribution relative to the nominal distribution. Smoothing is applied separately to each variation in each category and is only applied to the jet, muon, electron and $E_{\mathrm{T}}^{\mathrm{miss}}$ uncertainties that have a shape changing effect.

To reduce the complexity of the fit nuisance parameters that have a negligible effect on the final result are removed from the fit. A nuisance parameter is considered "negligible" if the variation in any given bin, considering all categories, does not exceed 0.5%. This procedure was adopted from the standard ATLAS $V(H \to b\bar{b})$ analysis as well.

8.8 Results

The results of the search for the SM $H \to b\bar{b}$ decay obtained from a fit to the data are detailed in section 8.8.1. This result is validated with diboson events, which is explained in section 8.8.2. Eventually the results of different jet correction methods are compared in section 8.8.3.

8.8.1 $Z(H \to b\bar{b})$ Results

The observed significance of the $V(H \to b\bar{b})$ signal, which is dominated by $Z(H \to b\bar{b})$ events, is 2.9 σ compared to a hypothesis assuming only the existence of the SM background processes. This significance corresponds to a probability of smaller than 0.4% that the observed signal is caused by a fluctuation in the background. It is close to an "evidence", which by convention corresponds to a measured significance of 3 σ. The measured significance is compared to an expected significance of 2.8 σ, which is given by the simulated events fixed to the best prediction of their normalisation and taking into account all analysis uncertainties. The measured signal strength and its statistical (stat.) and systematic (syst.) errors are:

$$\mu = 1.15 \pm 0.29(\text{stat.})^{+0.36}_{-0.30}(\text{syst.}) \tag{8.5}$$

Thus the observed result is well compatible with the predictions of the SM including a SM Higgs boson. To obtain the result separately in the 0 and 2 lepton channel two free floating μ were introduced and extracted in a single fit to data. The summary of the 0 lepton channel, 2 lepton channel and combined measured signal strength is shown in figure 8.9. The results are $\mu = 0.7^{+0.54}_{-0.51}$ and $\mu = 1.83^{+0.79}_{-0.65}$ for the 0 lepton and 2 lepton channel respectively. This means the 0 lepton channel is well compatible with the SM expectation within its error and that the obtained signal strength in the 2 lepton channel is compatible with the SM within 1.3 σ. The result of this analysis is further supported by the good agreement of the data and simulated events in the BDTs output distributions, shown in figure 8.10. In those distributions the simulated events of each signal and background component are scaled to the normalisations obtained in the fit. The tables of the signal and background yields and the distributions of the $e\mu$ CRs are included in appendix B.4.

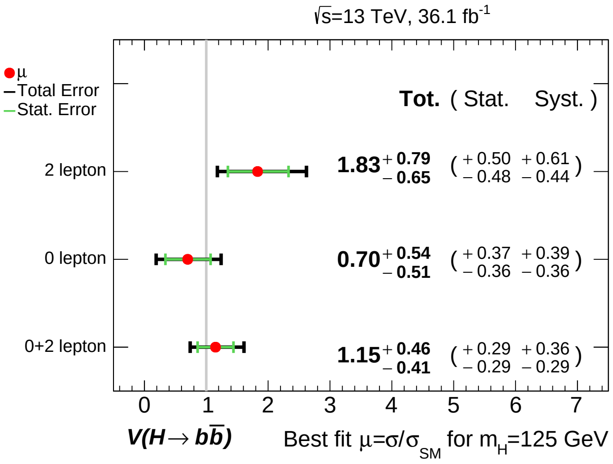
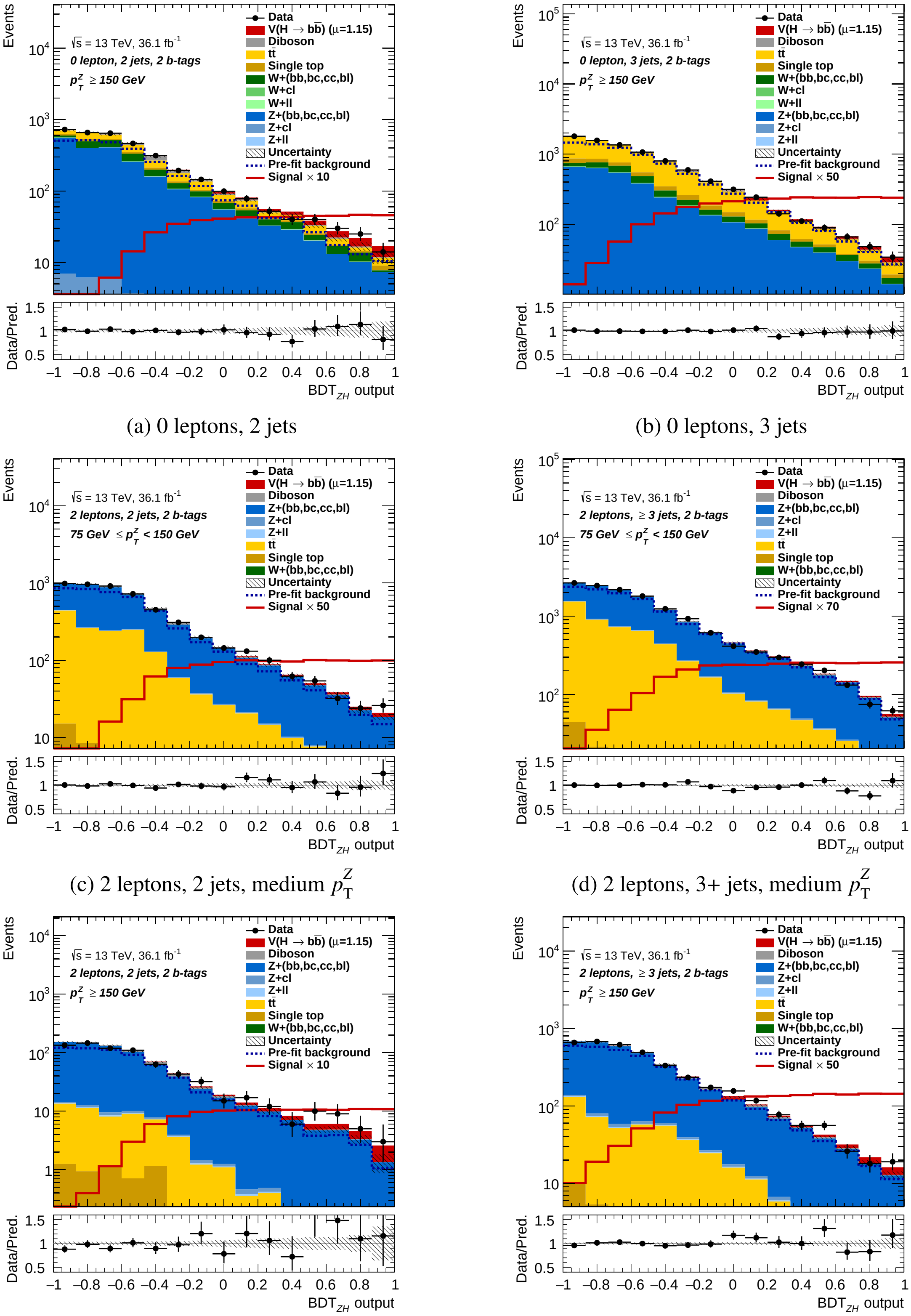


Figure 8.9: The measured signal strength in the 0 lepton channel, 2 lepton channel and of the whole analysis. The 0 and 2 lepton channel results were obtained in a single fit to the data but μ being not correlated between the 0 and 2 lepton channel.

The size of the statistical and systematic error on the obtained results is of the same order. This means the precision of the $Z(H \to b\bar{b})$ analysis is already limited by the impact of the systematic uncertainties. The collection of more LHC data in the future will decrease the statistical error but not necessarily the systematic error thus making it hard to eventually claim a discovery of the $H \to b\bar{b}$ decay. A closer look at the most dominant systematic uncertainties could reveal potential for improvements in future analyses. Table 8.11 shows the impact of different groups of uncertainties on the error of the measured μ. A new fit is performed for each listed group by fixing the nuisance parameters of this group to their values determined in the standard fit. All other nuisance parameters are still allowed to vary within their given external constraints. The impact on the error is determined by quadratically subtracting the error of the new fit from the total error of the standard fit.

The three largest sources of systematic uncertainties are: 1. the uncertainties of the signal prediction and its modelling in the simulations, 2. the uncertainties connected to the utilised b-tagging techniques, 3. the statistical uncertainty of the simulated MC data sets. In order to determine which improvement in



(a) 0 leptons, 2 jets

(b) 0 leptons, 3 jets

(c) 2 leptons, 2 jets, medium p_T^Z

(d) 2 leptons, 3+ jets, medium p_T^Z

(e) 2 leptons, 2 jets, high p_T^Z

(f) 2 leptons, 3+ jets, high p_T^Z

Figure 8.10: The BDTs output distributions of the $Z(H \to b\bar{b})$ SRs. All signal and background components are scaled to their normalisations as determined in the fit. The normalisation of the sum of all background components predicted by the simulated events is given by the dashed blue line. The shaded bands represent the total uncertainty.

terms of systematic uncertainties the analysis would benefit most from, the impact of each single nuisance parameter on the measured μ is investigated as well. Figure 8.12 shows the 15 nuisance parameters that have the largest influence on the measured μ, thus called "nuisance parameter ranking". The procedure to determine the impact is similar to the one used for the break down by groups but instead of whole groups a single nuisance parameter is fixed in each separate fit. The signal uncertainty that has the largest impact is the VH acceptance uncertainty due to the modelling of the PS and UE. It is also the nuisance parameter amongst all nuisance parameters with the largest overall impact. In addition, signal uncertainties connected to the choice of the QCD scale are sizeable as well. The b-tagging uncertainties that have the largest impact are the ones connected to the efficiency calibration for b-jets. The nuisance parameter ranking also reveals that three uncertainties of the Z+jets background model are amongst the 7 most impactful ones. In conclusion, the largest part of the limitations of the analysis due to the systematic uncertainties originates from the utilised simulation models. Those are also the ones that dominate the uncertainties of the b-tagging efficiency calibration. To improve the analysis in the future a larger amount of more precise simulated events are necessary. Another option is to change the analysis strategy in two ways: less reliance on simulated models or decrease of the sensitivity to these uncertainties, e.g. through the selection of the phase space.

Set of uncertainties	Impact on error	
Total	+0.46	-0.41
Stat.	+0.29	-0.29
Syst.	+0.36	-0.30
Floating normalisations	+0.08	-0.10
Jets and E_T^{miss}	+0.09	-0.06
***b*-tagging**	**+0.15**	**-0.15**
Electrons and muons	+0.02	-0.02
Luminosity	+0.05	-0.03
Diboson modelling	+0.07	-0.06
Z+jets modelling	+0.10	-0.12
W+jets modelling	+0.03	-0.03
$t\bar{t}$ modelling	+0.08	-0.09
Single top modelling	+0.07	-0.07
Signal Uncertainties	**+0.22**	**-0.10**
MC stat.	**+0.14**	**-0.14**

Figure 8.11: The break down of the sources of uncertainties. Given is the error on the measured μ that is caused by each group. The bold entries are the three largest sources. "MC stat." refers to the statistical uncertainty of the simulated data sets. "Modelling" refers to the uncertainties of the description of the given process in the utilised simulated data set. "Signal uncertainties" includes modelling and theory uncertainties.

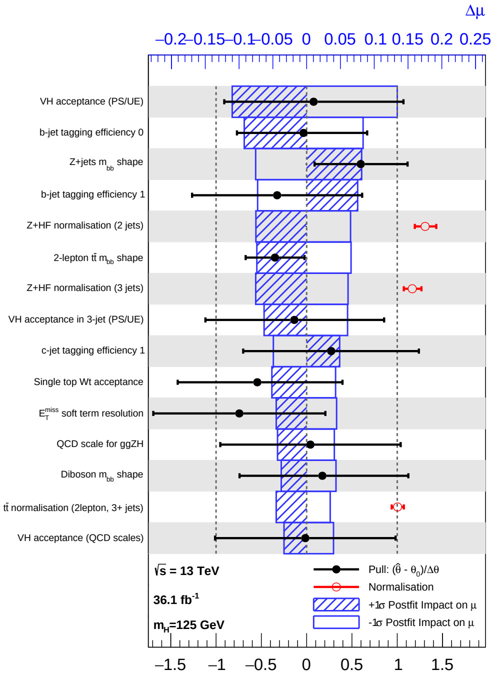
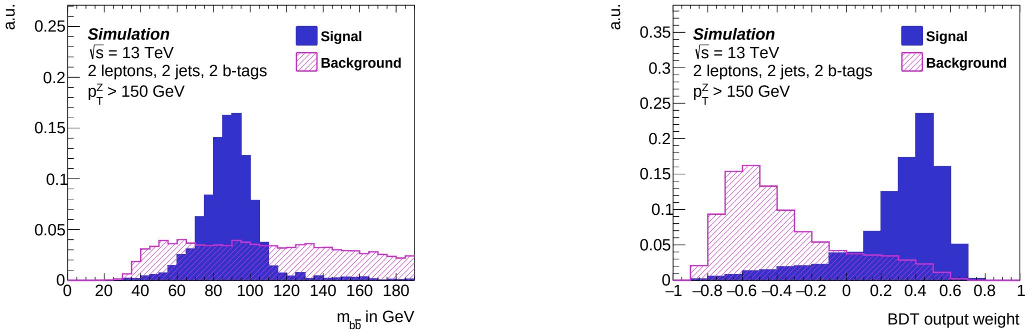


Figure 8.12: The ranking of the nuisance parameters based on their impact on the measured μ (blue boxes). The "pull", i.e. how much the fit has to vary this nuisance parameter away from its nominal value, of each nuisance parameter is given as well (black points). "Z+HF" refers to the Z + (bb, bc, bl, cc) background component.

8.8.2 Diboson Cross Check

The diboson process VZ, as aforementioned is the only background contribution that contains a $b\bar{b}$ resonance. Therefore it is an excellent process to validate the $Z(H \to b\bar{b})$ analysis process. The idea is that in a robust analysis, the techniques that are used to search for the $H \to b\bar{b}$ decay should also be able to be used in a search for the $Z \to b\bar{b}$ decay. The advantage with respect to the SM $Z(H \to b\bar{b})$ process is the approximately 3 times larger production cross section for diboson processes, see table 8.5. To search for this decay the same object selections, event selection and analysis categories as for the $Z(H \to b\bar{b})$ analysis are used. Approximately 80% of all selected diboson events are $V(Z \to b\bar{b})$ decays. The b-jets are corrected with the jet energy regression as well. New BDTs are trained with the diboson processes as the signal process. The set-up for the training and the input variables are the same as for the $Z(H \to b\bar{b})$ analysis since the diboson processes exhibit similar features compared to the $Z(H \to b\bar{b})$ processes. Therefore it can be assumed that the diboson BDTs training performance is similar to the ZH training performance. Figure 8.13 shows the $m_{b\bar{b}}$ and BDTs output distributions of the diboson signal and the sum of all backgrounds in the 2 lepton 2 jets high p_{T}^{Z} category. The diboson signal exhibits the expected peak in the $m_{b\bar{b}}$ distribution around the Z boson mass and the BDTs achieve a good separation of the diboson signal events from the background events. All distributions for all analysis categories are shown in appendix B.5.



(a) $m_{b\bar{b}}$

(b) BDTs output

Figure 8.13: The a) $m_{b\bar{b}}$ and b) BDTs output distributions for the 2 lepton 2 jets high p_{T}^{Z} category of the diboson validation analysis. Simulated events are shown and the distributions are normalised to the same area. "Signal" refers to the sum of all diboson processes.

The fit set-up, including all systematic uncertainties as described in section 8.6.2 and categories and distributions as listed in table 8.9, is the same as for the $Z(H \to b\bar{b})$ analysis with two small exceptions:

- **The signal strength μ**, is defined as the ratio of the observed to expected number of diboson events. The overall normalisation uncertainties of the diboson predictions are not included in the fit.

- The $V(H \to b\bar{b})$ component is fixed to the normalisation expected in the SM and an overall 50% uncertainty on this normalisation is added to the set of VH nuisance parameters

The production of WZ/ZZ boson pairs in the decay to leptons or $E_{\mathrm{T}}^{\mathrm{miss}}$ and two b-jets is observed with a significance of 5.6 σ which is compared to an expected significance of 6.2 σ. The measured signal strength in the combined fit of the 0 and 2 lepton channel is:

$$\mu = 0.97 \pm 0.12(\mathrm{stat.})^{+0.19}_{-0.16}(\mathrm{syst.}) \tag{8.6}$$

which is well compatible with the SM expectation. It is noticeable that the error of the diboson analysis is approximately two times smaller than the error of the $Z(H \to b\bar{b})$ result. A look at the break down of the uncertainties' impact on the total error, included in appendix B.5, traces this difference down to a reduced impact of the b-tagging uncertainties (±0.05 compared to ±0.15) and the prediction of the VH process does not impact the diboson result since it is a a background process with a small contribution. The prediction of the diboson production, which is described by the Sherpa MC generator, has a smaller impact on the diboson result than the prediction of the VH process, which is described by the Powheg +Pythia MC generators, has on the VH result. The largest contributors to the error of the diboson result are the modelling uncertainties of the Z+jets process.

As an additional check the diboson result is obtained separately in the 0 lepton and 2 lepton channel by decorrelating μ in those two channels. Figure 8.14 displays the measured μ in the 0 lepton, 2 lepton and 0+2 lepton combined analysis. The results are $\mu = 1.21^{+0.30}_{-0.28}$ and $\mu = 0.60^{+0.30}_{-0.28}$ for the 0 lepton and 2 lepton channel respectively. Overall the compatibility with the SM is at a similar level compared to the results observed in the $Z(H \to b\bar{b})$ analysis. This enhances the confidence in the obtained $H \to b\bar{b}$ result. The BDTs output distributions comparing data and simulated events are shown in figure 8.15 for the 0 lepton 2 jets and 2 lepton 2 jets high p_{T}^{Z} categories. Similar histograms for the other analysis categories are included in appendix B.5. The histograms show that a scenario of background events and diboson events is favoured by the data.

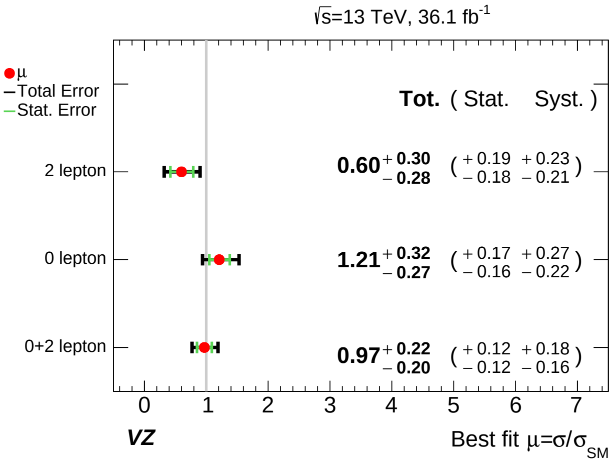
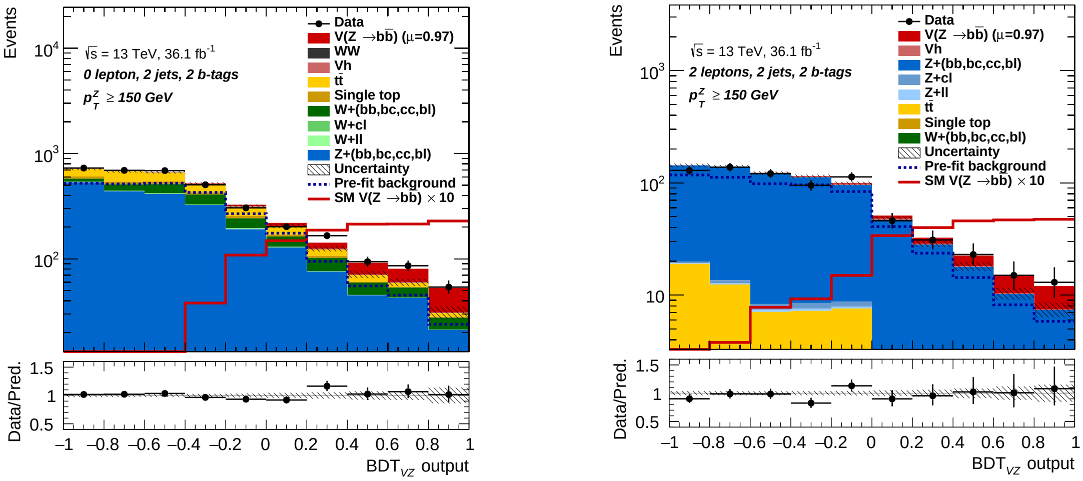


Figure 8.14: The measured signal strength in the 0 lepton channel, 2 lepton channel and of the combined analysis. The 0 and 2 lepton channel results were obtained in a single fit to data but μ being not correlated between the 0 and 2 lepton channel.

8.8.3 Comparison with Alternative $Z(H \to b\bar{b})$ Analyses

The obtained $Z(H \to b\bar{b})$ results used the jet energy regression to correct the $m_{b\bar{b}}$ distribution. The $m_{b\bar{b}}$ distribution is the most powerful variable of the employed MVA whose output distribution is used as the discriminant of the analysis. This b-jet energy correction technique was never used in an ATLAS analysis. The good agreement of the obtained results with the SM prediction and the validation in diboson processes is a good indicator that the jet energy regression is a suitable method to correct the b-jet energy without introducing any bias towards the Higgs boson signature. In this section this result is compared to alternative correction techniques and the published results of the ATLAS and CMS $H \to b\bar{b}$ searches in



(a) 0 leptons, 2 jets

(b) 2 leptons, 2 jets, high p_{T}^{Z}

Figure 8.15: The BDTs output distributions of the diboson validation analysis for the a) 0 lepton 2 jets and b) 2 lepton 2 jets high p_{T}^{Z} category. All signal and background components are scaled to their normalisations as determined in the fit. The normalisation of the sum of all background components predicted by the simulated events is given by the dashed blue line. The shaded bands represent the total uncertainty.

the VH production channel.

To compare different methods the same analysis is performed with two alternative BDTs trainings as explained in section 8.5: 1. the input variables to the BDTs are not corrected (nominal), 2. the input variables are corrected using the standard ATLAS $V(H \to b\bar{b})$ corrections for b-jets (*default* corrections), see table 8.6. The same analysis regions, systematic uncertainties and fit set-up are used as for the results obtained with the jet energy regression. The expected significances based on the prediction of the simulations for each analysis including systematic uncertainties are shown in figure 8.16. It shows that the advantage of the jet energy regression is largest in the 0 lepton channel and comparable to the *default* corrections in the 2 lepton channel. In general, the sensitivity of the analysis increases if b-jet corrections are applied. All expected and measured results are summarised in table 8.10. Overall the analysis using the jet energy regression yields the best expected significance of 2.8 σ including all analysis categories, compared to 2.7 σ for the *default* corrections and 2.6 σ for the nominal case without any correction. This corresponds to an improvement in the sensitivity of 4% with respect to the *default* corrections and 8% with respect to nominal case. This improvement is also reflected in the expected errors on the expected signal strength ($\mu = 1$) which are smaller for the jet energy regression compared to the other scenarios. Although the uncertainties of the $Z(H \to b\bar{b})$ analysis are large compared to these improvements a consistent improvement is observed in the $m_{b\bar{b}}$ resolutions and the discriminative power of the BDTs output across analysis categories. The observed results of the combined analysis of the 0 and 2 lepton channel are compatible with the SM expectation for all three scenarios. If the signal strengths are decorrelated between the 0 and 2 lepton channel all three correction methods yield very similar results in the 0 lepton channel. In the 2 lepton channel the results of the nominal case and the jet energy regression are still close together and compatible with the SM within 1.3 σ. The result with the *default* corrections exhibits a larger deviation of 1.7σ from the expected SM value with a measured μ of 2.10 ± 0.72. It is surprising that the measured μ of the analysis using the *default* corrections deviates more from the SM prediction than the jet energy regression and nominal result, which are very similar. In general, the measured $H \to b\bar{b}$ signal should not depend on the utilised correction since the exact same events and uncertainties are used. The only difference is the shape of the final discriminant. Therefore the jet energy corrections should only have an influence on the observed significance not the measured μ. No obvious reason is found to explain this difference but the results hint that the origin is in the 2 lepton channel. The standard ATLAS $V(H \to b\bar{b})$ analysis [75], which uses the *default* corrections, observes a 1.4 σ deviation in the 2 lepton channel with a measured μ of $1.90^{+0.78}_{-0.64}$(total). However the comparability of those results with the results obtained in this thesis is limited since the standard ATLAS $V(H \to b\bar{b})$ analysis includes a 1 lepton channel targeting the $W^{\pm}H \to \ell^{\pm}\overset{(-)}{\nu}b\bar{b}$ decay. All in all, due to the large error of the measurement it is difficult to draw any concrete conclusions about the behaviour of jet energy corrections based on the observed signal strengths at this point in time but the consistent improvement in the expected results using the jet energy regression is worth mentioning. Figure 8.17 shows the $m_{b\bar{b}}$ distribution in data and simulated events after all background distributions, except for diboson, have been subtracted from data. It provides a visual impression of the effect of the b-jet corrections on the two $b\bar{b}$ resonances in the analysis: $Z \to b\bar{b}$ and $H \to b\bar{b}$. After the application of b-jet corrections the $V(H \to b\bar{b})$ peak is visually more resolved with respect to the diboson peak. No visual difference is observable between the *default* corrections and the jet energy regression. In all three cases the diboson peak is clearly visible in data.

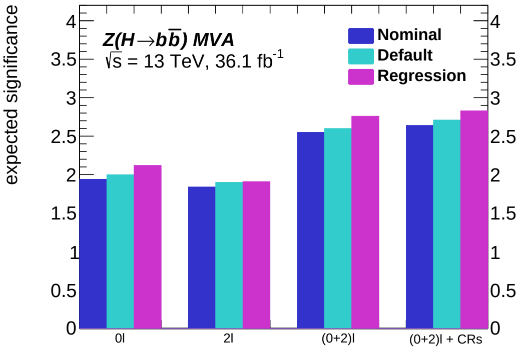
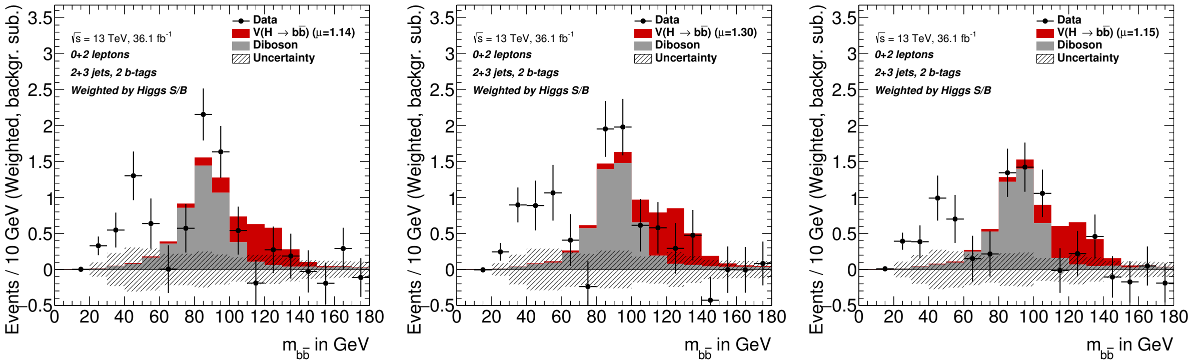


Figure 8.16: The expected significances of the 0 lepton (0l) SRs, 2 lepton (2l) SRs, 0 and 2 lepton ((0 + 2)l) SRs and 0 and 2 lepton SRs plus the $e\mu$ CRs ((0 + 2)l + CRs). Three different b-jet correction methods utilised in the analysis are compared: nominal (blue), *default* corrections (cyan) and jet energy regression (pink). Systematic uncertainties are considered in the estimate of the significance.

The overall result of this analysis is well compatible with the standard ATLAS $V(H \to b\bar{b})$ analysis, which analyses the same data set as this thesis but includes the additional $W^{\pm}H \to \ell^{\pm}\overset{(-)}{\nu}b\bar{b}$ channel. In this analysis, the expected significances of the channels targeting the $W^{\pm}H \to \ell^{\pm}\overset{(-)}{\nu}b\bar{b}$, $ZH \to \nu\bar{\nu}b\bar{b}$ and $ZH \to \ell^{-}\ell^{+}b\bar{b}$ decays are all of the same size (approximately 2 σ). The measured signal strength of the standard ATLAS analysis is $\mu = 1.20^{+0.42}_{-0.36}$(total) with an observed(expected) significance of 3.5 σ(3.0 σ). Last but not least, the CMS collaboration measured the $V(H \to b\bar{b})$ signal using the 2015+2016 LHC data as well. Their analysis targets the $V(H \to b\bar{b})$ signal process as well and measures it with a signal strength of $\mu = 1.19^{+0.40}_{-0.38}$(total) and an observed(expected) significance of 3.3σ(2.8σ). This is compatible with the results of the analysis presented here and further supports evidence for the $H \to b\bar{b}$ decay.

Correction	Exp. Significance	Expected μ	Obs. Significance	Measured μ
Regression	**2.8 σ**	$\mathbf{1.00^{+0.27}_{-0.26}}$**(stat.)**$\mathbf{^{+0.33}_{-0.26}}$**(syst.)**	**2.9 σ**	**1.15 ± 0.29(stat.)**$\mathbf{^{+0.36}_{-0.30}}$**(syst.)**
Default	2.7 σ	$1.00^{+0.27}_{-0.27}$(stat.)$^{+0.33}_{-0.27}$(syst.)	3.2 σ	1.30 ± 0.29(stat.)$^{+0.37}_{-0.30}$(syst.)
Nominal	2.6 σ	$1.00^{+0.28}_{-0.27}$(stat.)$^{+0.34}_{-0.28}$(syst.)	2.7 σ	1.14 ± 0.29(stat.)$^{+0.37}_{-0.30}$(syst.)
Regression 0 lepton	2.1 σ	$1.00^{+0.35}_{-0.34}$(stat.)$^{+0.37}_{-0.32}$(syst.)	1.4 σ	$0.70^{+0.37}_{-0.36}$(stat.)$^{+0.38}_{-0.35}$(syst.)
Default 0 lepton	2.0 σ	$1.00^{+0.36}_{-0.34}$(stat.)$^{+0.38}_{-0.33}$(syst.)	1.3 σ	$0.68^{+0.38}_{-0.37}$(stat.)$^{+0.40}_{-0.39}$(syst.)
Nominal 0 lepton	1.9 σ	$1.00^{+0.37}_{-0.36}$(stat.)$^{+0.40}_{-0.35}$(syst.)	1.2 σ	$0.67^{+0.38}_{-0.37}$(stat.)$^{+0.41}_{-0.39}$(syst.)
Regression 2 lepton	1.9 σ	$1.00^{+0.43}_{-0.41}$(stat.)$^{+0.42}_{-0.32}$(syst.)	2.8 σ	$1.83^{+0.50}_{-0.48}$(stat.)$^{+0.61}_{-0.44}$(syst.)
Default 2 lepton	1.9 σ	$1.00^{+0.44}_{-0.41}$(stat.)$^{+0.41}_{-0.31}$(syst.)	3.2 σ	$2.10^{+0.50}_{-0.48}$(stat.)$^{+0.62}_{-0.44}$(syst.)
Nominal 2 lepton	1.8 σ	$1.00^{+0.44}_{-0.42}$(stat.)$^{+0.42}_{-0.32}$(syst.)	2.7 σ	$1.73^{+0.49}_{-0.47}$(stat.)$^{+0.57}_{-0.42}$(syst.)

Table 8.10: The obtained results for the $Z(H \rightarrow b\bar{b})$ analysis with different jet correction methods. The expected results based on the simulated data and measured results are provided. The significance of the $V(H \rightarrow b\bar{b})$ signal with respect to the background distribution and the signal strength of the background+signal hypothesis are determined.



(a) no b-jet corrections (b) *default* corrections (c) jet energy regression

Figure 8.17: The $m_{b\bar{b}}$ distributions after the background distributions, except for the diboson process, have been subtracted from the data. Shown are the distributions with a) no dedicated b-jet correction applied, b) the *default* corrections applied and c) the jet energy regression applied. The distributions are normalised according to the values determined by the fit using the BDTs output distributions and additionally weighted by the $V(H \rightarrow b\bar{b})$ signal to background ratio of each bin. The signal distributions are each scaled by the measured μ of each individual fit. The data excess at low $m_{b\bar{b}}$ is due to problems of the background simulation to model this regime and is similarly observed in the $ZH \rightarrow \ell^- \ell^+ c\bar{c}$ analysis.

CHAPTER 9

Search for the Standard Model $H \rightarrow c\bar{c}$ Decay

The LHC will deliver a huge amount of data in the upcoming years — several hundred fb^{-1} — allowing to study physics processes with small cross sections and branching ratios. This is an opportunity to study Higgs boson decays to light fermions, which have reduced BRs compared to heavy fermions since the coupling to the Higgs boson is proportional to the fermion's mass. Therefore most Higgs boson measurements focus on couplings to heavy fermions: top quarks, bottom quarks and τ-leptons. This chapter introduces an analysis for a direct search of the Higgs boson decay to charm quarks. The expected BR of this decay is 2.9% [4]. Only loose experimental constraints on the Higgs boson to charm quark coupling exist. The upper limits obtained in a search for the $H \rightarrow J/\psi\gamma$ decay correspond to an indirect limit on the $H \rightarrow c\bar{c}$ BR of approximately 220 times the SM expectation [92, 93]. Based on the observed Higgs boson decay channels and their measurements the BR of unobserved decay channels, such as $H \rightarrow c\bar{c}$, can be as high as 20% [94] for a SM Higgs boson. The only direct limit is set by the LHCb experiment which uses a 2 fb^{-1} data set of $\sqrt{s} = 8$ TeV LHC data. This analysis sets an upper limit on the product of VH production cross section and $H \rightarrow c\bar{c}$ BR of 6400 times the SM expectation [95]. All these limits do not exclude scenarios in which substantial modifications of the Higgs boson to charm quark couplings could exist.

The direct search for the $H \rightarrow c\bar{c}$ decay presented here relies on the ZH production channel to reduce the amount of multi-jet background events and to sufficiently trigger these events using the Z boson decay products as detailed in the previous chapter. It uses novel c-tagging techniques, as explained in section 5.5.2. The same background processes are relevant for the $H \rightarrow c\bar{c}$ analysis as for the $H \rightarrow b\bar{b}$ analysis, though the background composition differs due to usage of the c-tagging requirements. In addition, due to the high mis-identification probabilities especially the accepted amount of multi-jet events is larger. This is the reason, why only the $ZH \rightarrow \ell^-\ell^+c\bar{c}$ decay channel is investigated at this point in time. This analysis uses the 2015+2016 ATLAS good quality data set of $\sqrt{s} = 13$ TeV which corresponds to 36.1 fb^{-1} and is published in [96]. Overall, due to the challenges of the $H \rightarrow c\bar{c}$ decay channel — the small cross section and BR, high amount of background events and c-jet identification — this analysis is also regarded a feasibility study to proof the concept of direct $H \rightarrow c\bar{c}$ searches using c-tagging techniques.

9.1 Signal and Background Processes

The considered signal is the $ZH \rightarrow \ell^-\ell^+c\bar{c}$ decay. The SM cross section times branching ratio for this process is $\sigma(pp \rightarrow ZH) \times \mathrm{BR}(Z \rightarrow \ell^-\ell^+) \times \mathrm{BR}(H \rightarrow c\bar{c}) = 1.7$ fb, including gluon induced and quark induced ZH production. The experimental signature of this channel are two leptons and two jets. The simulated samples are produced with the same generator and prescriptions as the simulated $V(H \to b\bar{b})$ events, see table 8.2.

The background processes, which are V+jets, $t\bar{t}$, diboson and single top production, are the same as for the $Z(H \to b\bar{b})$ analysis since the same final states are probed. The requirement of c-jets does not introduce other background processes in the analysis. Only the flavour composition of the background processes is expected to change. The SM $ZH \to \ell^-\ell^+b\bar{b}$ production is considered as a background in this analysis. Its SM $\sigma\times$BR is 20 times larger than for the signal process. The same simulated data sets as for the $V(H \to b\bar{b})$ analysis are used. A description of all background processes is given in section 8.1 and summarised in table 8.2.

9.2 Object and Event Selection

The object and event selections are based on the selection of the $Z(H \to b\bar{b})$ analysis. The object definitions summarised in table 8.3 are the same for both analyses. Only an additional jet definition is introduced: c-jets. A *signal* jet is identified as a c-jet if it passes the c-tagging requirements, as described in section 5.5.2. This requirement is optimised for this analysis and details are given below. The overlap removal procedure is the same as for the $Z(H \to b\bar{b})$ analysis. In the following the event selection, which is summarised in table 9.1, and its optimisation are described.

Events of interest are identified with a single lepton trigger. The Z-boson candidate is reconstructed from the two leptons in the event. The $m_{\ell^-\ell^+}$ window and $p_{\mathrm{T}}^{Z} > 75\,\mathrm{GeV}$ requirements are imposed as well. In addition, at least two *signal* jets have to be present in the event and the leading p_{T} jet has to have a p_{T} of larger than 45 GeV. In contrast to the $Z(H \to b\bar{b})$ analysis, the Higgs boson candidate is reconstructed from the leading and sub-leading *signal* jet of the event and the c-tagging requirements are optimised as described in section 5.5.2. The details of the optimisation procedure are explained in section 9.2.1. To increase the amount of selected signal events, events with 1 c-tag are considered in this analysis as well. In the following the di-jet system that represents the Higgs boson candidate is referred to as $c\bar{c}$-system for simplicity. Since this analysis does not make use of a MVA additional selection criteria are assigned based on the ΔR between the two Higgs candidate jets. These distributions exhibit distinct shapes for signal events compared to background events similar to the $Z(H \to b\bar{b})$ analysis. The $\Delta R(c_1, c_2)$ requirements depend on p_{T}^{Z} and decrease for increasing p_{T}^{Z}, as listed in table 9.1. Their optimisation is detailed in section 9.2.1. Due to statistical limitations of the simulated data sets after the application of c-tagging requirements, especially in the 2 c-tags categories, truth tagging is used for all simulated background data sets throughout this analysis. An additional uncertainty is considered to account for differences in the $\Delta R(c_1, c_2)$ distributions between truth tagging and the direct usage of c-tagging requirements. All other kinematic distributions that are relevant for the analysis exhibit a good agreement between truth and direct c-tagging.

Table 9.2 lists the amount of expected and accepted $ZH \to \ell^-\ell^+c\bar{c}$ events in the analysis phase space. Since the $ZH \to \ell^-\ell^+b\bar{b}$ events exhibit a peak in the spectrum of the invariant mass of the Higgs candidate jets than the signal, the amount of selected $ZH \to \ell^-\ell^+b\bar{b}$ is given in the table as well. In total, there are approximately 61 $ZH \to \ell^-\ell^+c\bar{c}$ events expected in the 2015+2016 ATLAS data set. This analysis selects 5 of these events, which corresponds to an acceptance of 8.1%. In comparison, the acceptance of $ZH \to \ell^-\ell^+b\bar{b}$ events is only 4.6%. Nevertheless, there are still 10 times more $ZH \to \ell^-\ell^+b\bar{b}$ events in the analysis phase space than $ZH \to \ell^-\ell^+c\bar{c}$ events. The compositions of the analysis categories are shown in figure 9.1. All categories are dominated by Z+jets events with varying flavour compositions between the categories. The 1 c-tag categories contain more than 50% $Z + ll$ events and sizeable contributions from $Z + cl$ and $Z + (bb, bc, bl)$. Although $Z + ll$ is still the largest contributor in the 2 c-tags

Z boson candidate	Higgs boson candidate
2 *VH loose* leptons $\geq$ 1 *VH tight* lepton $p_{\mathrm{T}}^{Z} > 75\,\mathrm{GeV}$ $81\,\mathrm{GeV} < m_{\ell^-\ell^+} < 101\,\mathrm{GeV}$	$\geq$ 2 *signal* jets $\geq$ 1 c-jet $p_{\mathrm{T}}^{c_1} > 45\,\mathrm{GeV}$
$\Delta R(c_1, c_2) < 2.2$ ($75\,\mathrm{GeV} < p_{\mathrm{T}}^{Z} < 150\,\mathrm{GeV}$) $\Delta R(c_1, c_2) < 1.5$ ($150\,\mathrm{GeV} < p_{\mathrm{T}}^{Z} < 200\,\mathrm{GeV}$) $\Delta R(c_1, c_2) < 1.3$ ($p_{\mathrm{T}}^{Z} > 200\,\mathrm{GeV}$)	

Table 9.1: Event selection for the $ZH \to \ell^-\ell^+c\bar{c}$ analysis.

categories its size is reduced and $Z + cl$, $Z + cc$ and $Z + (bb, bc, bl)$ have similarly large contributions in the order of 25% to 15%. In the $ZH \to \ell^-\ell^+c\bar{c}$ analysis $t\bar{t}$ is a sub-dominant contributor with less than 5.5% in any given region. The diboson contribution is of a similar size as the $t\bar{t}$ background events. Single top events only have a sizeable contribution in the 2 c-tags high p_{T}^{Z} category and contributes less than 1% to the amount of events in the other categories.

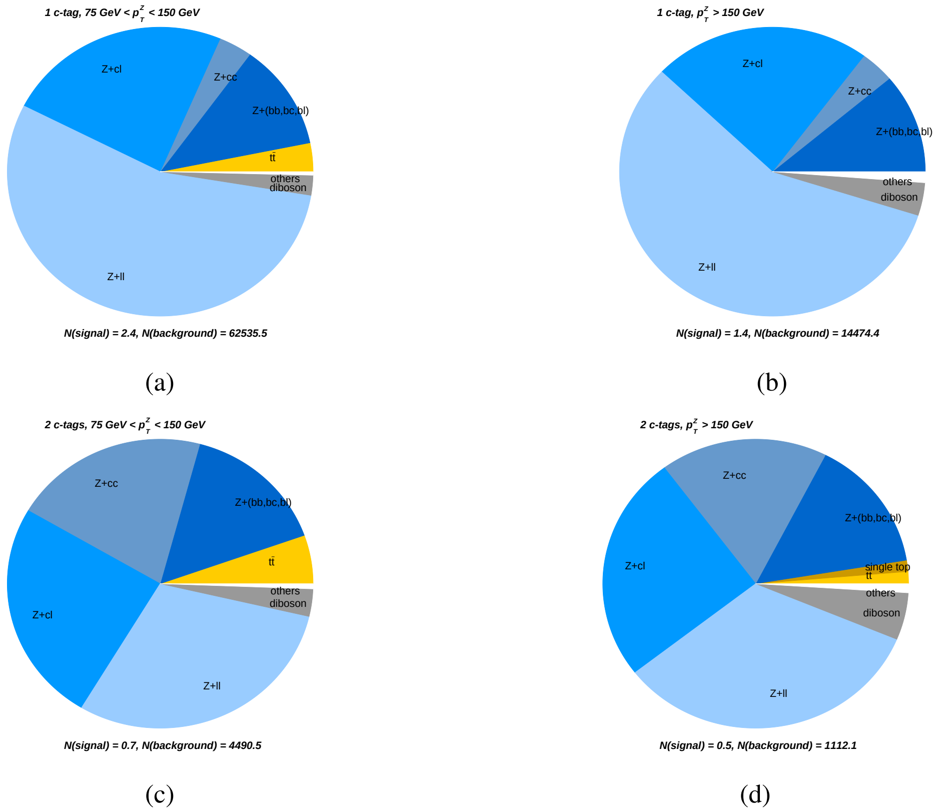
Process	Cross section × BR	Expected N(events)	Selected N(events)	Acceptance
$ZH \to \ell^-\ell^+c\bar{c}$	1.7 fb	61.4	5.0	8.1%
$ZH \to \ell^-\ell^+b\bar{b}$	34.7 fb	1252.7	58.1	4.6%

Table 9.2: The cross section × BR — as predicted by the SM — for the $ZH \to \ell^-\ell^+c\bar{c}$ and $ZH \to \ell^-\ell^+b\bar{b}$ processes, as well as the amount of expected and selected number of events (N(events)) and the overall acceptance. The expected amount of events are calculated using an integrated luminosity of $36.1\,\mathrm{fb}^{-1}$ which corresponds to the 2015+2016 ATLAS data set. The selected amount of events were determined from the amount of simulated signal events that pass the selection criteria. The acceptance is calculated as the fraction of the number of selected to expected events.

The analysis phase is further split into 4 categories of different signal purities:

- **1 c-tag, medium p_{T}^{Z}:** events with $75\,\mathrm{GeV} < p_{\mathrm{T}}^{Z} < 150\,\mathrm{GeV}$ and the leading *signal* jet has to be a c-jet
- **1 c-tag, high p_{T}^{Z}:** events with $p_{\mathrm{T}}^{Z} > 150\,\mathrm{GeV}$ and the leading *signal* jet has to be a c-jet
- **2 c-tags, medium p_{T}^{Z}:** events with $75\,\mathrm{GeV} < p_{\mathrm{T}}^{Z} < 150\,\mathrm{GeV}$ and the leading and sub-leading *signal* jets have to be c-jets
- **2 c-tags, high p_{T}^{Z}:** events with $p_{\mathrm{T}}^{Z} > 150\,\mathrm{GeV}$ and the leading and sub-leading *signal* jets have to be c-jets

The final discriminant in each category is the invariant mass of the leading and sub-leading *signal* jet $m_{c\bar{c}}$. To improve the $m_{c\bar{c}}$ resolution the muon-in-jet correction is applied to the leading and sub-leading *signal* jet. The semi-muonic BR is as high as 6.7% for some c-hadrons [7]. No further jet corrections are applied since the c-jet energy measurement is on average not as much underestimated as the b-jet energy.



(a)

(b)

(c)

(d)

Figure 9.1: The composition of the $ZH \to \ell^-\ell^+c\bar{c}$ analysis categories: a) 1 c-tag, medium p_T^Z, b) 1 c-tag, high p_T^Z, c) 2 c-tags, medium p_T^Z and d) 2 c-tags, high p_T^Z. Processes that contribute less than 1% and summed together in "others" (white). The contribution from signal events is too low and thus included in "others" as well. The amount of signal and total amount of background events are given as N(signal) and N(background) underneath the pie charts.

9.2.1 Event Selection Optimisation

The $ZH \to \ell^-\ell^+c\bar{c}$ event selection differs from the $Z(H \to b\bar{b})$ analysis due to the usage of c-tagging techniques and the $\Delta R(c_1, c_2)$ requirements. Both of them are optimised for this analysis. All optimisation studies include events with $p_\mathrm{T}^Z < 75\,\mathrm{GeV}$.

The c-tagging selection consists of a maximum requirement on the MV2c100 tagger output (b-jet vs. c-jet training) and a minimum requirement on the MV2cl100 tagger output (c-jet vs. light jet training). The optimisation procedure scans through values of "maximum MV2cl100" and "minimum MV2cl100" and imposes these requirements on the leading and sub-leading *signal* jets or only on the leading *signal* jet in the event. The full analysis is repeated for each pair of MV2c100 and MV2cl100 criteria and the expected limit on the signal strength is used as a measure to find optimal cut values. The expected limit is calculated from simulated events based on the invariant mass distribution of the leading and sub-leading *signal* jet. Only events with exactly two *signal* jets are considered for this study since those events allow an unambiguous identification of the Higgs boson candidate jets. Figure 9.2 displays the expected limits as well as the signal and background efficiencies for this parameter scan for events with $p_\mathrm{T}^Z > 150\,\mathrm{GeV}$. The figure shows that a cluster of optimal performance exists around values of MV2c100<0.4 and MV2cl100>0.3, which is the value that is chosen. Jets that pass these requirements are referred to as c-jets. A similar cluster is found for events with $p_{\mathrm{T}}^{Z} < 150\,\mathrm{GeV}$ and the corresponding figures are included in appendix C. The chosen c-tagging requirements yield a signal efficiency of approximately 20% if 2 c-tags are required and 45% if only 1 c-tag is required in the event. This is why events with 1 c-tag are considered in this analysis although they yield a worse limit due to the higher background efficiency. The efficiencies for the total amount of selected background events are approximately 1% and 8% for 2 c-tags and 1 c-tag events, respectively. The described effects are similarly observed in events with $p_{\mathrm{T}}^{Z} < 150\,\mathrm{GeV}$ as well. Besides for the leading and sub-leading *signal* jet the $ZH \to \ell^{-}\ell^{+}c\bar{c}$ event selection does not impose any c- or b-tagging requirements on the additional *signal* jets. This means they are allowed to pass c-jet or b-jet identification requirements. The scenario that requires exactly two c-jets in the events of the 2 c-tags categories and no b-jets in any events deteriorated the expected limit by approximately 9% due to the rejection of additional signal events.

The $\Delta R(c_1, c_2)$ requirements are optimised applying the selection requirements listed in table 9.1, except for $p_{\mathrm{T}}^{Z} > 75\,\mathrm{GeV}$ and $\Delta R(c_1, c_2)$ requirements. They are optimised separately for events with $p_{\mathrm{T}}^{Z} < 150\,\mathrm{GeV}$ and $p_{\mathrm{T}}^{Z} > 150\,\mathrm{GeV}$. Each p_{T}^{Z} regime is further divided into "sub-regimes" based on p_{T}^{Z}:

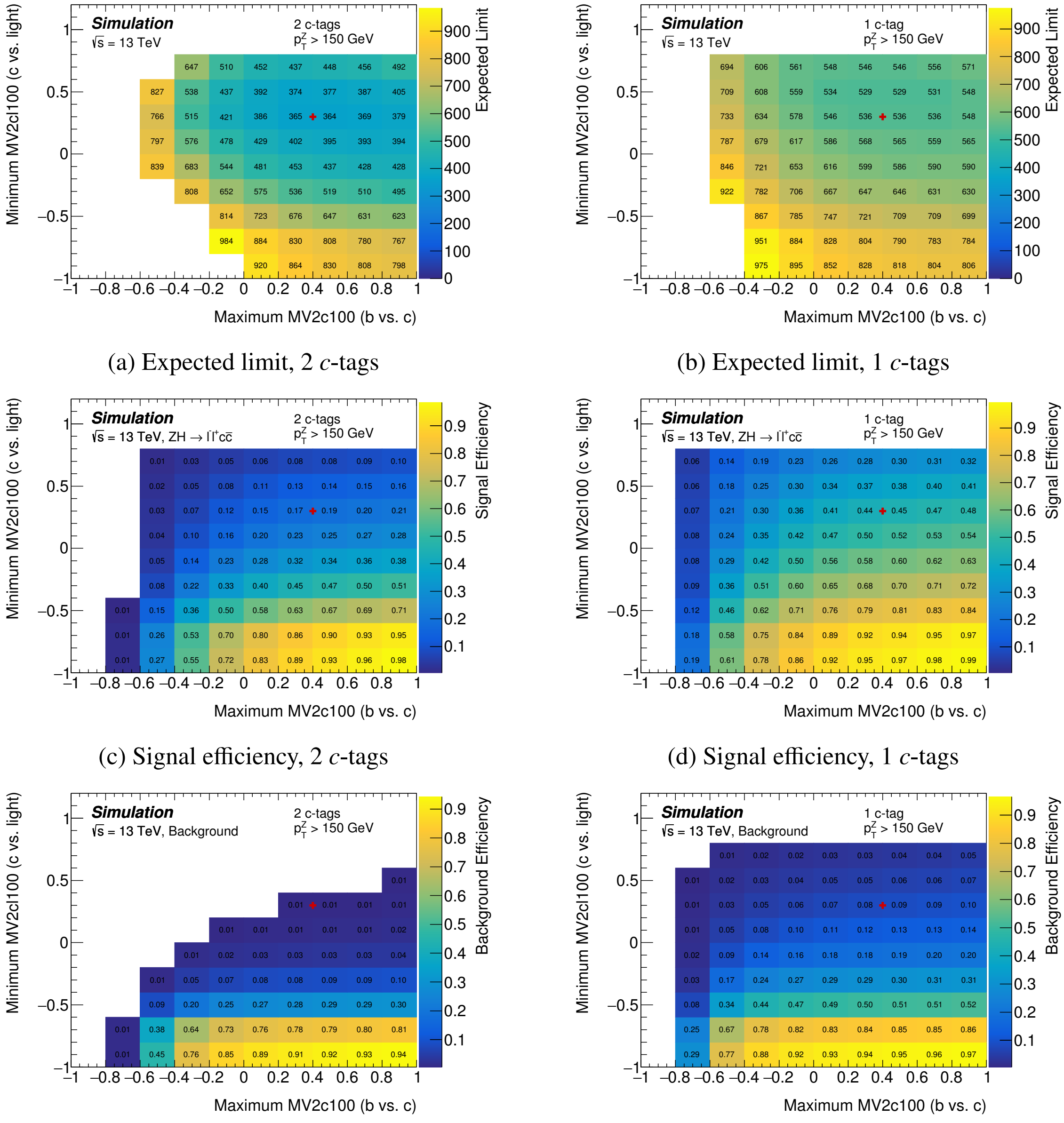
- $p_{\mathrm{T}}^{Z} < 150\,\mathrm{GeV}$: $p_{\mathrm{T}}^{Z} < 75\,\mathrm{GeV}$ and $75\,\mathrm{GeV} < p_{\mathrm{T}}^{Z} < 150\,\mathrm{GeV}$
- $p_{\mathrm{T}}^{Z} > 150\,\mathrm{GeV}$: $150\,\mathrm{GeV} < p_{\mathrm{T}}^{Z} < 200\,\mathrm{GeV}$ and $p_{\mathrm{T}}^{Z} > 200\,\mathrm{GeV}$

The $\Delta R(c_1, c_2)$ distributions for each "sub-regime" are shown in figure 9.3 for the signal and background processes. For $p_{\mathrm{T}}^{Z} < 75\,\mathrm{GeV}$ the signal and background processes both have maxima at high $\Delta R(c_1, c_2)$. The background distribution is flat in all other "sub-regimes" whereas the distinct maxima of the signal distributions move towards low $\Delta R(c_1, c_2)$ for increasing p_{T}^{Z}. In each p_{T}^{Z} regime a scan through different maximum $\Delta R(c_1, c_2)$ requirements for the "sub-regimes" is performed and the performance is assessed based on the expected statistical limit of the p_{T}^{Z} regime based on the $m_{c\bar{c}}$ distribution. The limit scans are displayed in figure 9.3 as well. In both p_{T}^{Z} regimes a cluster of optimal limits exists based on which the final requirements are chosen. Those correspond to the ones listed in table 9.1.

9.3 Systematic Uncertainties

The same set of experimental uncertainties are used in this analysis compared to the $Z(H \to b\bar{b})$ analysis for electrons, muons, jets, pile-up and the luminosity, compare section 8.6.1. An exception are the uncertainties related to the jet flavour identification. To calibrate the c-tagging algorithms and to estimate its uncertainties, the same calibration methods as for the b-tagging algorithms are used. The final set of uncertainties is decomposed into 3 uncertainties for c-jets and b-jets each and 5 uncertainties for light jets. The size of the uncertainties is in the order of 20% for c-jets, 5% for b-jets, and 20% for light jets [46].

The estimate of the $ZH \to \ell^{-}\ell^{+}c\bar{c}$ modelling uncertainties is done in a similar but more simplified way as for the $Z(H \to b\bar{b})$ analysis. The reason for the simplification are the statistical limitations of this analysis, which are present in the simulated data sets as well and thus limit the modelling studies. Therefore the assigned uncertainties are more inclusive in analysis categories and their size represent an upper bound. The modelling uncertainties are derived from comparisons of alternative simulation models. An exception are the $t\bar{t}$ normalisation uncertainties, which are estimated from $e\mu$ CRs. The strategy to design these CRs is similar to the $Z(H \to b\bar{b})$ analysis. The nominal $ZH \to \ell^{-}\ell^{+}c\bar{c}$ selection requirements, see table 9.1, are applied but the two leptons in the event have to be a muon and an electron of opposite charge. The difference between data and MC simulation is assigned as an uncertainty. Since the $e\mu$ CRs contain small contributions from other processes, that do not contain a $Z \to \ell^{-}\ell^{+}$ decay, e.g. single top, this uncertainty includes normalisation uncertainties for these processes as well. Shape



(a) Expected limit, 2 *c*-tags

(b) Expected limit, 1 *c*-tags

(c) Signal efficiency, 2 *c*-tags

(d) Signal efficiency, 1 *c*-tags

(e) background efficiency, 2 *c*-tags

(f) Background efficiency, 2 *c*-tags

Figure 9.2: The expected limits on the signal strength (top), signal efficiencies (middle) and background efficiencies (bottom) obtained for combinations of MV2c100 and MV2cl100 requirements applied to the leading *signal* and sub-leading *signal* jet (left) or just the leading *signal* jet (right). Only events with exactly two *signal* jets are considered. The expected limit corresponds only to the expected limit of the shown analysis category and not the full analysis phase space. The red cross marks the requirement that is chosen for the analysis. Bins of MV2c100 and MV2cl100 that yield limits larger than 1000 or efficiencies smaller than 0.005 are not shown.

uncertainties are assigned to the $m_{c\bar{c}}$ distribution as it is the final discriminant of the analysis. In contrast to the $Z(H \rightarrow b\bar{b})$ analysis, no uncertainties are assigned to the Z+jets flavour fractions since the shape of the $m_{c\bar{c}}$ distribution changes by less than 2% in any given bin of the distribution if a certain flavour component is increased or decreased by 100%. All modelling uncertainties are summarised in table 9.3.

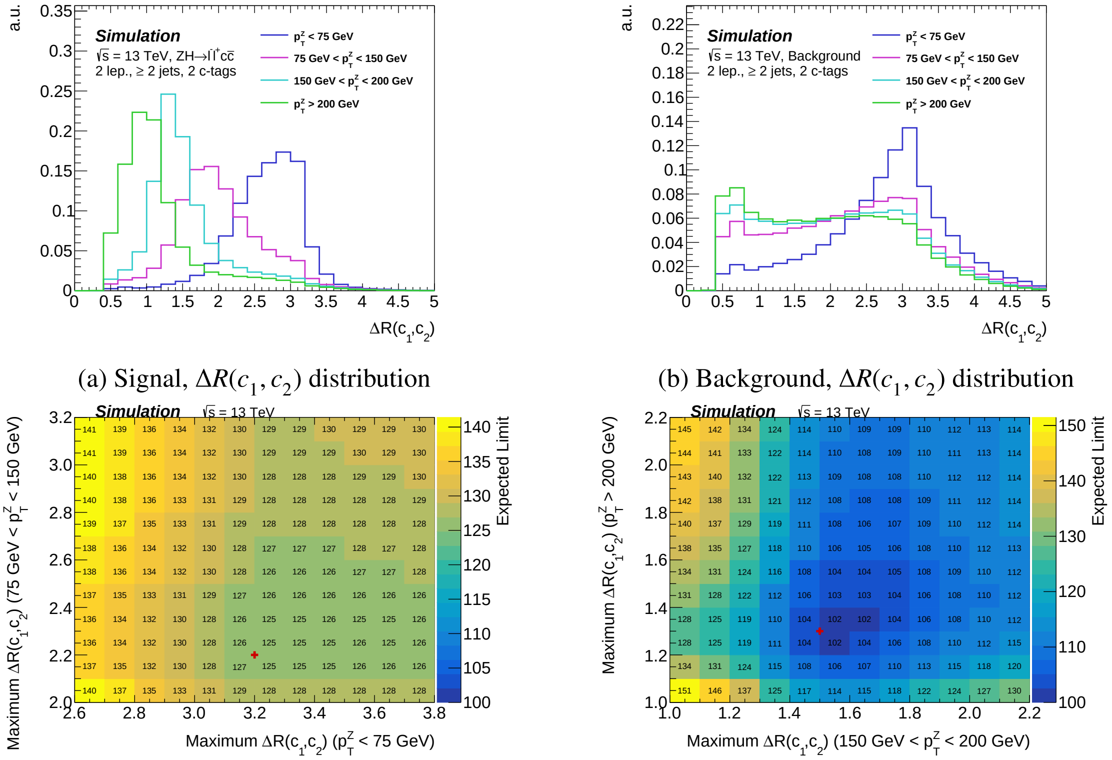
Uncertainty	Source	$ZH \to \ell^- \ell^+ c\bar{c}$		$ZH \to \ell^- \ell^+ b\bar{b}$	
		medium p_T^Z	high p_T^Z	medium p_T^Z	high p_T^Z
$\Delta\sigma(pp \to ZH)$	QCD,PDF, α_S	$^{+4.1\%}_{-3.5\%}$			
ΔBR($H \to c\bar{c}$,)	QCD, EW, m_c,α_S	$^{+5.5\%}_{-2.0\%}$		-	
ΔBR($H \to b\bar{b}$)	QCD, EW, m_b,α_S	-		$^{+1.2\%}_{-1.3\%}$	
$\Delta_\mathrm{acc.}$(analysis phase space)	ME, PS	5%	5%	5%	5%
$\Delta_\mathrm{acc.}$(high p_T^Z w.r.t. medium p_T^Z)	missing higher orders (EW)	-	3%	-	3%
$\Delta_\mathrm{shape}(m_{c\bar{c}})$	PS	S	S	S	S

Uncertainty	Source	Z+jets			
		1 c-tag, medium p_T^Z	2 c-tags, medium p_T^Z	1 c-tag, high p_T^Z	2 c-tags, high p_T^Z
$\Delta_\mathrm{norm.}(Z + (bb, bc, bl, cc, cl))$	-	float	float	float	float
$\Delta_\mathrm{norm.}(Z + ll)$	-	float	float	float	float
$\Delta_\mathrm{shape}(m_{c\bar{c}})$, $Z + (bb, bc, bl, cc, cl)$	various	S	S	S	S
$\Delta_\mathrm{shape}(m_{c\bar{c}})$, $Z + ll$	various	S	S	S	S

Uncertainty	Source	ZZ		WZ	
		medium p_T^Z	high p_T^Z	medium p_T^Z	high p_T^Z
$\Delta_\mathrm{norm.}$	various	5%		5%	
$\Delta_\mathrm{acc.}$(analysis phase space)	various	13%	13%	12%	12%
$\Delta_\mathrm{shape}(m_{c\bar{c}})$	various	S	S	S	S

Uncertainty	Source	$t\bar{t}$			
		1 c-tag, medium p_T^Z	2 c-tags, medium p_T^Z	1 c-tag, high p_T^Z	2 c-tags, high p_T^Z
$\Delta_\mathrm{norm.}$	various	14%		38%	
$\Delta_\mathrm{acc.}$(high p_T^Z w.r.t. medium p_T^Z)	various	-	-	7%	7%
$\Delta_\mathrm{shape}(m_{c\bar{c}})$	PS	S	S	S	S

Table 9.3: Modelling uncertainties assigned to the simulated signal and background processes. Given is the type of the systematic uncertainty, the source and the size for each analysis category and background process. Shape effects are labelled as "S" and normalisations that are determined from data as "float". If the source is given as "various" the uncertainty was derived from the comparison to a simulation from a different MC generator which incorporates different PDF sets, tunes, PS models, scales, merging schemes etc. or from a comparison to data. If the source is given as "ME" the uncertainty was derived from a comparison to a model of a MC generator that uses the same PS model as the nominal model but has a different ME generator which also includes PDF differences, merging between ME and PS differences etc. If not further specified the same uncertainties are used in the 1 c-tag and 2 c-tags categories.



(a) Signal, $\Delta R(c_1, c_2)$ distribution

(b) Background, $\Delta R(c_1, c_2)$ distribution

(c) Expected limit scan, $p_{\mathrm{T}}^Z < 150\,\mathrm{GeV}$

(d) Expected limit scan, $p_{\mathrm{T}}^Z > 150\,\mathrm{GeV}$

Figure 9.3: The $\Delta R(c_1, c_2)$ distributions in signal and background events for different p_{T}^Z ranges (top) and the expected limits for a scan through maximum $\Delta R(c_1, c_2)$ requirements for $p_{\mathrm{T}}^Z < 75\,\mathrm{GeV}$ and $p_{\mathrm{T}}^Z > 150\,\mathrm{GeV}$ events (bottom). The expected limit corresponds only to the displayed region itself and not the full analysis phase space. The red cross marks the requirements chosen for the analysis.

9.4 Statistical Analysis

A fit to data is performed and no signal is found. Therefore an upper limit on the signal strength of the $Z(H \to c\bar{c})$ process, which can be excluded at a 95% confidence level, is set. The limit calculation uses the CL_S method — a modified frequentist method [52, 54]. To extract it a profile likelihood fit, implemented in the RooStats Framework, is employed.

The following components, as given by their simulated data, are considered in the fit:

- $\boldsymbol{ZH \to \ell^-\ell^+ c\bar{c}}$ **signal**
- **SM** $\boldsymbol{ZH \to \ell^-\ell^+ b\bar{b}}$
- ***Z*+jets**: $Z + (bb, bc, bl, cc, cl)$, $Z + ll$; significant differences in the shape of the $m_{c\bar{c}}$ distribution are only observed for those two components. Therefore no finer split of the flavour components is introduced in the fit model
- $\boldsymbol{t\bar{t}}$: the uncertainties on the $t\bar{t}$ template are chosen such that it also encompasses the contributions from single top background processes
- **diboson**: ZZ, WZ

The normalisation of the $Z + (bb, bc, bl, cc, cl)$ and $Z + ll$ components are free floating in the fit separately in each analysis category. Their normalisation is determined from data, which is allowed by the large amount of those events in the analysis categories.

The uncertainties described in section 9.3 are introduced as nuisance parameters that act as Gaussian priors. The experimental uncertainties are treated as fully correlated across categories and processes. Each uncertainty listed in table 9.3 is treated in a correlated way across regions but not across processes. An exception are the $ZH \to \ell^-\ell^+c\bar{c}$ and $ZH \to \ell^-\ell^+b\bar{b}$ uncertainties which are correlated apart from the BR uncertainties. The latter is motivated by the difference in their size originating from the uncertainty on the charm quark mass. The statistical error due to the limited size of the MC simulated data sets are taken into account as nuisance parameters with Poissonian priors. The nuisance parameters are smoothed using Gaussian kernel density estimation [97]. The threshold for pruning nuisance parameters is 2%, i.e. a nuisance parameter that does not cause a variation larger than 2% in any given bin in any analysis category is pruned.

The final discriminant are the binned $m_{c\bar{c}}$ distributions in the four analysis categories. The bin width is 10 GeV. The $m_{c\bar{c}}$ distributions are restricted to a window around the expected Higgs boson mass: $50\,\text{GeV} < m_{c\bar{c}} < 200\,\text{GeV}$.

9.5 $ZH \to \ell^-\ell^+c\bar{c}$ Results

The $m_{c\bar{c}}$ distributions normalised according to the fit results are shown in figure 9.4. A good agreement between data and the simulated events is achieved after the fit. The histograms demonstrate that the 2 c-tags high p_{T}^{Z} category has the highest signal to background ratio. The observed upper limit on the signal strength is:

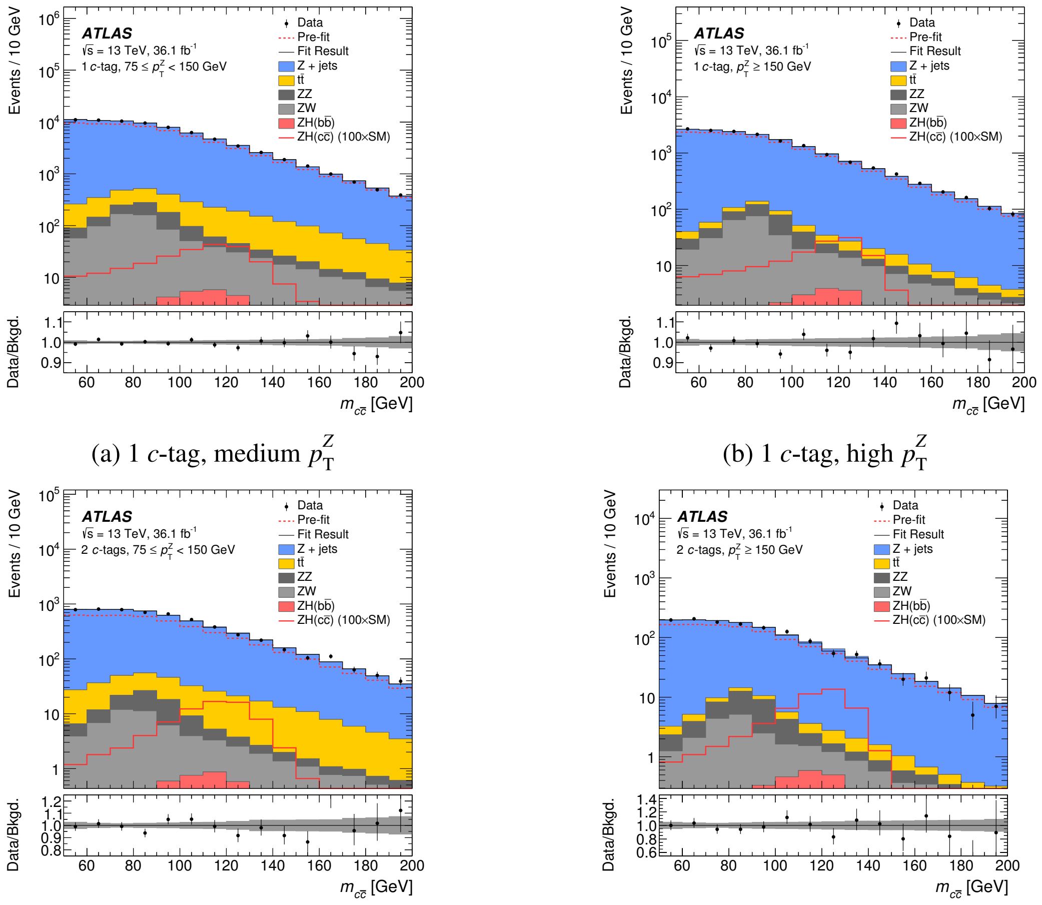
$$\mu < 110 \tag{9.1}$$

compared to an expected limit of 150^{+80}_{-40}. To assess the uncertainties that impact the analysis result the most, the break down of uncertainties is determined in the same way as for the $Z(H \to b\bar{b})$ analysis. The signal strength is measured and the nuisance parameters of a given set of uncertainties are fixed to their values determined by the fit and the fit is performed again only including the remaining uncertainties as nuisance parameters. The uncertainty break down is listed in table 9.4. It shows that the analysis is already limited by systematic uncertainties and the c-tagging uncertainties are the most dominant ones. Nevertheless it is important to keep in mind that the c-tagging uncertainties are partially impacted by statistical uncertainties themselves since the c-tagging efficiency limits the data set that is used for the efficiency calibration.

9.5.1 Diboson Cross Check

Similar to the $Z(H \to b\bar{b})$ analysis the idea is to validate the $ZH \to \ell^-\ell^+c\bar{c}$ analysis using the diboson process as the signal. This diboson analysis uses the same event selection as the $ZH \to \ell^-\ell^+c\bar{c}$ analysis. The fraction of selected ZZ and WZ events in the $ZH \to \ell^-\ell^+c\bar{c}$ analysis phase space is roughly equal in total. The flavour composition of the V-boson candidate, given by the leading and sub-leading *signal* jet with either one or two c-tags, is shown in figure 9.5. Similar histograms for the other analysis categories are included in appendix C. The WZ events contribute mainly in the 1 c-tag categories and 65% of the WZ event in those categories are $W \to cs, cd$ decays. The ZZ events contribute mainly in the 2 c-tags categories with 55% of ZZ events being $Z \to c\bar{c}$ decays.

The diboson analysis uses the same fit model as the $ZH \to \ell^-\ell^+c\bar{c}$ analysis including the same nuisance parameters, except for the overall diboson normalisation nuisance parameters which are removed from



(a) 1 c-tag, medium p_{T}^{Z}

(b) 1 c-tag, high p_{T}^{Z}

(c) 2 c-tag, medium p_{T}^{Z}

(d) 2 c-tags, high p_{T}^{Z}

Figure 9.4: The $m_{c\bar{c}}$ distributions of the $ZH \to \ell^{-}\ell^{+}c\bar{c}$ analysis regions [96]. All signal and background components are scaled to their normalisations as determined by the fit. The normalisation of the sum of all background components as predicted by the simulated events is given by the dashed red line. The solid red line shows an overlay of the $ZH \to \ell^{-}\ell^{+}c\bar{c}$ signal distribution scaled by a factor 100. The shaded bands represent the total uncertainty. A version of the same histograms without the logarithmic y-axis is shown in appendix C.

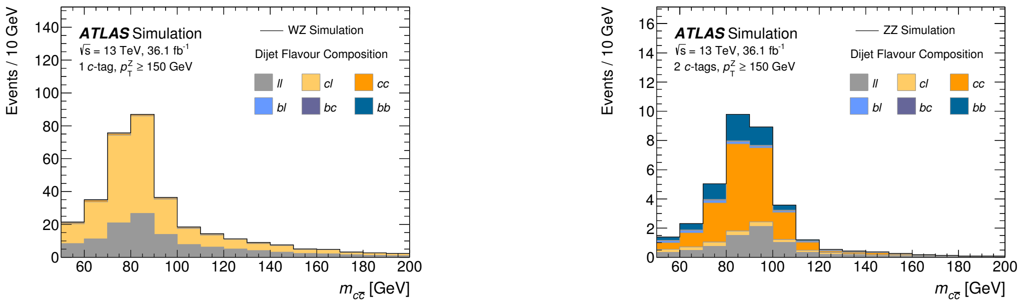
the fit. The $ZH \to \ell^{-}\ell^{+}c\bar{c}$ process is fixed to its SM expectation and the parameter of interest is instead the signal strength of the diboson, sum of ZZ and WZ, events. The final discriminant and analysis categories remain unchanged. The diboson process is found with a observed (expected) significance of 1.4(2.2) σ. The measured signal strength is:

$$\mu = 0.6^{+0.5}_{-0.4} \tag{9.2}$$

Given the small obtained significance it is not possible to employ the diboson analysis as a validation for the $ZH \to \ell^{-}\ell^{+}c\bar{c}$ analysis at this point in time. Nevertheless it is a potential cross check for future analysis. In addition, it probes W and Z decays that are usually not probed in other ATLAS analyses and the measurement of these decay channels is interesting by itself.

Set of uncertainties	Relative impact on error
Stat.	49%
Floating Z+jets normalisations	31%
Syst.	87%
c-tagging	73%
Background modelling	47%
Electrons, muons, jets and luminosity	28%
Signal uncertainties	28%
MC stat.	6%

Table 9.4: The break down of the sources of uncertainties [96]. Given is the relative error on the measured μ that is caused by each group. The sum of all exceeds 100% since the defined groups have correlations amongst each other. The floating Z+jets normalisations, which are extracted from data, are included in the statistical uncertainties since their impact is influenced by the available data statistics. "MC statistical" refers to the statistical uncertainty of the simulated data sets. "Modelling" refers to the uncertainties of the description of the given process in the utilised simulated data set. "Signal uncertainties" includes modelling and theory uncertainties.



(a) WZ, 1 c-tag, high p_{T}^{Z}

(b) ZZ, 2 c-tags, high p_{T}^{Z}

Figure 9.5: The flavour composition of the a) WZ events in the 1 c-tag high ptz category and the b) ZZ events in the 2 c-tags high p_{T}^{Z} category [96].

9.6 Outlook

The presented $ZH \rightarrow \ell^-\ell^+c\bar{c}$ is a proof of concept for direct $H \rightarrow c\bar{c}$ searches. However, the analysis is far from observing the SM $H \rightarrow c\bar{c}$ decay. If the $ZH \rightarrow \ell^-\ell^+c\bar{c}$ signal and background processes are scaled approximately by a factor 10 and 100, to an equivalent luminosity of 300 fb^{-1} and 3 000 fb^{-1}, the expected statistical limits are approximately $\mu < 25$ and $\mu < 7$, respectively. Even lower limits may be achieved if for future analysis the $ZH \rightarrow \nu\bar{\nu}c\bar{c}$ and $W^{\pm}H \rightarrow \ell^{\pm}\overset{(-)}{\nu}c\bar{c}$ channel are included, provided that the challenging backgrounds, especially multi-jet production, and their compositions can be modelled correctly. This chapter discusses two other aspects that will become important as well: the correlation and combination with $V(H \rightarrow b\bar{b})$ searches and the possibility to exploit a MVA.

9.6.1 MVA

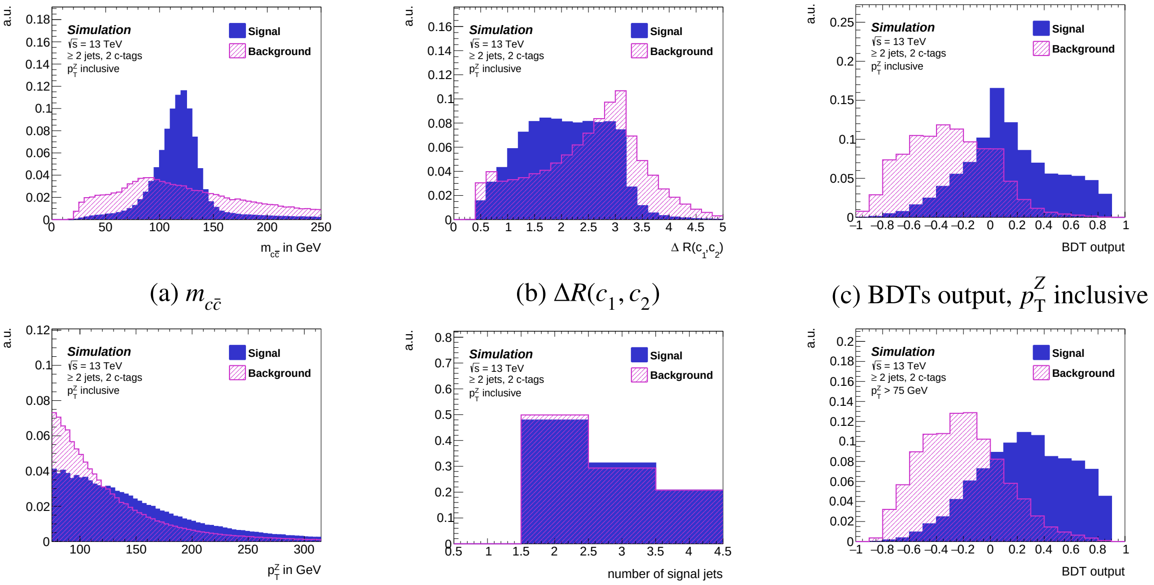
One possible way to increase the sensitivity of the analysis is to exploit a MVA. The possible gain is studied based on the current analysis. The main focus is to set-up a simple MVA utilising a small amount of input variables and a reduced amount of analysis categories. Therefore as a first approach the most sensitive variables of the $ZH \to \ell^-\ell^+c\bar{c}$ analysis are used as input variables for a BDTs training: $m_{c\bar{c}}$, $\Delta R(c_1, c_2)$ and p_{T}^{Z}. Introducing the latter two allows the BDTs to make use of the correlation between them in a more exhaustive way than the standard set of $\Delta R(c_1, c_2)$ requirements. It is found that the number of *signal* jets as bins of 2 jets, 3 jets and 4 or more jets increases the performance, measured as the binned significance of the BDTs output distribution, by approximately 15%. Thus it is introduced as the fourth input variable. The distributions of these four variables in simulated signal and background events are shown in figure 9.6. The BDTs are trained using simulated events that passed the $ZH \to \ell^-\ell^+c\bar{c}$ event selection with a few exceptions: the training uses the full p_{T}^{Z} range, only 2 c-tag events are used and the $\Delta R(c_1, c_2)$ requirements are not applied. Different sets of training parameters — number of trees, depth of the trees and number of cuts allowed on the input variables' distributions — are tested to find the optimal one amongst them, which proved to have a large influence on the performance with differences up to 50% between parameters sets. The BDTs output distributions are shown in figure 9.6 for the training phase space (p_{T}^{Z} inclusive) and the standard $ZH \to \ell^-\ell^+c\bar{c}$ analysis phase space that only includes events with $p_{\mathrm{T}}^{Z} > 75\,\mathrm{GeV}$. The p_{T}^{Z} inclusive BDTs output distributions exhibit a distinct peak at around 0.1. This is caused by events with $p_{\mathrm{T}}^{Z} < 75\,\mathrm{GeV}$ since the $\Delta R(c_1, c_2)$ distributions for the signal and background processes are very similar, compare figure 9.3, and the BDTs can only make use of the $m_{c\bar{c}}$ information. This effect is mitigated if only the standard analysis phase space of $p_{\mathrm{T}}^{Z} > 75\,\mathrm{GeV}$ is considered.

Using the BDTs output distribution as the discriminant in the fit improves the expected statistical limit by 15% with respect to the standard analysis if 2 c-tags events with $p_{\mathrm{T}}^{Z} > 75\,\mathrm{GeV}$ are considered. Another advantage is that no split in p_{T}^{Z} categories is introduced, which reduces the number of analysis categories by a factor of two. The same BDTs training, although not fully optimal since this phase space is not included in the training, is also applied to 1 c-tag events. The combined performance of the 2 c-tags BDTs output and 1 c-tag BDTs output also predicts a 15% better statistical limit compared to the standard $ZH \to \ell^-\ell^+c\bar{c}$ analysis that uses $m_{c\bar{c}}$ as the discriminant and includes four analysis categories. In conclusion, a simple multivariate analysis which uses the 4 most sensitive variables of the $ZH \to \ell^-\ell^+c\bar{c}$ analysis provides an improvement in the order of 15% in the expected sensitivity compared to the baseline analysis and the amount of analysis categories is halved.

9.6.2 Correlation with $V(H \to b\bar{b})$

The $V(H \to b\bar{b})$ and $V(H \to c\bar{c})$ signal processes have the same final state signature. They may only be distinguished to a certain extent by utilising b- and c-tagging techniques. Since the $V(H \to b\bar{b})$ signal events peak at the Higgs boson mass as well it is important to know how this influences the $V(H \to c\bar{c})$ result. This test is performed for the current analysis: The $ZH \to \ell^-\ell^+b\bar{b}$ signal is scaled to 0 and 2 times the SM expectation and the observed $ZH \to \ell^-\ell^+c\bar{c}$ limit is computed, as documented in [96]. At the current sensitivity the observed limit changes by 5%. Nevertheless this impact will grow if the sensitivity of the $ZH \to \ell^-\ell^+c\bar{c}$ analysis grows. Thus finding ways to suppress the $V(H \to b\bar{b})$ contribution or to separate the $V(H \to b\bar{b})$ and $V(H \to c\bar{c})$ signal is a crucial part in the future. Currently, new flavour tagging algorithms are studied and developed based on neural nets that improve the c-tagging efficiency. Another option is to directly train a multivariate algorithm to distinguish $V(H \to b\bar{b})$ and $V(H \to c\bar{c})$ events. Possible features that may be exploited are the output of the c-tagging and b-tagging tagger variables for different jet flavours together with other kinematic information of the event. However, this would require an efficiency calibration of the full distributions instead of calibrations for a specific tagging requirement.

Furthermore, for future analyses it is important to avoid overlap between the $V(H \to b\bar{b})$ and $V(H \to c\bar{c})$ analysis phase space such that a combination of the results is possible. At the current level



(a) $m_{c\bar{c}}$

(b) $\Delta R(c_1, c_2)$

(c) BDTs output, p_T^Z inclusive

(d) p_T^Z

(e) number of *signal* jets

(f) BDTs output, $p_T^Z > 75\,\text{GeV}$

Figure 9.6: Distribution of the $ZH \to \ell^-\ell^+c\bar{c}$ MVA input variables for simulated signal and background events: a) $m_{c\bar{c}}$, b) $\Delta R(c_1, c_2)$, d) p_T^Z and e) number of *signal* jets. The BDTs output distributions are shown inclusive in p_T^Z (c) and for events with $p_T^Z > 75\,\text{GeV}$ (f). All events are normalised to the same area. The peak at BDTs output around 1 in (c) is caused by events with $p_T^Z < 75\,\text{GeV}$.

the $V(H \to b\bar{b})$ analysis contains events of the $ZH \to \ell^-\ell^+c\bar{c}$ analysis and vice versa. To achieve orthogonality a combination of b-tagging and c-tagging requirements has to be applied. In order to do so, the correlation between both tagging techniques have to be taken into account. Another option is to extract the $V(H \to b\bar{b})$ and $V(H \to c\bar{c})$ signal strength from a simultaneous fit. Nevertheless the challenges remain that both signal processes are difficult to disentangle and that the $V(H \to b\bar{b})$ signal is the dominating signal contribution.

CHAPTER 10

Search for a Heavy Scalar Boson $A \rightarrow ZH$

When the LHC run 2 started in 2015 it was the first time that pp collisions were recorded at an energy of $\sqrt{s} = 13\,\text{TeV}$. This increase in the collision energy compared to run 1 increased the probability to produce new, especially heavy particles. The analysis presented here searches for a new CP-odd Higgs boson A. Such a boson A is predicted by theories which include an extended Higgs sector where the discovered Higgs Boson H is only one of many other Higgs bosons. The assumption for this analysis is that A decays into a SM Z boson and a CP-even Higgs boson that corresponds to the discovered Higgs boson with a mass of 125 GeV. The analysis is designed to probe the same final states as the SM $Z(H \rightarrow b\bar{b})$ analysis, i.e. $ZH \rightarrow \nu\bar{\nu}b\bar{b}$ (0 lepton channel) and $ZH \rightarrow \ell^-\ell^+b\bar{b}$ (2 lepton channel). Besides the increase in the collision energy for run 2, an increase in the luminosity also delivered more pile-up events. These new conditions made it necessary to carefully study trigger performance, object reconstruction, event kinematics and new MC simulations. Therefore the search for the $A \rightarrow ZH$ decay provides a good opportunity to study the involved objects and background processes early on to perform the SM $Z(H \rightarrow b\bar{b})$ analysis later with a larger dataset as detailed in chapter 8. Many choices for the SM analysis are based on the experience from the $A \rightarrow ZH$ analysis.

The search for the $A \rightarrow ZH$ decay uses the 2015 ATLAS data set which corresponds to an integrated luminosity of $3.2\,\text{fb}^{-1}$. Masses of the A boson of 220 GeV up to 2 TeV are investigated. Similar searches in the same final state were already carried out at LEP [98] as well as with ATLAS and with CMS using LHC data obtained at $\sqrt{s} = 8\,\text{TeV}$ [99, 100]. No new effects were discovered so far.

10.1 Signal and Background Processes

The signal process for this analysis is the gluon induced production of a new CP odd boson A, as shown in figure 10.1. The selection is optimised to target the A boson decay into ZH and subsequent decays into $ZH \rightarrow \nu\bar{\nu}b\bar{b}$ and $ZH \rightarrow \ell^-\ell^+b\bar{b}$. Hence the signature of the signal process are two b-jets and either a larger amount of $E_{\text{T}}^{\text{miss}}$ or two muons or two electrons. The decay into τ-lepton is not considered for the same reasons as aforementioned. No specific production cross section, branching ratio for the $A \rightarrow ZH$ decay or mass is assumed for the A boson. The analysis probes A bosons with masses between 220 GeV and 2 TeV. The BR of the investigated Z boson and Higgs boson decay channels are fixed to their SM values, as listed in table 8.1. To study the expected signal signatures various data sets of simulated events with different A mass hypotheses are used. The $A \rightarrow ZH$ process is simulated with the MadGraph5_aMC@NLO generator, which is interfaced to Pythia for the simulation of the PS and UE. The PS/UE uses the A14 set of tunes. The NNPDF2.3LO [101] PDF set it used for the ME calculation. This simulation provides event weights to assess the effect of varied μ_{F}, μ_{R} and tuning parameters. In addition, the event weights allow to asses the effect of different PDF sets. The narrow width approximation is used in all simulations assuming that the decay width of the A boson is much smaller than its mass and the detector resolution.

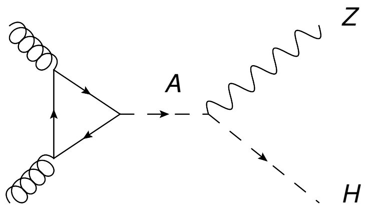


Figure 10.1: Feynman diagram for the gluon induced production of a new boson A decaying to ZH.

Since the same final states are probed as for the SM $Z(H \rightarrow b\bar{b})$ search the processes that contribute as background events are the same. Details can be found in section 8.1. In addition, the SM $V(H \rightarrow b\bar{b})$ process is considered as a background process and the cross sections and branching ratios are fixed to the values predicted by the SM. The choice of MC generators used for the simulation of background events differs between the $A \rightarrow ZH$ and SM $Z(H \rightarrow b\bar{b})$ analysis. The MC generators used for the $A \rightarrow ZH$ are listed in table 10.1. In many cases the simulation models were improved to better describe the run 2 data based on experience gained with the early run 2 analyses, such as the $A \rightarrow ZH$ analysis.

10.2 Object and Event Selection

The object selection criteria are the same as defined in section 8.2 with two small exceptions: First, the p_{T} threshold for leptons to be defined as *VH tight* is 25 GeV since the thresholds of the single lepton trigger is lower for 2015 data. Second, the algorithm used to identify b-jets is the MV2c20 algorithm. The chosen b-tagging requirement also corresponds to a 70% identification efficiency but the mis-identification efficiencies are larger, see section 5.5.1. All other object definitions are the same as summarised in table 8.3.

Overall the analysis phase spaces of the $A \rightarrow ZH$ and $Z(H \rightarrow b\bar{b})$ analysis are very similar since the same final states are probed. However, the kinematic distributions of the final state particles are expected to depend on the A boson mass and thus also differ from the SM $Z(H \rightarrow b\bar{b})$ scenario. Nevertheless no A boson mass dependent selection criteria are applied since the analysis is expected to be limited by the statistical uncertainty of the data, which requires a less restrictive selection of events. Therefore differences in the event selection with respect to the SM $Z(H \rightarrow b\bar{b})$ analysis originate from the necessity to widen the phase space to increase the amount of accepted signal events since the investigated data set is 10 times smaller.

The selection of the Higgs boson candidate and additional *signal* jets is different from the SM $Z(H \rightarrow b\bar{b})$ search. The phase space of the 0 lepton channel is increased by allowing any amount of *signal* jets in addition to the minimum required amount of 2 *signal* jets. As a result, the b-tagging requirements are imposed on the leading *signal* jets — instead of searching for any two b-jets amongst the *signal* jets — to minimise the increase of background events due to mis-identified jets. To increase the amount of accepted signal events only the leading *signal* jet has to pass the b-tagging requirement, optionally also the sub-leading *signal* jet may pass it. However, besides the leading and sub-leading *signal* jet, no other b-jets are allowed in the events. The Higgs boson candidate in the $A \rightarrow ZH$ search is reconstructed from the leading and sub-leading *signal* jet and its mass, referred to as $m_{b\bar{b}}$, is restricted to be within 110 GeV and 140 GeV. Events that do not fall inside this mass window are used to define

Process	PDF	ME	PS+UE	tune	variations
$gg \to A \to (Z \to \nu\bar{\nu}/\ell^-\ell^+)(H \to b\bar{b})$	**NNPDF2.3LO**	**MadGraph5_aMC@NLO**	**Pythia 8**	**A14**	μ_F, μ_R, PDF, A14
$q\bar{q} \to VH$	**NNPDF2.3LO**	**Pythia**	**Pythia**	**A14**	
$gg \to ZH$	**CT10**	**Powheg**	**Pythia**	**AZNLO**	
V+jets	**CT10**	**Sherpa**	**Sherpa**	**Sherpa**	μ_F, μ_R, resummation scale, merging scale
	NNPDF2.3LO	MadGraph5_aMC@NLO	Pythia	A14	
$t\bar{t}$	**CT10**	**Powheg**	**Pythia**	**Perugia 2012**	high radiation, low radiation
	CT10	Powheg	Herwig	UE-EE-5	
	CT10	MadGraph5_aMC@NLO	Herwig	UE-EE-5	
singletop	**CT10**	**Powheg**	**Pythia**	**Perugia 2012**	
VV	**CT10**	**Sherpa**	**Sherpa**	**Sherpa**	

Table 10.1: Monte Carlo generators and their parameters utilised in the $A \to ZH$ analysis. The nominal MC generator recommendation for ATLAS analysis of 2015 data are given in bold letters for each physics process. All other MC generators are alternative simulations and are used to assess systematic uncertainties.

control regions (CRs). The $m_{b\bar{b}}$ resolution is improved with the muon-in-jet and a preliminary version of the *PtReco* correction which is only applied to the b-jets of the di-jet system[1]. The 4-momentum of the di-jet system is additionally scaled by $m_H/m_{b\bar{b}}$. The outlined strategy is adopted for the 0 and 2 lepton channel.

The $A \to ZH$ analysis targets a large range of the A boson masses. For A boson masses above approximately 1 TeV, its decay products could have large transverse momenta and the b-jets from the Higgs boson decay could overlap. A dedicated analysis is developed to reconstruct this event topology, as demonstrated in [102]. These strategy is not considered for this analysis since the main focus is to study final states similar to the ones expected for the SM $Z(H \to b\bar{b})$ analysis. To exclude events that tend to contain overlapping jets an additional requirement is imposed on the p_{T} of the Z boson (represented by $E_{\mathrm{T}}^{\mathrm{miss}}$ in the 0 lepton channel), which has to be smaller than 500 GeV. The 0 and 2 lepton specific event selection criteria are described below and all selection criteria are summarised in table 10.2.

<table>
<tr><th colspan="2">Common selection</th></tr>
<tr><td colspan="2">≥ 2 <i>signal</i> jets
1 or 2 b-jets
≥ 1 b-jet with $p_{\mathrm{T}} > 45\,\mathrm{GeV}$
$110\,\mathrm{GeV} < m_{b\bar{b}} < 140\,\mathrm{GeV}$</td></tr>
<tr><th>0 lepton selection</th><th>2 lepton selection</th></tr>
<tr><td>0 <i>VH loose</i> leptons</td><td>2 <i>VH loose</i> leptons</td></tr>
<tr><td>0 <i>VH tight</i> leptons</td><td>≥ 1 <i>VH tight</i></td></tr>
<tr><td>$150\,\mathrm{GeV} < E_{\mathrm{T}}^{\mathrm{miss}} < 500\,\mathrm{GeV}$</td><td>$p_{\mathrm{T}}^{Z} < 500\,\mathrm{GeV}$</td></tr>
<tr><td>$\sum_{i=1}^{N_{\mathrm{jet}}=2(n)} p_{\mathrm{T}}^{i} > 120\,\mathrm{GeV}(150\,\mathrm{GeV})$</td><td>$70\,\mathrm{GeV} < m_{\ell^-\ell^+} < 100\,\mathrm{GeV}$</td></tr>
<tr><td>no hadronic τ-leptons</td><td>$E_{\mathrm{T}}^{\mathrm{miss}}/\sqrt{H_{\mathrm{T}}} < 3.5\ \sqrt{\mathrm{GeV}}$</td></tr>
<tr><td>$p_{\mathrm{T}}^{\mathrm{miss}} > 30\,\mathrm{GeV}$</td><td></td></tr>
<tr><td>multijet suppression cuts</td><td></td></tr>
</table>

Table 10.2: Event selection criteria for the SRs of the $A \to ZH$ analysis. Numbers in brackets for the 0 lepton selection indicate the requirements if there are 3 or more instead of 2 *signal* jets present in the event.

The remaining 0 lepton channel specific selections are very similar to the SM $Z(H \to b\bar{b})$ search. A minimum measured $E_{\mathrm{T}}^{\mathrm{miss}}$ of 150 GeV is required since the same trigger requirements as for the SM search are used to identify 0 lepton events. An additional requirement to suppress non-collision background events is introduced: $p_{\mathrm{T}}^{\mathrm{miss}} > 30\,\mathrm{GeV}$, which is the missing transverse momentum reconstructed from tracks instead of calorimeter clusters. Furthermore, since τ-leptons are not considered in the overlap removal procedure for this analysis, the events in the 0 lepton channel are required not to contain any hadronically decaying τ-lepton.

The 2 lepton channel, in contrast to the SM analysis, does not impose a lower limit on p_{T}^{Z} to increase the amount of accepted signal events. The requirement on the invariant mass of the dilepton system $m_{\ell^-\ell^+}$ is loosened to $70\,\mathrm{GeV} < m_{\ell^-\ell^+} < 110\,\mathrm{GeV}$ and the mass of the dilepton system is rescaled by $m_Z/m_{\ell^-\ell^+}$. An additional requirement is applied using the quasi-significance of the measured $E_{\mathrm{T}}^{\mathrm{miss}}$[2], which has to be smaller than $3.5\ \sqrt{\mathrm{GeV}}$, to suppress $t\bar{t}$ events. This is necessary since no MVA is employed that could

[1] The jet energy regression was briefly tested for this analysis but not further pursued as a jet energy correction for this analysis.

[2] It is defined as $E_{\mathrm{T}}^{\mathrm{miss}}/\sqrt{H_{\mathrm{T}}}$ and H_{T} is the scalar sum of the p_{T} of the leptons and jets in the event.

make use of the differences in the $E_{\mathrm{T}}^{\mathrm{miss}}$ distribution between signal and background events.

The events that pass the outlined selection criteria are split into the following categories to enhance the sensitivity of the analysis to specific background components. The following categories are defined as SRs, which all have to have 2 or more *signal* jets:

- **SR, 0 lepton, 1 *b*-tag**: events in the 0 lepton channel with an invariant mass of the di-jet system between 110 GeV and 140 GeV; the leading jet has to be a *b*-jet
- **SR, 0 lepton, 2 *b*-tags**: events in the 0 lepton channel with an invariant mass of the di-jet system between 110 GeV and 140 GeV; the leading and sub-leading jet have to be *b*-jets
- **SR, 2 lepton, 1 *b*-tag**: events in the 2 lepton channel with an invariant mass of the di-jet system between 110 GeV and 140 GeV; the leading jet has to be a *b*-jet
- **SR, 2 lepton, 2 *b*-tags**: events in the 2 lepton channel with an invariant mass of the di-jet system between 110 GeV and 140 GeV; the leading and sub-leading jet have to be *b*-jets

The following CRs are defined using the events that fail to pass the di-jet mass window criteria:

- **sideband CR, 0 lepton, 1 *b*-tag**: events in the 0 lepton channel with an invariant mass of the di-jet system of smaller than 110 GeV or larger than 140 GeV; the leading jet has to be a *b*-jet
- **sideband CR, 0 lepton, 2 *b*-tags**: events in the 0 lepton channel with an invariant mass of the di-jet system of smaller than 110 GeV or larger than 140 GeV; the leading and sub-leading jet have to be *b*-jets
- **sideband CR, 2 lepton, 1 *b*-tag**: events in the 2 lepton channel with an invariant mass of the di-jet system of smaller than 110 GeV or larger than 140 GeV; the leading jet has to be a *b*-jet
- **sideband CR, 2 lepton, 2 *b*-tags**: events in the 2 lepton channel with an invariant mass of the di-jet system of smaller than 110 GeV or larger than 140 GeV; the leading and sub-leading jet have to be *b*-jets

For the 2 lepton channel additional $t\bar{t}$ enriched CRs are defined. Events are required to contain exactly one muon and one electron. In addition, the $E_{\mathrm{T}}^{\mathrm{miss}}$ significance requirement is removed and the $m_{\ell^-\ell^+}$ requirement is changed to $m_{e\mu} > 40$ GeV and the $m_{b\bar{b}}$ mass window criteria is not applied. The two $e\mu$ CRs categories are:

- ***eμ* CR, 2 lepton, 1 *b*-tag**: events in the 2 lepton channel that contain exactly one muon and one electron; the leading jet has to be a *b*-jet
- ***eμ* CR, 2 lepton, 2 *b*-tags**: events in the 2 lepton channel that contain exactly one muon and one electron; the leading jet and the sub-leading jet have to be *b*-jets

The compositions of the background events in the SRs and CRs are shown in figure 10.2. The 2 lepton channel is dominated by Z+jets events with sizeable contributions from $t\bar{t}$ production. The composition in the SRs and sideband CRs are almost identical. The exception is the small contribution of diboson events which are only present in the sideband CRs since the Z boson mass is outside the selected $m_{b\bar{b}}$ window of the SRs. The main Z+jets component in the 2 b-tags regions is the production of Z bosons in association with heavy flavour jets. In the 1 b-tag region the Z+jets events are dominated by events with one heavy flavour jet and one light jet. This is similar in the 0 lepton categories and also observed for the W+jets events that are a sizeable contributor to the 0 lepton channel. Overall the 0 lepton channel exhibits a much more diverse composition compared to the 2 lepton channel. In the 2 b-tags categories $t\bar{t}$ events are dominating whereas the dominant contributor to the 1 b-tags categories are V+jets events. Differences between the SRs and sideband CRs are visible as well: $t\bar{t}$ events are a larger contributor to the SRs compared to the corresponding sideband CRs. In addition, a sizeable relative amount of single top events and a small amount of diboson events are present in all 0 lepton categories. The two $e\mu$ CRs have a high purity of $t\bar{t}$ events and small contributions of single top events. Other background processes contribute less around 2% in the 1 b-tag category and less than 0.5% in the 2 b-tags category in this CRs.

10.3 The $m(ZH)$ discriminant

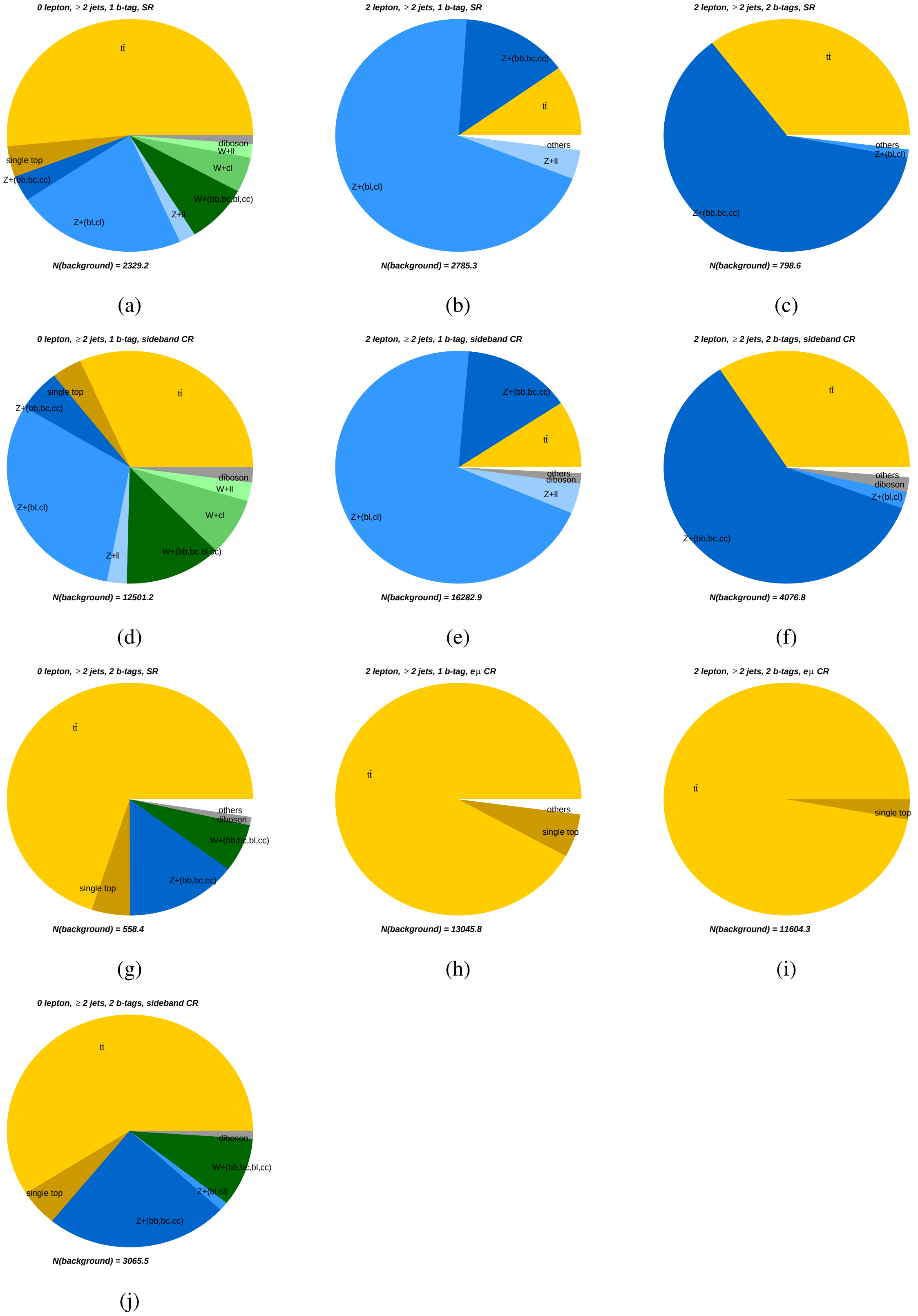
The final discriminant of the $A \to ZH$ analysis is the invariant mass of the Z and Higgs boson candidates $m(ZH)$. This analysis does not utilise a MVA since a large range of signal kinematics, which depend on the A boson mass, has to be covered. The development of a MVA strategy for this scenario is beyond the scope of this analysis. Thus $m(ZH)$ is used as the final discriminant. In the 2 lepton channel the full event is reconstructed and therefore the reconstructed A mass is the invariant mass of the two leptons and the di-jet system that represents the Higgs boson candidate. In the 0 lepton channel only the transverse mass of the ZH system $m_{\mathrm{T}}(ZH)$ is accessible due to the escaping neutrinos. The transverse mass is reconstructed from the di-jet system and the measured $E_{\mathrm{T}}^{\mathrm{miss}}$ of the event:

$$m_{\mathrm{T}}(ZH) = \sqrt{(E_{\mathrm{T}}^{\mathrm{dijet}} + E_{\mathrm{T}}^{\mathrm{miss}})^2 - (\vec{p}_{\mathrm{T}}^{\,\mathrm{dijet}} + \vec{E}_{\mathrm{T}}^{\mathrm{miss}})^2} \tag{10.1}$$

For simplicity the final discriminant is referred to as $m(ZH)$ although it means $m_{\mathrm{T}}(ZH)$ in the 0 lepton channel. Figure 10.3 displays the $m(ZH)$ distributions for different A boson mass hypotheses and the sum of all backgrounds in the signal regions. The $m(ZH)$ resolution is very good in the 2 lepton categories compared to the 0 lepton categories. There is no significant difference between the 1 b-tag and 2 b-tags category. The resolution is best for low A boson masses due to the decreasing momentum resolution for leptons with high p_{T}. However, the bulk of the background events accumulate at low $m(ZH)$ as well. The signal $m_{\mathrm{T}}(ZH)$ resolution in the 0 lepton channel is deteriorated due to the missing momentum information in the z-direction. The $m_{\mathrm{T}}(ZH)$ distributions are similar in the 1 b-tag and 2 b-tags categories. For low A boson masses the wide $m_{\mathrm{T}}(ZH)$ distribution of the signal is spread over the the bins that are populated by a high amount of background events. Although the $m_{\mathrm{T}}(ZH)$ resolution is even worse for high masses the advantage is the small amount of background events expected in that mass regime. The 0 lepton channel has a statistical advantage with respect to the 2 lepton channel due to the more than 3 times higher BR of the $Z \to \nu\bar{\nu}$ decay compared to the $Z \to \ell^-\ell^+$ decay. This is crucial for high A boson masses since the cross section for the production of the A boson decreases with increasing mass. The distributions of the CRs are shown in appendix D.1.

10.4 Systematic Uncertainties

Experimental uncertainties are assigned to the objects used in the analysis, i.e. electrons, muons, $E_{\mathrm{T}}^{\mathrm{miss}}$, jets, as well as the utilised b-tagging techniques and the recorded luminosity of the 2015 data set. The sources of experimental uncertainties for those are the same as described in section 8.6.1. However, the size of the individual uncertainties may vary between the SM $Z(H \to b\bar{b})$ analysis and the $A \to ZH$ analysis since some of them were re-evaluated once more data was recorded and experimental techniques were refined for the run 2 conditions.



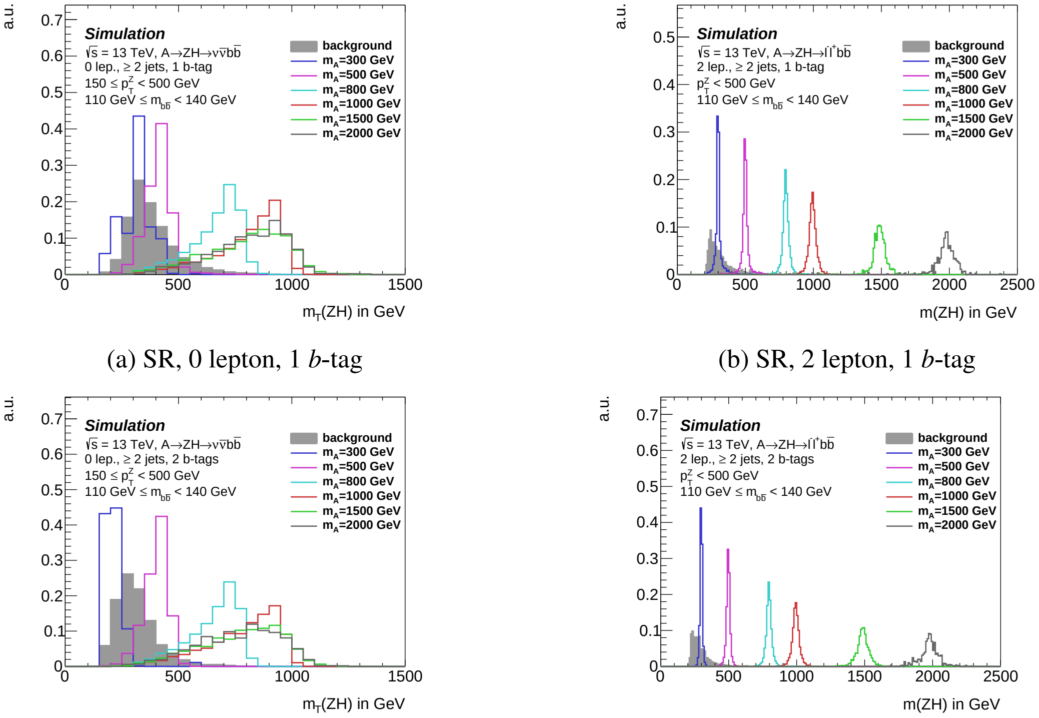
(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

(j)

Figure 10.2: The background processes that contribute to the $A \rightarrow ZH$ analysis categories. Shown are the relative contributions for each analysis category. Components that are smaller than 1% are grouped together in "others" (white).



(a) SR, 0 lepton, 1 b-tag

(b) SR, 2 lepton, 1 b-tag

(c) SR, 0 lepton, 2 b-tags

(d) SR, 2 lepton, 2 b-tags

Figure 10.3: The $m_{\mathrm{T}}(ZH)$ and $m(ZH)$ distributions for different A boson mass hypotheses and the sum of all background processes in the SRs of the $A \to ZH$ analysis: a) SR, 0 lepton, 1 b-tag, b) SR, 2 lepton, 1 b-tag, c) SR, 0 lepton, 2 b-tags, d) SR, 2 lepton, 2 b-tags. All distributions are normalised to unity.

The modelling uncertainties for the $A \to ZH$ analysis are specifically derived for this analysis. Normalisation, acceptance and shape uncertainties are assigned to the signal, $t\bar{t}$ and V+jets normalisation. A similar strategy as for the SM $Z(H \to b\bar{b})$ analysis is followed: normalisation and acceptance uncertainties are derived from the sum in quadrature of all variations whereas shape uncertainties are derived by fitting an analytic function to the largest observed variation.

The V+jets uncertainties are assigned in a similar way as for the SM $Z(H \to b\bar{b})$ analysis. In the following only the differences in the derivation of the modelling uncertainties of the other simulated data sets are described. Since the comparison with alternative simulated $t\bar{t}$ data sets are suffering from large statistical errors additional shape uncertainties from the run 1 ATLAS $V(H \to b\bar{b})$ analysis [66] are assigned. They are parametrised as a function of the average top quark p_{T} and the p_{T} of the $t\bar{t}$ system using MC truth information. Due to missing alternative MC data sets for the sub-leading single top and diboson background processes only overall single top normalisation uncertainties and the ATLAS $V(H \to b\bar{b})$ run 1 diboson uncertainties are assigned. The diboson uncertainties are parametrised as functions of $m_{b\bar{b}}$ and p_{T}^{Z}. The uncertainty on the $A \to ZH$ simulation model are assigned in dependence of the A boson mass. These uncertainties are largest for low masses and smallest for masses around 1 TeV, above 1 TeV they increase again. The same uncertainties are assigned to the 1 b-tag and 2 b-tags categories since the $m(ZH)$ distributions are similar between these (compare figure 10.3). All modelling uncertainties are summarised in table 10.3.

Uncertainty	Source	$A \to ZH$ all
$\Delta_{\text{norm.}}$	μ_F, μ_R, PDF, PS	1.8% up to 6.9% (depending on m_A)

Uncertainty	Source	$ZH \to \nu\bar{\nu}b\bar{b}$ all	$W^{\pm}H \to \ell^{\pm}\overset{(-)}{\nu}b\bar{b}$ all	$ZH \to \ell^{-}\ell^{+}b\bar{b}$ all
$\Delta_{\text{norm.}}$	experimental constraints	50%		

Uncertainty	Source	Z+jets				W+jets	
		0ℓ SR	0ℓ CR	2ℓ SR	2ℓ CR	0ℓ SR	0ℓ CR
$\Delta_{\text{norm.}}(Z + (bb, bc, cc))$	various	float				-	-
$\Delta_{\text{norm.}}(Z + (bl, cl))$	various	float				-	-
$\Delta_{\text{norm.}}(W + (bb, bc, bl, cc))$	μ_F, μ_R, merging & resummation scale	-	-	-	-	30%	
$\Delta_{\text{norm.}}(W + cl)$	μ_F, μ_R, merging & resummation scale	-	-	-	-	30%	
$\Delta_{\text{norm.}}(V + ll)$	μ_F, μ_R, merging & resummation scale	26%				10%	
$\Delta_{\text{acc.}}(V + cl \text{ w.r.t } V + bl)$	μ_F, μ_R, merging & resummation scale	13%				-	-
$\Delta_{\text{acc.}}(V + bl \text{ w.r.t } V + bb)$	μ_F, μ_R, merging & resummation scale	-	-	-	-	35%	
$\Delta_{\text{acc.}}(V + (bb, bc, cc),\ 0\ell \text{ w.r.t } 2\ell)$	various	17%		-	-	-	-
$\Delta_{\text{acc.}}(V + (bl, cl),\ 0\ell \text{ w.r.t } 2\ell)$	various	12%		-	-	-	-
$\Delta_{\text{acc.}}(V + ll,\ 0\ell \text{ w.r.t } 2\ell)$	various	9%		-	-	-	-
$\Delta_{\text{shape}}(m(ZH))$, $V + (bb, bc, bl, cc)$	various	S	S	S	S	S	S
$\Delta_{\text{shape}}(m(ZH))$, $V + cl$	various	S	S	S	S	S	S
$\Delta_{\text{shape}}(m(ZH))$, $V + ll$	various	S	S	S	S	S	S

Uncertainty	Source	$t\bar{t}$				
		0ℓ SR	0ℓ CR	2ℓ SR	2ℓ CR	2ℓ $e\mu$ CR
$\Delta_{\text{norm.}}$	-	float		float		
$\Delta_{\text{acc.}}$(SR w.r.t. sideband CR)	ME, PS, μ_F, μ_R	12%	-	9 %	-	-
$\Delta_{\text{acc.}}$($e\mu$ CR w.r.t. SR)	ME, PS, μ_F, μ_R	-	-	-	-	2.5%
$\Delta_{\text{shape}}(m(ZH))$	ME	S		S		S
$\Delta_{\text{shape}}(m(ZH))$	PS	S		S		S
$\Delta_{\text{shape}}(p_T^t)$	NNLO effects	S				
$\Delta_{\text{shape}}(p_T^{t\bar{t}})$	NNLO effects	S				

Uncertainty	Source	t-channel all	Wt all	s-channel all
$\Delta_{\text{norm.}}$	μ_F, μ_R, PDF, α_S	4%	7%	4%

Uncertainty	Source	ZZ all	WZ all	WW all
$\Delta_{\text{shape}}(m_{b\bar{b}})$	PS	S		
$\Delta_{\text{shape}}(p_T^Z)$	PDF, α_S	S		
$\Delta_{\text{shape}}(p_T^Z)$	μ_F	S		
$\Delta_{\text{shape}}(p_T^Z)$	μ_R	S		

Table 10.3: Modelling uncertainties assigned to the simulated signal and background processes. Given is the type of the systematic uncertainty, the source and the size for each analysis category and background process. Shape effects are labelled as "S" and normalisations that will be determined from data as "float". The following abbreviations are used: 0ℓ = 0 lepton channel, 2ℓ = 2 lepton channel, "all"=all analysis categories. "CR" refers to the sideband CR if not further specified. If the source is given as "various" the uncertainty was derived from the comparison to a simulation from a different MC generator which incorporates different PDF sets, tunes, PS models, scales, merging schemes etc. If the source is given as "ME" the uncertainty was derived from a comparison to a model of a MC generator that uses the same PS model as the nominal model but has a different ME generator which also includes PDF differences, merging between ME and PS differences etc.

10.5 Statistical Analysis

A fit to data is performed to set an upper limit on $\sigma(gg \to A) \times \mathrm{BR}(A \to ZH) \times \mathrm{BR}(H \to b\bar{b})$, which can be excluded in the data at a 95% confidence level. These limits are determined using the CL_S method [52, 54]. A profile likelihood fit, implemented in the RooStats Framework, is used.

The following components are considered in the fit:

- **Signal**: $gg \to A \to ZH \to \nu\bar{\nu}b\bar{b}$, $gg \to A \to ZH \to \ell^-\ell^+b\bar{b}$
- **SM $V(H \to b\bar{b})$**: $gg \to ZH \to \nu\bar{\nu}b\bar{b}$, $qq \to ZH \to \nu\bar{\nu}b\bar{b}$, $W^\pm H \to \ell^\pm \overset{(-)}{\nu} b\bar{b}$, $gg \to ZH \to \ell^-\ell^+b\bar{b}$, $qq \to ZH \to \ell^-\ell^+b\bar{b}$
- **V+jets**: $Z+bb$, $Z+bc$, $Z+bl$, $Z+cc$, $Z+cl$, $Z+ll$, $W+bb$, $W+bc$, $W+bl$, $W+cc$, $W+cl$, $W+ll$
- **$t\bar{t}$**
- **single top**: t-channel, Wt, s-channel
- **diboson**: ZZ, WZ, WW

The normalisations of four of these components are allowed to vary freely ("free floating") in the fit. The information provided in the fit is sufficient to extract these normalisations from data. These are also components that exhibit disagreements in the normalisation between data and simulated events. The free floating normalisations are:

- Z + (bb, bc, cc) normalisation: These events are dominating the 2 lepton 2 b-tags categories.
- $Z + (bl, cl)$ normalisation: Events of those components are present in the 2 lepton 1 b-tag categories with a high purity.
- $t\bar{t}$ 2 lepton normalisation: The 2 lepton channel includes the $e\mu$ CRs which have an excellent purity in $t\bar{t}$ events
- $t\bar{t}$ 0 lepton normalisation: Due to the different $t\bar{t}$ final states probed in the 0 lepton channel this component is free floating separately from the $t\bar{t}$ 2 lepton normalisation. The 0 lepton 2 b-tags categories pre-dominantly contain $t\bar{t}$ events.

In addition to these free floating normalisations, all systematic uncertainties as described in section 10.4 are introduced as nuisance parameters in the fit. They are implemented as Gaussian priors. Each experimental uncertainty is correlated across categories and components whereas the modelling uncertainties are correlated across categories but not correlated across components. The $t\bar{t}$ component is an exception and it is not correlated between the 0 and 2 lepton channel due to the very different $t\bar{t}$ signatures in these channels. The modelling uncertainties that are assigned to groups of V+jets components are treated as correlated across these components. In addition, the statistical error of the simulated events in each bin is introduced as a nuisance parameter, which acts as a Poissonian priors. The same nuisance parameter smoothing and pruning is used as for the $Z(H \to b\bar{b})$ analysis 8.7.3, to ensure reliable performance.

The final discriminants are the binned $m(ZH)$ distributions, which are used for all analysis categories. Since there are almost no data, and respective simulated, events expected beyond 1 TeV the range of the distributions is restricted to below 1 TeV. This is necessary to ensure a stable performance of the fit. All events greater than 1 TeV are included in the last bin of each distribution. To ensure that the analysis categories that are provided in the fit yield close to optimal sensitivity several different strategies are tested. For these studies two mass points were considered: $m_A = 300\,\text{GeV}$ and $m_A = 800\,\text{GeV}$, of which one is located in the bulk of the distribution of the background events and the other is located in an almost background free mass range. To enhance the sensitivity to certain background components, especially increase of the separation of V+jets and $t\bar{t}$ events, a split based on the jet multiplicities is tested. Although an improvement of up to 20% is achieved for $m_A = 800\,\text{GeV}$ with a split of each category in three categories — 2 or 3 jets, 4 jets and ≥ 5 jets — the small amount of expected events in some analysis categories due to this aggressive splitting scheme might introduce large statistical uncertainties in some bins of the $m(ZH)$ distributions. Thus an alternative strategy is explored, which does not introduce additional splits of the categories but decreases the bin size in the 2 lepton SRs, to exploit the shape information of the $m(ZH)$. If bins of 10 GeV width are used in the 2 lepton SRs an improvement of 20% on the limit for $m_A = 300\,\text{GeV}$ and of 10% for the limit of $m_A = 300\,\text{GeV}$. A second study is performed to determine if the fit model may be simplified by excluding categories. This is not the case since the $m_{b\bar{b}}$ sideband CRs and the 1b-tag categories add 7% and an additional 5% improvement in the expected limit for $m_A = 300\,\text{GeV}$. The improvement is smaller for $m_A = 800\,\text{GeV}$: 6% in total. Since a non-negligible amount of the signal events are accepted in the $m_{b\bar{b}}$ sideband and 1b-tag CRs these improvements are not surprising. The $e\mu$ CRs do not improve the expected limit but they are kept since they are important to extract the $t\bar{t}$ normalisation. A summary of all categories that are used in the fit is given in table 10.4 and figures showing the evolution of the expected limits for the outlined fit model studies are given in appendix D.2. The bin width for the 2 lepton SRs is set to 10 GeV as detailed before. In the 0 lepton SRs the bin width is set to 100 GeV. A very coarse bin width of 200 GeV is used in all CRs.

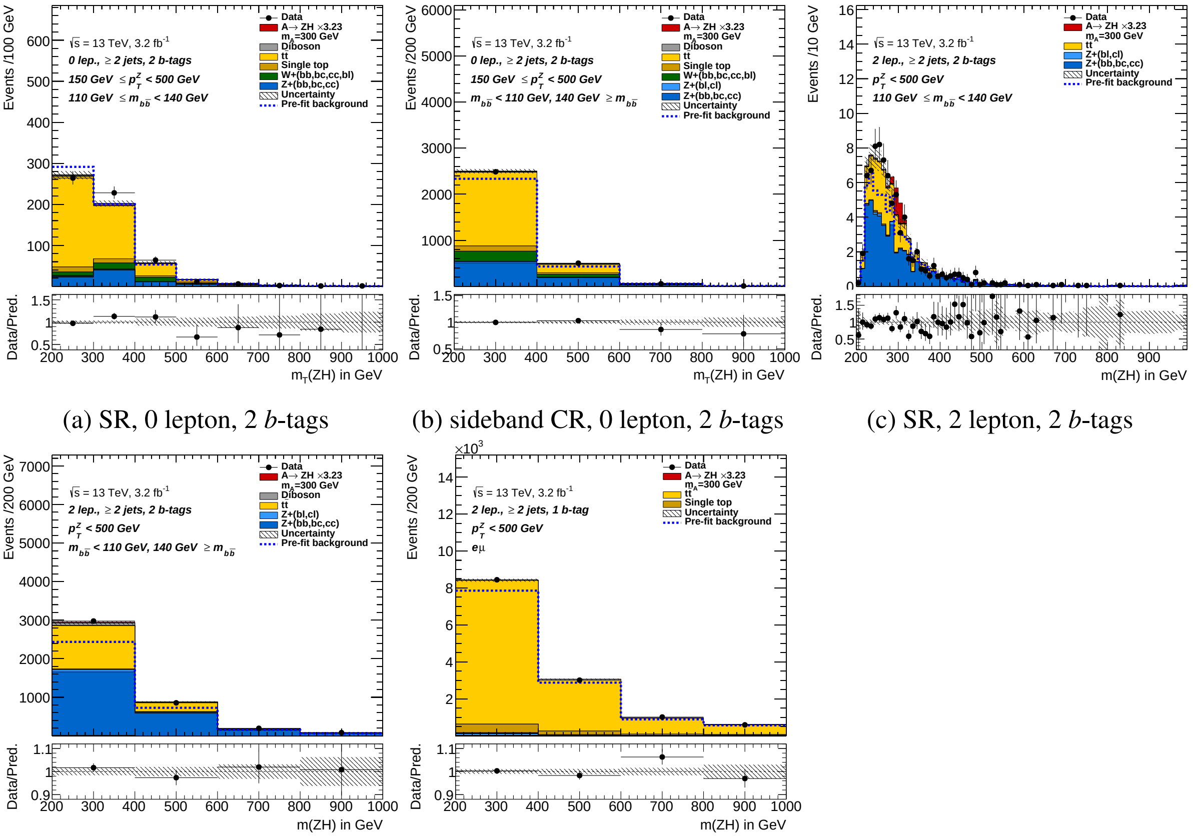
		0 leptons		2 leptons	
		$m_{b\bar{b}}$ window	$m_{b\bar{b}}$ sideband	$m_{b\bar{b}}$ window	$m_{b\bar{b}}$ sideband
	1 b-tag	SR	CR	SR	CR
	2 b-tags	SR	CR	SR	CR
$e\mu$ CRs	1 b-tag			CR	
	2 b-tags			CR	

Table 10.4: The definition of the SRs and CRs that are included in the $A \to ZH$ fit. The $m_{b\bar{b}}$ window is defined as $110\,\text{GeV} \leq m_{b\bar{b}} < 140\,\text{GeV}$ and the $m_{b\bar{b}}$ sideband as $m_{b\bar{b}} < 110\,\text{GeV}, m_{b\bar{b}} \geq 140\,\text{GeV}$.

10.6 Results

Upper limits are calculated for each mass point that is available as a simulated data set. The probed masses range from 220 GeV in steps of 20 GeV up to 500 GeV, in step sizes of 50 GeV up to 1 TeV and in steps of 100 GeV up to 2 TeV. The 900 GeV mass point is excluded due to the missing simulated sample in the 0 lepton channel. The $m(ZH)$ distributions of the 2 b-tags categories are displayed in figure 10.4 for data and simulated events. All components are scaled as determined by the fit. An exemplary signal with $m_A = 300\,\text{GeV}$ that is scaled to its upper exclusion limit is shown as well. Due to the limited amount of events a lot of statistical fluctuations are observed. The distributions of the 1 b-tag categories are included in appendix D.3.

Figure 10.5a) display the expected limits for all A boson mass points for the combination of the 0+2 lepton channel as well as for separate fits using either only the 0 lepton channel or 2 lepton channel. For low A boson masses the sensitivity of the 2 lepton channel dominates the expected limit. The deteriorated $m(ZH)$ resolution in the 0 lepton channel does not provide enough shape information for the signal to



(a) SR, 0 lepton, 2 b-tags

(b) sideband CR, 0 lepton, 2 b-tags

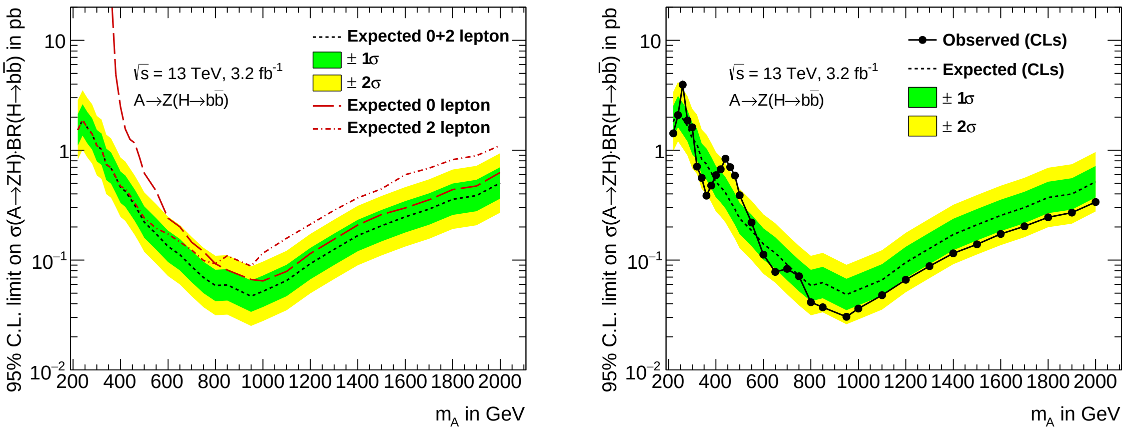
(c) SR, 2 lepton, 2 b-tags

(d) sideband CR, 2 lepton, 2 b-tags

(e) $e\mu$ CR, 2 lepton, 2 b-tags

Figure 10.4: The $m(ZH)$ distributions of the 2 b-tags categories of the $A \to ZH$ analysis. All simulated components are scaled to their postfit values. The simulated $A \to ZH$ signal process with a mass of 300 GeV is shown exemplary. Its normalisation is scaled to the upper excluded limit for that mass, which is 3.23 pb.

stand out over the large amount of background events in those regimes. In contrast, the $m(ZH)$ resolution is very good in those regimes. At masses larger than 500 GeV the 0 lepton channel gains importance and takes over as the driving channel for the limits above 800 GeV. This behaviour is expected since there are more signal events expected in the amount of expected signal events based on the larger BR of the $Z \to \nu\bar{\nu}$ decay compared to the $Z \to \ell^-\ell^+$ decay. The worse $m(ZH)$ resolution in the 0 lepton channel is not a limiting factor any more since only a small amount of background events is expected at high masses. The expected limits increase again above 1 TeV since the amount of selected signal events is decreasing due to the upper requirement on p_{T}^{Z}. Additionally, the $m(ZH)$ distributions are limited to 1 TeV. Thus the information of the 2 lepton channel about masses higher than this is condensed in the last bin of the $m(ZH)$ information which does not provide any shape information. Some shape information is conserved due to the long tails of the 0 lepton channel and the tails of the 2 lepton signal distributions in the CRs. Nevertheless these tails do not offer a large sensitivity and become less and less sensitive for higher masses. The observed limits are compatible with a background-only hypothesis to a large extent. The excluded values for $\sigma(gg \to A) \times \mathrm{BR}(A \to ZH) \times \mathrm{BR}(H \to b\bar{b})$ vary between 4 pb for $m_A = 260$ GeV and 0.03 pb for $m_A = 950$ GeV. Two excesses are observed at $m(ZH) = 260$ GeV and $m(ZH) = 460$ GeV and the two neighbouring mass points. The local significances of those are 2.3 σ and 1.8 σ respectively, which are not significant. An ATLAS analysis that searches for the $A \to Z(H \to b\bar{b})$ decay using the full 2015+2016 LHC data set observes an excess at 440 GeV in another A boson production channel — associated production with b-quarks — with a global significance of 2.4 σ [103]. Nevertheless the same analysis also probes the same production channel as presented here, $gg \to A$, and does not observe any excess at the same mass. Both investigated production channels do not confirm the $m(ZH) = 260$ GeV excess observed here. Hence, given the smallness of the excesses and the additional information from the analysis with the full data set, it is concluded that no new effects are observed.



(a) Expected limits: 0, 2 and 0+2 lepton channel

(b) Expected and observed limits

Figure 10.5: The obtained limits on $\sigma(gg \to A) \times \mathrm{BR}(A \to ZH) \times \mathrm{BR}(H \to b\bar{b})$ for A boson masses from 220 GeV up to 2 TeV: a) expected limits for the 0 lepton channel (long red dashed line), 2 lepton channel (finely red dashed line) and combined fit of both (black dashed line), b) the expected combined limit (black dashed line) and the observed limit (points and black solid line). The green and yellow bands represent the $\pm 1\sigma$ and $\pm 2\sigma$ error bands of the combined expected limit.

CHAPTER 11

Summary

In this thesis a search for the Standard Model (SM) $H \to b\bar{b}$ and $H \to c\bar{c}$ decays is presented, which are important Higgs boson decay channels to establish the Yukawa nature of the Higgs boson to fermion couplings. In addition, their observation would provide more confidence that the discovered scalar boson with a mass of 125 GeV is indeed the SM Higgs boson. The predicted SM branching ratio for the $H \to b\bar{b}$ decay is approximately 58% whereas it is 3% for the $H \to c\bar{c}$ decay. Both of these decay channels have hadronic signatures, i.e. they manifest as jets, and the challenge is the separation of these events from SM processes that exhibit similar signatures.

The analyses presented in this thesis use events of proton-proton collisions, which are provided by the Large Hadron Collider (LHC) at a collision energy of 13 TeV and are recorded by the ATLAS experiment. The integrated luminosity of the analysed data set is 36.1 fb^{-1} and corresponds to the data that was taken in 2015 and 2016.

The production of multi-jet events is an abundant signature in pp collisions. To suppress the amount of multi-jet events, events are selected that probe the Higgs boson production in association with a Z boson. In addition, the leptonic signature of Z boson decays into pairs of neutrinos, electrons or muons is used to identify events of interest. This allows for much lower p_{T} trigger thresholds — $O(150\,\text{GeV})$ for $E_{\text{T}}^{\text{miss}}$ from neutrinos and $O(25\,\text{GeV})$ for single electrons and muons — than jets. Thus these signal events are recorded more efficiently compared to signal events with a fully hadronic signature. In addition, the amount of multi-jet events in the total amount of events selected for the analysis is suppressed to less than 1% in the ZH search channel. Furthermore, jets that originate from the fragmentation of b-quarks (or c-quarks for the $H \to c\bar{c}$ search) are identified with multivariate b-jet identification algorithms to further suppress background events containing light jets. Therefore the main contributors to the background events is the production of Z bosons in association with additional (heavy flavour) jets and top quark pairs. However, the production cross section for these events is still orders of magnitude higher than the production cross section for ZH events.

Since the LHC operated at $\sqrt{s} = 13\,\text{TeV}$ for the first time in 2015 the desired final states — 2 jets and 2 leptons or missing transverse energy — were first studied at these high energies in the search for a heavy CP odd scalar boson A, such as predicted by Higgs doublet models, decaying into ZH. The requirements for the selection of $ZH \to \nu\bar{\nu}b\bar{b}$ and $ZH \to \ell^-\ell^+b\bar{b}$ final states are defined separately. After a dedicated selection of events, several signal enriched and background enriched phase space regions are identified. In a simultaneous fit to all regions, using the transverse mass of $E_{\text{T}}^{\text{miss}}$ and the Higgs boson candidate jets and the invariant mass of the di-lepton and Higgs boson candidate jets as the discriminant, upper limits for the $gg \to A$ cross section times the $A \to ZH$ and $H \to b\bar{b}$ branching ratios are obtained for a dataset of 3.2 fb^{-1}. Hypothetical masses for the A boson between 220 GeV and 2 TeV are probed and signals cross sections ranging between 4 pb and 0.03 pb are excluded at a 95% confidence level.

To search for the SM Higgs boson in the $ZH \to \nu\bar{\nu}b\bar{b}$ and $ZH \to \ell^-\ell^+b\bar{b}$ decay channel a multivariate analysis (MVA) using Boosted Decision Trees (BDTs) is employed to increase the signal-to-background separation. Since the $b\bar{b}$ resonance structure of the signal process distinguishes it from the non-resonant backgrounds the invariant mass of the two b-jets is the most discriminating variable that is used in the MVA. To improve the b-jet momentum resolution, which is deteriorated due to semi-leptonic b-hadron decays, out-of-cone leakage and the intrinsic calorimeter resolution, a multivariate b-jet momentum correction is developed in the context of this thesis. It uses variables connected to the jet kinematics, leptons inside the jet, tracks, secondary vertices, pile-up and close-by jets and improves the transverse momentum resolution by approximately 20% with respect to the standard ATLAS jet calibration. The correction method is successfully validated in $t\bar{t}$ and Zb events. This improvement translates to an improvement in the $m_{b\bar{b}}$ resolution of 15% to 25% depending on the kinematic region of the $Z(H \to b\bar{b})$ phase space. The improvement of the sensitivity is 8%, based on the BDTs output distribution in 6 different kinematic regions as the final discriminant, yielding an expected significance of 2.8 σ. The obtained observed significance is 2.9 σ, including the multivariate b-jet momentum correction, which is close to an evidence for the $H \to b\bar{b}$ decay. The signal is measured with a strength of:

$$\mu = 1.15 \pm 0.29(\text{stat.})^{+0.36}_{-0.30}(\text{syst.}) \tag{11.1}$$

which is compatible with the SM expectation. A similar result in the $V(H \to b\bar{b})$ decay channel is obtained by the CMS experiment as well. The analysis strategy is further verified by extracting a diboson, i.e. WZ and ZZ, signal strength of $\mu = 0.97 \pm 0.12(\text{stat.})^{+0.19}_{-0.16}(\text{syst.})$ with an observed (expected) significance of 5.6 σ (6.2 σ) from the same final state using the same analysis phase space. The sources of the error of the $Z(H \to b\bar{b})$ result is further investigated. Given that the analysis is limited by the systematic uncertainties it is important to understand these sources to improve them in future analysis leading towards the discovery of the $H \to b\bar{b}$ decay. The three largest sources of systematic uncertainties are the theory uncertainties and simulation uncertainties of the signal process, the experimental uncertainties of the utilised b-jet identification techniques and the statistical uncertainties due to the limited amount of simulated events for the background models.

Furthermore, the same data set is analysed to set an upper limit on the cross section for ZH production times the branching ratio for the $H \to c\bar{c}$ decay. In this analysis only the $ZH \to \ell^-\ell^+c\bar{c}$ final state is investigated since it has a better signal-to-background ratio and only one dominant, Z+jets, and two sub-dominant, $t\bar{t}$ and diboson, background contributions compared to $ZH \to \nu\bar{\nu}c\bar{c}$. Novel multivariate c-jet identification techniques are used and the optimal requirement for the analysis is studied. Using the invariant mass of the two Higgs boson candidate jets as a discriminant, signal strengths of $\mu > 110$ are excluded at a 95% confidence level. This represents the best direct limit on Higgs boson to charm quark couplings to date. The limit exhibits a slight dependence on the $V(H \to b\bar{b})$ signal strength, 5% change in the $ZH \to \ell^-\ell^+c\bar{c}$ limit if $V(H \to b\bar{b})$ is varied by ±100%, since the $V(H \to b\bar{b})$ signal peaks at the same invariant mass as the $ZH \to \ell^-\ell^+c\bar{c}$ signal. However, this dependence will become larger once the analysis becomes more sensitive and the strategy has to be evolved for the treatment of the interplay between the two Higgs boson decay signatures. To validate the current analysis strategy the diboson signal is measured with a significance of 2.2 σ and a signal strength of $\mu = 0.6^{+0.5}_{-0.4}$. This cross check analysis is not yet sensitive to the diboson signal but will become an important tool to validate future $V(H \to c\bar{c})$ analyses of larger data sets. The prospects of a simple BDTs set-up using 4 input variables are studied as well and suggest that the limit may be improved by up to 15% with a multivariate analysis.

In conclusion, this thesis measured Higgs boson decays into heavy flavour jets. First, a search for a new boson $A \to ZH$ is performed using the ATLAS data set from the 2015 data taking period. This analysis sets exclusion limits for A boson masses between 220 GeV and 2 TeV ranging between 4 pb and 0.03 pb. Second, a search for the SM $Z(H \to b\bar{b})$ decay is performed using the ATLAS data set from the 2015 and 2016 data taking periods. For this analysis a dedicated multivariate b-jet momentum correction is developed which improves the b-jet momentum resolution by 20% and the $m_{b\bar{b}}$ resolution accordingly. A multivariate analysis is employed to extract the $Z(H \to b\bar{b})$ signal. This analysis measures the $H \to b\bar{b}$ decay with a significance of 2.9 σ and the measured signal strength is compatible with the SM expectation. Last, the same data set is analysed using novel c-jet identification techniques to search for the $H \to c\bar{c}$ decay in the $ZH \to \ell^{-}\ell^{+}c\bar{c}$ decay channel. An upper limit of 110 times the SM expectation is set on the signal strength, which corresponds to the best direct upper limit on the Higgs boson to charm quark coupling to date.

Bibliography

[1] A. Zee, *Quantum Field Theory in a Nutshell*, 1st ed., Princeton University Press, 2003 (cit. on pp. 4, 6).

[2] F. Halzen and A. D. Martin, *Quarks and Leptons: An Introductory Course in Modern Particle Physics*, Wiley, 1984, ISBN: 9780471887416 (cit. on pp. 4, 6).

[3] R. Ellis, W. Stirling and B. Webber, *QCD and Collider Physics*, 1st ed., Cambridge University Press, 2003 (cit. on pp. 4, 15).

[4] D. de Florian et al., *Handbook of LHC Higgs Cross Sections: 4. Deciphering the Nature of the Higgs Sector*, (2016), arXiv: `1610.07922 [hep-ph]` (cit. on pp. 7, 51, 72, 85, 162).

[5] I. Brock and T. Schörner-Sadenius, eds., *Physics at the Terascale*, 1st ed., Wiley-VCH, 2011 (cit. on pp. 9, 16–18).

[6] V. D. Barger and R. J. N. Phillips, *Collider Physics*, 1st ed., Westview Press, 1996 (cit. on p. 9).

[7] C. Patrignani et al., *Review of Particle Physics*, Chin. Phys. **C40** (2016) 100001 (cit. on pp. 9, 13, 26, 35, 51, 87).

[8] O. S. Broening et al., eds., *LHC Design Report. 1. The LHC Main Ring*, CERN-2004-003-V-1, CERN-2004-003, 2004, URL: `https://cdsweb.cern.ch/record/782076` (cit. on pp. 9, 10).

[9] M. Benedikt et al., *LHC Design Report. 3. The LHC injector chain*, (2004) (cit. on p. 10).

[10] L. Evans and P. Bryant, *LHC Machine*, Journal of Instrumentation **3** (2008) S08001, URL: `http://stacks.iop.org/1748-0221/3/i=08/a=S08001` (cit. on p. 10).

[11] M. Aaboud et al., *Measurement of the total cross section from elastic scattering in pp collisions at $\sqrt{s} = 8$ TeV with the ATLAS detector*, Phys. Lett. **B761** (2016) 158, arXiv: `1607.06605 [hep-ex]` (cit. on p. 10).

[12] *Luminosity Public Results Run 2*, ATLAS, 2017, URL: `https://twiki.cern.ch/twiki/bin/view/AtlasPublic/LuminosityPublicResultsRun2` (visited on 10/01/2017) (cit. on p. 10).

[13] W. Stirling, private communication, 2012 (cit. on p. 11).

[14] ATLAS Collaboration, *Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC*, Physics Letters B **716** (2012) 1, ISSN: 0370-2693, URL: `http://www.sciencedirect.com/science/article/pii/S037026931200857X` (cit. on pp. 11, 49).

[15] ATLAS Collaboration, *Expected Performance of the ATLAS Experiment – Detector, Trigger and Physics*, 2009, arXiv: `0901.0512` (cit. on pp. 12–14, 21, 22, 24, 25).

[16] *ATLAS: technical proposal for a general-purpose pp experiment at the Large Hadron Collider at CERN*, LHC Tech. Proposal, CERN, 1994, URL: `https://cds.cern.ch/record/290968` (cit. on pp. 12, 13).

[17] T. A. Collaboration et al., *The ATLAS Experiment at the CERN Large Hadron Collider*, Journal of Instrumentation **3** (2008) S08003, URL: `http://stacks.iop.org/1748-0221/3/i=08/a=S08003` (cit. on pp. 12, 13).

[18] M. Capeans et al., *ATLAS Insertable B-Layer Technical Design Report*, tech. rep. CERN-LHCC-2010-013. ATLAS-TDR-19, 2010, URL: `https://cds.cern.ch/record/1291633` (cit. on p. 12).

[19] J. Pequenao, “Computer generated image of the whole ATLAS detector”, 2008, URL: `https://cds.cern.ch/record/1095924` (cit. on p. 14).

[20] M. Aaboud et al., *Performance of the ATLAS Trigger System in 2015*, Eur. Phys. J. **C77** (2017) 317, arXiv: `1611.09661 [hep-ex]` (cit. on pp. 14, 49).

[21] G. Dissertori, I. Knowles and M. Schmelling, *Quantum Chromodynamics*, 1st ed., Oxford University Press, 2009 (cit. on pp. 15, 17).

[22] P. Skands, “Introduction to QCD”, *Proceedings, Theoretical Advanced Study Institute in Elementary Particle Physics: Searching for New Physics at Small and Large Scales (TASI 2012): Boulder, Colorado, June 4-29, 2012*, 2013 341, arXiv: `1207.2389 [hep-ph]`, URL: `https://inspirehep.net/record/1121892/files/arXiv:1207.2389.pdf` (cit. on pp. 15–19).

[23] T. Sjostrand, “Monte Carlo Generators”, *High-energy physics. Proceedings, European School, Aronsborg, Sweden, June 18-July 1, 2006*, 2006 51, arXiv: `hep-ph/0611247 [hep-ph]`, URL: `http://weblib.cern.ch/abstract?CERN-LCGAPP-2006-06` (cit. on pp. 17–19).

[24] T. Sjostrand, S. Mrenna and P. Z. Skands, *A Brief Introduction to PYTHIA 8.1*, Comput. Phys. Commun. **178** (2008) 852, arXiv: `0710.3820 [hep-ph]` (cit. on p. 19).

[25] M. Bähr et al., *Herwig++ physics and manual*, The European Physical Journal C **58** (2008) 639, ISSN: 1434-6052, URL: `https://doi.org/10.1140/epjc/s10052-008-0798-9` (cit. on p. 19).

[26] J. Bellm et al., *Herwig 7.0/Herwig++ 3.0 release note*, Eur. Phys. J. **C76** (2016) 196, arXiv: `1512.01178 [hep-ph]` (cit. on pp. 19, 54).

[27] T. Gleisberg et al., *Event generation with SHERPA 1.1*, JHEP **02** (2009) 007, arXiv: `0811.4622 [hep-ph]` (cit. on pp. 19, 20).

[28] S. Alioli et al., *A general framework for implementing NLO calculations in shower Monte Carlo programs: the POWHEG BOX*, JHEP **06** (2010) 043, arXiv: `1002.2581 [hep-ph]` (cit. on pp. 19, 51).

[29] J. Alwall et al., *The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations*, JHEP **07** (2014) 079, arXiv: `1405.0301 [hep-ph]` (cit. on p. 19).

[30] S. Frixione and B. R. Webber, *Matching NLO QCD computations and parton shower simulations*, JHEP **06** (2002) 029, arXiv: `hep-ph/0204244 [hep-ph]` (cit. on p. 19).

[31] G. Aad et al., *The ATLAS Simulation Infrastructure*, The European Physical Journal C **70** (2010) 823, ISSN: 1434-6052, URL: `http://dx.doi.org/10.1140/epjc/s10052-010-1429-9` (cit. on pp. 19, 20).

[32] G. Aad et al., *Measurement of the Z/γ^* boson transverse momentum distribution in pp collisions at $\sqrt{s}$ = 7 TeV with the ATLAS detector*, JHEP **09** (2014) 145, arXiv: `1406.3660 [hep-ex]` (cit. on p. 19).

[33] *ATLAS Run 1 Pythia8 tunes*, tech. rep. ATL-PHYS-PUB-2014-021, CERN, 2014, URL: `https://cds.cern.ch/record/1966419` (cit. on p. 19).

[34] D. J. Lange, *The EvtGen particle decay simulation package*, Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment **462** (2001) 152, BEAUTY2000, Proceedings of the 7th Int. Conf. on B-Physics at Hadron Machines, ISSN: 0168-9002, URL: `http://www.sciencedirect.com/science/article/pii/S0168900201000894` (cit. on p. 20).

[35] S. Agostinelli et al., *Geant4—a simulation toolkit*, Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment **506** (2003) 250, ISSN: 0168-9002, URL: `http://www.sciencedirect.com/science/article/pii/S0168900203013688` (cit. on p. 20).

[36] *The Optimization of ATLAS Track Reconstruction in Dense Environments*, tech. rep. ATL-PHYS-PUB-2015-006, CERN, 2015, URL: `https://cds.cern.ch/record/2002609` (cit. on p. 21).

[37] *Electron efficiency measurements with the ATLAS detector using the 2015 LHC proton-proton collision data*, tech. rep. ATLAS-CONF-2016-024, CERN, 2016, URL: `https://cds.cern.ch/record/2157687` (cit. on pp. 22, 38, 69).

[38] G. Aad et al., *Muon reconstruction performance of the ATLAS detector in proton-proton collision data at $\sqrt{s}$ =13 TeV*, Eur. Phys. J. **C76** (2016) 292, arXiv: `1603.05598 [hep-ex]` (cit. on pp. 22, 37, 38, 69).

[39] *Reconstruction, Energy Calibration, and Identification of Hadronically Decaying Tau Leptons in the ATLAS Experiment for Run-2 of the LHC*, tech. rep. ATL-PHYS-PUB-2015-045, CERN, 2015, URL: `https://cds.cern.ch/record/2064383` (cit. on p. 22).

[40] *Performance of missing transverse momentum reconstruction for the ATLAS detector in the first proton-proton collisions at at $\sqrt{s}$= 13 TeV*, tech. rep. ATL-PHYS-PUB-2015-027, CERN, 2015, URL: `https://cds.cern.ch/record/2037904` (cit. on pp. 23, 69).

[41] M. Cacciari, G. P. Salam and G. Soyez, *The Anti-k(t) jet clustering algorithm*, JHEP **04** (2008) 063, arXiv: `0802.1189 [hep-ph]` (cit. on pp. 23, 24).

[42] *Optimisation of the ATLAS b-tagging performance for the 2016 LHC Run*, tech. rep. ATL-PHYS-PUB-2016-012, CERN, 2016, URL: `https://cds.cern.ch/record/2160731` (cit. on pp. 25, 38).

[43] *Expected performance of the ATLAS b-tagging algorithms in Run-2*, tech. rep. ATL-PHYS-PUB-2015-022, CERN, 2015, URL: `https://cds.cern.ch/record/2037697` (cit. on p. 25).

[44] *Calibration of b-tagging using dileptonic top pair events in a combinatorial likelihood approach with the ATLAS experiment*, tech. rep. ATLAS-CONF-2014-004, CERN, 2014, URL: `http://cds.cern.ch/record/1664335` (cit. on p. 25).

[45] *Calibration of the performance of b-tagging for c and light-flavour jets in the 2012 ATLAS data*, tech. rep. ATLAS-CONF-2014-046, CERN, 2014, URL: `http://cds.cern.ch/record/1741020` (cit. on p. 25).

[46] *Search for the decay of the Higgs boson to charm quarks with the ATLAS experiment*, tech. rep. ATLAS-CONF-2017-078, CERN, 2017, URL: `https://cds.cern.ch/record/2292061` (cit. on pp. 26, 89).

[47] C. M. Bishop, *Pattern Recognition and Machine Learning*, 1st ed., Springer, 2006 (cit. on p. 30).

[48] T. Hastie, R. Tibshirani and J. Friedman, *The Elements of Statistical Learning*, 2nd ed., Springer, 2009 (cit. on pp. 30, 31).

[49] G. James et al., *An Introduction to Statistical Learning*, 1st ed., Springer, 2013 (cit. on pp. 30, 31).

[50] A. Hoecker et al., *TMVA - Toolkit for Multivariate Data Analysis*, (2007), eprint: `physics/0703039`, URL: `https://github.com/root-project/root/blob/master/documentation/tmva/UsersGuide/TMVAUsersGuide.pdf`, TMVA was updated for newer ROOT versions, the URL links to the documentation of the updated version (cit. on pp. 30, 31, 39).

[51] P. J. Huber, *Robust estimation of a location parameter*, Ann. Math. Statist. **35** (1964) 73, ISSN: 0003-4851, URL: `https://doi.org/10.1214/aoms/1177703732` (cit. on p. 31).

[52] G. Cowan et al., *Asymptotic formulae for likelihood-based tests of new physics*, Eur. Phys. J. **C71** (2011) 1554, [Erratum: Eur. Phys. J.C73,2501(2013)], arXiv: `1007.1727 [physics.data-an]` (cit. on pp. 32, 33, 92, 108).

[53] E. Gross and A. Klier, "Higgs statistics for pedestrians", *Supersymmetry and unification of fundamental interactions. Proceedings, 10th International Conference, SUSY'02, Hamburg, Germany, June 17-23, 2002*, 2002 4, arXiv: `hep-ex/0211058 [hep-ex]`, URL: `http://www-library.desy.de/preparch/desy/proc/proc02-02/Proceedings/pl.1a/gross_pr.pdf` (cit. on pp. 32, 33).

[54] A. L. Read, *Presentation of search results: the CL s technique*, Journal of Physics G: Nuclear and Particle Physics **28** (2002) 2693, URL: `http://stacks.iop.org/0954-3899/28/i=10/a=313` (cit. on pp. 33, 92, 108).

[55] M. Aaboud et al., *Jet energy scale measurements and their systematic uncertainties in proton-proton collisions at* $\sqrt{s} = 13$ *TeV with the ATLAS detector*, (2017), arXiv: `1703.09665 [hep-ex]` (cit. on pp. 35, 69).

[56] T. Aaltonen et al., *Improved b-jet Energy Correction for* $H \rightarrow b\bar{b}$ *Searches at CDF*, (2011), arXiv: `1107.3026 [hep-ex]` (cit. on p. 35).

[57] CMS Collaboration, *Search for the standard model Higgs boson produced through vector boson fusion and decaying to bb with proton-proton collisions at sqrt(s) = 13 TeV*, tech. rep. CMS-PAS-HIG-16-003, CERN, 2016, URL: `https://cds.cern.ch/record/2160154` (cit. on p. 35).

[58] ATLAS Collaboration,
Proposal for truth particle observable definitions in physics measurements,
tech. rep. ATL-PHYS-PUB-2015-013, CERN, 2015,
URL: `https://cds.cern.ch/record/2022743` (cit. on p. 36).

[59] M. Cacciari, G. P. Salam and G. Soyez, *The Catchment Area of Jets*, JHEP **04** (2008) 005,
arXiv: `0802.1188 [hep-ph]` (cit. on p. 37).

[60] ATLAS Collaboration,
Selection of jets produced in 13TeV proton-proton collisions with the ATLAS detector,
tech. rep. ATLAS-CONF-2015-029, CERN, 2015,
URL: `https://cds.cern.ch/record/2037702` (cit. on pp. 38, 55).

[61] *Commissioning of the ATLAS high-performance b-tagging algorithms in the 7 TeV collision data*,
tech. rep. ATLAS-CONF-2011-102, CERN, 2011,
URL: `https://cds.cern.ch/record/1369219` (cit. on p. 38).

[62] *Tagging and suppression of pileup jets with the ATLAS detector*,
tech. rep. ATLAS-CONF-2014-018, CERN, 2014,
URL: `https://cds.cern.ch/record/1700870` (cit. on pp. 38, 55).

[63] A. Bukin, *Fitting function for asymmetric peaks*, 2007, arXiv: `0711.4449v2` (cit. on p. 41).

[64] CMS Collaboration,
Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC,
Physics Letters B **716** (2012) 30, ISSN: 0370-2693,
URL: `http://www.sciencedirect.com/science/article/pii/S0370269312008581`
(cit. on p. 49).

[65] G. Aad et al.,
Evidence for the Higgs-boson Yukawa coupling to tau leptons with the ATLAS detector,
Journal of High Energy Physics **2015** (2015) 117, ISSN: 1029-8479,
URL: `https://doi.org/10.1007/JHEP04(2015)117` (cit. on p. 49).

[66] ATLAS Collaboration, *Search for the $b\bar{b}$ decay of the Standard Model Higgs boson in associated (W/Z)H production with the ATLAS detector*, Journal of High Energy Physics **2015** (2015) 69,
ISSN: 1029-8479, URL: `http://dx.doi.org/10.1007/JHEP01(2015)069`
(cit. on pp. 49, 65, 75, 106, 139).

[67] C. Collaboration, *Search for the standard model Higgs boson produced in association with a W or a Z boson and decaying to bottom quarks*, Phys. Rev. D **89** (1 2014) 012003,
URL: `http://link.aps.org/doi/10.1103/PhysRevD.89.012003` (cit. on p. 49).

[68] M. Aaboud et al., *Search for the Standard Model Higgs boson produced by vector-boson fusion and decaying to bottom quarks in $\sqrt{s} = 8$ TeV pp collisions with the ATLAS detector*,
JHEP **11** (2016) 112, arXiv: `1606.02181 [hep-ex]` (cit. on p. 50).

[69] M. Aaboud et al., *Search for the Standard Model Higgs boson produced in association with top quarks and decaying into a $b\bar{b}$ pair in pp collisions at $\sqrt{s}$ = 13 TeV with the ATLAS detector*,
Submitted to: Phys. Rev. D (2017), arXiv: `1712.08895 [hep-ex]` (cit. on p. 50).

[70] G. Cullen et al., *Automated One-Loop Calculations with GoSam*, Eur. Phys. J. **C72** (2012) 1889,
arXiv: `1111.2034 [hep-ph]` (cit. on p. 51).

[71] K. Hamilton, P. Nason and G. Zanderighi, *MINLO: Multi-Scale Improved NLO*,
JHEP **10** (2012) 155, arXiv: `1206.3572 [hep-ph]` (cit. on p. 51).

[72] G. Luisoni et al., *$HW^{\pm}$/HZ + 0 and 1 jet at NLO with the POWHEG BOX interfaced to GoSam and their merging within MiNLO*, JHEP **10** (2013) 083, arXiv: `1306.2542 [hep-ph]` (cit. on p. 51).

[73] J. Butterworth et al., *Single Boson and Diboson Production Cross Sections in pp Collisions at sqrts=7 TeV*, tech. rep. ATL-COM-PHYS-2010-695, CERN, 2010, URL: `https://cds.cern.ch/record/1287902` (cit. on p. 52).

[74] M. Aaboud et al., *Measurements of the production cross section of a Z boson in association with jets in pp collisions at $\sqrt{s}$ = 13 TeV with the ATLAS detector*, Eur. Phys. J. **C77** (2017) 361, arXiv: `1702.05725 [hep-ex]` (cit. on p. 52).

[75] M. Aaboud et al., *Evidence for the $H \to b\bar{b}$ decay with the ATLAS detector*, (2017), arXiv: `1708.03299 [hep-ex]` (cit. on pp. 53, 56, 57, 62, 83).

[76] J. Butterworth et al., *PDF4LHC recommendations for LHC Run II*, J. Phys. **G43** (2016) 023001, arXiv: `1510.03865 [hep-ph]` (cit. on p. 54).

[77] R. D. Ball et al., *Parton distributions for the LHC Run II*, JHEP **04** (2015) 040, arXiv: `1410.8849 [hep-ph]` (cit. on p. 54).

[78] S. Gieseke, C. Rohr and A. Siodmok, *Colour reconnections in Herwig++*, Eur. Phys. J. **C72** (2012) 2225, arXiv: `1206.0041 [hep-ph]` (cit. on p. 54).

[79] *Selection of jets produced in proton-proton collisions with the ATLAS detector using 2011 data*, tech. rep. ATLAS-CONF-2012-020, CERN, 2012, URL: `https://cds.cern.ch/record/1430034` (cit. on p. 55).

[80] G. Aad et al., *Electron reconstruction and identification efficiency measurements with the ATLAS detector using the 2011 LHC proton-proton collision data*, Eur. Phys. J. **C74** (2014) 2941, arXiv: `1404.2240 [hep-ex]` (cit. on p. 57).

[81] S. van der Meer, *Calibration of the effective beam height in the ISR*, tech. rep. CERN-ISR-PO-68-31. ISR-PO-68-31, CERN, 1968, URL: `https://cds.cern.ch/record/296752` (cit. on p. 69).

[82] M. Aaboud et al., *Luminosity determination in pp collisions at $\sqrt{s}$ = 8 TeV using the ATLAS detector at the LHC*, Eur. Phys. J. **C76** (2016) 653, arXiv: `1608.03953 [hep-ex]` (cit. on p. 69).

[83] G. Aad et al., *Jet energy resolution in proton-proton collisions at $\sqrt{s}$ = 7 TeV recorded in 2010 with the ATLAS detector*, Eur. Phys. J. **C73** (2013) 2306, arXiv: `1210.6210 [hep-ex]` (cit. on p. 69).

[84] *Expected performance of missing transverse momentum reconstruction for the ATLAS detector at $\sqrt{s}$ = 13 TeV*, tech. rep. ATL-PHYS-PUB-2015-023, CERN, 2015, URL: `https://cds.cern.ch/record/2037700` (cit. on p. 69).

[85] G. Aad et al., *Performance of b-Jet Identification in the ATLAS Experiment*, JINST **11** (2016) P04008, arXiv: `1512.01094 [hep-ex]` (cit. on p. 70).

[86] J. R. Andersen et al., *Handbook of LHC Higgs Cross Sections: 3. Higgs Properties*, (2013), ed. by S. Heinemeyer et al., arXiv: `1307.1347 [hep-ph]` (cit. on pp. 72, 162).

[87] A. Djouadi, J. Kalinowski and M. Spira, *HDECAY: A Program for Higgs boson decays in the standard model and its supersymmetric extension*, Comput. Phys. Commun. **108** (1998) 56, arXiv: `hep-ph/9704448 [hep-ph]` (cit. on pp. 72, 162).

[88] I. W. Stewart and F. J. Tackmann, *Theory uncertainties for Higgs mass and other searches using jet bins*, Phys. Rev. D **85** (3 2012) 034011, URL: https://link.aps.org/doi/10.1103/PhysRevD.85.034011 (cit. on pp. 72, 163).

[89] W. Verkerke and D. Kirkby, *The RooFit toolkit for data modeling*, (2003), arXiv: physics/0306116 (cit. on p. 72).

[90] L. Moneta et al., *The RooStats project*, (2010), arXiv: 1009.1003 (cit. on p. 72).

[91] R. Barlow and C. Beeston, *Fitting using finite Monte Carlo samples*, Computer Physics Communications **77** (1993) 219, ISSN: 0010-4655, URL: http://www.sciencedirect.com/science/article/pii/001046559390005W (cit. on p. 76).

[92] G. Aad et al., *Search for Higgs and Z Boson Decays to J/ψ and $\Upsilon(nS)$ with the ATLAS Detector*, Phys. Rev. Lett. **114** (2015) 121801, arXiv: 1501.03276 [hep-ex] (cit. on p. 85).

[93] G. Perez et al., *Constraining the charm Yukawa and Higgs-quark coupling universality*, Phys. Rev. D **92** (3 2015) 033016, URL: https://link.aps.org/doi/10.1103/PhysRevD.92.033016 (cit. on p. 85).

[94] C. Delaunay et al., *Enhanced Higgs boson coupling to charm pairs*, Phys. Rev. D **89** (3 2014) 033014, URL: https://link.aps.org/doi/10.1103/PhysRevD.89.033014 (cit. on p. 85).

[95] *Search for $H^0 \to b\bar{b}$ or $c\bar{c}$ in association with a W or Z boson in the forward region of pp collisions*, (2016), URL: https://cds.cern.ch/record/2209531 (cit. on p. 85).

[96] M. Aaboud et al., *Search for the Decay of the Higgs Boson to Charm Quarks with the ATLAS Experiment*, (2018), arXiv: 1802.04329 [hep-ex] (cit. on pp. 85, 94–96, 179).

[97] K. S. Cranmer, *Kernel estimation in high-energy physics*, Comput. Phys. Commun. **136** (2001) 198, arXiv: hep-ex/0011057 [hep-ex] (cit. on p. 93).

[98] S. Schael et al., *Search for neutral MSSM Higgs bosons at LEP*, Eur. Phys. J. **C47** (2006) 547, arXiv: hep-ex/0602042 [hep-ex] (cit. on p. 99).

[99] G. C. Branco et al., *Theory and phenomenology of two-Higgs-doublet models*, Phys. Rept. **516** (2012) 1, arXiv: 1106.0034 [hep-ph] (cit. on p. 99).

[100] G. Aad et al., *Search for a CP-odd Higgs boson decaying to Zh in pp collisions at $\sqrt{s} = 8$ TeV with the ATLAS detector*, Phys. Lett. **B744** (2015) 163, arXiv: 1502.04478 [hep-ex] (cit. on p. 99).

[101] R. D. Ball et al., *Parton distributions with LHC data*, Nucl. Phys. **B867** (2013) 244, arXiv: 1207.1303 [hep-ph] (cit. on p. 99).

[102] *Search for a CP-odd Higgs boson decaying to Zh in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector*, tech. rep. ATLAS-CONF-2016-015, CERN, 2016, URL: http://cds.cern.ch/record/2141003 (cit. on p. 102).

[103] M. Aaboud et al., *Search for heavy resonances decaying into a W or Z boson and a Higgs boson in final states with leptons and b-jets in 36 fb^{-1} of $\sqrt{s} = 13$ TeV pp collisions with the ATLAS detector*, (2017), arXiv: 1712.06518 [hep-ex] (cit. on p. 111).

APPENDIX A

Appendix - Jet Energy Regression

A.1 Regression Training Parameters

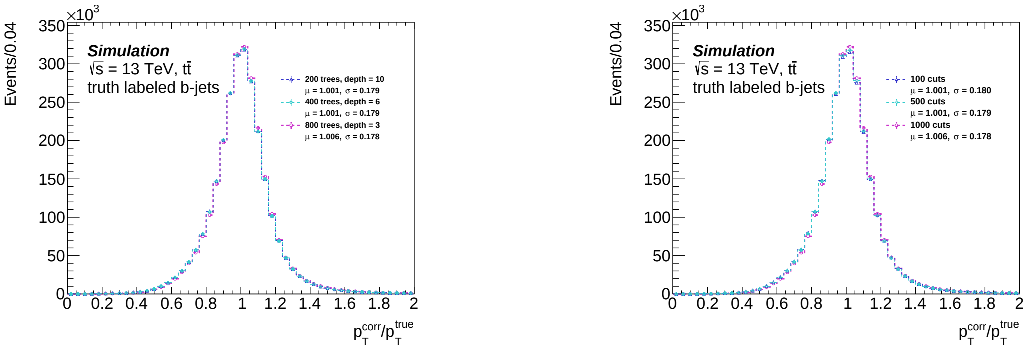
TMVA allows several training parameters of the BDTs to be modified. For the jet energy regression only the granularity of the cuts imposed on the input variables, was modified. All other parameters were taken with their default values. Table A.1 summarises the parameters and their values used for the regression. Modifying the number of trees and their individual depth was tested but no differences in the performance could be observed as shown in fig A.1.

Parameter	Description	Value
NTrees	number of decision trees	800
MaxDepth	maximum allowed depth of each decision tree	3
MinNodeSize	minimum amount of events in each node (as percentage of total events)	0.2%
nCuts	granularity of the cuts allowed on the input variables	1000
BoostType	algorithm used for boosting	gradient boost
Shrinkage	learning rate	1

Table A.1: BDTs training parameters that can be modified for the training and their explanation. The last column indicates which values were used to train the jet energy regression.

A.2 Input Variable Optimisation

The aim of the optimisation of the jet energy regression input variables is to find the smallest possible set of variables without compromising the performance. In total 24 variables were considered. For simplicity not all possible combinations of subsets were investigated. Instead subsets of highly correlated variables were identified and within these subsets of 3 to 5 variables all possible combinations were tested and the best one was identified. Then this best combination was the baseline set for the next subset to be investigated, i.e. new variables from the subset under investigation were added to this best combination. This test procedure is shown in fig. A.2. All variables and subsets are listed in tab. A.2. The measure for performance is the standard deviation of the distribution of the jet energy regression corrected p_{T} divided by the truth jet p_{T}. If several subsets yield the same performance, within their statistical uncertainties, the one with the smallest amount of variables was chosen. The final subset includes 9 variables which are marked in tab A.2 and further explained in section 7.1.



(a) number of trees + depth

(b) number of cuts

Figure A.1: Distribution of the jet energy regression corrected p_{T} divided by the truth jet p_{T} for variations of a) the number of decision trees and their maximum depth and b) the number of cuts allowed on the regression variables. The mean and standard deviation of the distributions are given and there is no observable difference in performance.

A.3 Jet Energy Regression in the $t\bar{t}$ Validation Region

The description of the input variables in the MC in comparison to data is validated. A dedicated event selection is applied which targets the $tt \to bb(W \to e\nu)(W \to \mu\nu)$ decay and selects $t\bar{t}$ events with a very large purity. This is achieved with the requirement that there is exactly one muon, exactly one electron (this region will also be referred to as $e\mu$ region) and exactly two jets that pass the 70% b-tagging requirement present in the event. The events are required to pass the standard single lepton trigger and the objects have to pass standard quality requirements and have to pass the standard overlap removal procedure as described in chapter 4. The leptons have to pass the *VH loose* lepton requirement and at least one of them also has to pass the *VH tight* requirement as they are defined in sec. 5. In addition, the leptons have to be of opposite electrical charge and their invariant mass $m_{e\mu}$ has to be at least 50 GeV to reduce multi-jet events. The jets have to fulfil the *signal* jet criteria which are also defined in sec. 5 and pass the aforementioned b-tagging requirement. The agreement between data and MC is checked in a kinematic region similar to the kinematic region of the jet energy regression training sample (see fig. A.3). The same $t\bar{t}$ MC as utilised for the jet energy regression training is used — which is Powheg interfaced to Pythia for parton showering and underlying event description. The data set that is used corresponds to 36.1 fb^{-1} of ATLAS data recorded in the years 2015 and 2016. The comparisons of the jet energy regression input variables in data and simulated events are shown in figure A.4.

A.4 Zb Reweighting

To compensate for discontinuities in the Sherpa 2.2.1 samples due to the slicing the events for the Zb selection are weighted per event based on their p_{T}^{Z}. The corresponding reweighting histogram is extracted from the discrepancy between data and MC simulation in the p_{T}^{Z} distribution (fig. A.7). The bins have variable sizes to make sure that the statistical uncertainty in each bin never exceeds 3%. For the reweighting procedure the data and MC events were normalised to the same number of events and therefore only differences in the shape are corrected. In an additional step a global normalisation factor of 1.35 is extracted by fitting a straight line through the remaining difference between data and MC in the p_{T}^{Z} distribution. The normalisation discrepancy stems from the b-tagging requirement. Since the shape and normalisation discrepancies originate from different effect they are corrected separately. The nominal p_{T}^{Z} distribution, the p_{T}^{Z} distribution after the reweighting and the p_{T}^{Z} distribution after the scale factor was

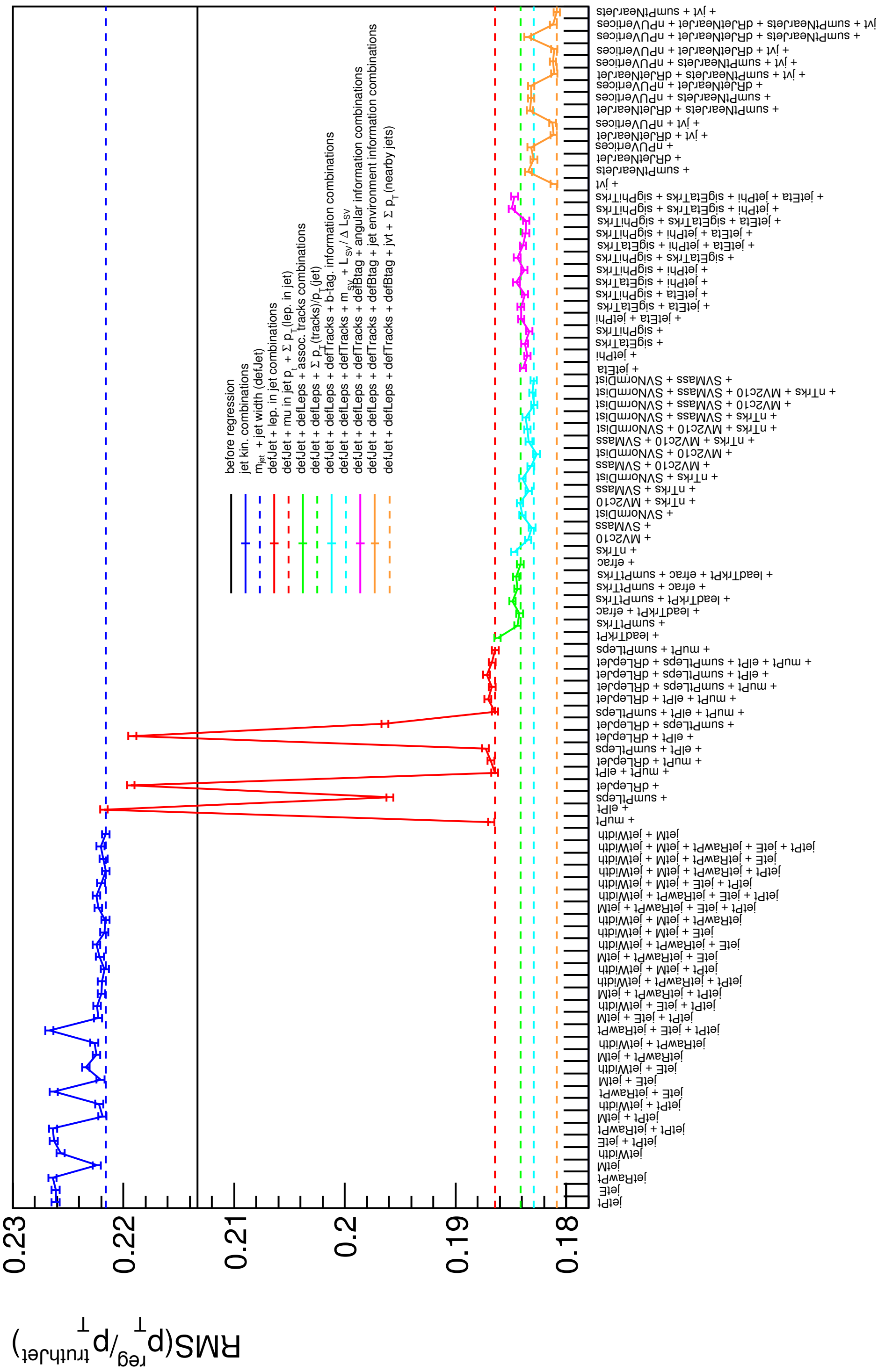


Figure A.2: Evolution of the standard deviation of the jet energy regression corrected p_T divided by the truth jet p_T for all tested variable sets. Each colour corresponds to a subset that was tested independently. The dashed horizontal lines correspond to the standard deviation of the combination of input variables that was chosen for each corresponding subset. The chosen combination of a previous subset was the starting point for adding new variables from the following subset.

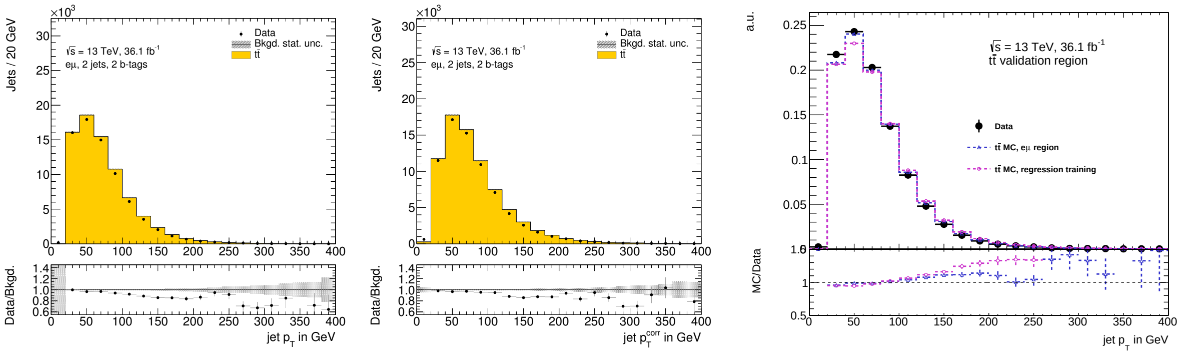
Name	Description	Used?
jetPt	transverse momentum p_T of the jet	-
jetE	energy E of the jet	-
jetM	mass m of the jet	✓
jetWidth	width of the jet	✓
jetRawPt	p_T of the jet at the electro-magnetic scale ($\equiv$ not calibrated)	-
muPt	sum of p_T of muons inside the jet	✓
elPt	sum of p_T of electrons inside the jet	-
sumPtLeps	sum of p_T of electrons and muons inside the jet	✓
dRLepJet	distance ΔR between the jet axis and the nearest lepton inside the jet	-
leadTrkPt	p_T of the leading track associated to the jet	-
sumPtTrks	sum of p_T of all tracks associated to the jet	-
efrac	fraction of p_T carried by the tracks associated to the jet ($\equiv \sum\limits_{\mathrm{track}_i} p_\mathrm{T}^{\mathrm{track}_i}/p_\mathrm{T}^{\mathrm{jet}}$)	✓
nTrks	number of tracks associated to the jet	-
MV2c10	b-tagging weight of the jet obtained with the MV2c10 algorithm	-
SVMass	mass m of the secondary vertex associated to the jet	✓
SVNormDist	ratio of 3D distance between primary and secondary vertex and its error	✓
jetEta	pseudo-rapidity η of the jet	-
jetPhi	azimuthal angle ϕ of the jet	-
sigEtaTrks	standard deviation in η of the associated tracks	-
sigPhiTrks	standard deviation in ϕ of the associated tracks	-
jvt	jet vertex tagger weight of the jet	✓
sumPtNearJets	sum of p_T of jets close to the jet (within $\Delta R \leq 1$)	✓
dRJetNearJet	distance ΔR between the jet axis and the axis of the nearest jet (within $\Delta R \leq 1$)	-
nPUVertices	number of pile-up vertices	-

Table A.2: Full list of input variables that were considered for the regression including their names and a short description. The last column indicates if the input variables was used in the final training of the jet energy regression. The horizontal lines indicate the subsets that were defined for the variable set optimisation.

applied are shown in fig. A.5. Figure A.6 shows a comparison between data and MC for the distribution of the transverse momentum of the b-jet before and after the reweighting + scaling was applied and before and after the application of the jet energy regression. No major differences are observed between the reconstructed jets and the jets after they were additionally corrected using the jet energy regression. The reweighting has a consistent effect on $p_\mathrm{T}^\mathrm{jet}$ and p_T^Z which is demonstrated in figure A.6 since the measure $p_\mathrm{T}^\mathrm{jet}/p_\mathrm{T}^Z$ of the Zb cross check is not affected.

A.5 Zb cross check: $p_\mathrm{T}^\mathrm{jet}/p_\mathrm{T}^Z$ distributions

The comparisons of the $p_\mathrm{T}^\mathrm{jet}/p_\mathrm{T}^Z$ distributions in data and simulated events for different phase spaces are shown in figure A.8 for the nominal jets, i.e. after SJC, and in figure A.9 for the regression corrected jets. The distributions comparing the nominal and regression corrected jets by fitting a Bukin function are



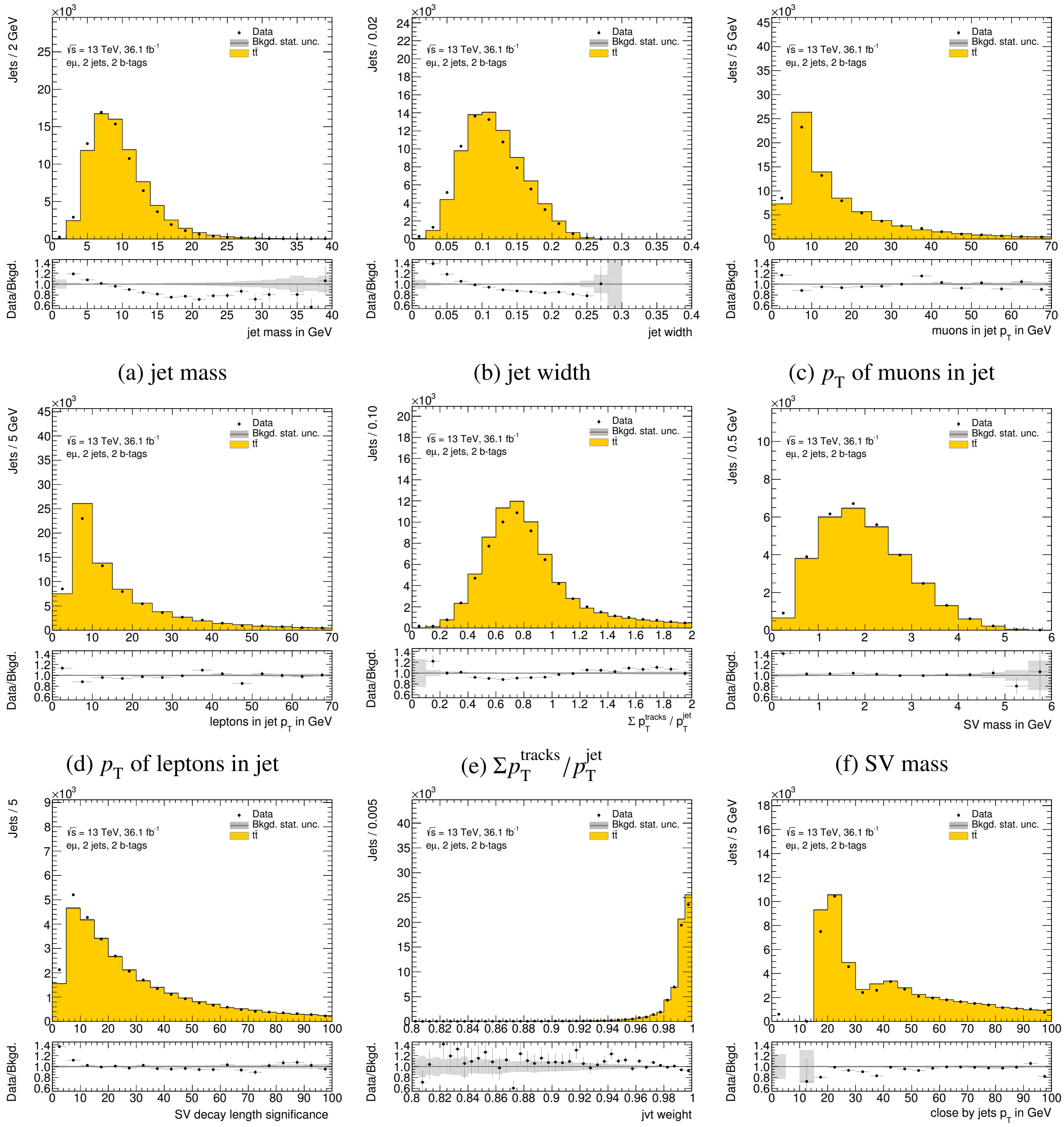
(a) $p_{\mathrm{T}}^{\mathrm{reco}}$

(b) $p_{\mathrm{T}}^{\mathrm{corr}}$

(c)

Figure A.3: Data to MC comparison for $p_{\mathrm{T}}^{\mathrm{jet}}$ a) before and b) after the jet energy regression was applied in the $t\bar{t}$ validation region. c) Compares $p_{\mathrm{T}}^{\mathrm{jet}}$ for data, the $t\bar{t}$ validation region and the actual $p_{\mathrm{T}}^{\mathrm{jet}}$ distribution present in the training. The distributions were all normalised to the same area.

shown in figure A.10 for data events and in figure A.11. The distribution for truth labelled c-jets and light jets is shown in figure A.12.



(a) jet mass

(b) jet width

(c) p_{T} of muons in jet

(d) p_{T} of leptons in jet

(e) $\Sigma p_{\mathrm{T}}^{\mathrm{tracks}}/p_{\mathrm{T}}^{\mathrm{jet}}$

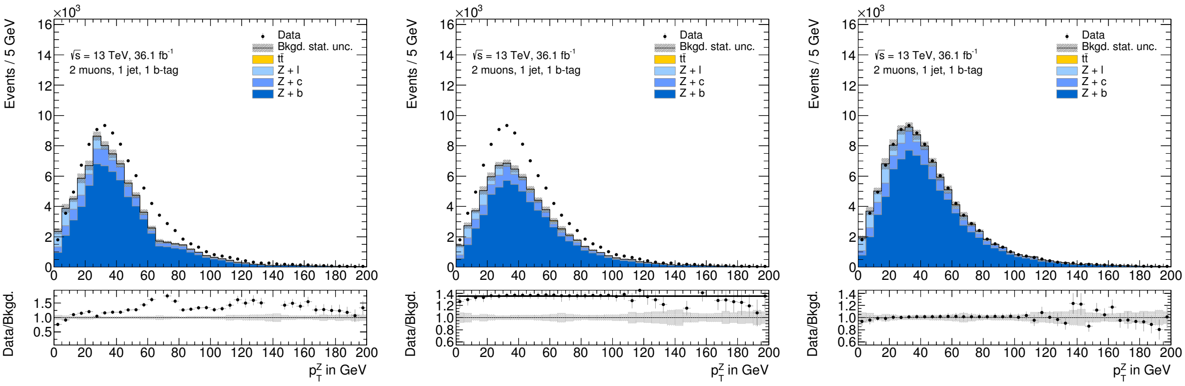
(f) SV mass

(g) SV L_{3D} significance

(h) jvt weight

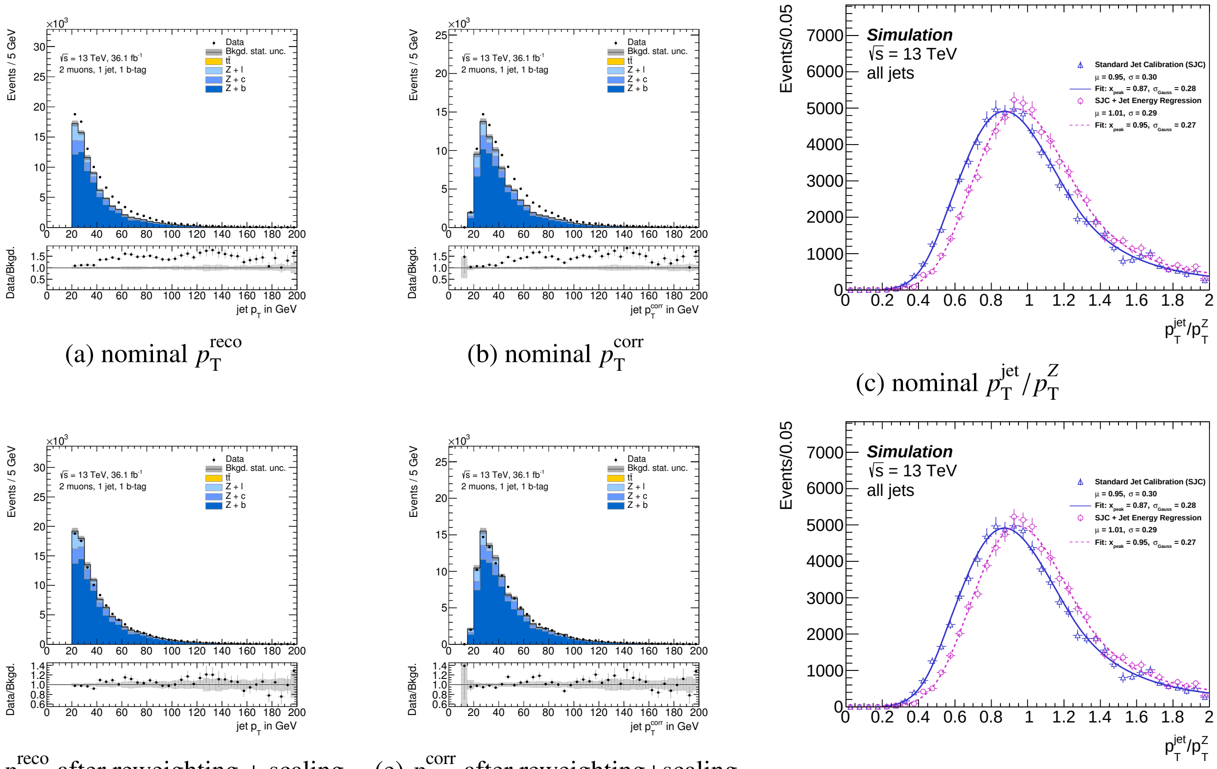
(i) p_{T} of close by jets

Figure A.4: Data to MC comparison of the jet energy regression input variables in the $t\bar{t}$ validation region. The kinematic selection mimics the kinematics present during the jet energy regression training.



(a) nominal p_T^Z (b) p_T^Z after reweighting (c) p_T^Z after reweighting+scaling

Figure A.5: Comparison of data and MC of the p_T^Z distribution. (a) Due to the Sherpa 2.2.1 Z+jets slicing discontinuities appear for selection requiring exactly one b-tagged jet. (b) After a bin wise reweighting as a function of p_T^Z is applied an additional normalisation factor is extracted from a straight line fit. (c) Good data and MC is achieved after the reweighting and scaling is applied.



(a) nominal p_T^{reco} (b) nominal p_T^{corr} (c) nominal p_T^{jet}/p_T^Z

(d) p_T^{reco} after reweighting + scaling (e) p_T^{corr} after reweighting+scaling (f) p_T^{jet}/p_T^Z after reweighting+scaling

Figure A.6: Comparison of data and MC of the nominal p_T^{jet} distribution a) after the SJC and b) after the jet energy regression is applied on top of it. The same distributions are shown in d) and e) respectively after the Sherpa 2.2.1 reweighting and scaling is applied. After the corrections good agreement between data and MC is observed with no major differences between p_T^{reco} and p_T^{corr}. The reweighting and scaling is consistently applied and has no influence on the measure p_T^{jet}/p_T^Z of the Zb validation which is shown in c) nominal and f) after the reweighting and scaling.

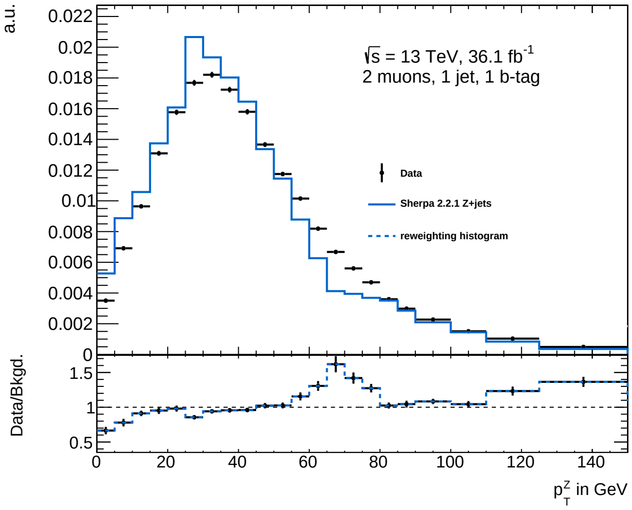
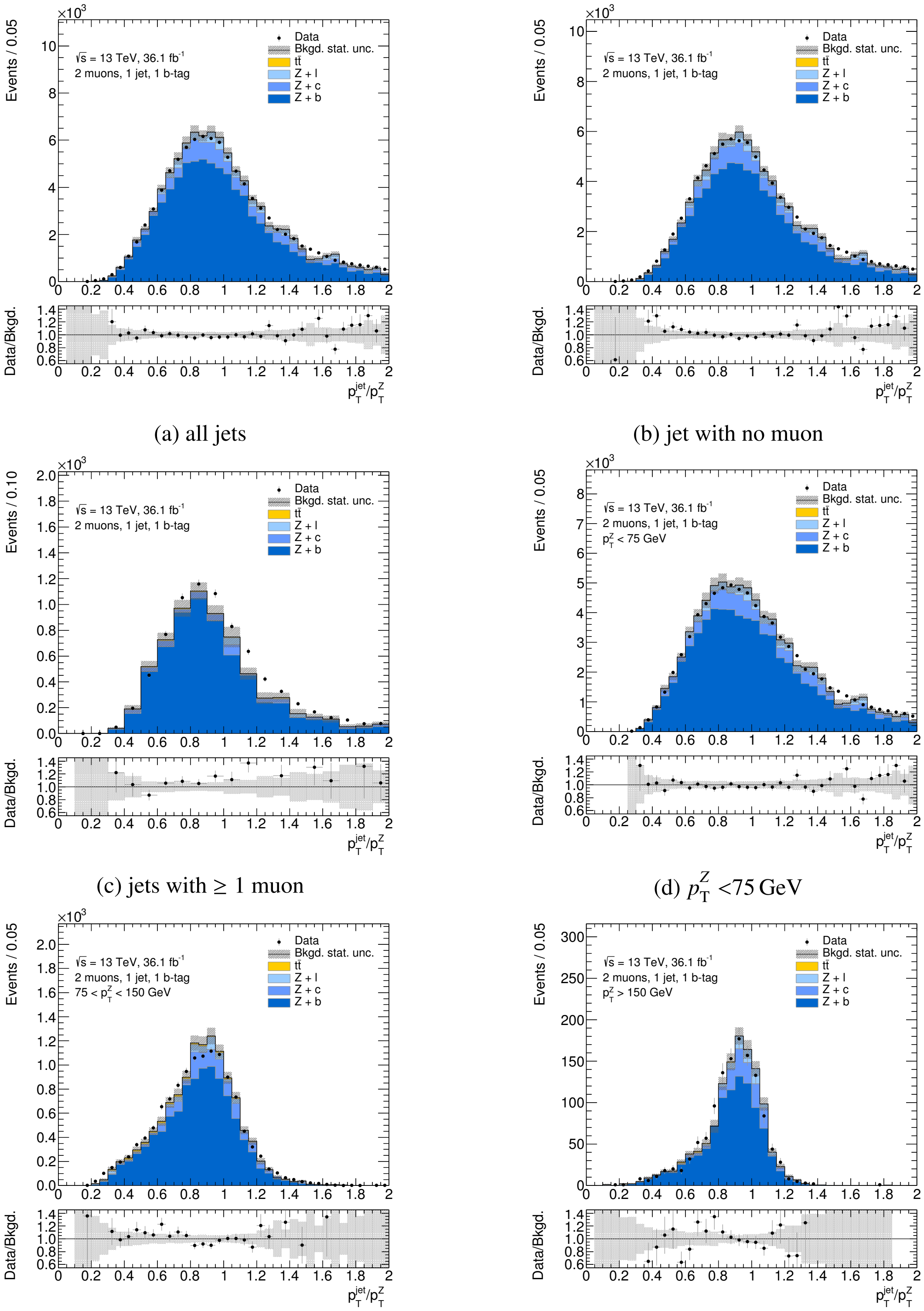


Figure A.7: The p_{T}^{Z} distribution in the SHERPA 2.2.1 Z+jets MC and in data normalised to the same number of events. The bin size varies and ensures that the statistical uncertainty in any bin does not exceed 3%. A reweighting histogram is extracted from the ratio of data to Z+jets MC events.



(a) all jets

(b) jet with no muon

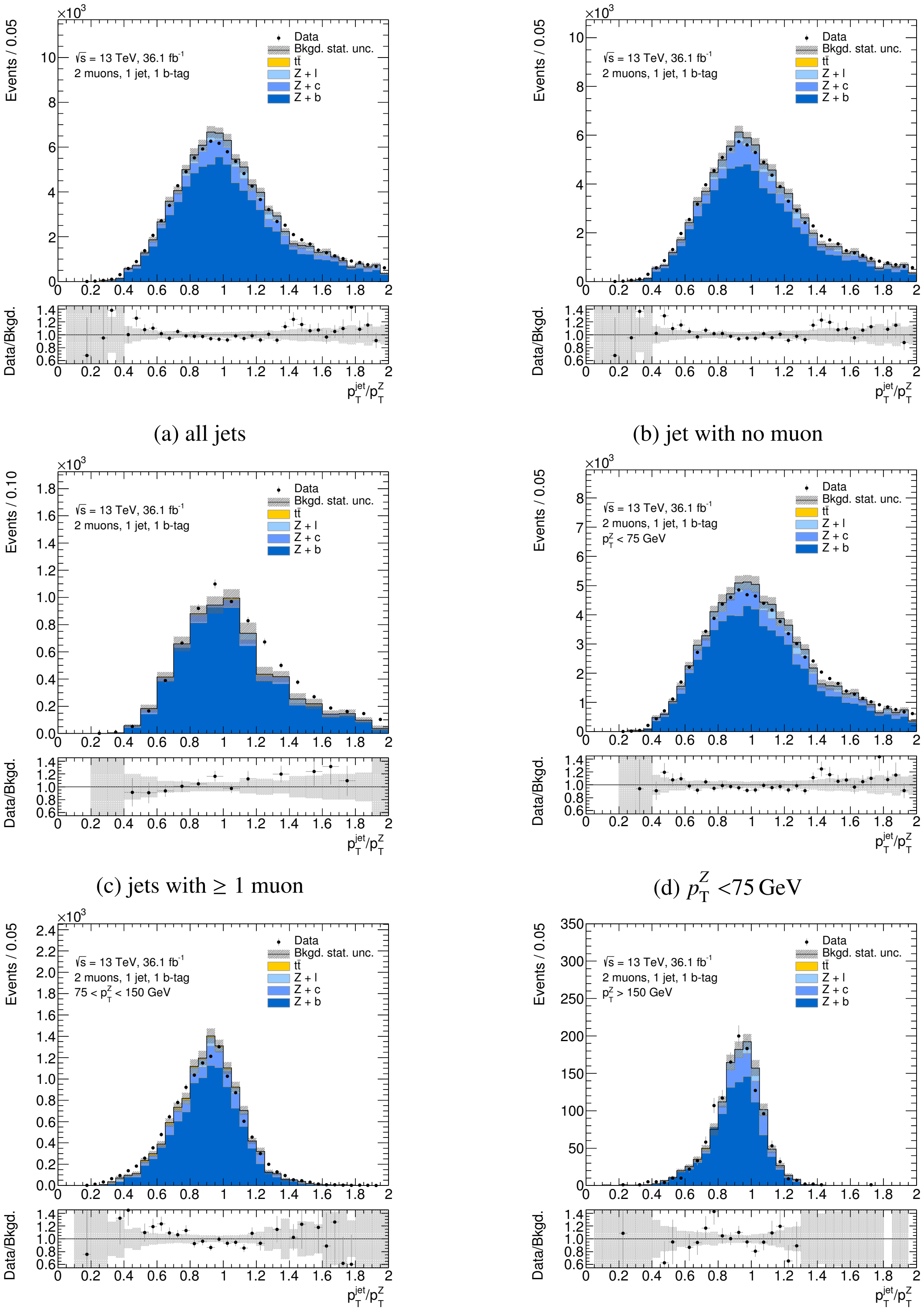
(c) jets with ≥ 1 muon

(d) p_{T}^{Z} <75 GeV

(e) 75 GeV< p_{T}^{Z} <150 GeV

(f) p_{T}^{Z} >150 GeV

Figure A.8: Comparisons of the $p_{\mathrm{T}}^{\mathrm{jet}}/p_{\mathrm{T}}^{Z}$ distributions in data and simulated events for different phase spaces for the *b*-jets that were calibrated with the SJC.



(a) all jets

(b) jet with no muon

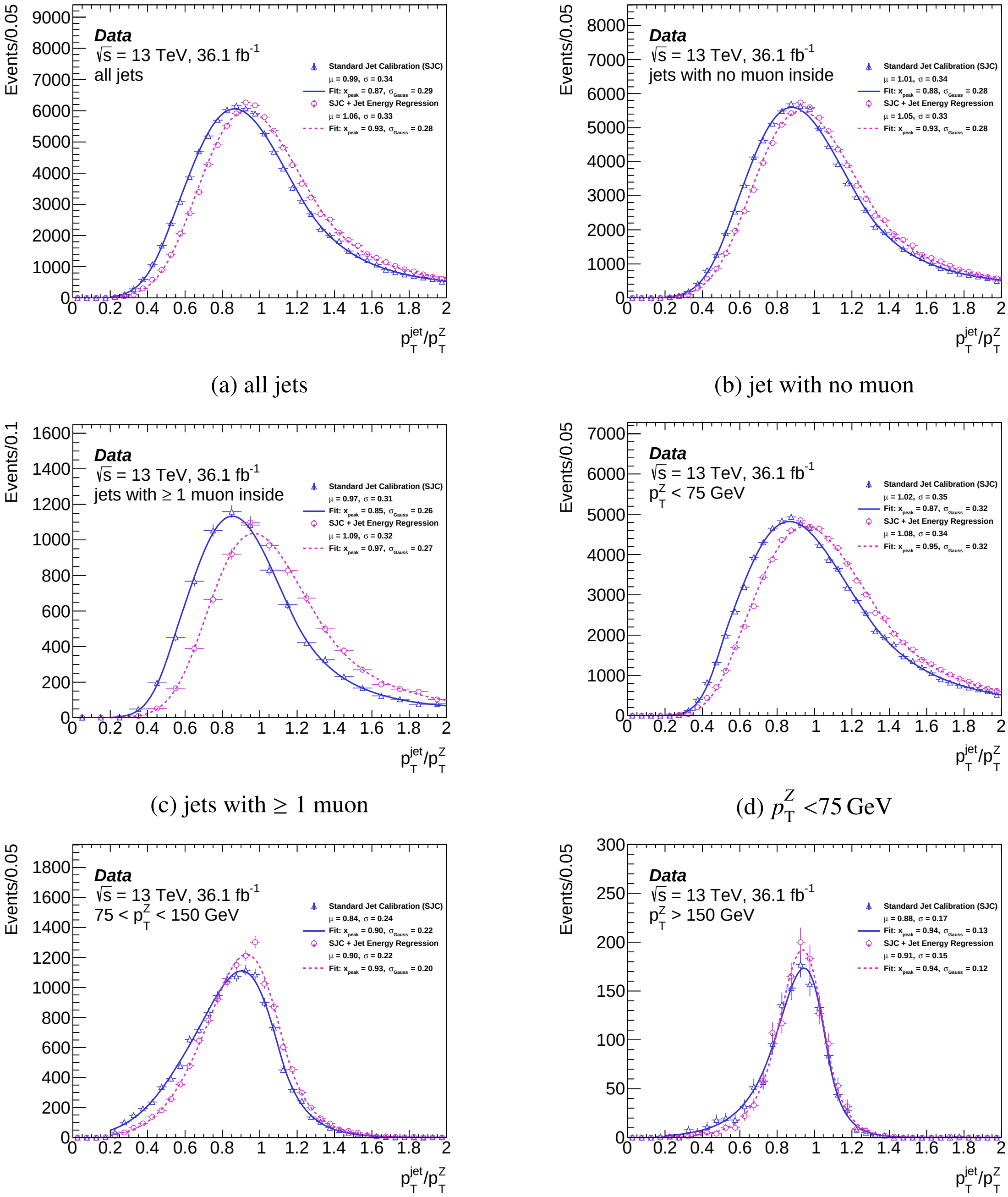
(c) jets with ≥ 1 muon

(d) p_{T}^{Z} <75 GeV

(e) 75 GeV< p_{T}^{Z} <150 GeV

(f) p_{T}^{Z} >150 GeV

Figure A.9: Comparisons of the $p_{\mathrm{T}}^{\mathrm{jet}}/p_{\mathrm{T}}^{Z}$ distributions in data and simulated events for different phase spaces for the *b*-jets that were corrected with the jet energy regression.



(a) all jets

(b) jet with no muon

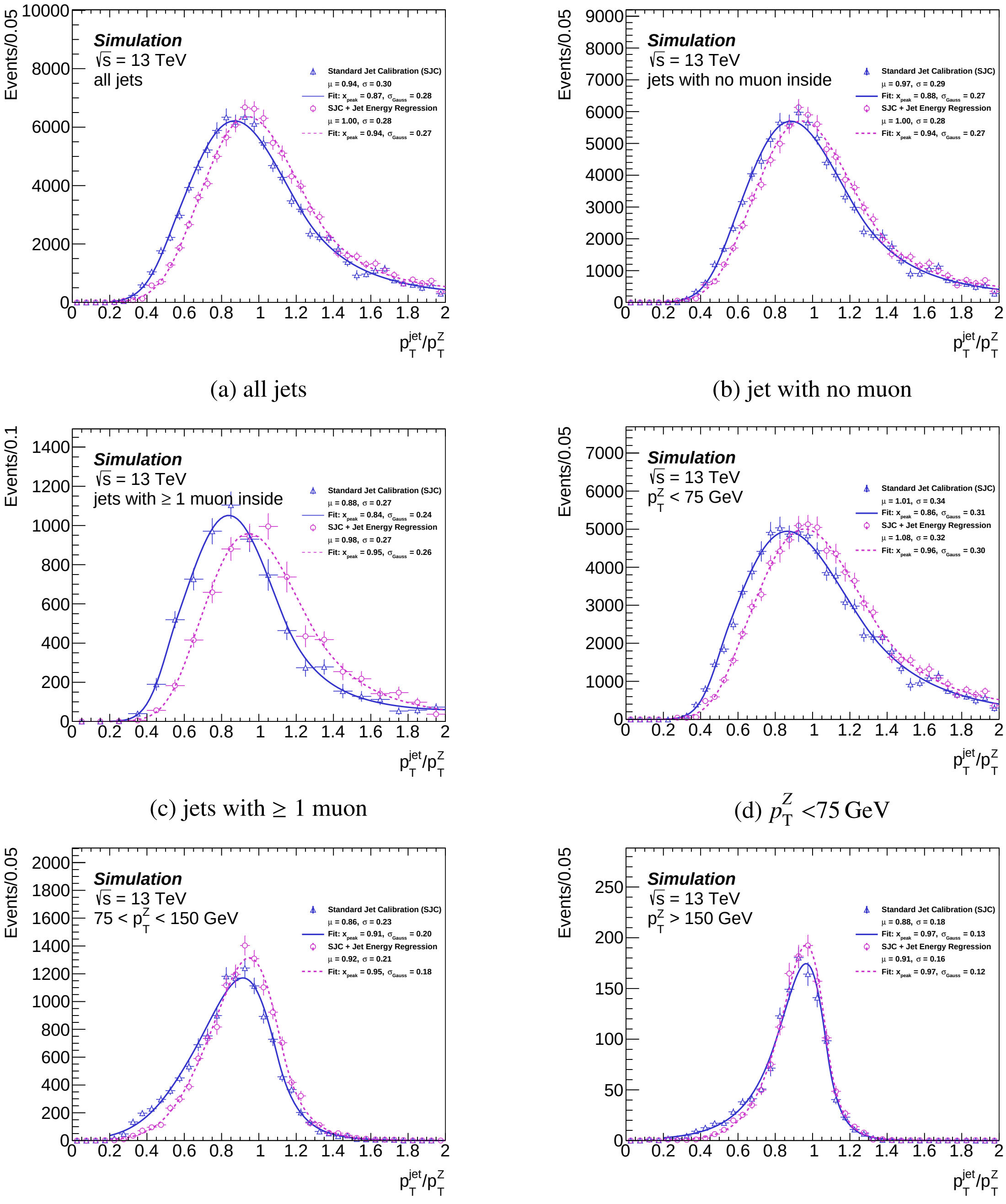
(c) jets with ≥ 1 muon

(d) p_T^Z <75 GeV

(e) 75 GeV< p_T^Z <150 GeV

(f) p_T^Z >150 GeV

Figure A.10: The p_T^{jet}/p_T^Z distributions in data events comparing the nominal and regression corrected jets by fitting a Bukin function.



(a) all jets

(b) jet with no muon

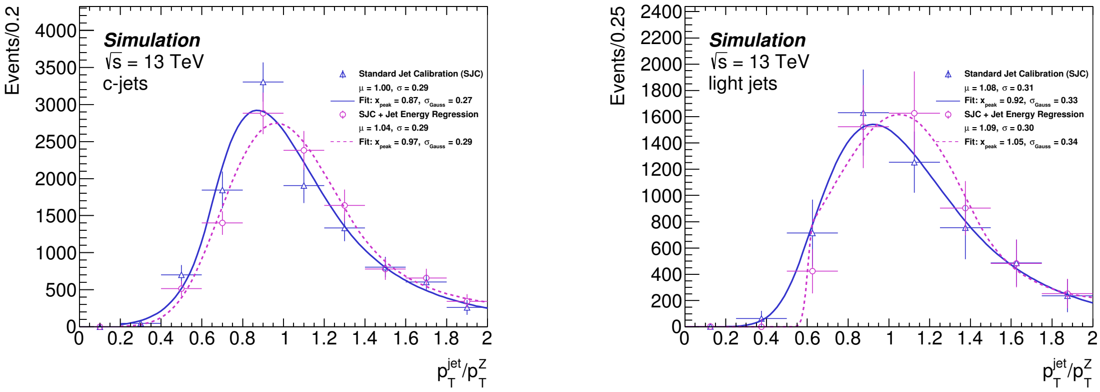
(c) jets with ≥ 1 muon

(d) p_T^Z <75 GeV

(e) 75 GeV< p_T^Z <150 GeV

(f) p_T^Z >150 GeV

Figure A.11: The p_T^{jet}/p_T^Z distributions in simulated events comparing the nominal and regression corrected jets by fitting a Bukin function.



(a) *c*-jets

(b) light jets

Figure A.12: The $p_{\mathrm{T}}^{\mathrm{jet}}/p_{\mathrm{T}}^{Z}$ distributions for truth labelled *c*-jets and light jets in simulated *Z*+jets events comparing nominal and regression corrected *b*-jets. A Bukin function is fitted to all distributions.

APPENDIX B

Appendix - SM $Z(H \rightarrow b\bar{b})$ Search

B.1 MVA Setup

The training parameters of the BDTs that were set for the training of the $Z(H \rightarrow b\bar{b})$ BDTs. They were optimised for the $V(H \rightarrow b\bar{b})$ search with run 1 ATLAS data [66] by scanning the parameter phase space in order to get the best separation between signal and background. These optimised parameters were found to be suitable for the analysis with run 2 data as well and they are listed in table B.1.

Parameter	Description	Value
NTrees	number of decision trees	200
MaxDepth	maximum allowed depth of each decision tree	4
MinNodeSize	minimum amount of events in each node (as percentage of total events)	5%
nCuts	granularity of the cuts allowed on the input variables	100
BoostType	algorithm used for boosting	adaptive boost
AdaBoostBeta	learning rate	0.15
SeparationType	separation criterion for node splitting	Gini Index

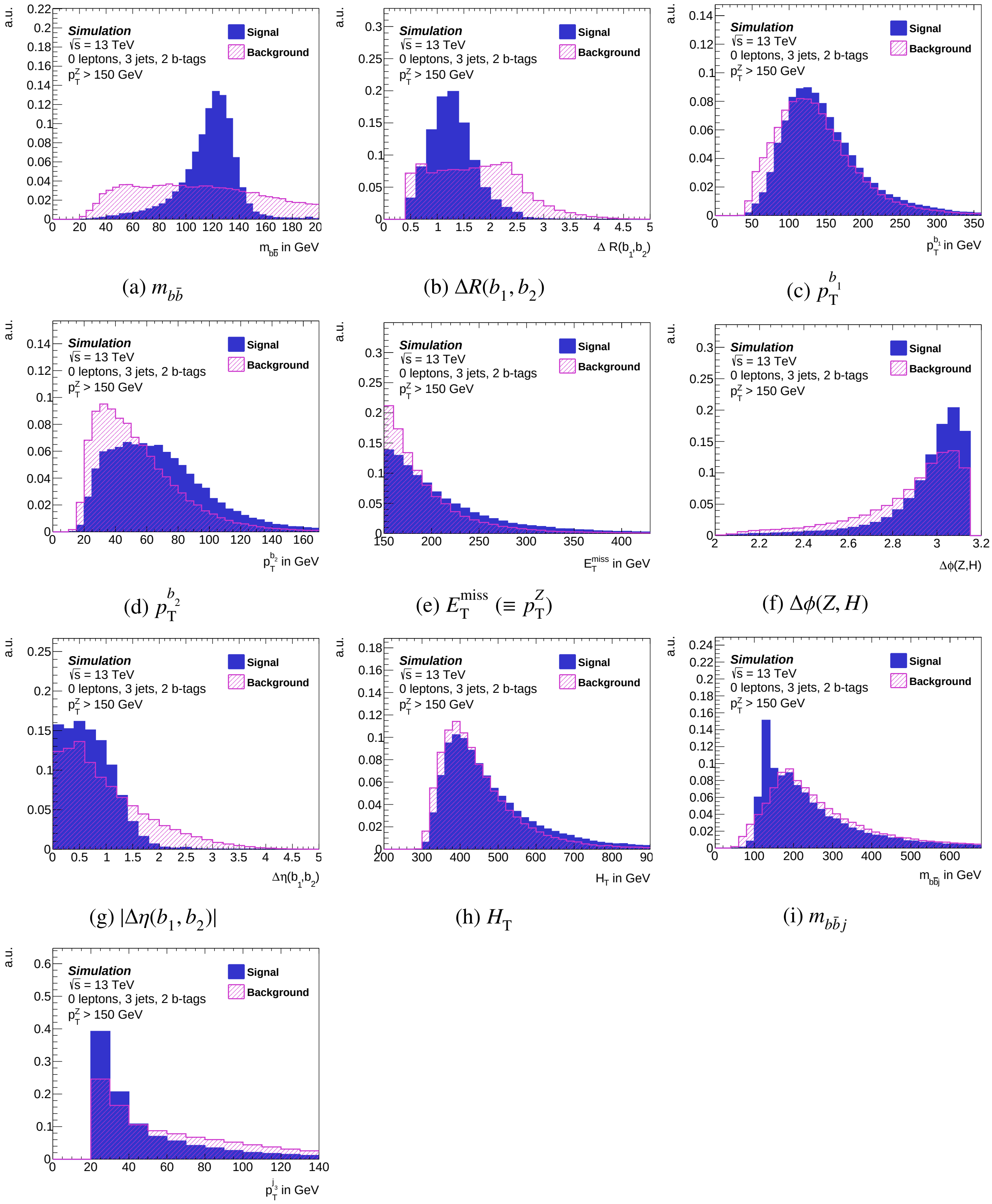
Table B.1: BDTs training parameters that can be modified for the training of the $Z(H \rightarrow b\bar{b})$ BDTs and their explanation. The last column indicates which values were used in the training.

The training performance is deteriorated if variables have long tails since the granularity of the cuts the algorithm can set becomes worse. This is not a problem for angular variables. The ranges of the remaining BDTs input variables are restricted to a range that contains 99% of all signal events (modulo 5 GeV). If a variable of an event has a larger value then the value is set to the maximum value. Table B.2 summarises the maximum values for all input variables in all MVA regions.

B.1.1 MVA signal and background distributions

The comparison of the distributions of the BDTs input variables in simulated signal and background events are given in figure B.1 for the 0 lepton 3 jets category, in B.2 for the 2 lepton 2 jets medium p_{T}^{Z} category, in B.3 for the 2 lepton 3+ jets medium p_{T}^{Z} category and in figure B.4 for the 2 lepton 3+ jets high p_{T}^{Z} category.

The BDT output for signal and background for each analysis category for the trainings using events with an even event number B.5 and an odd event number B.6 including an *over training* check. For the *over training* check the BDTs output distribution for the training data set is compared to a statistically independent test data set.



(a) $m_{b\bar{b}}$

(b) $\Delta R(b_1, b_2)$

(c) $p_{\mathrm{T}}^{b_1}$

(d) $p_{\mathrm{T}}^{b_2}$

(e) $E_{\mathrm{T}}^{\mathrm{miss}}$ $(\equiv p_{\mathrm{T}}^{Z})$

(f) $\Delta\phi(Z, H)$

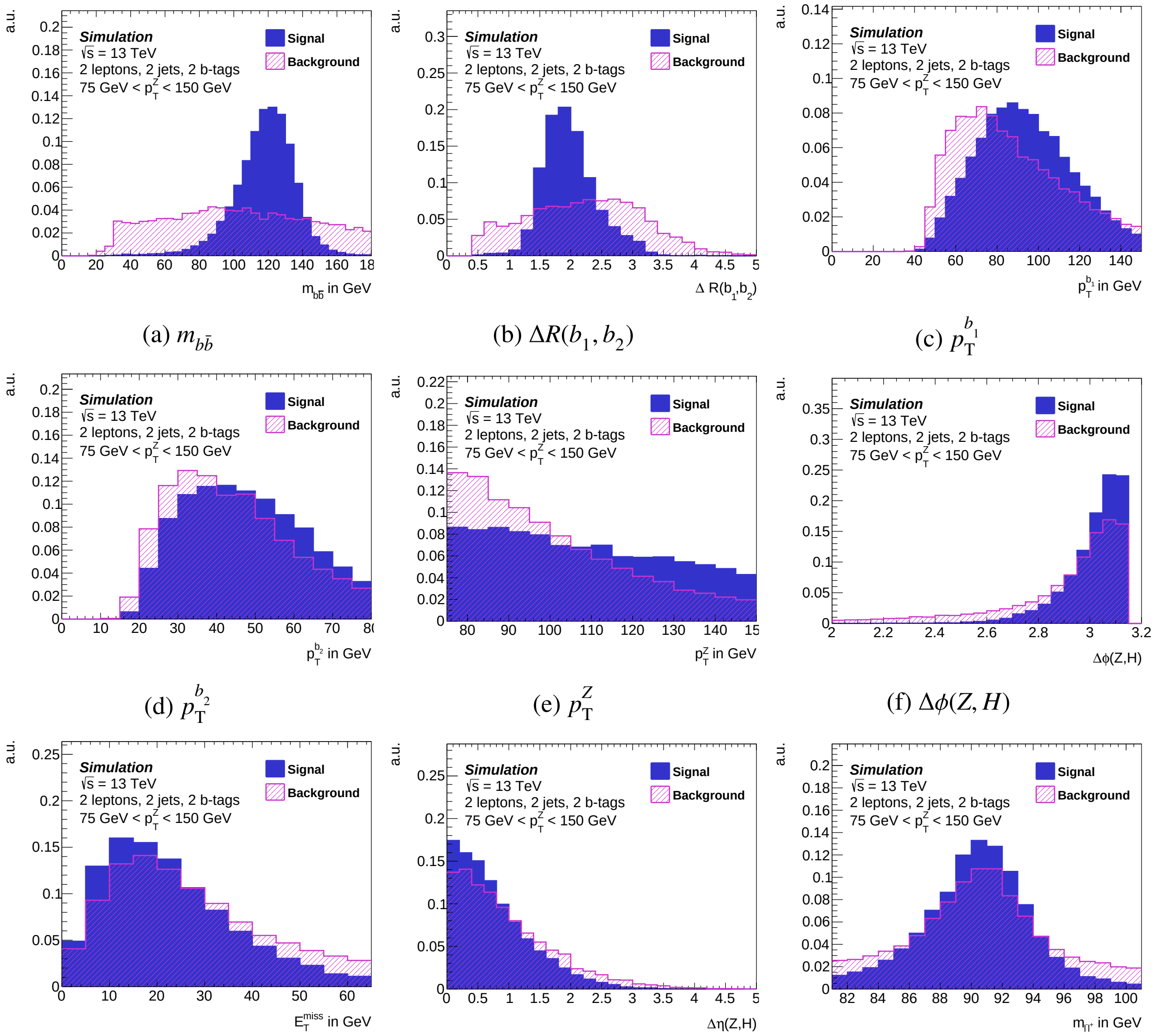
(g) $|\Delta\eta(b_1, b_2)|$

(h) H_{T}

(i) $m_{b\bar{b}j}$

(j) $p_{\mathrm{T}}^{j_3}$

Figure B.1: Input variables for the 0 lepton 3 jet MVA. Shown are the distributions in simulated signal events and the sum of all simulated background events. All distributions are normalised to the same area.



(a) $m_{b\bar{b}}$

(b) $\Delta R(b_1, b_2)$

(c) $p_{\mathrm{T}}^{b_1}$

(d) $p_{\mathrm{T}}^{b_2}$

(e) p_{T}^{Z}

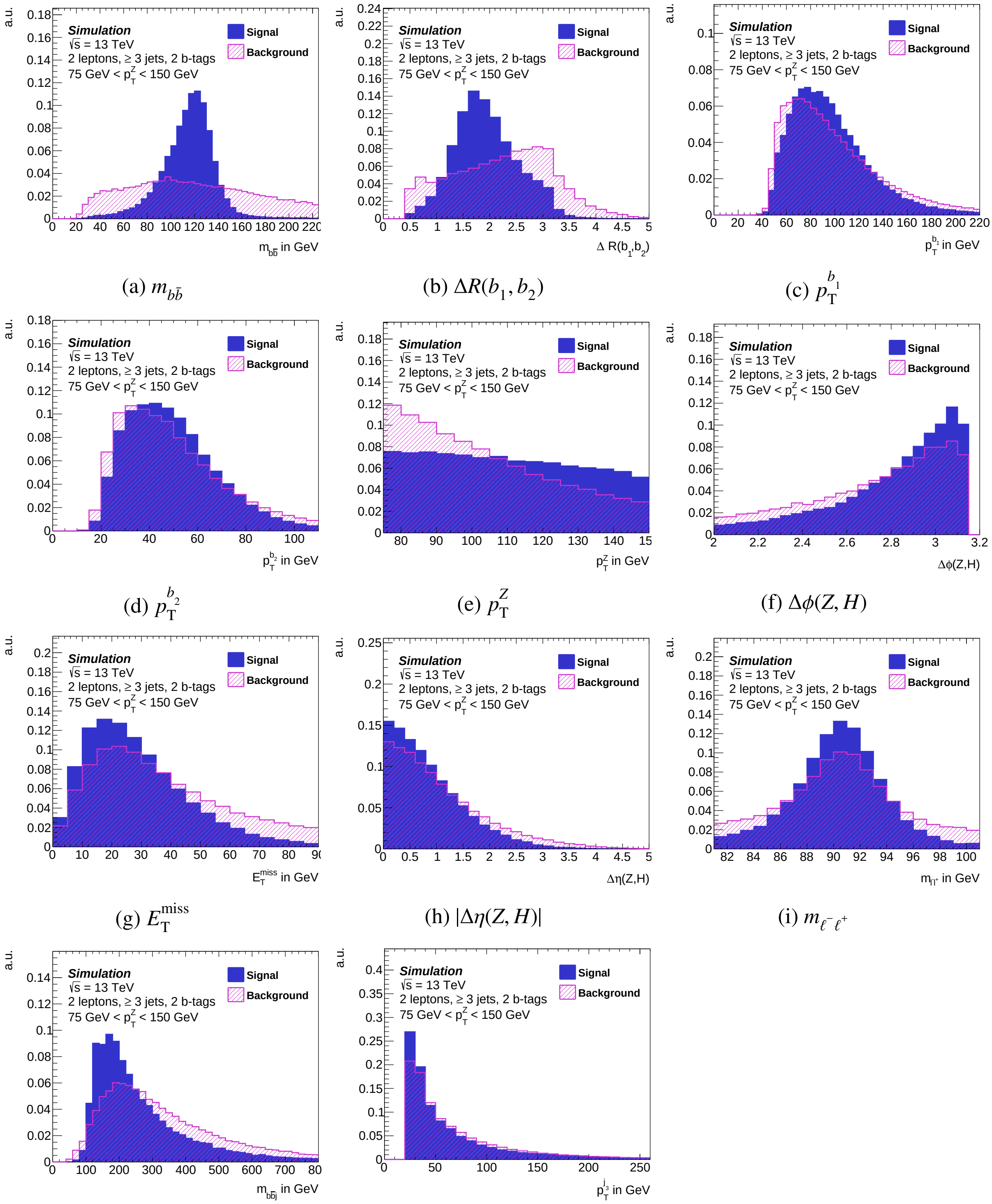
(f) $\Delta\phi(Z, H)$

(g) $E_{\mathrm{T}}^{\mathrm{miss}}$

(h) $|\Delta\eta(Z, H)|$

(i) $m_{\ell^-\ell^+}$

Figure B.2: Input variables for the 2 lepton 2 jet medium p_{T}^{Z} MVA. Shown are the distributions in simulated signal events and the sum of all simulated background events. All distributions are normalised to the same area.



(a) $m_{b\bar{b}}$

(b) $\Delta R(b_1, b_2)$

(c) $p_\mathrm{T}^{b_1}$

(d) $p_\mathrm{T}^{b_2}$

(e) p_T^{Z}

(f) $\Delta\phi(Z, H)$

(g) $E_\mathrm{T}^\mathrm{miss}$

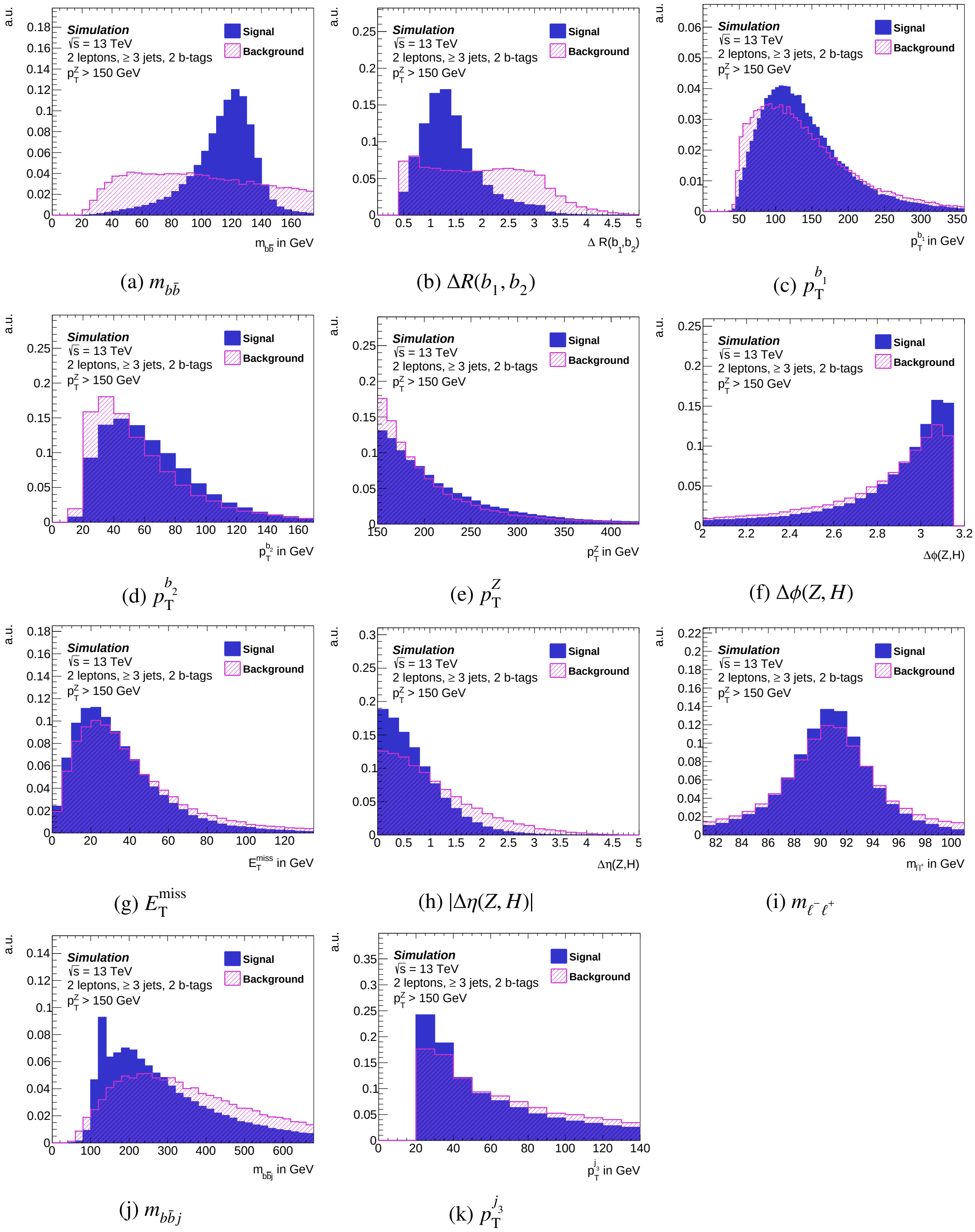
(h) $|\Delta\eta(Z, H)|$

(i) $m_{\ell^-\ell^+}$

(j) $m_{b\bar{b}j}$

(k) $p_\mathrm{T}^{j_3}$

Figure B.3: Input variables for the 2 lepton 3 jet medium p_T^Z MVA. Shown are the distributions in simulated signal events and the sum of all simulated background events. All distributions are normalised to the same area.



(a) $m_{b\bar{b}}$

(b) $\Delta R(b_1, b_2)$

(c) $p_T^{b_1}$

(d) $p_T^{b_2}$

(e) p_T^Z

(f) $\Delta\phi(Z, H)$

(g) E_T^{miss}

(h) $|\Delta\eta(Z, H)|$

(i) $m_{\ell^-\ell^+}$

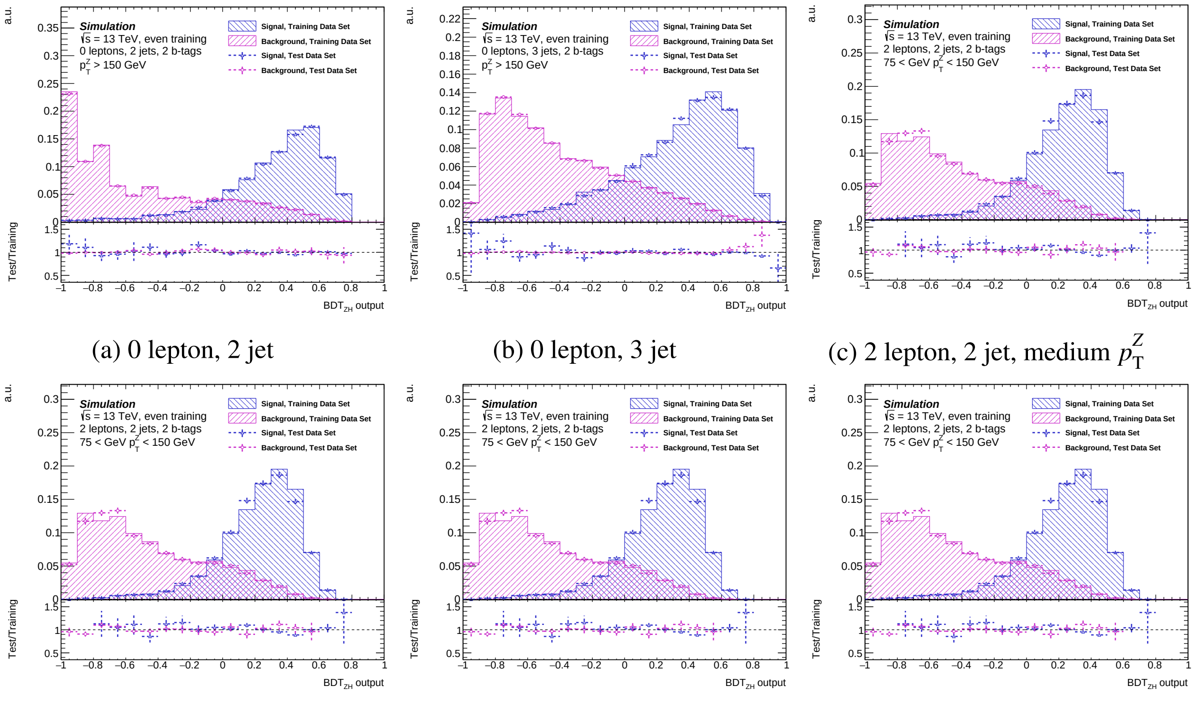
(j) $m_{b\bar{b}j}$

(k) $p_T^{j_3}$

Figure B.4: Input variables for the 2 lepton 3 jet high p_T^Z MVA. Shown are the distributions in simulated signal events and the sum of all simulated background events. All distributions are normalised to the same area.

Variable	Maximum value in GeV for region:					
	$0\ell 2j$	$0\ell 3j$	$2\ell 2j$ medium p_{T}^{Z}	$2\ell 2j$ high p_{T}^{Z}	$2\ell \geq 3j$ medium p_{T}^{Z}	$2\ell \geq 3j$ medium p_{T}^{Z}
$m_{b\bar{b}}$	175	200	180	190	225	280
$p_{\mathrm{T}}^{b_1}$	380	365	150	360	220	370
$p_{\mathrm{T}}^{b_2}$	190	175	85	170	115	175
$E_{\mathrm{T}}^{\mathrm{miss}}$	440	435	65	135	90	135
H_T	895	900	–	–	–	–
p_{T}^{Z}	$\equiv E_{\mathrm{T}}^{\mathrm{miss}}$	$\equiv E_{\mathrm{T}}^{\mathrm{miss}}$	medium p_{T}^{Z}	490	medium p_{T}^{Z}	500
$m_{b\bar{b}j}$	–	680	–	–	815	890
$p_{\mathrm{T}}^{j_3}$	–	140	–	–	260	370

Table B.2: Ranges of the MVA input variables (in GeV) for each analysis region. (−) indicates that this variable is not used in the corresponding BDTs. The range is chosen such that it contains 99% of all signal events. Events with variables with larger values are set to these maximum values. 0ℓ refers to the 0 lepton channel and 2ℓ to the 2 lepton channel. $2j$ refers to events with exactly 2 *signal* jets, $3j$ to events with exactly 3 *signal* jets and $\geq 3j$ to events with 3 or more *signal* jets. In the 2 lepton medium p_{T}^{Z} category the range for p_{T}^{Z} is given by the category definition.



(a) 0 lepton, 2 jet

(b) 0 lepton, 3 jet

(c) 2 lepton, 2 jet, medium p_{T}^{Z}

(d) 2 lepton, 2 jet, high p_{T}^{Z}

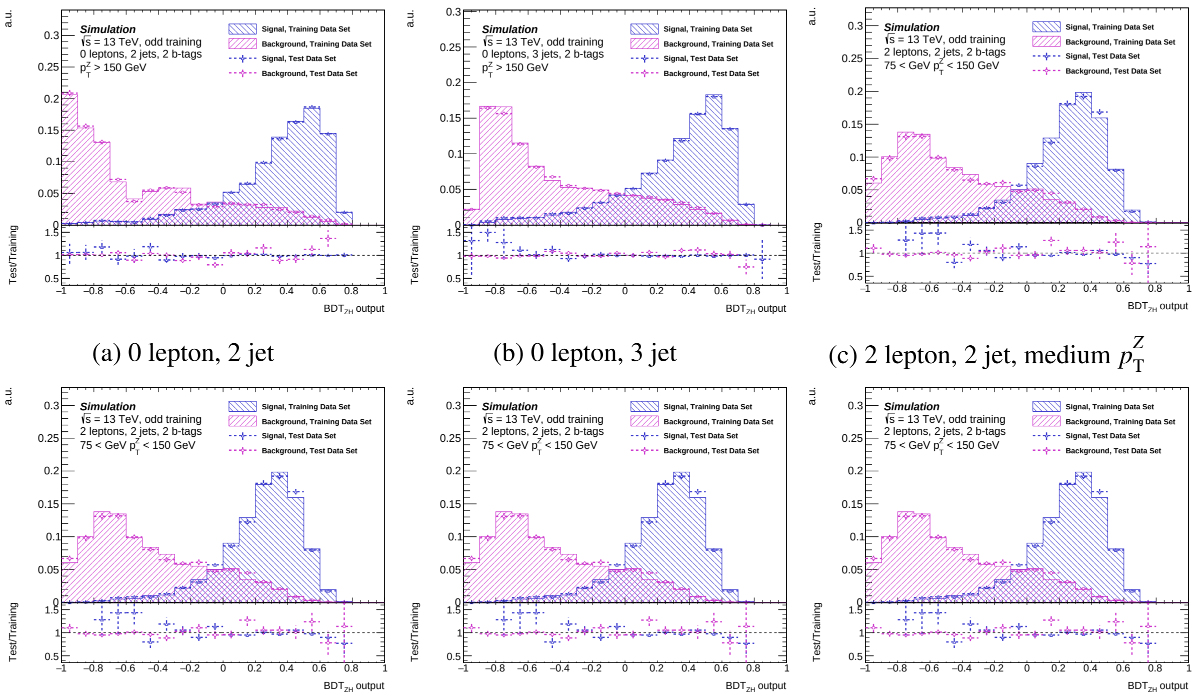
(e) 2 lepton, 3+ jet, medium p_{T}^{Z}

(f) 2 lepton, 3+ jet, high p_{T}^{Z}

Figure B.5: BDT output distributions for the $Z(H \to b\bar{b})$ MVA categories for the even training. The ratio plot shows the agreement between the training data set and a statistically independent test data set.

B.1.2 Comparison of MVA for different jet corrections

Comparison of the BDT output distribution for trainings using different b-jet correction methods — no correction, *default* ATLAS $V(H \to b\bar{b})$ corrections, jet energy regression — for the 2 jet categories B.7 and 3(3+) jet categories B.8. The distributions for the simulated signal events and background events are shown separately.



(a) 0 lepton, 2 jet

(b) 0 lepton, 3 jet

(c) 2 lepton, 2 jet, medium p_{T}^{Z}

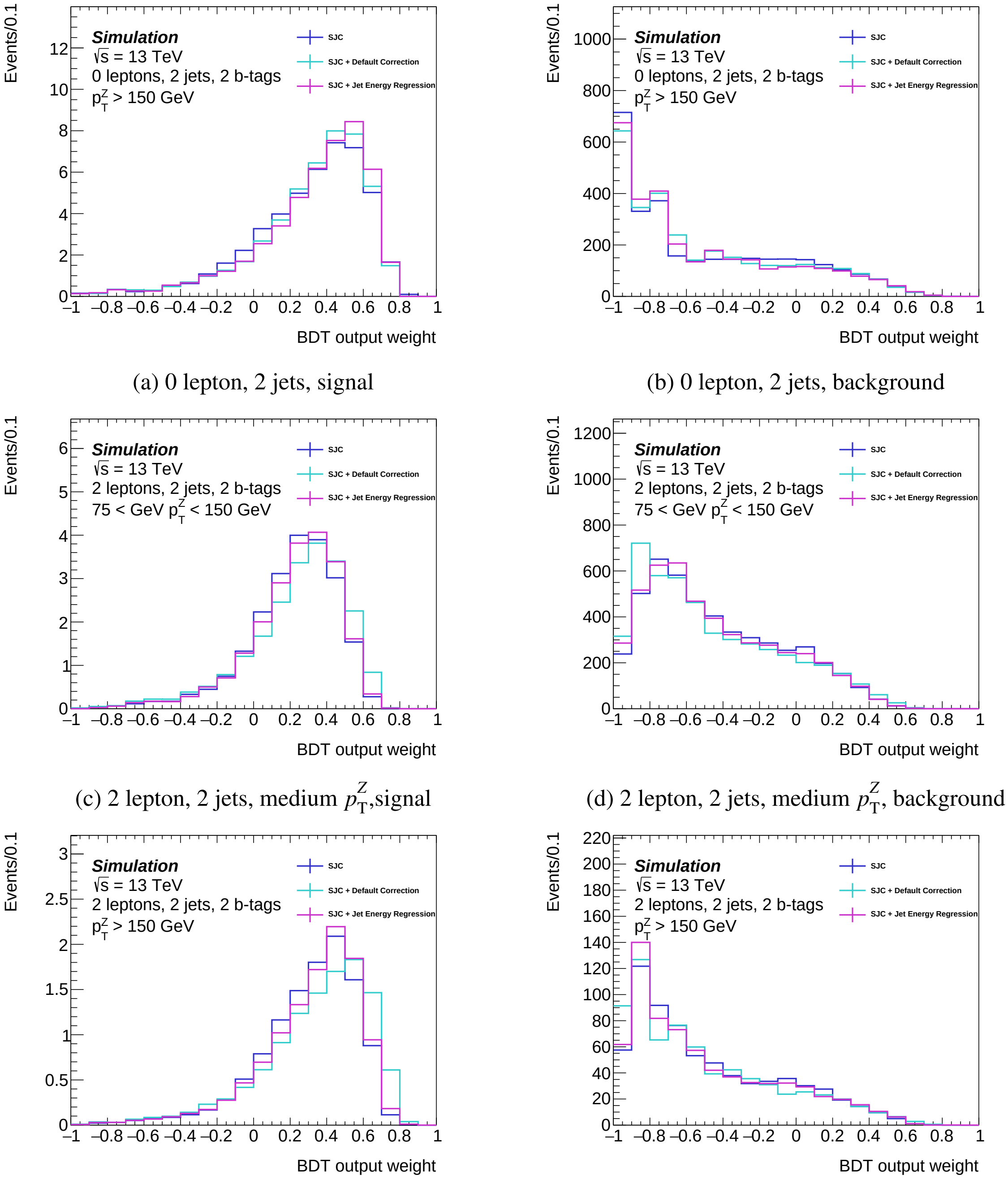
(d) 2 lepton, 2 jet, high p_{T}^{Z}

(e) 2 lepton, 3+ jet, medium p_{T}^{Z}

(f) 2 lepton, 3+ jet, high p_{T}^{Z}

Figure B.6: BDT output distributions for the $Z(H \to b\bar{b})$ MVA categories for the odd training. The ratio plot shows the agreement between the training data set and a statistically independent test data set.

Comparison of the MVA input variables for trainings using different b-jet correction methods: no correction, *default* ATLAS $V(H \to b\bar{b})$ corrections, jet energy regression. Only those input variables that are affected by the corrections are shown separately for simulated signal and background events:
0 lepton 2 jets category in figure B.9,
0 lepton 3 jets category in figure B.10,
2 lepton 2 jets medium p_{T}^{Z} category in figure B.11,
2 lepton 3+ jets medium p_{T}^{Z} category in figure B.13,
2 lepton 2 jets high p_{T}^{Z} category in figure B.12 and
2 lepton 3+ jets high p_{T}^{Z} category in figure B.14.



(a) 0 lepton, 2 jets, signal

(b) 0 lepton, 2 jets, background

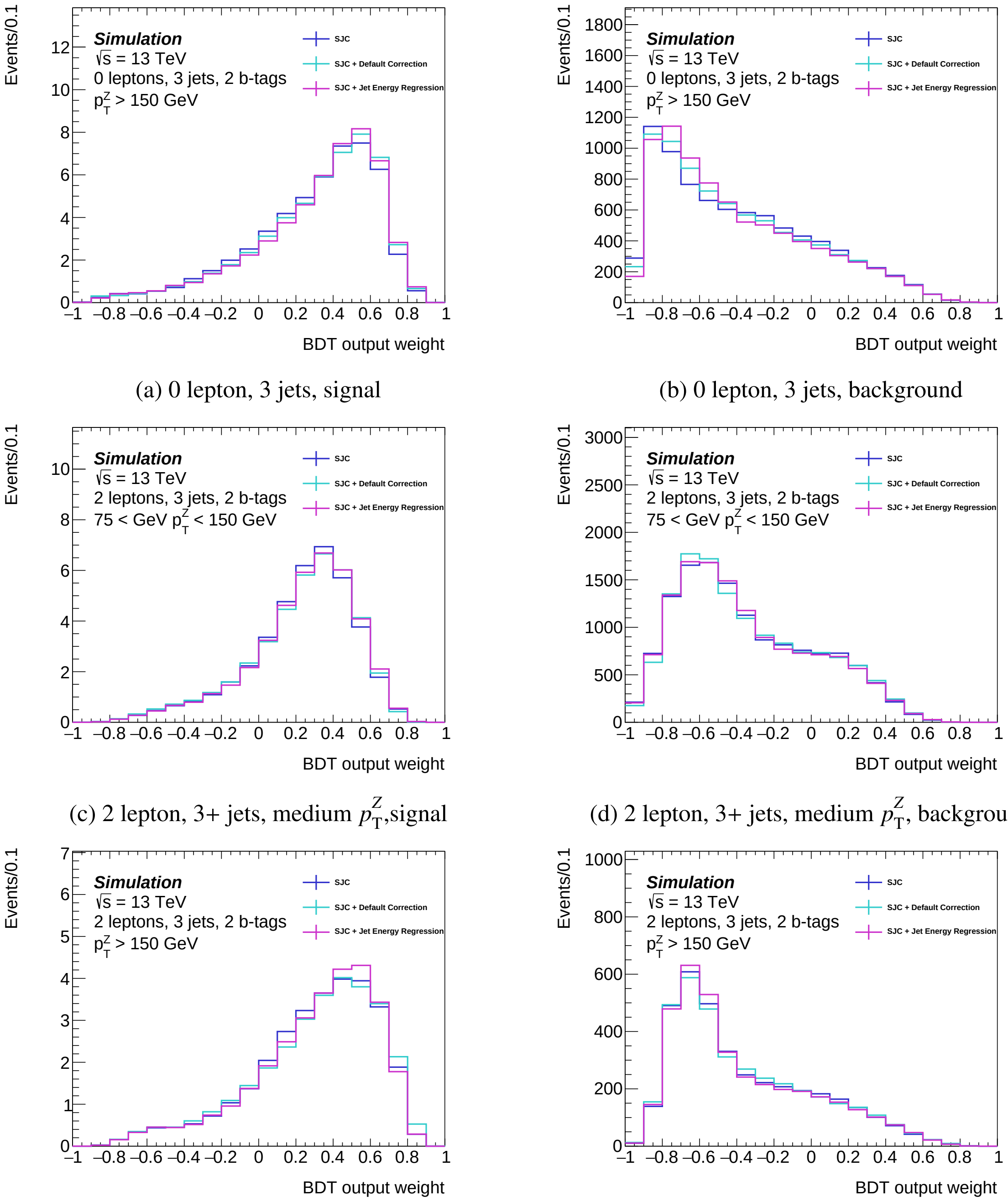
(c) 2 lepton, 2 jets, medium p_T^Z,signal

(d) 2 lepton, 2 jets, medium p_T^Z, background

(e) 2 lepton, 2 jets, high p_T^Z, signal

(f) 2 lepton, 2 jets, high p_T^Z, background

Figure B.7: BDT Output distributions for different jet energy correction methods separate for signal and background events



(a) 0 lepton, 3 jets, signal

(b) 0 lepton, 3 jets, background

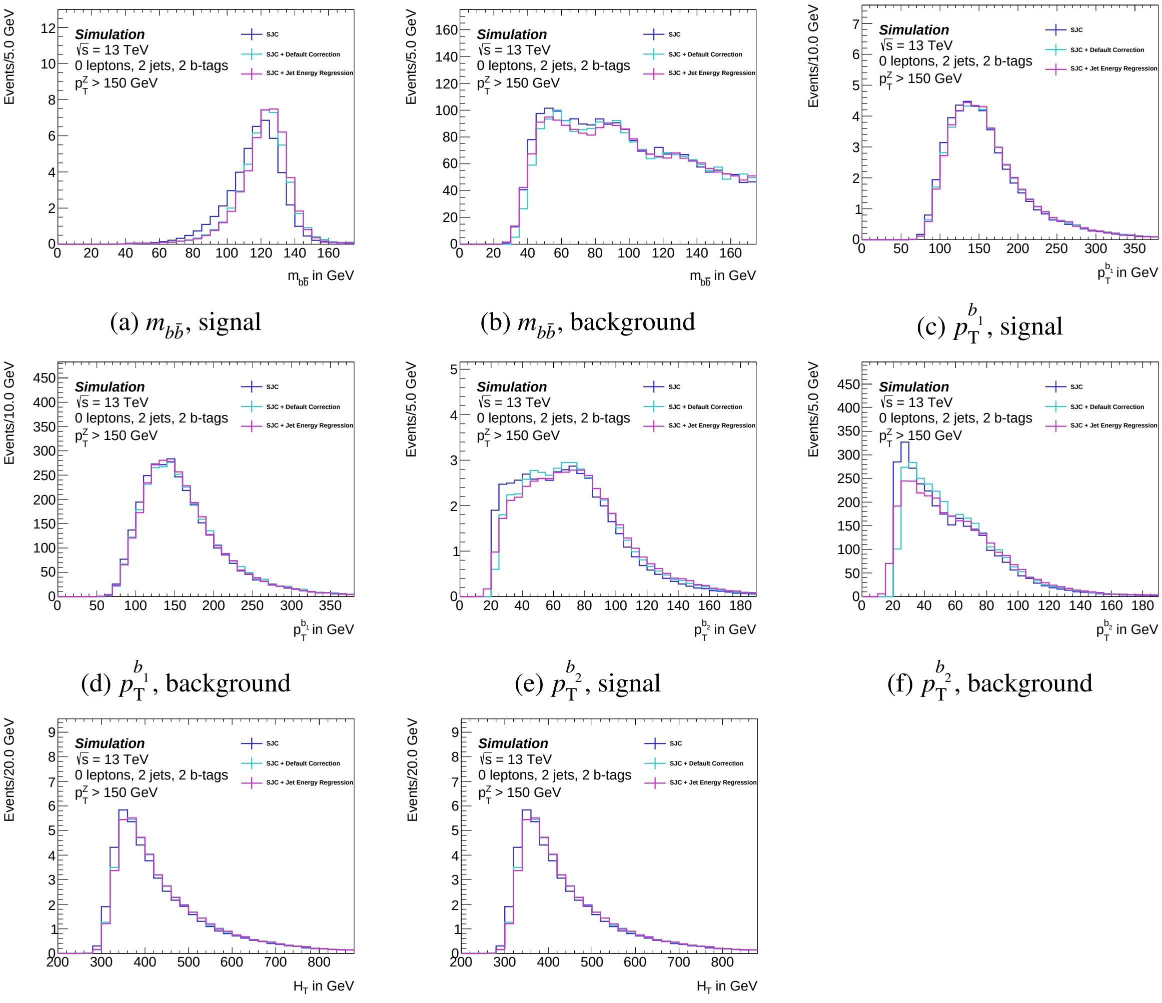
(c) 2 lepton, 3+ jets, medium p_T^Z,signal

(d) 2 lepton, 3+ jets, medium p_T^Z, background

(e) 2 lepton, 3+ jets, high p_T^Z, signal

(f) 2 lepton, 3+ jets, high p_T^Z, background

Figure B.8: BDT Output distributions for different jet energy correction methods separate for signal and background events



(a) $m_{b\bar{b}}$, signal

(b) $m_{b\bar{b}}$, background

(c) $p_T^{b_1}$, signal

(d) $p_T^{b_1}$, background

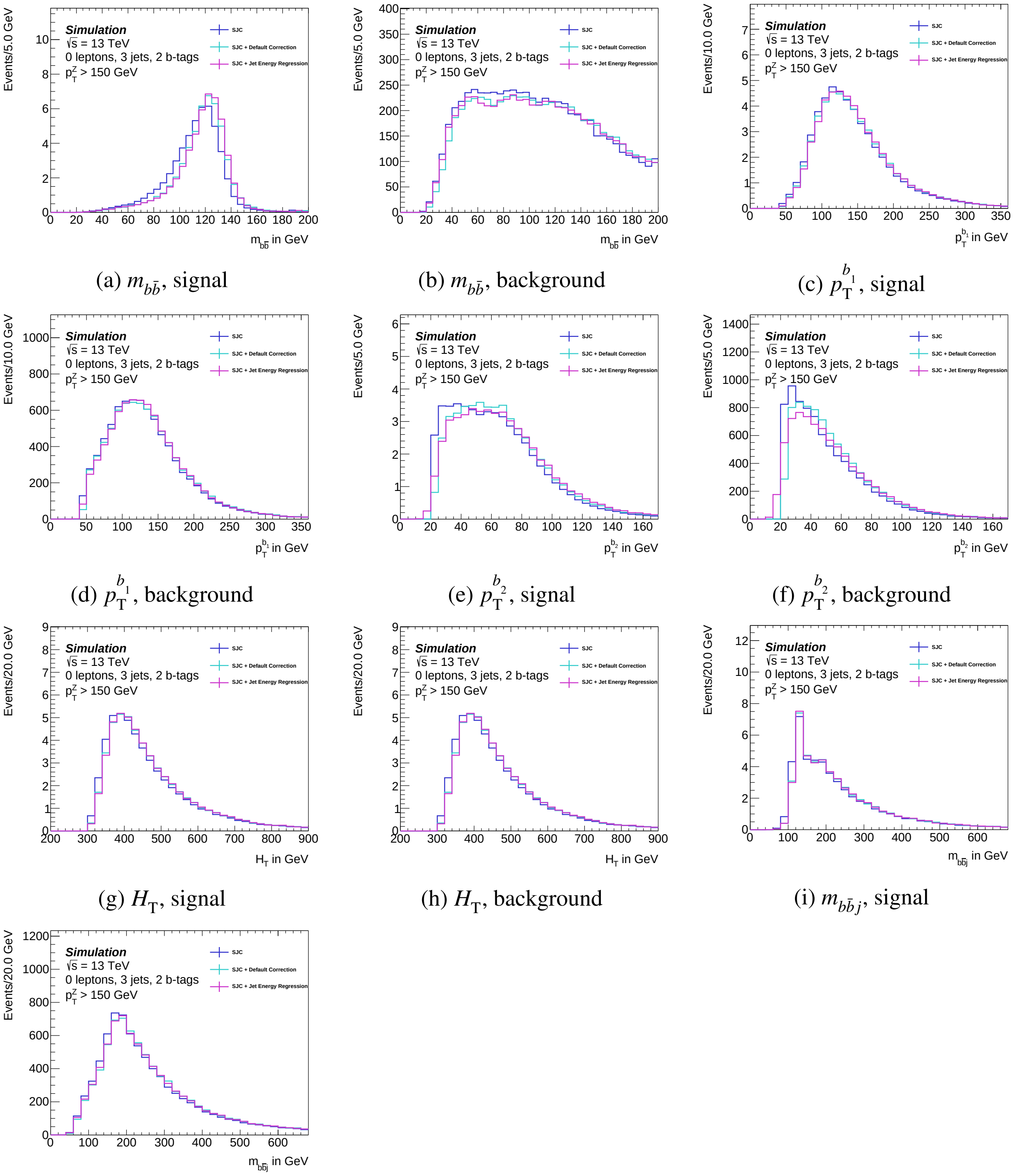
(e) $p_T^{b_2}$, signal

(f) $p_T^{b_2}$, background

(g) H_T, signal

(h) H_T, background

Figure B.9: Input variables distributions for different jet correction methods, 0 lepton 2 jets



(a) $m_{b\bar{b}}$, signal

(b) $m_{b\bar{b}}$, background

(c) $p_T^{b_1}$, signal

(d) $p_T^{b_1}$, background

(e) $p_T^{b_2}$, signal

(f) $p_T^{b_2}$, background

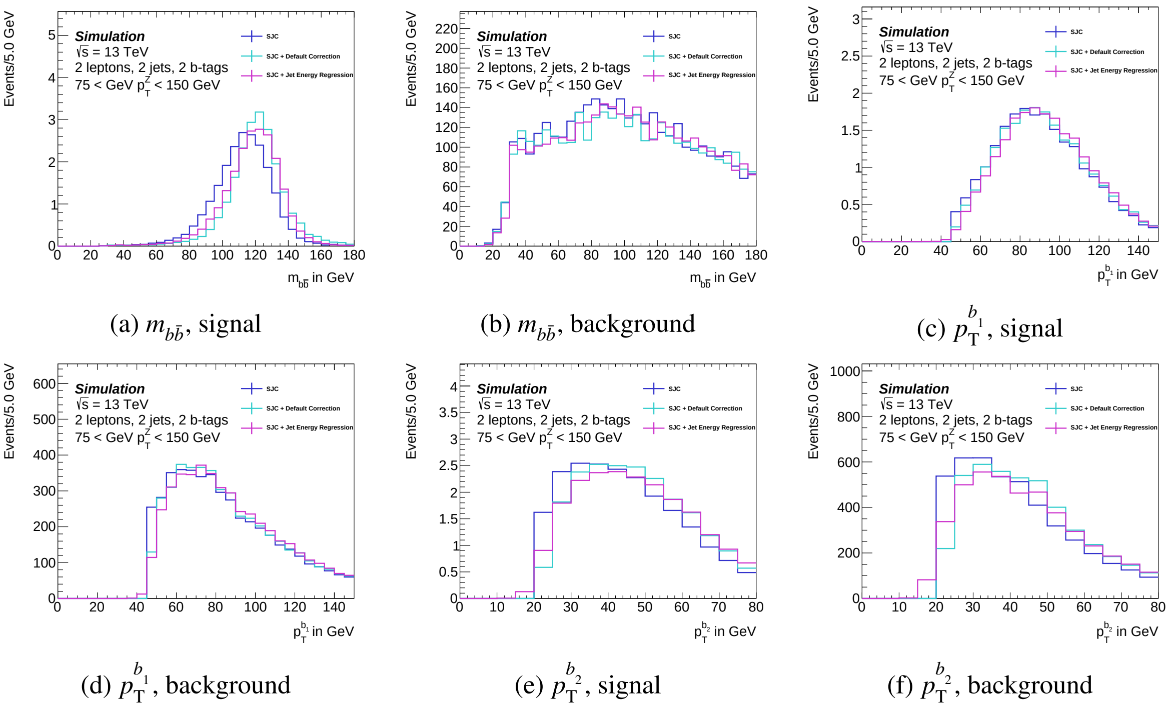
(g) H_T, signal

(h) H_T, background

(i) $m_{b\bar{b}j}$, signal

(j) $m_{b\bar{b}j}$, background

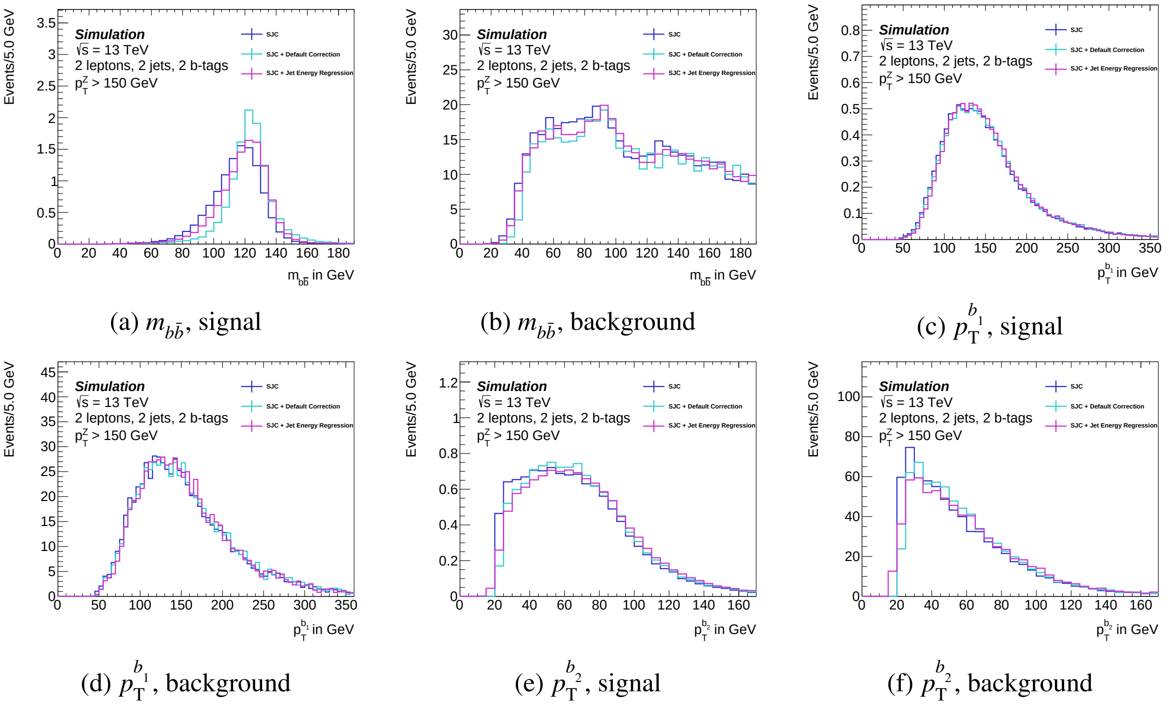
Figure B.10: Input variables distributions for different jet correction methods, 0 lepton 3 jets



(a) $m_{b\bar{b}}$, signal (b) $m_{b\bar{b}}$, background (c) $p_T^{b_1}$, signal

(d) $p_T^{b_1}$, background (e) $p_T^{b_2}$, signal (f) $p_T^{b_2}$, background

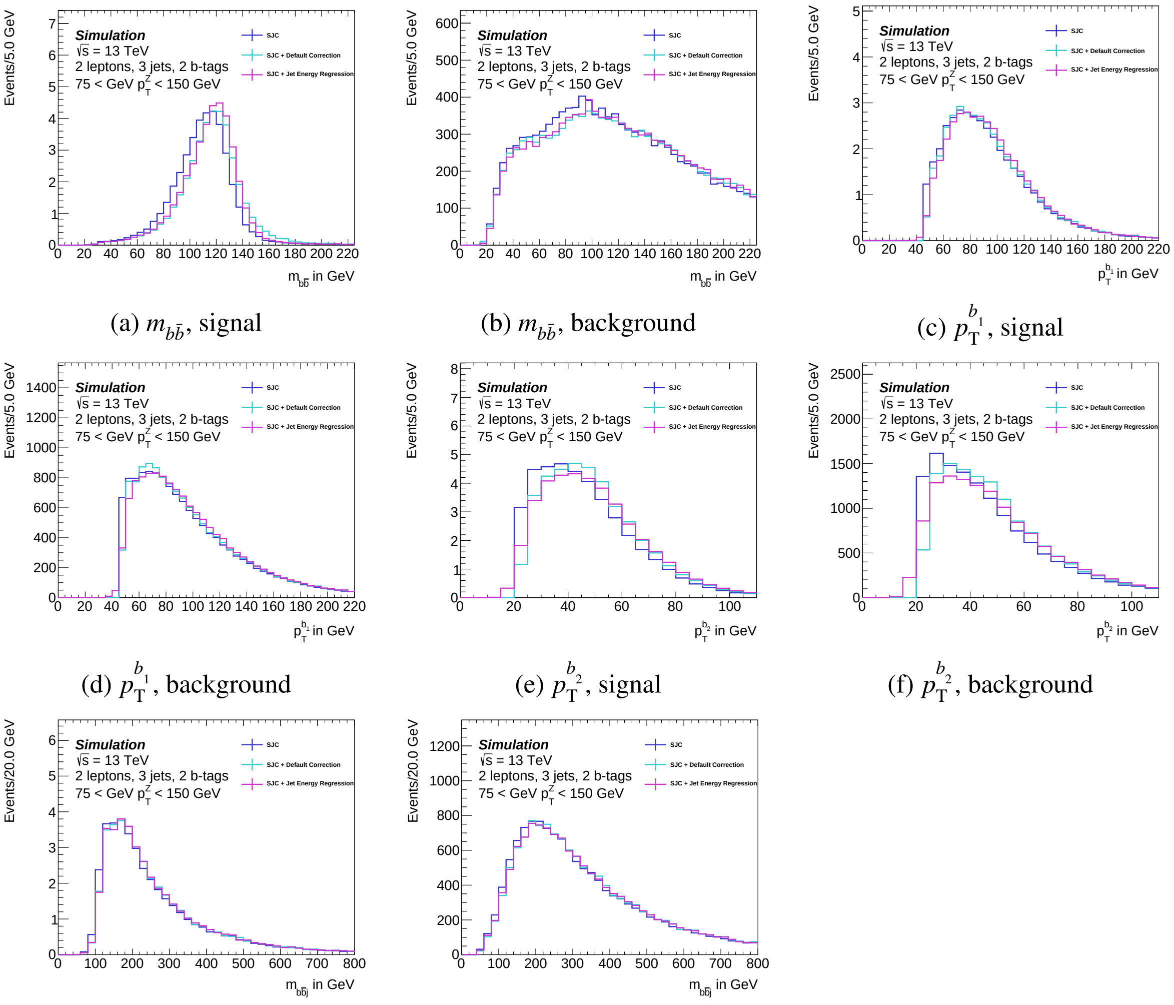
Figure B.11: Input variables distributions for different jet correction methods, 2 lepton 2 jets medium p_T^Z



(a) $m_{b\bar{b}}$, signal (b) $m_{b\bar{b}}$, background (c) $p_T^{b_1}$, signal

(d) $p_T^{b_1}$, background (e) $p_T^{b_2}$, signal (f) $p_T^{b_2}$, background

Figure B.12: Input variables distributions for different jet correction methods, 2 lepton 2 jets high p_T^Z



(a) $m_{b\bar{b}}$, signal

(b) $m_{b\bar{b}}$, background

(c) $p_{\mathrm{T}}^{b_1}$, signal

(d) $p_{\mathrm{T}}^{b_1}$, background

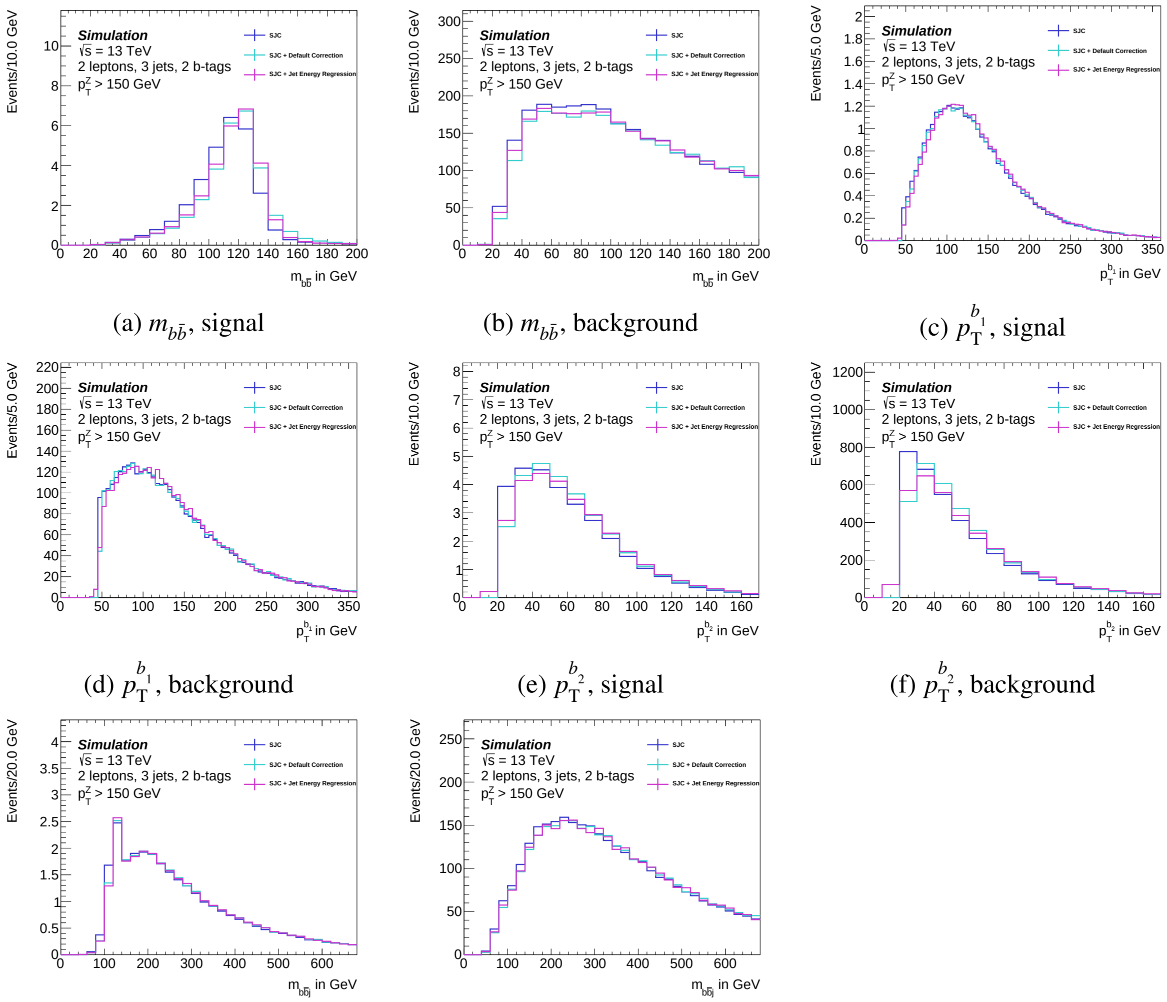
(e) $p_{\mathrm{T}}^{b_2}$, signal

(f) $p_{\mathrm{T}}^{b_2}$, background

(g) $m_{b\bar{b}j}$, signal

(h) $m_{b\bar{b}j}$, background

Figure B.13: Input variables distributions for different jet correction methods, 2 lepton 3+ jets medium p_{T}^{Z}



(a) $m_{b\bar{b}}$, signal

(b) $m_{b\bar{b}}$, background

(c) $p_{\mathrm{T}}^{b_1}$, signal

(d) $p_{\mathrm{T}}^{b_1}$, background

(e) $p_{\mathrm{T}}^{b_2}$, signal

(f) $p_{\mathrm{T}}^{b_2}$, background

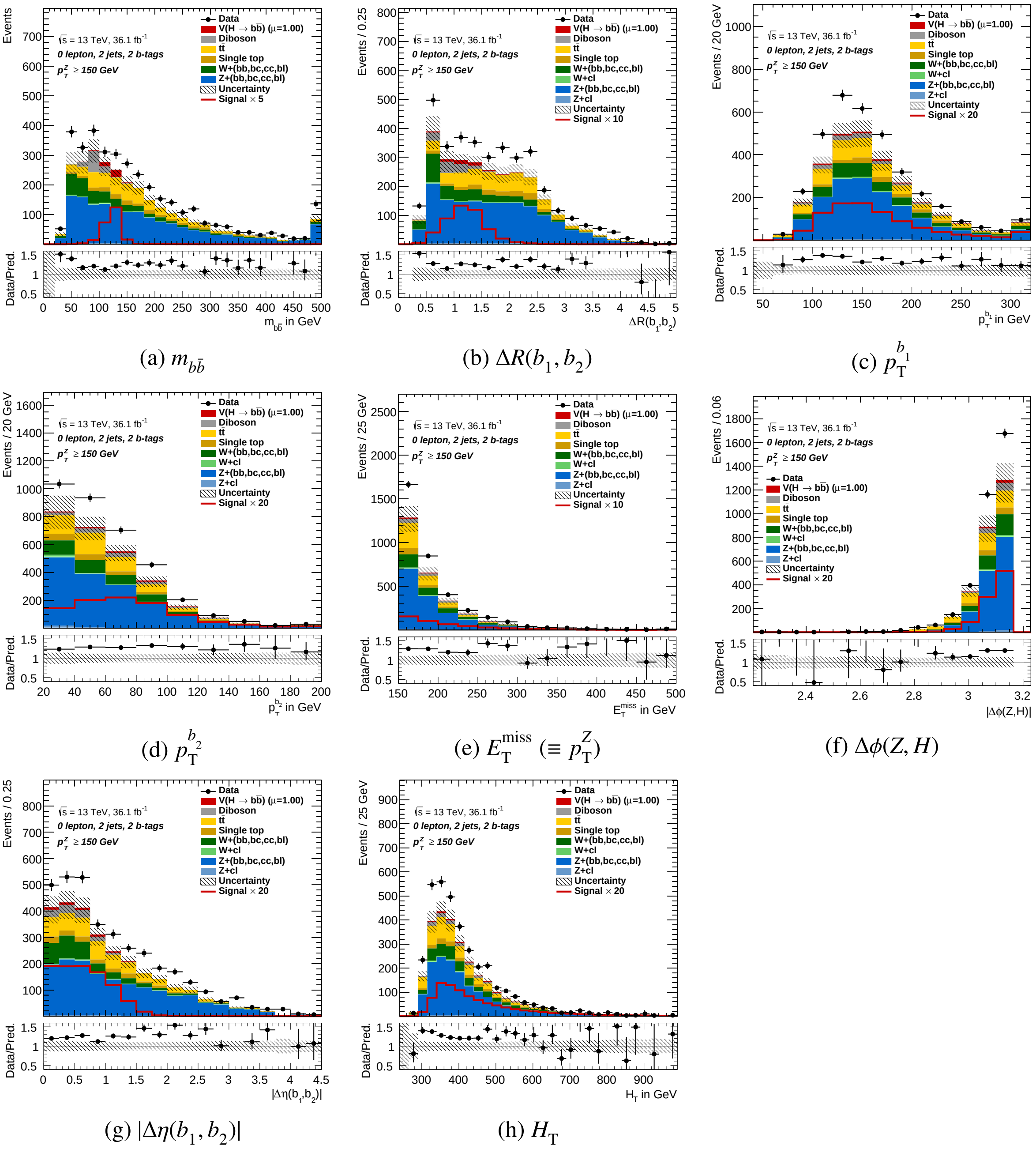
(g) $m_{b\bar{b}j}$, signal

(h) $m_{b\bar{b}j}$, background

Figure B.14: Input variables distributions for different jet correction methods, 2 lepton 3+ jets high p_{T}^{Z}

B.1.3 Data and MC comparisons for input variables

Comparison of the "prefit" MVA input variables for in data and simulated events:
0 lepton 2 jets category in figure B.15,
0 lepton 3 jets category in figure B.16,
2 lepton 2 jets medium p_{T}^{Z} category in figure B.17,
2 lepton 3+ jets medium p_{T}^{Z} category in figure B.18,
2 lepton 2 jets high p_{T}^{Z} category in figure B.19 and
2 lepton 3+ jets high p_{T}^{Z} category in figure B.20.



(a) $m_{b\bar{b}}$

(b) $\Delta R(b_1, b_2)$

(c) $p_{\mathrm{T}}^{b_1}$

(d) $p_{\mathrm{T}}^{b_2}$

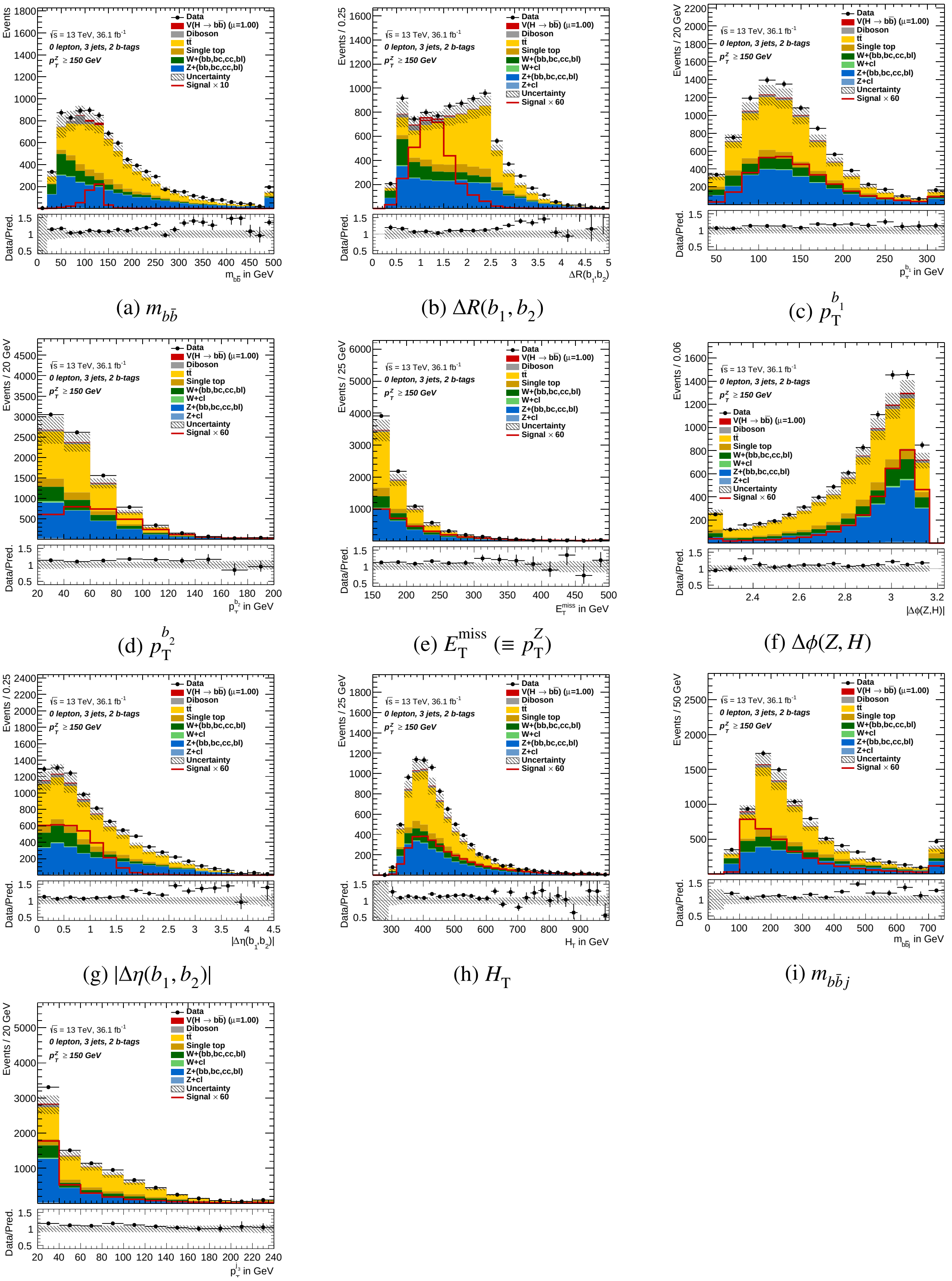
(e) $E_{\mathrm{T}}^{\mathrm{miss}}$ ($\equiv p_{\mathrm{T}}^{Z}$)

(f) $\Delta\phi(Z, H)$

(g) $|\Delta\eta(b_1, b_2)|$

(h) H_{T}

Figure B.15: Comparison of data and simulated events of input variables distributions, 0 lepton 2 jets. The distributions are "prefit", i.e. not normalised according to the parameters determined by the fit. The distributions in simulated events corresponds to the ones used in the training.



(a) $m_{b\bar{b}}$

(b) $\Delta R(b_1, b_2)$

(c) $p_{\mathrm{T}}^{b_1}$

(d) $p_{\mathrm{T}}^{b_2}$

(e) $E_{\mathrm{T}}^{\mathrm{miss}}$ ($\equiv p_{\mathrm{T}}^{Z}$)

(f) $\Delta\phi(Z, H)$

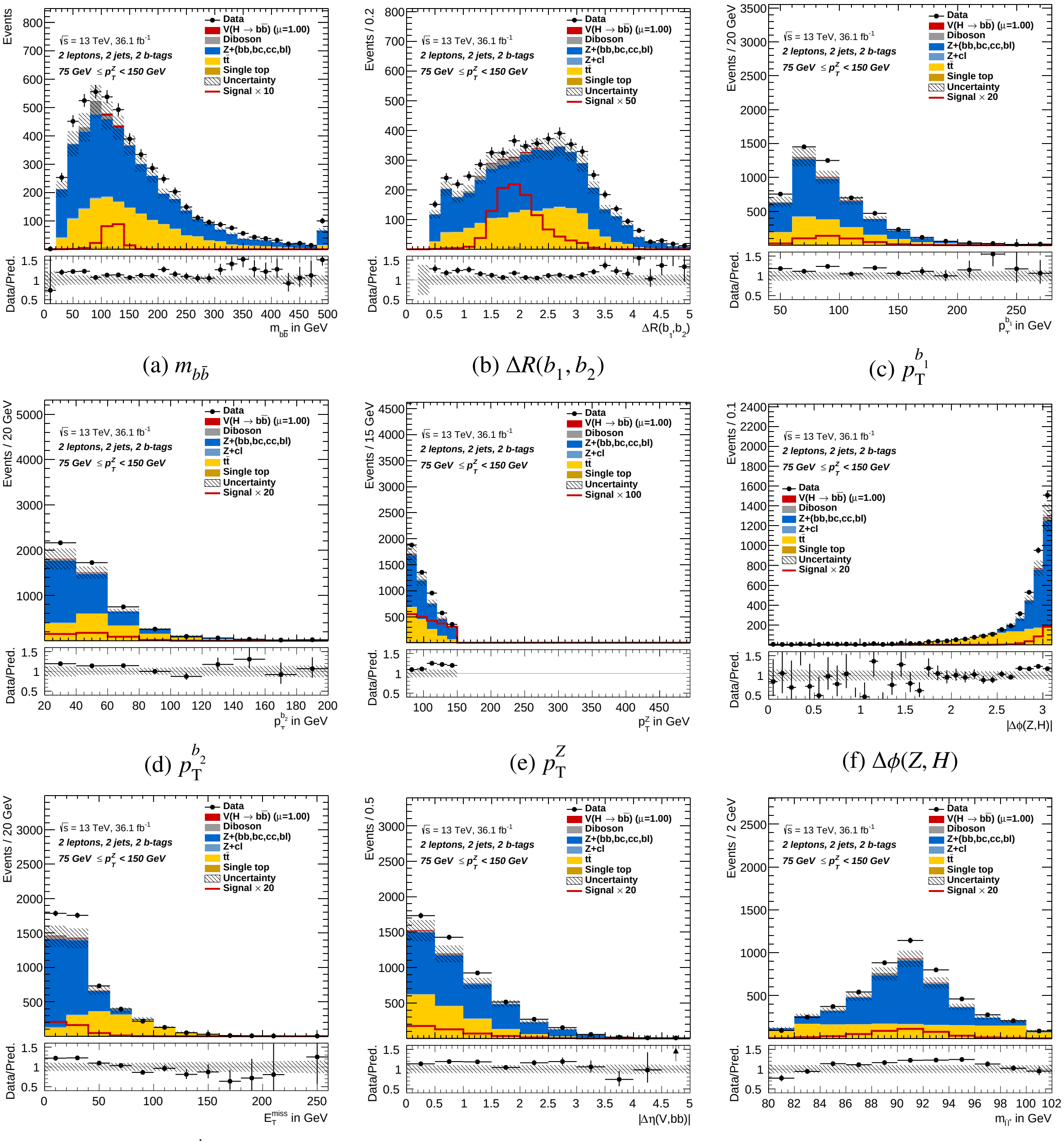
(g) $|\Delta\eta(b_1, b_2)|$

(h) H_{T}

(i) $m_{b\bar{b}j}$

(j) $p_{\mathrm{T}}^{j_3}$

Figure B.16: Comparison of data and simulated events of input variables distributions, 0 lepton 3 jets. The distributions are "prefit", i.e. not normalised according to the parameters determined by the fit. The distributions in simulated events corresponds to the ones used in the training.



(a) $m_{b\bar{b}}$

(b) $\Delta R(b_1, b_2)$

(c) $p_{\mathrm{T}}^{b_1}$

(d) $p_{\mathrm{T}}^{b_2}$

(e) p_{T}^{Z}

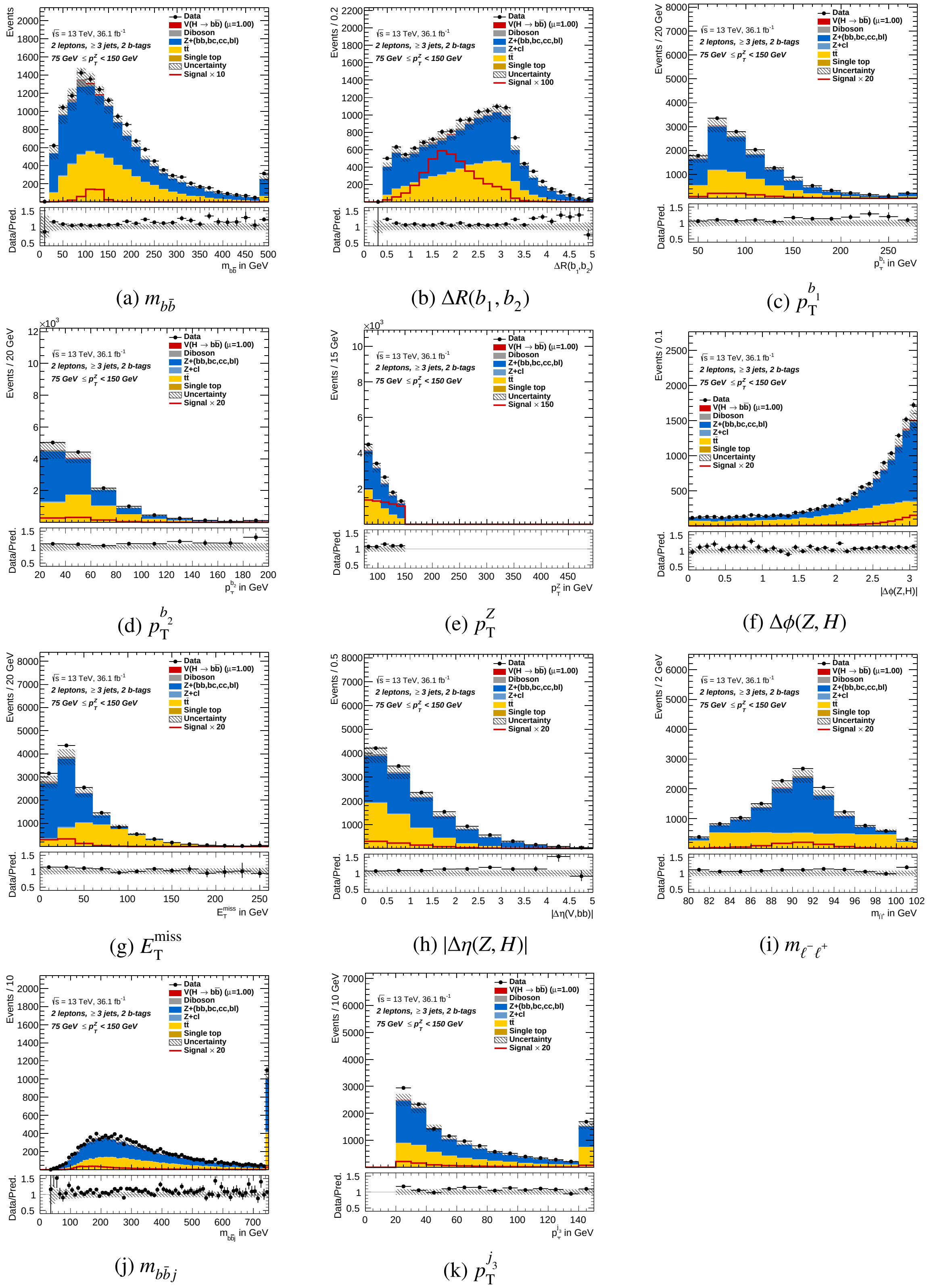
(f) $\Delta\phi(Z, H)$

(g) $E_{\mathrm{T}}^{\mathrm{miss}}$

(h) $|\Delta\eta(Z, H)|$

(i) $m_{\ell^- \ell^+}$

Figure B.17: Comparison of data and simulated events of input variables distributions, 2 lepton 2 jets medium p_{T}^{Z}. The distributions are "prefit", i.e. not normalised according to the parameters determined by the fit. The distributions in simulated events corresponds to the ones used in the training.



(a) $m_{b\bar{b}}$

(b) $\Delta R(b_1, b_2)$

(c) $p_{\mathrm{T}}^{b_1}$

(d) $p_{\mathrm{T}}^{b_2}$

(e) p_{T}^{Z}

(f) $\Delta\phi(Z, H)$

(g) $E_{\mathrm{T}}^{\mathrm{miss}}$

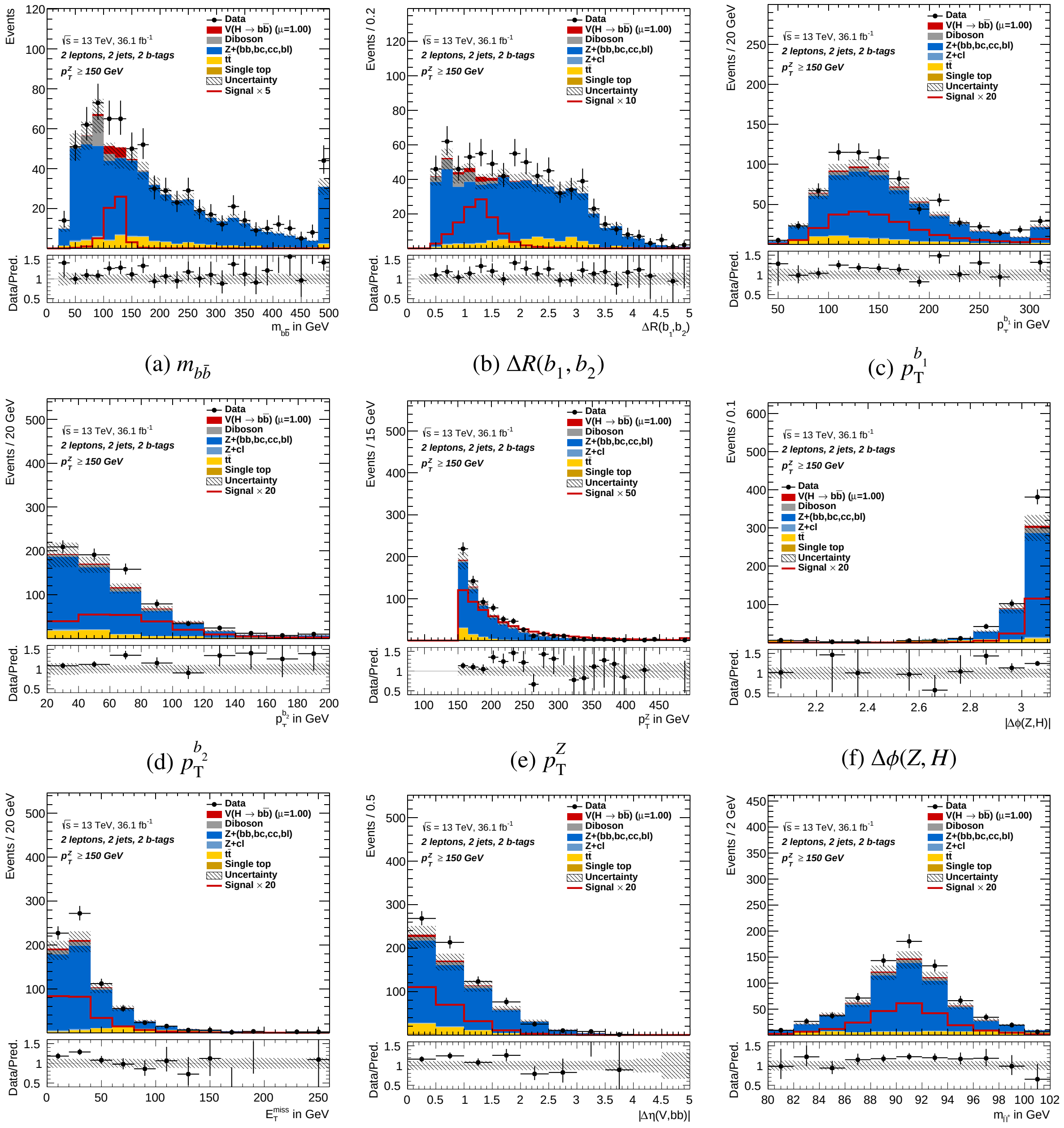
(h) $|\Delta\eta(Z, H)|$

(i) $m_{\ell^-\ell^+}$

(j) $m_{b\bar{b}j}$

(k) $p_{\mathrm{T}}^{j_3}$

Figure B.18: Comparison of data and simulated events of input variables distributions, 2 lepton 3+ jets medium p_{T}^{Z}. The distributions are "prefit", i.e. not normalised according to the parameters determined by the fit. The distributions in simulated events corresponds to the ones used in the training.



(a) $m_{b\bar{b}}$

(b) $\Delta R(b_1, b_2)$

(c) $p_T^{b_1}$

(d) $p_T^{b_2}$

(e) p_T^Z

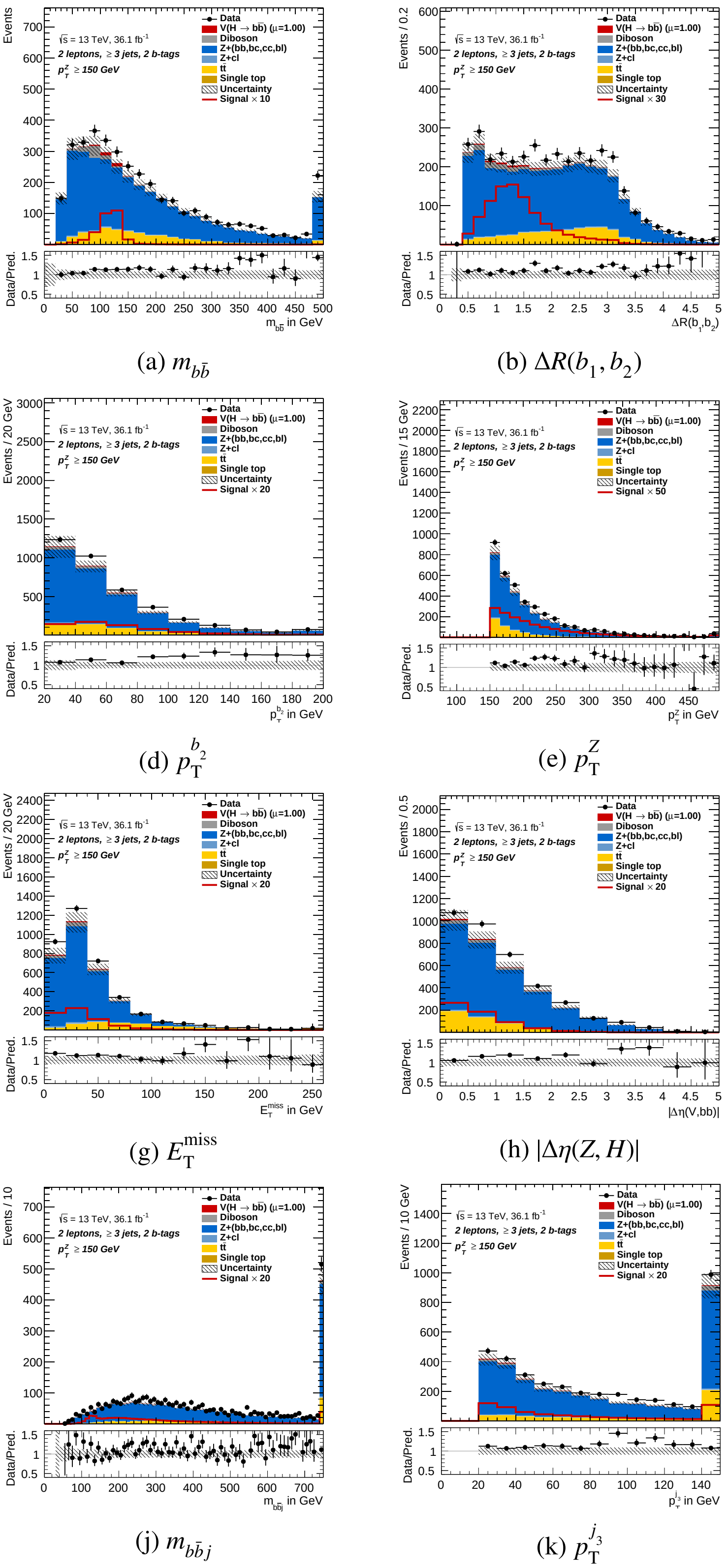
(f) $\Delta\phi(Z,H)$

(g) E_T^{miss}

(h) $|\Delta\eta(Z,H)|$

(i) $m_{\ell^-\ell^+}$

Figure B.19: Comparison of data and simulated events of input variables distributions, 2 lepton 2 jets high p_T^Z. The distributions are "prefit", i.e. not normalised according to the parameters determined by the fit. The distributions in simulated events corresponds to the ones used in the training.



(a) $m_{b\bar{b}}$

(b) $\Delta R(b_1, b_2)$

(c) $p_T^{b_1}$

(d) $p_T^{b_2}$

(e) p_T^Z

(f) $\Delta\phi(Z, H)$

(g) E_T^{miss}

(h) $|\Delta\eta(Z, H)|$

(i) $m_{\ell^-\ell^+}$

(j) $m_{b\bar{b}j}$

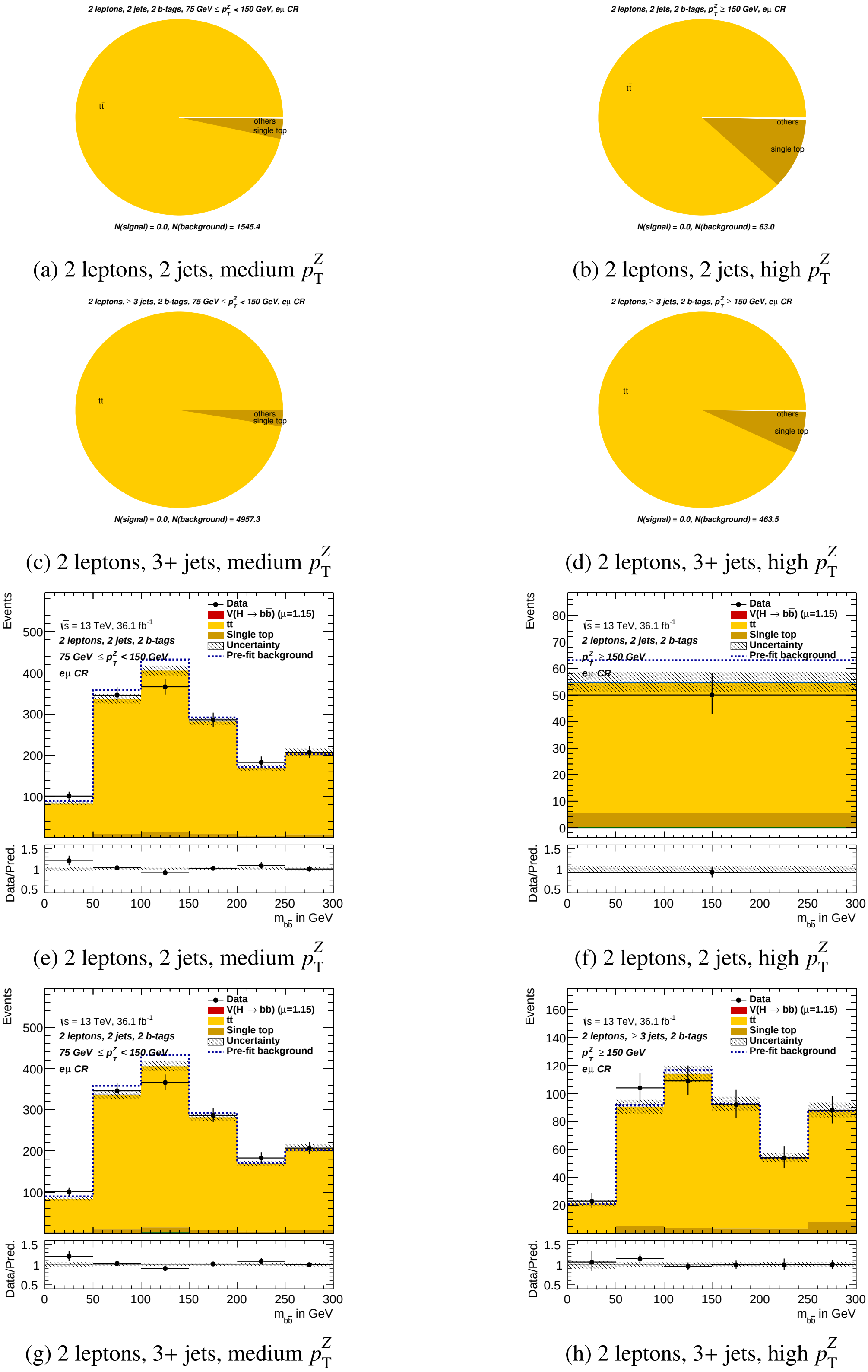
(k) $p_T^{j_3}$

Figure B.20: Comparison of data and simulated events of input variables distributions, 2 lepton 3+ jets high p_T^Z. The distributions are "prefit", i.e. not normalised according to the parameters determined by the fit. The distributions in simulated events corresponds to the ones used in the training.

B.2 $e\mu$ CRs

The $e\mu$ CRs are introduced in the 2 lepton channel to provide additional information about the $t\bar{t}$ but also single top backgrounds in that channel. The only change in the event selection with respect to the default 2 lepton channel selection is the requirement of exactly one muon and one electron instead of two muons or two electrons. The composition of the $e\mu$ CRs is shown in figure B.21. The $t\bar{t}$ events are dominating these CRs in all categories with a significance contribution of single top events. Other processes contribute less than 0.5% of the total amount of events.

Figure B.21 shows the distributions in the categories of the $e\mu$ CRs that are used in the fit. In those distributions the simulated events of each signal and background component are scaled to the normalisations obtained in the fit. Good agreement between data and simulated events is observed in these CRs.



(a) 2 leptons, 2 jets, medium p_T^Z

(b) 2 leptons, 2 jets, high p_T^Z

(c) 2 leptons, 3+ jets, medium p_T^Z

(d) 2 leptons, 3+ jets, high p_T^Z

(e) 2 leptons, 2 jets, medium p_T^Z

(f) 2 leptons, 2 jets, high p_T^Z

(g) 2 leptons, 3+ jets, medium p_T^Z

(h) 2 leptons, 3+ jets, high p_T^Z

Figure B.21: The composition (a to d) of the 2 lepton $e\mu$ CRs. Contributions that are smaller than 1% are grouped together in "others". Comparisons of the data and simulated events (e to h) in the $e\mu$ CRs. All components are scaled to their normalisations as determined in the fit. The normalisation of the sum of all background components as predicted by simulated events ("prefit") is given by the dashed blue line. The shaded bands represent the total uncertainty.

B.3 Modelling Uncertainties

A detailed description of the source of each modelling uncertainty assigned in the $Z(H \to b\bar{b})$ analysis and how they were derived.

B.3.1 Signal: $V(H \to b\bar{b})$

The first set of systematic uncertainties arise from the precision of the theoretical prediction of the VH production cross section and the BR of the $H \to b\bar{b}$ decay. These uncertainties are not analysis specific and affect all signal events alike. The following normalisation uncertainties are assigned based on the official recommendations of the LHC Higgs cross section working group:

- VH cross section, missing higher orders in QCD: Separate uncertainties are assigned for quark and gluon induced ZH production. Since the recommendations do not distinguish between them it is assumed that the uncertainty for $q\bar{q} \to ZH$ is as large as for WH production which is exclusively produced via $q\bar{q} \to WH$. The uncertainty for $gg \to ZH$ is then derived such that the cross section weighted sum in quadrature of the $q\bar{q} \to ZH$ and $gg \to ZH$ uncertainties yields the total ZH uncertainty as given in [4].

- VH cross section, variations in the PDF set and α_{S}: Separate uncertainties are assigned for the $ZH \to \nu\bar{\nu}b\bar{b}$, $W^{\pm}H \to \ell^{\pm}\overset{(-)}{\nu}b\bar{b}$ and $ZH \to \ell^{-}\ell^{+}b\bar{b}$ decay. The uncertainty for the $q\bar{q} \to ZH$ cross section is taken from the latest recommendation [4] whereas it is taken from an earlier recommendation for $gg \to ZH$ production [86] since it is not given separately in the latest one. The aforementioned assumption that the uncertainty for WH and $q\bar{q} \to ZH$ are the same cannot be used in this case since WH and $q\bar{q} \to ZH$ production have different PDF contributions.

- p_{T}^{Z} shape uncertainty, missing higher orders in EW: This uncertainty as parametrised as a function of p_{T}^{Z}. It is determined as:

  $$\Delta_{\mathrm{shape}} = \max(1\%, \delta_{\mathrm{EW}}(p_{\mathrm{T}}^{Z}), \Delta_{\gamma})$$

  with the maximum expected impact of missing higher order EW contributions (1%), the size of the NLO EW correction for a given p_{T}^{Z} ($\delta_{\mathrm{EW}}(p_{\mathrm{T}}^{Z})$) and the uncertainty of the photon induced cross section relative to the total cross section (Δ_{γ}) [4]. The latter only becomes sizeable for WH production.

- $H \to b\bar{b}$ branching ratio: This uncertainty is assigned to all signal processes. It takes into account uncertainties on the b-quark mass, α_{S} as well as effects from missing higher order contributions in QCD and EW theory [87].

Another large set of analysis specific uncertainties is determined based on comparisons of the simulation models listed in table 8.2. Uncertainties for the $ZH \to \nu\bar{\nu}b\bar{b}$ decay are determined in the 0 lepton channel and in the 2 lepton channel for the $ZH \to \ell^{-}\ell^{+}b\bar{b}$ decay. In all cases the same uncertainties are assigned to the quark and gluon induced ZH production. Since a non-negligible fraction of $W(H \to b\bar{b})$ events is selected in the 0 lepton channel systematic uncertainties for WH production are taken from the 1 lepton channel in the standard ATLAS $V(H \to b\bar{b})$ analysis. This channel implements an event selection that targets the $W^{\pm}H \to \ell^{\pm}\overset{(-)}{\nu}b\bar{b}$ decay. The following analysis specific uncertainties are derived:

- Normalisation uncertainty, variations in the PS and UE event model: Normalisation uncertainties to account for PS and UE modelling variation are assigned for each signal process. They are derived from a comparison of the nominal simulation that uses the PYTHIA PS to the alternative simulation that uses HERWIG instead. In addition, the nominal simulation is varied by different variations of the A14 PS and UE tune but the effect from the HERWIG comparison is found to be much larger. Therefore the difference between the PYTHIA and HERWIG model is assigned as a systematic uncertainty.

- Acceptance uncertainty, variations in the PS and UE event model: Additional PS and UE acceptance uncertainties are assigned that affect the normalisation of the 3 and 3+ jets category with respect to the 2 jet categories. They are derived in the same way as the normalisation uncertainty.
- $m_{b\bar{b}}$ and p_{T}^{Z} shape uncertainties, variations in the PS and UE event model: In contrast to the normalisation and acceptance uncertainties the shape variations due to variations in the PS and UE model are governed by the A14 variations. Therefore the shape uncertainties are derived from the envelope of the A14 variations observed in the $m_{b\bar{b}}$ and p_{T}^{Z} distributions. The observed shape variations are similar for all signal processes and regions. Hence the same uncertainty is applied to all of them.
- Normalisation uncertainty, variations in the factorisation μ_{F} and renormalisation scale μ_{R}: The jet multiplicity of the signal processes is very sensitive to changes in μ_{F} and μ_{R}. The recommended procedure to calculate acceptance uncertainties for separate jet multiplicities is to use the Stewart-Tackmann method [88]. It allows to calculate normalisation uncertainties for each jet category taking into account correlations between the categories. They were derived from the envelope of the μ_{F} and μ_{R} variations that are available for the nominal simulation. In the 0 lepton 3 jets category additional uncertainties are assigned that originate from the exclusion of events with 4 or more jets.
- $m_{b\bar{b}}$ and p_{T}^{Z} shape uncertainties, variations in μ_{F} and μ_{R}: The shape uncertainties are determined from the maximum difference between the nominal model and the μ_{F} and μ_{R} variations. Separate uncertainties are assigned to each signal process and the 2 jets and 3 (3+) jets categories.
- Normalisation uncertainty, variations in the PDF set and α_{S}: These uncertainties are derived using the available variations of the PDF set and α_{S} in the nominal simulation. The final uncertainty is derived by the sum in quadrature of these variations. Since the number of expected events varied consistently in the 2 jets and 3 (3+) jets categories the same uncertainties are used for both.
- $m_{b\bar{b}}$ and p_{T}^{Z} shape uncertainties, variations in PDF set and α_{S}: No variations are observed in the shape of the $m_{b\bar{b}}$ distribution. Therefore a shape uncertainty is only assigned as a function of p_{T}^{Z} which is determined from the envelope of the PDF and α_{S} variations. The size of the p_{T}^{Z} variation is similar in all regions.

B.3.2 V+jets

The first set of V+jets modelling uncertainties are analysis specific normalisation and acceptance uncertainties. The Z+jets and W+jets processes are treated separately. Modelling uncertainties for the W+jets process are only derived for the 0 lepton channel because the contribution of these events to the 2 lepton channel is negligible. Due to the event selection bias and differences in the contributing production mechanisms depending on the "flavour" of the two *signal* jets, that were identified as b-jets, uncertainties are derived depending on this "flavour". The largest component is $V + (bb, bc, bl, cc)$ whose normalisation is treated separately:

- Normalisation uncertainty V + (bb, bc, bl, cc): Comparisons of V+jets simulated events with data, see e.g. appendix A.4, with b-tagging requirements show a large discrepancy in their normalisations. Therefore the normalisation of the simulated $Z + (bb, bc, bl, cc)$ events is a free parameter (*floating*) in the profile likelihood fit. Since especially the 2 lepton channel has a high purity of these events the normalisation can be determined in this way. The same procedure cannot be applied to the W + (bb, bc, bl, cc) simulated events since it is a less dominant background process in the 0 lepton channel. Therefore a global normalisation uncertainty of 33% is assigned based on the maximum observed difference between data and simulated events in the 1 lepton channel of the standard ATLAS analysis.

Additional uncertainties are derived from comparisons of the numerous variations, see table 8.2, that are available for the Sherpa MC generator. Uncertainties are also determined from the comparison of the Sherpa model with the model of an alternative MC generator, MadGraph5_aMC@NLO interfaced with Pythia for the description of the PS. The final uncertainties are calculated as the sum in quadrature of all these variations. Therefore their size represent an upper bound on the V+jets modelling uncertainties encompassing various possible sources of modelling inaccuracies:

- Acceptance uncertainties, flavour fraction differences: To account for differences in the flavour composition of the V + (bb, bc, bl, cc) component acceptance uncertainties are derived for the bc, bl and cc component each with respect to the bb component. They are derived separately for the 0 lepton Z+jets, 2 lepton Z+jets and 0 lepton W+jets process. The size of these uncertainties in the 2 lepton 2 jets and 2 lepton 3+ jets category are different.

- Acceptance uncertainty, Z+jets 0 and 2 lepton differences: All assigned Z+jets modelling uncertainties are treated in a correlated way between the 0 lepton and 2 lepton channel in the profile likelihood fit. Therefore an additional uncertainty is derived to account for differences in the 0 lepton channel with respect to the 2 lepton channel.

- Normalisation uncertainty V + cl: Separate normalisation uncertainties are derived for the Z+jets and W+jets process. The same uncertainties is used for the 0 and 2 lepton channel.

- Normalisation uncertainty V+light: Separate normalisation uncertainties are derived for the Z+jets and W+jets process. The same uncertainties is used for the 0 and 2 lepton channel.

Additional shape uncertainties are derived separately for the Z+jets and W+jets process:

- $m_{b\bar{b}}$ and p_{T}^{Z} shape uncertainty, Z+jets data to simulation differences: The shape uncertainties of the simulated Z+jets events were derived from comparisons of data with the nominal MC model in the 2 lepton channel. This channel provides a high purity of Z+jets events. To minimise the amount of $t\bar{t}$ events an additional selection criteria on the $E_{\mathrm{T}}^{\mathrm{miss}}$ significance[1] of $< 3.5\ \sqrt{\mathrm{GeV}}$ is applied. It is only applied when the modelling uncertainties are assessed. To assess different jet "flavour" compositions the shapes are studied in regions with 2 b-jets and in regions where only one or none of the *signal* jets is a b-jets. The two latter being enriched in c-jets and light jets. The size of the observed variations in these distributions is similar for the 2 jets and 3+ jets category independent if there are 0, 1 or 2 b-jets present. Therefore a single p_{T}^{Z} and a single $m_{b\bar{b}}$ shape uncertainty are derived for all Z+jets components and assigned to all analysis categories. Since no fundamental difference in terms of the underlying physics is expected between the $Z \to \nu\bar{\nu}$ and $Z \to \ell^{-}\ell^{+}$ decay the same shape uncertainties are used for the Z+jets simulated events in the 0 lepton channel as well.

[1] The $E_{\mathrm{T}}^{\mathrm{miss}}$ significance is defined as: $E_{\mathrm{T}}^{\mathrm{miss}} / \sqrt{H_{\mathrm{T}}}$, with H_{T} the scalar sum of the p_{T} of the leptons and jets in the event

- $m_{b\bar{b}}$ and p_{T}^{Z} shape uncertainties, Sherpa to MadGraph5_aMC@NLO differences: The W+jets shape uncertainties are derived from comparisons of two different MC generators in the 1 lepton channel of the standard ATLAS analysis. The simulated events used for this comparison were generated including ATLAS detector simulation and reconstruction. The W+jets uncertainties cannot be derived from data since no pure W+jets analysis region exists. The same shape uncertainties are assigned to the 2 jets and 3 jets category. The shape differences in the available variations of the Sherpa model are investigated as well but found to be negligible with respect to the size of the variation from the comparison to the MadGraph5_aMC@NLO model.

B.3.3 $t\bar{t}$

The 0 lepton and 2 lepton channel probe different $t\bar{t}$ final states. Therefore separate uncertainties are derived for those two channels. Due to the large amount of $t\bar{t}$ events in the 0 lepton 3 jets category and the dedicated $t\bar{t}$ control region in the 2 lepton channel the overall $t\bar{t}$ normalisation is treated separately:

- Normalisation uncertainty: The normalisation of the simulated $t\bar{t}$ events is a free parameter (*floating*) in the profile likelihood fit. Since different $t\bar{t}$ phase spaces are probed in the 0 lepton, 2 lepton 2 jets and 2 lepton 3+ jets category independent floating normalisation parameters are introduced for each of them.

Additional uncertainties are derived from the comparison of the nominal MC simulation model with the available alternative simulation models listed in table 8.2:

- Acceptance uncertainties: Since a common floating normalisation parameter is introduced for the 0 lepton 2 jets and 0 lepton 3 jets category an additional acceptance uncertainty is determined from the sum in quadrature of all simulation model differences. This uncertainty accounts for differences in the 2 jet category with respect to the 3 jet category, which is more enriched in $t\bar{t}$ events.

- $m_{b\bar{b}}$ and p_{T}^{Z} shape uncertainties: Separate shape uncertainties are derived for the 0 lepton, 2 lepton 2 jets and 2 lepton 3+ jets categories. The variation in the shape observed by comparing the nominal model, simulated with Powheg, to a model simulated with MadGraph5_aMC@NLO is significantly larger than the variations of other MC models. Therefore the shape uncertainty is derived from the MadGraph5_aMC@NLO variation.

B.3.4 Diboson

A similar strategy as for the VH modelling uncertainties is used to determine the diboson modelling uncertainties. Since the amount of WZ events is small in the 0 and 2 lepton channel the uncertainties from the 1 lepton channel in the standard ATLAS analysis are used. Besides an overall normalisation uncertainty no modelling uncertainties are derived for the WW process. The WW process contributes less than 0.1% to the total background and is small compared to ZZ and WZ. The following analysis specific uncertainties are assigned to simulated diboson events:

- normalisation uncertainties: A normalisation uncertainty for the total amount of expected ZZ, WZ and WW events in the analysis phase space is calculated from the sum in quadrature of the variations observed in all available alternative MC models, see table 8.2. Therefore it represents an upper bound on the maximum modelling uncertainty.

- acceptance uncertainties, PS and UE variations: Additional acceptance uncertainties are assigned to account for variations of the number of expected events in each lepton channel with respect to the total amount of diboson events. They originate from uncertainties in the PS and UE event model. To derive them the sum in quadrature of all available PS tune variations is compared to the difference between the Pythia and Herwig parton shower model. The larger variation from the nominal is considered as the uncertainty. The same strategy is used to derive an uncertainty for the 3 (3+) jets categories to account for differences in the acceptance between them and the 2 jet categories. Additional uncertainties for the 0 lepton channel are derived to account for differences in the ZZ normalisation with respect to the 2 lepton channel and WZ normalisation with respect to the 1 lepton channel. Since the 1 lepton channel is not considered for this analysis the uncertainty is treated as a normalisation uncertainty.

- normalisation uncertainties, μ_{F} and μ_{R} variations: Normalisation uncertainties are derived from the available variations of μ_{F} and μ_{R}. The maximum observed differences from the nominal model is used as the uncertainty. Similar to the signal uncertainties, these uncertainties are assigned using the Stewart-Tackman method.
- $m_{b\bar{b}}$ and p_{T}^{Z} shape uncertainties, PS and UE variations: The shape uncertainties due to the PS and UE model are derived from the difference of the diboson simulation that was generated using Powheg +Pythia and Powheg +Herwig. The shape differences from the variations in the PS tune are negligible.
- $m_{b\bar{b}}$ and p_{T}^{Z} shape uncertainties, μ_{F} and μ_{R} variations: To assess these shape uncertainties the nominal model of Sherpa is compared to the model of Powheg +Pythia which encrypts various changes in the scale choices. This strategy is used since the shape variations observed using varied μ_{F} and μ_{R} are negligible in most analysis categories.

Normalisation, acceptance and scale changes due to variations in the PDF set are investigates as well and found to be negligible.

B.3.5 Single top

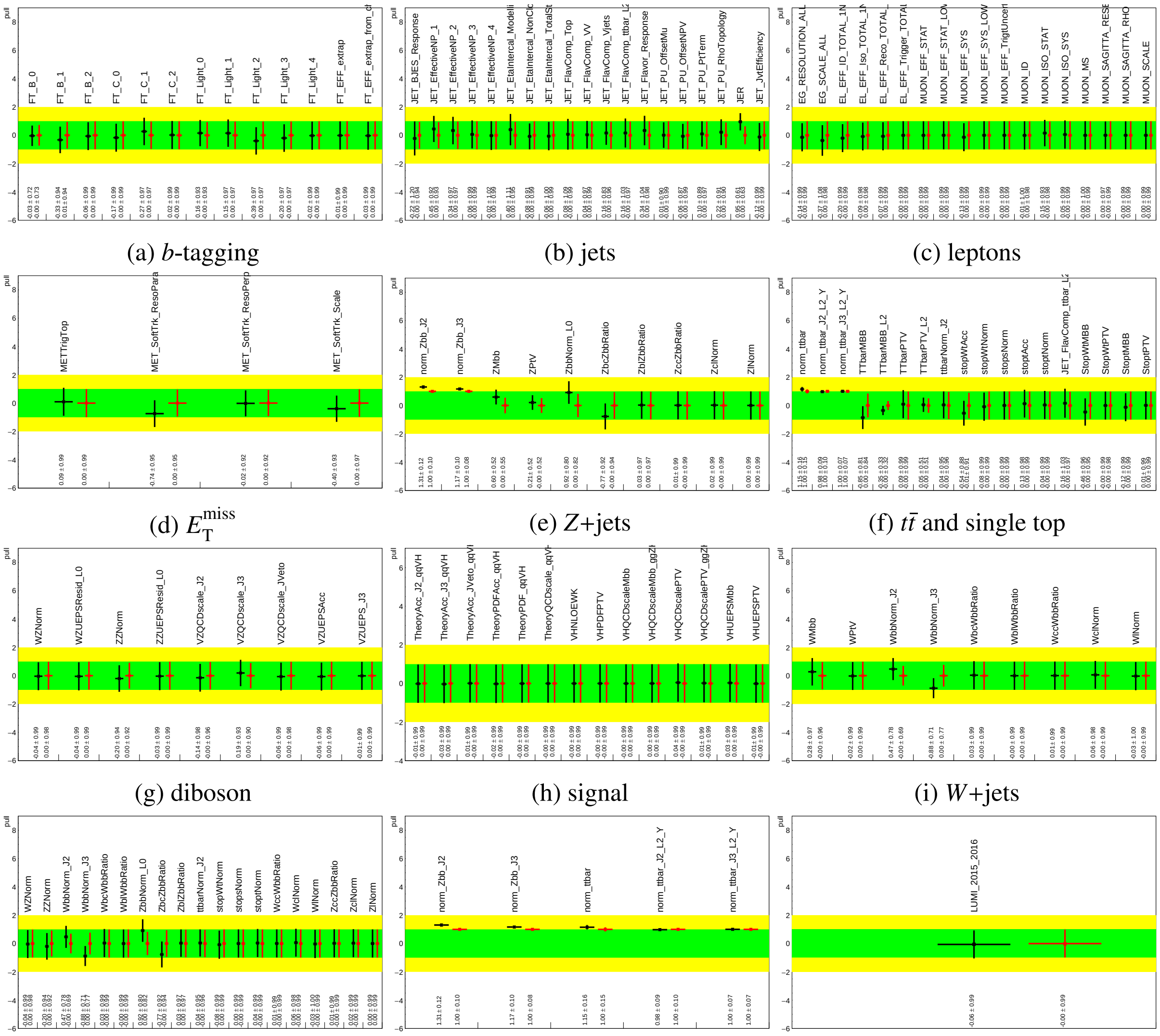
Single top modelling uncertainties are derived in a similar fashion as the $t\bar{t}$ uncertainties. Due to different production mechanisms and final states the t-channel, Wt and s-channel are considered separately. Overall normalisation uncertainties are assigned. Those uncertainties account for uncertainties in the PDF sets, α_S, μ_{F} and μ_{R}. Since the s-channel has a vanishing contribution in all analysis categories no further uncertainties are considered for this production channel. Additional analysis specific uncertainties are derived for the t-channel and Wt and the same uncertainties are used in the 0 lepton and 2 lepton channel:

- Normalisation uncertainties: Normalisation uncertainties are derived separately for the 2 jets and 3 (3+) jets channel from the sum in quadrature of the variations observed in the comparison between the nominal MC model and all available alternative MC models.
- $m_{b\bar{b}}$ and p_{T}^{Z} shape uncertainties: The shape uncertainties are assessed from the comparisons of the nominal model to all available alternative models. The shape differences in the t-channel are small and the envelope of all variations is assigned as a shape uncertainties. The shape variations in the Wt are dominated by the differences between the DR and DS scheme which are two different prescriptions to remove $t\bar{t}$ interferences. Therefore this variation is assigned as the shape variation in the Wt.

B.4 Postfit Diagnostics - Pulls and Yields

The nuisance parameter pulls and constraints are shown in figure B.22 for an Asimov data set and a fit to data.

The postfit yields of the signal and background processes are given in:
table B.3 for the 0 lepton categories,
table B.4 for the 2 lepton medium p_{T}^{Z} categories,
table B.5 for the 2 lepton high p_{T}^{Z} categories,
table B.6 for the $e\mu$ CR medium p_{T}^{Z} categories,
table B.7 for the $e\mu$ CR high p_{T}^{Z} categories.



(a) b-tagging

(b) jets

(c) leptons

(d) $E_{\mathrm{T}}^{\mathrm{miss}}$

(e) Z+jets

(f) $t\bar{t}$ and single top

(g) diboson

(h) signal

(i) W+jets

(j) acceptances

(k) floating normalisations

(l) Z+jets

Figure B.22: Nuisance parameter pulls of the $Z(H \to b\bar{b})$ fit for an Asimov data set (red) and a fit to data (black). The panels are sorted according to the source of uncertainties. Some nuisance parameters appear in more than one panel.

	0 lepton, 2 jets
Zl	7.73 ± 4.63
Zcl	23.74 ± 8.93
Zhf	2139.81 ± 113.65
Wl	7.03 ± 4.43
Wcl	21.29 ± 9.55
Whf	429.37 ± 101.38
stop	144.51 ± 24.88
ttbar	564.57 ± 68.74
diboson	109.01 ± 24.77
Bkg	3447.06 ± 62.85
Signal	54.75 ± 19.98
SignalExpected	47.63 ± 17.39
data	3520

	0 lepton, 3 jets
Zl	14.43 ± 7.97
Zcl	38.61 ± 13.13
Zhf	3120.45 ± 204.49
Wl	13.92 ± 6.73
Wcl	43.80 ± 18.73
Whf	623.85 ± 184.11
stop	522.33 ± 111.07
ttbar	4063.16 ± 305.21
diboson	141.49 ± 44.13
Bkg	8582.05 ± 95.53
Signal	59.54 ± 22.22
expected Signal	51.80 ± 19.33
data	8634

Table B.3: Postfit yields, 0 lepton 2 jet and 3 jet categories

	2 lepton, 2 jets, medium p_T^Z (SR)
Zl	8.62 ± 5.13
Zcl	23.88 ± 9.25
Zhf	3473.76 ± 80.90
Wl	0.00 ± 0.00
Wcl	0.04 ± 0.00
Whf	2.63 ± 0.17
stop	46.37 ± 17.10
ttbar	1447.30 ± 47.97
diboson	70.33 ± 18.00
Bkg	5072.93 ± 67.25
Signal	24.62 ± 9.08
SignalExpected	21.42 ± 7.90
data	5113

	2 lepton, 3+ jets, medium p_T^Z (SR)
Zl	28.97 ± 16.36
Zcl	88.68 ± 33.82
Zhf	8254.44 ± 146.73
Wl	0.02 ± 0.00
Wcl	0.14 ± 0.00
Whf	4.53 ± 0.15
stop	119.90 ± 52.37
ttbar	4925.21 ± 96.65
diboson	168.20 ± 35.31
Bkg	13590.09 ± 112.94
Signal	46.42 ± 17.77
expected Signal	40.38 ± 15.46
data	13640

Table B.4: Postfit yields, 2 lepton, 2 and 3+ jets, medium p_T^Z (SR)

	2 lepton, 2 jets, high p_T^Z (SR)
Zl	1.71 ± 1.02
Zcl	4.90 ± 1.86
Zhf	656.90 ± 20.52
Wl	0.00 ± 0.00
Wcl	0.01 ± 0.00
Whf	0.24 ± 0.01
stop	5.34 ± 2.03
ttbar	50.41 ± 3.01
diboson	23.28 ± 5.86
Bkg	742.80 ± 19.48
Signal	13.00 ± 4.74
expected Signal	11.31 ± 4.12
data	724

	2 lepton, 3+ jets, high p_T^Z (SR)
Zl	10.98 ± 6.63
Zcl	32.80 ± 12.65
Zhf	3050.89 ± 66.02
Wl	0.00 ± 0.00
Wcl	0.04 ± 0.00
Whf	1.70 ± 0.06
stop	26.80 ± 11.40
ttbar	437.09 ± 23.17
diboson	97.78 ± 22.25
Bkg	3658.09 ± 57.20
Signal	34.78 ± 13.22
expected Signal	30.26 ± 11.50
data	3708

Table B.5: Postfit yields, 2 lepton, 2 and 3+ jets, high p_T^Z (SR)

	$e\mu$ CR, 2 lepton, 2 jets, medium p_{T}^{Z}
Zl	0.00 ± 0.00
Zcl	0.01 ± 0.00
Zhf	1.33 ± 0.12
Whf	2.48 ± 0.09
stop	44.56 ± 16.60
ttbar	1436.52 ± 41.61
Bkg	1484.90 ± 37.16
Signal	0.00 ± 0.00
SignalExpected	0.00 ± 0.00
data	1489

	$e\mu$ CR, 2 lepton, 3+ jets, medium p_{T}^{Z}
Zl	0.01 ± 0.00
Zcl	0.03 ± 0.00
Zhf	0.89 ± 0.08
Whf	2.11 ± 0.07
stop	113.34 ± 50.78
ttbar	4873.03 ± 90.14
diboson	0.23 ± 0.01
Bkg	4989.63 ± 69.08
Signal	0.01 ± 0.01
expected Signal	0.01 ± 0.00
data	4967

Table B.6: Postfit yields, $e\mu$ CR, 2 lepton, 2 and 3+ jets, medium p_{T}^{Z}

	$e\mu$ CR, 2 lepton, 2 jets, high p_{T}^{Z}
Zl	0.00 ± 0.00
Zcl	0.00 ± 0.00
Zhf	0.11 ± 0.01
Whf	0.16 ± 0.01
stop	5.28 ± 2.03
ttbar	49.18 ± 3.85
Bkg	54.73 ± 3.81
Signal	0.00 ± 0.00
expected Signal	0.00 ± 0.00
data	50

	$e\mu$ CR, 2lepton, 3+ jets, high p_{T}^{Z}
Zl	0.00 ± 0.00
Zcl	0.00 ± 0.00
Zhf	0.72 ± 0.06
Whf	0.47 ± 0.02
stop	24.47 ± 10.84
ttbar	435.52 ± 22.05
Bkg	461.19 ± 18.83
Signal	0.01 ± 0.00
SignalExpected	0.01 ± 0.00
data	470

Table B.7: Postfit yields, $e\mu$ CR, 2 lepton, 2 and 3+ jets, high p_{T}^{Z}

B.5 Diboson Cross Check

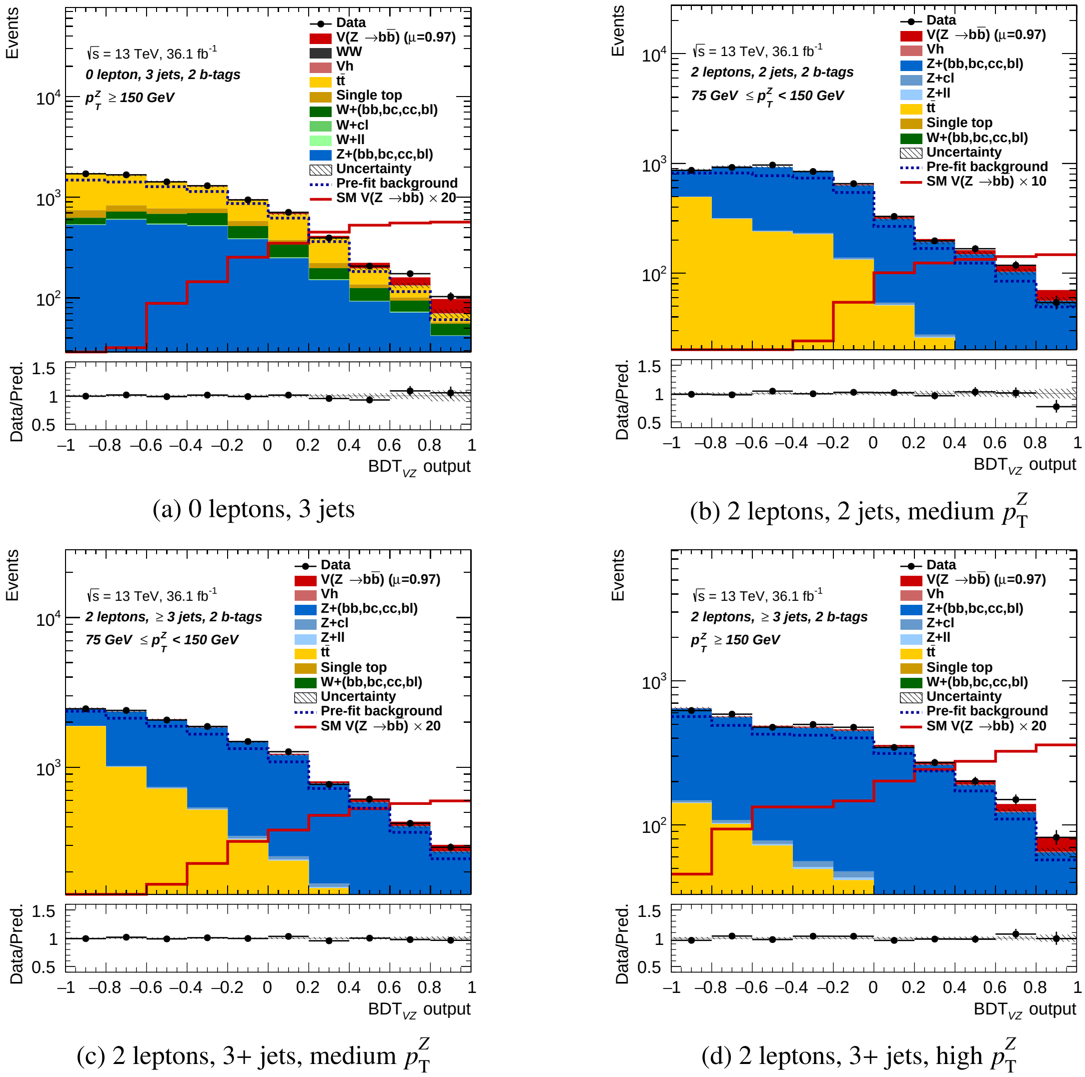
The postfit BDTs output distributions are shown in figure B.23.

The BDT output for signal and background for each analysis category for the trainings using events with an even event number B.24 and an odd event number B.25 including an *over training* check. For the *over training* check the BDTs output distribution for the training data set is compared to a statistically independent test data set.

The comparison of the distributions of the BDTs input variables in simulated signal and background events are given in:

figure B.26 for the 0 lepton 2 jets category,
figure B.27 for the 0 lepton 3 jets category,
figure B.28 for the 2 lepton 2 jets medium p_{T}^{Z} category,
figure B.29 for the 2 lepton 3+ jets medium p_{T}^{Z} category,
figure B.30 for the 2 lepton 2 jets high p_{T}^{Z} category,
figure B.31 for the 2 lepton 3+ jets high p_{T}^{Z} category.

The break down of the impact of the sources of uncertainties on the measured diboson μ is shown in table B.8.



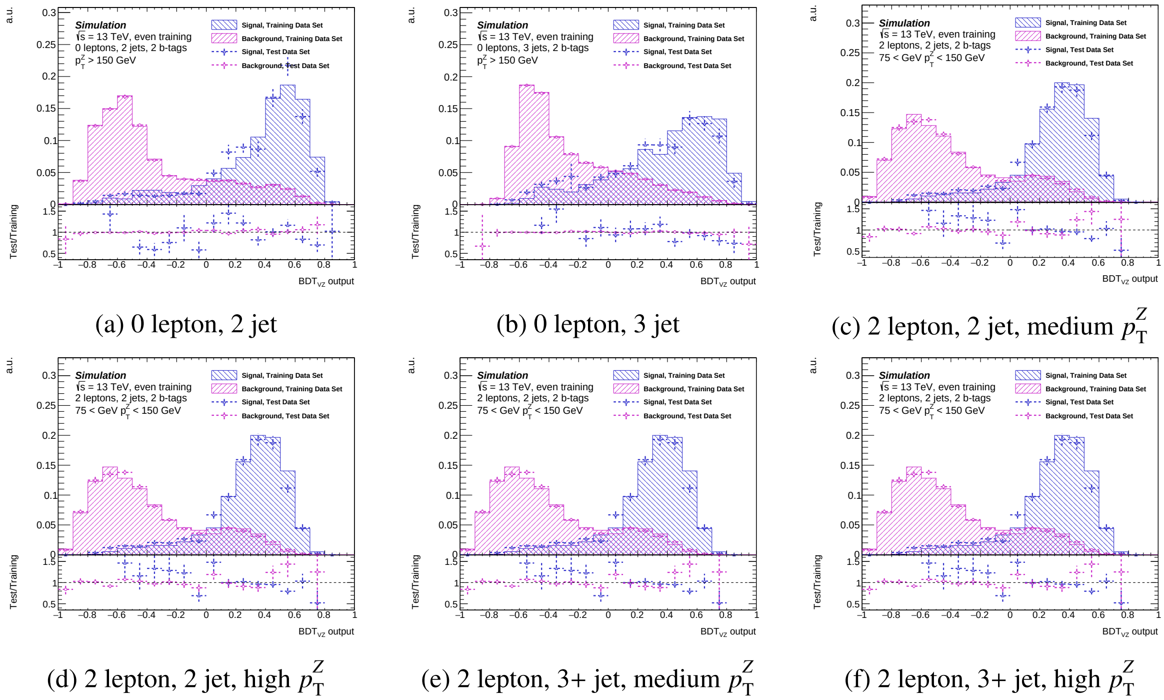
(a) 0 leptons, 3 jets

(b) 2 leptons, 2 jets, medium p_{T}^{Z}

(c) 2 leptons, 3+ jets, medium p_{T}^{Z}

(d) 2 leptons, 3+ jets, high p_{T}^{Z}

Figure B.23: The BDTs output distributions of the SRs of the diboson validation analysis. All signal and background components are scaled to their normalisations as determined in the fit. The normalisation of the sum of all background components predicted by the simulated events is given by the dashed blue line. The shaded bands represent the total uncertainty.



(a) 0 lepton, 2 jet

(b) 0 lepton, 3 jet

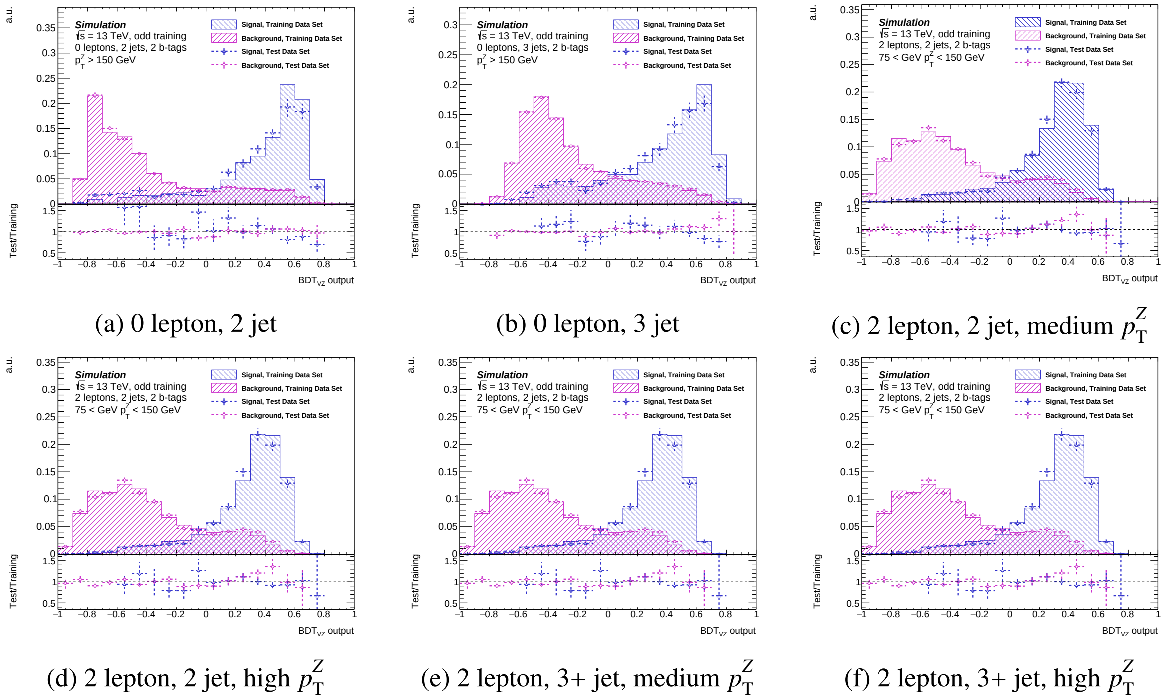
(c) 2 lepton, 2 jet, medium p_{T}^{Z}

(d) 2 lepton, 2 jet, high p_{T}^{Z}

(e) 2 lepton, 3+ jet, medium p_{T}^{Z}

(f) 2 lepton, 3+ jet, high p_{T}^{Z}

Figure B.24: BDT output distributions for the diboson MVA categories for the even training. The ratio plot shows the agreement between the training data set and an statistically independent test data set (over training check).



(a) 0 lepton, 2 jet

(b) 0 lepton, 3 jet

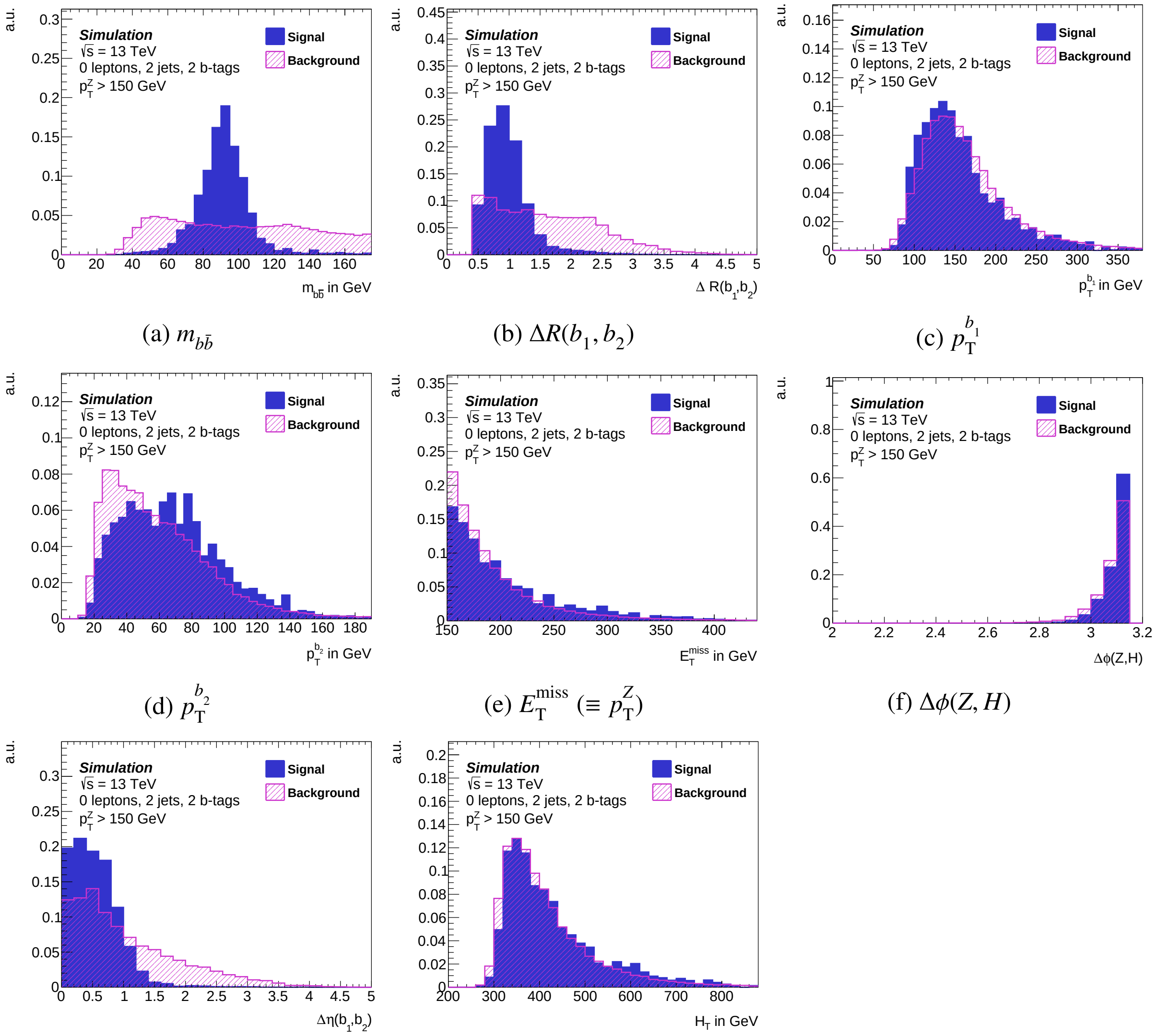
(c) 2 lepton, 2 jet, medium p_{T}^{Z}

(d) 2 lepton, 2 jet, high p_{T}^{Z}

(e) 2 lepton, 3+ jet, medium p_{T}^{Z}

(f) 2 lepton, 3+ jet, high p_{T}^{Z}

Figure B.25: BDT output distributions for the diboson MVA categories for the odd training. The ratio plot shows the agreement between the training data set and an statistically independent test data set (over training check).



(a) $m_{b\bar{b}}$

(b) $\Delta R(b_1, b_2)$

(c) $p_{\mathrm{T}}^{b_1}$

(d) $p_{\mathrm{T}}^{b_2}$

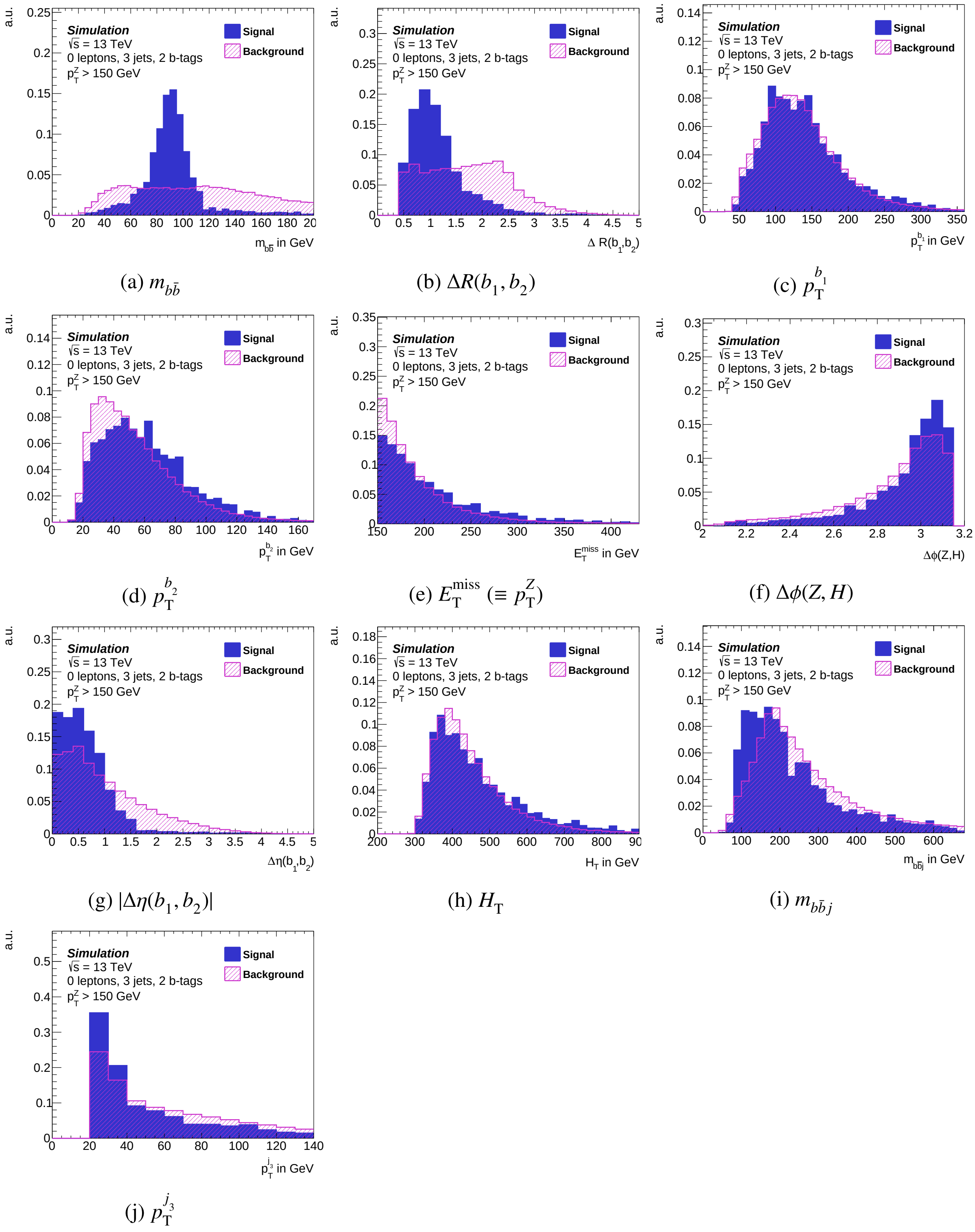
(e) $E_{\mathrm{T}}^{\mathrm{miss}}$ ($\equiv p_{\mathrm{T}}^{Z}$)

(f) $\Delta\phi(Z, H)$

(g) $|\Delta\eta(b_1, b_2)|$

(h) H_{T}

Figure B.26: Input variables for the 0 lepton 2 jet diboson MVA. Shown are the distributions in simulated signal events and the sum of all simulated background events. All distributions are normalised to the same area.



(a) $m_{b\bar{b}}$

(b) $\Delta R(b_1, b_2)$

(c) $p_{\mathrm{T}}^{b_1}$

(d) $p_{\mathrm{T}}^{b_2}$

(e) $E_{\mathrm{T}}^{\mathrm{miss}}$ $(\equiv p_{\mathrm{T}}^{Z})$

(f) $\Delta\phi(Z, H)$

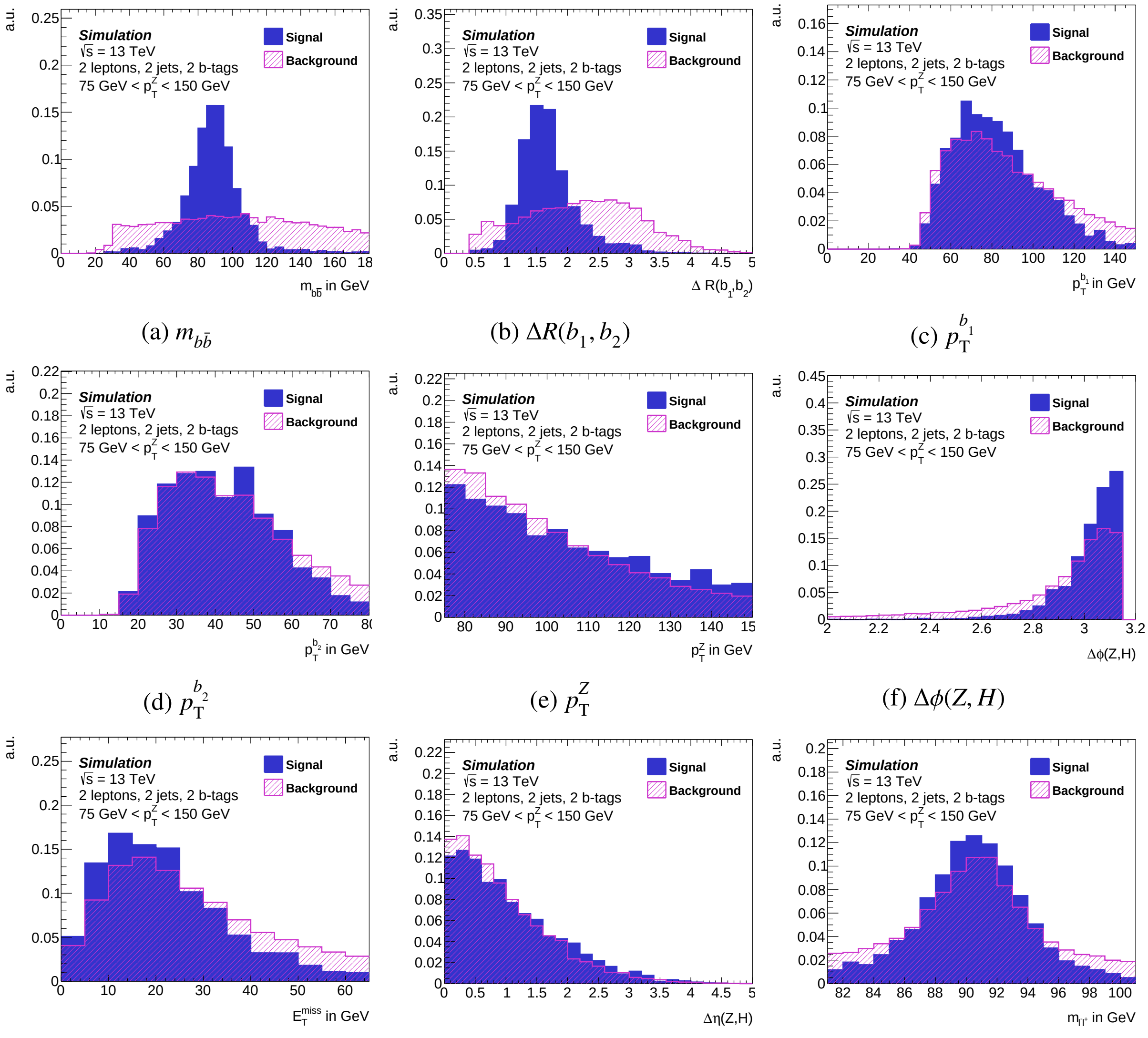
(g) $|\Delta\eta(b_1, b_2)|$

(h) H_{T}

(i) $m_{b\bar{b}j}$

(j) $p_{\mathrm{T}}^{j_3}$

Figure B.27: Input variables for the 0 lepton 3 jet diboson MVA. Shown are the distributions in simulated signal events and the sum of all simulated background events. All distributions are normalised to the same area.



(a) $m_{b\bar{b}}$

(b) $\Delta R(b_1, b_2)$

(c) $p_{\mathrm{T}}^{b_1}$

(d) $p_{\mathrm{T}}^{b_2}$

(e) p_{T}^{Z}

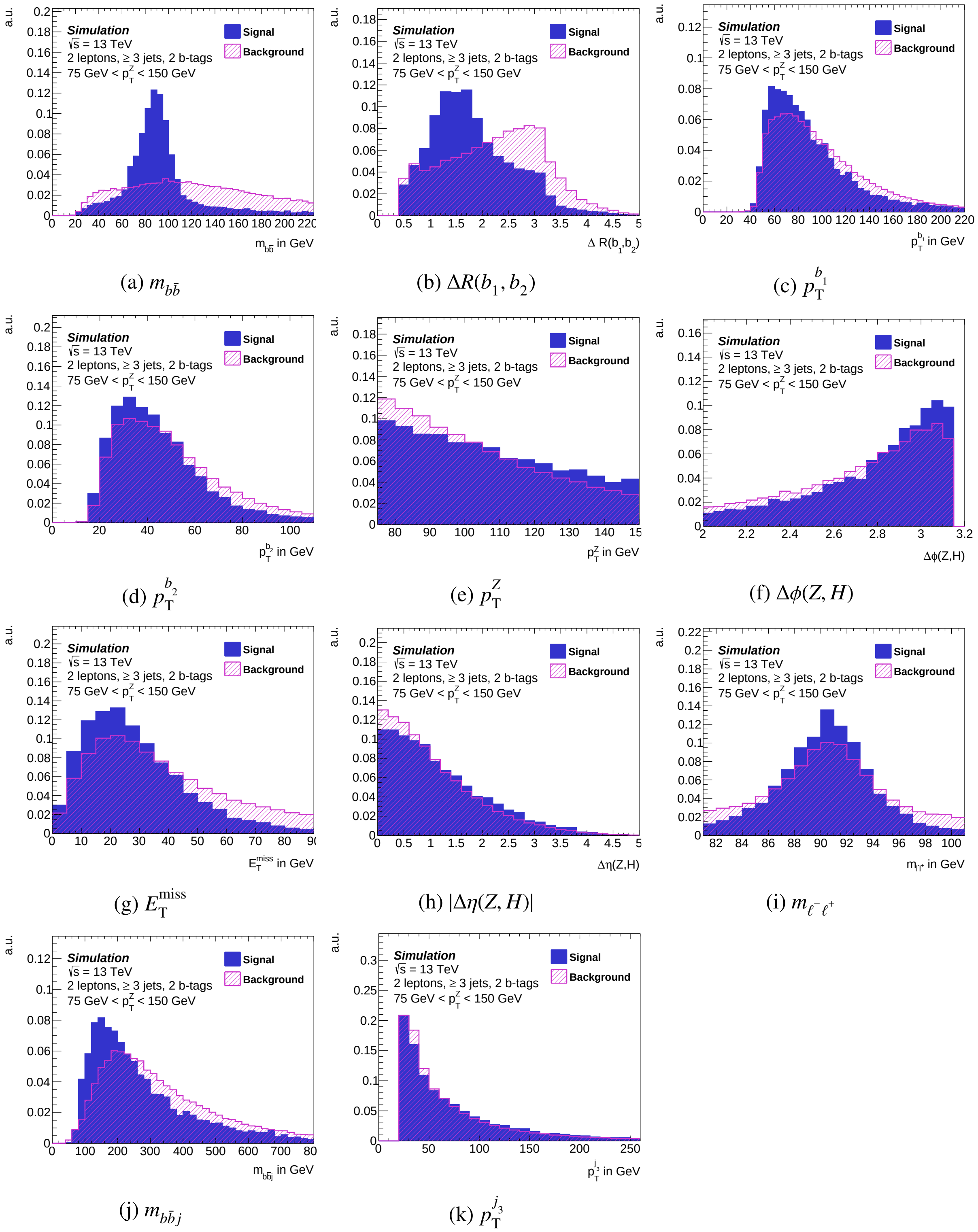
(f) $\Delta\phi(Z, H)$

(g) $E_{\mathrm{T}}^{\mathrm{miss}}$

(h) $|\Delta\eta(Z, H)|$

(i) $m_{\ell^-\ell^+}$

Figure B.28: Input variables for the 2 lepton 2 jet medium diboson p_{T}^{Z} MVA. Shown are the distributions in simulated signal events and the sum of all simulated background events. All distributions are normalised to the same area.



(a) $m_{b\bar{b}}$

(b) $\Delta R(b_1, b_2)$

(c) $p_T^{b_1}$

(d) $p_T^{b_2}$

(e) p_T^Z

(f) $\Delta\phi(Z,H)$

(g) E_T^{miss}

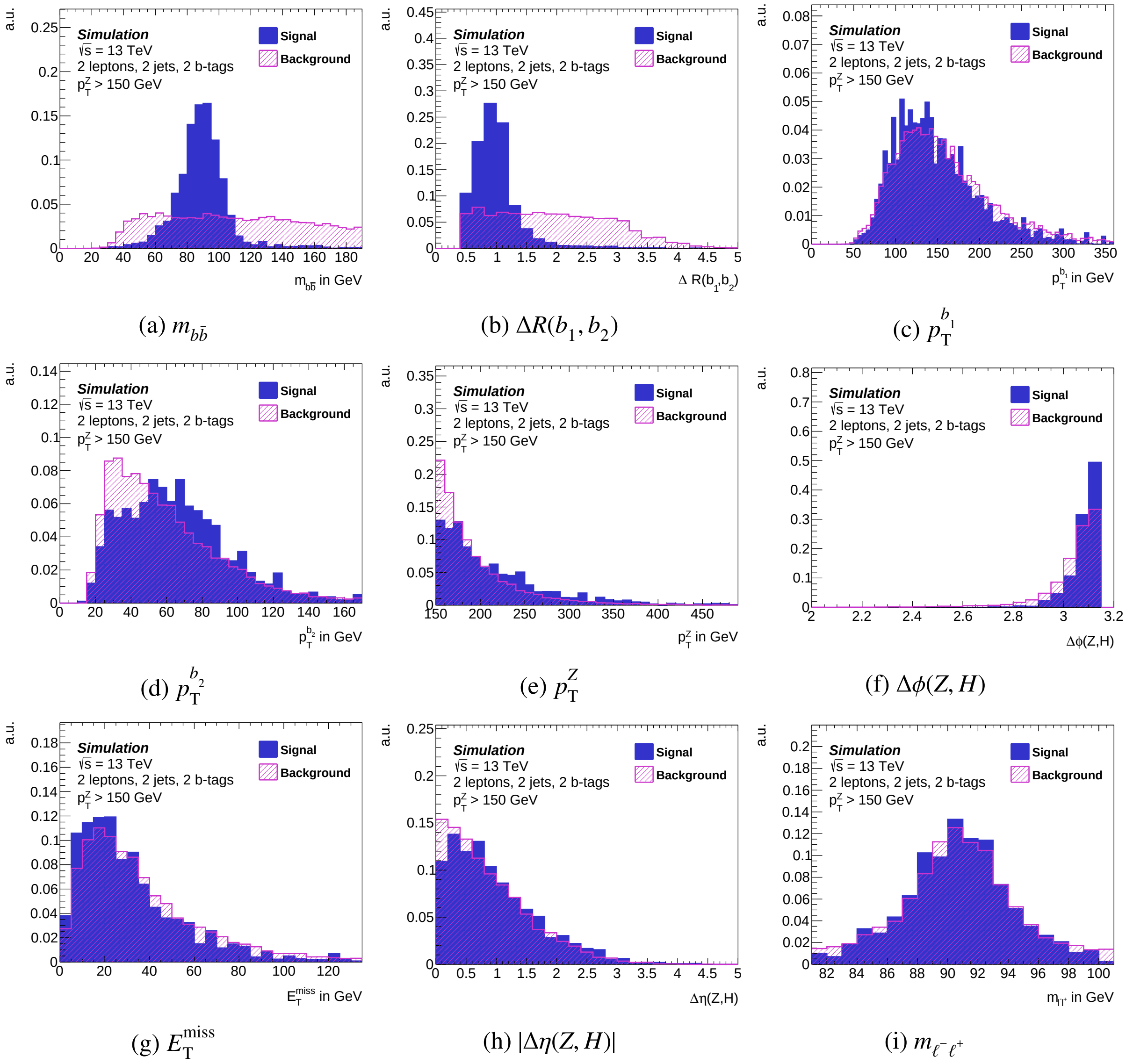
(h) $|\Delta\eta(Z,H)|$

(i) $m_{\ell^-\ell^+}$

(j) $m_{b\bar{b}j}$

(k) $p_T^{j_3}$

Figure B.29: Input variables for the 2 lepton 3 jet medium p_T^Z diboson MVA. Shown are the distributions in simulated signal events and the sum of all simulated background events. All distributions are normalised to the same area.



(a) $m_{b\bar{b}}$

(b) $\Delta R(b_1, b_2)$

(c) $p_{\mathrm{T}}^{b_1}$

(d) $p_{\mathrm{T}}^{b_2}$

(e) p_{T}^{Z}

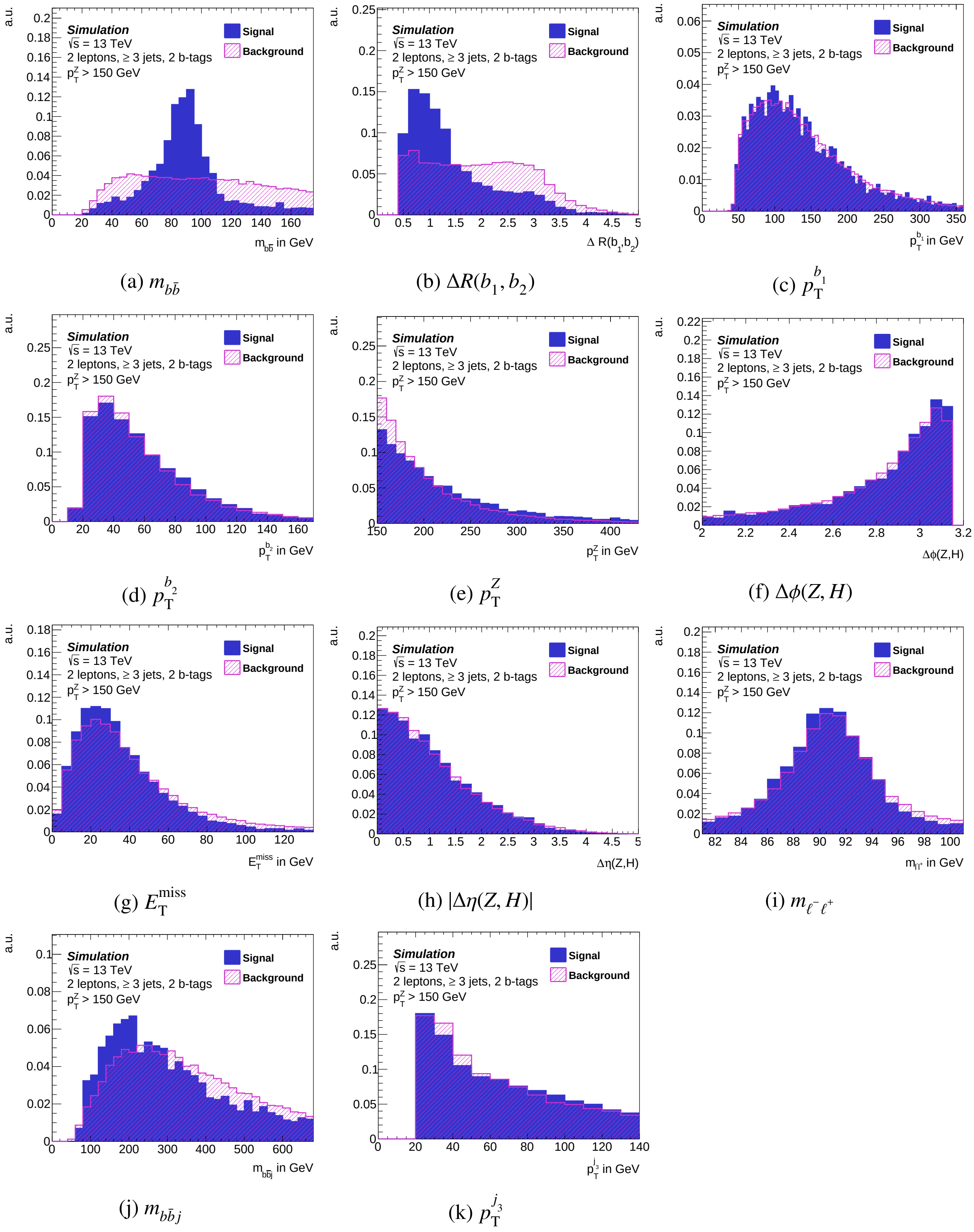
(f) $\Delta\phi(Z, H)$

(g) $E_{\mathrm{T}}^{\mathrm{miss}}$

(h) $|\Delta\eta(Z, H)|$

(i) $m_{\ell^-\ell^+}$

Figure B.30: Input variables for the 2 lepton 2 jet high diboson p_{T}^{Z} MVA. Shown are the distributions in simulated signal events and the sum of all simulated background events. All distributions are normalised to the same area.



(a) $m_{b\bar{b}}$

(b) $\Delta R(b_1, b_2)$

(c) $p_T^{b_1}$

(d) $p_T^{b_2}$

(e) p_T^Z

(f) $\Delta\phi(Z, H)$

(g) E_T^{miss}

(h) $|\Delta\eta(Z, H)|$

(i) $m_{\ell^-\ell^+}$

(j) $m_{b\bar{b}j}$

(k) $p_T^{j_3}$

Figure B.31: Input variables for the 2 lepton 3 jet high p_T^Z diboson MVA. Shown are the distributions in simulated signal events and the sum of all simulated background events. All distributions are normalised to the same area.

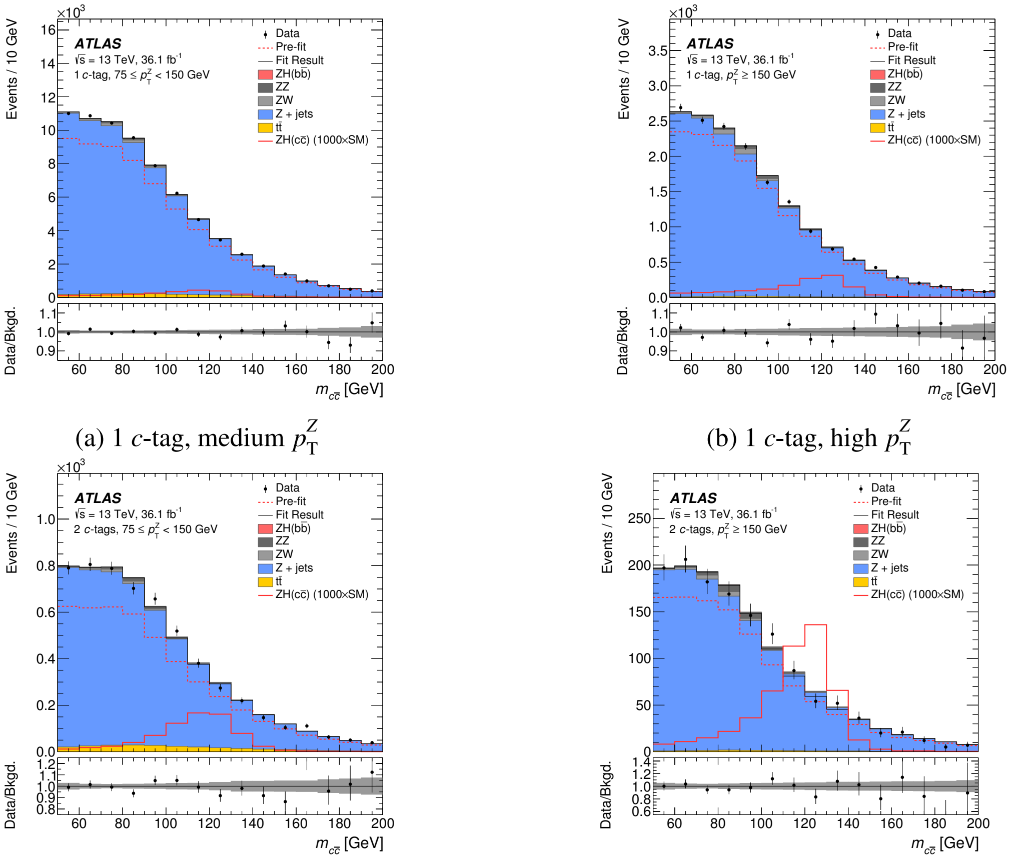
	measured μ: 0.968255
Set of nuisance	Impact on error
Total	+0.223 / -0.201
Stat.	+0.120 / -0.118
Syst.	+0.188 / -0.163
Floating normalisations	+0.049 / -0.048
Jets and $E_{\mathrm{T}}^{\mathrm{miss}}$	+0.052 / -0.054
b-tagging	+0.054 / -0.045
Electrons and muons	+0.007 / -0.006
Luminosity	+0.037 / -0.025
Diboson modelling	+0.047 / -0.031
Z+jets modelling	+0.080 / -0.079
W+jets modelling	+0.026 / -0.027
$t\bar{t}$ modelling	+0.044 / -0.050
Single top modelling	+0.008 / -0.007
VH uncertainties	+0.001 / -0.001
MC stat.	+0.068 / -0.065

Table B.8: Diboson $V(Z \rightarrow b\bar{b})$ break down of uncertainties

APPENDIX C

Appendix - SM $ZH \rightarrow \ell^{-}\ell^{+}c\bar{c}$ Search

The postfit $m_{c\bar{c}}$ distributions are shown in figure C.1.



(a) 1 c-tag, medium p_T^Z

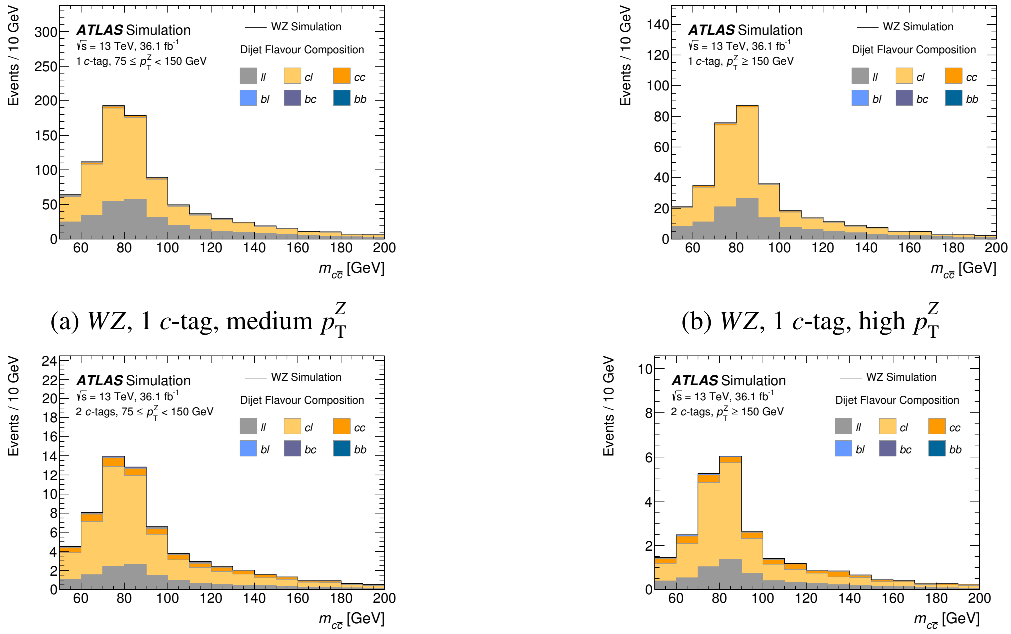
(b) 1 c-tag, high p_T^Z

(c) 2 c-tag, medium p_T^Z

(d) 2 c-tags, high p_T^Z

Figure C.1: The $m_{c\bar{c}}$ distributions of the $ZH \rightarrow \ell^{-}\ell^{+}c\bar{c}$ analysis regions [96]. All signal and background components are scaled to their normalisations as determined by the fit. The normalisation of the sum of all background components as predicted by the simulated events is given by the dashed red line. The solid red line shows an overlay of the $ZH \rightarrow \ell^{-}\ell^{+}c\bar{c}$ signal distribution scaled by a factor 1000. The shaded bands represent the total uncertainty.

The flavour composition of the WZ and ZZ events are shown in figure C.2 and C.3 respectively.



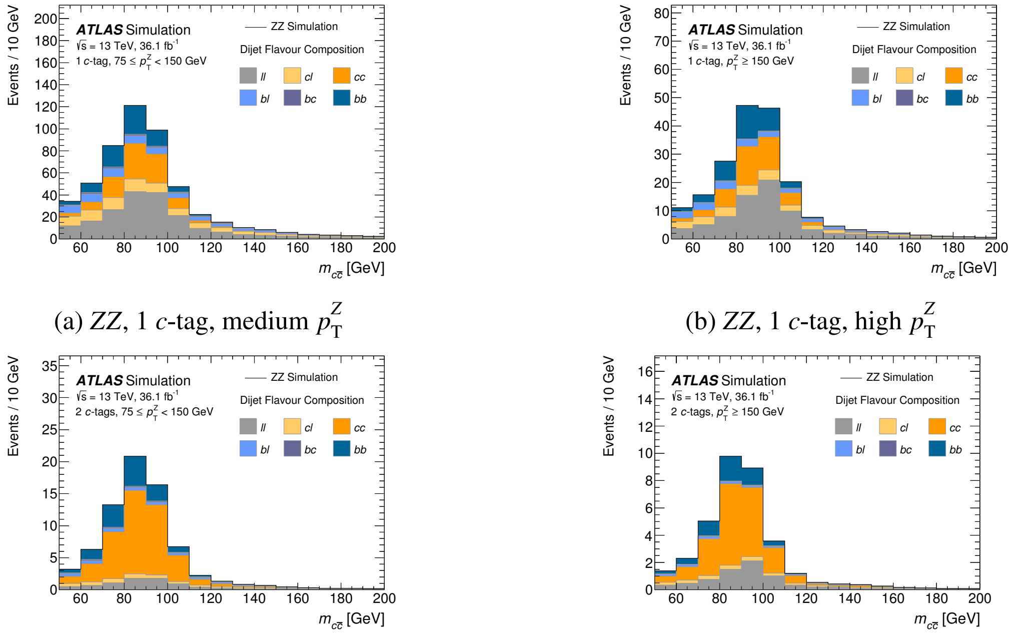
(a) WZ, 1 c-tag, medium p_{T}^{Z}

(b) WZ, 1 c-tag, high p_{T}^{Z}

(c) WZ, 2 c-tag, medium p_{T}^{Z}

(d) WZ, 2 c-tags, high p_{T}^{Z}

Figure C.2: The flavour composition of the WZ events in the $ZH \to \ell^- \ell^+ c\bar{c}$ analysis categories.



(a) ZZ, 1 c-tag, medium p_{T}^{Z}

(b) ZZ, 1 c-tag, high p_{T}^{Z}

(c) ZZ, 2 c-tag, medium p_{T}^{Z}

(d) ZZ, 2 c-tags, high p_{T}^{Z}

Figure C.3: The flavour composition of the ZZ events in the $ZH \to \ell^- \ell^+ c\bar{c}$ analysis categories.

APPENDIX D

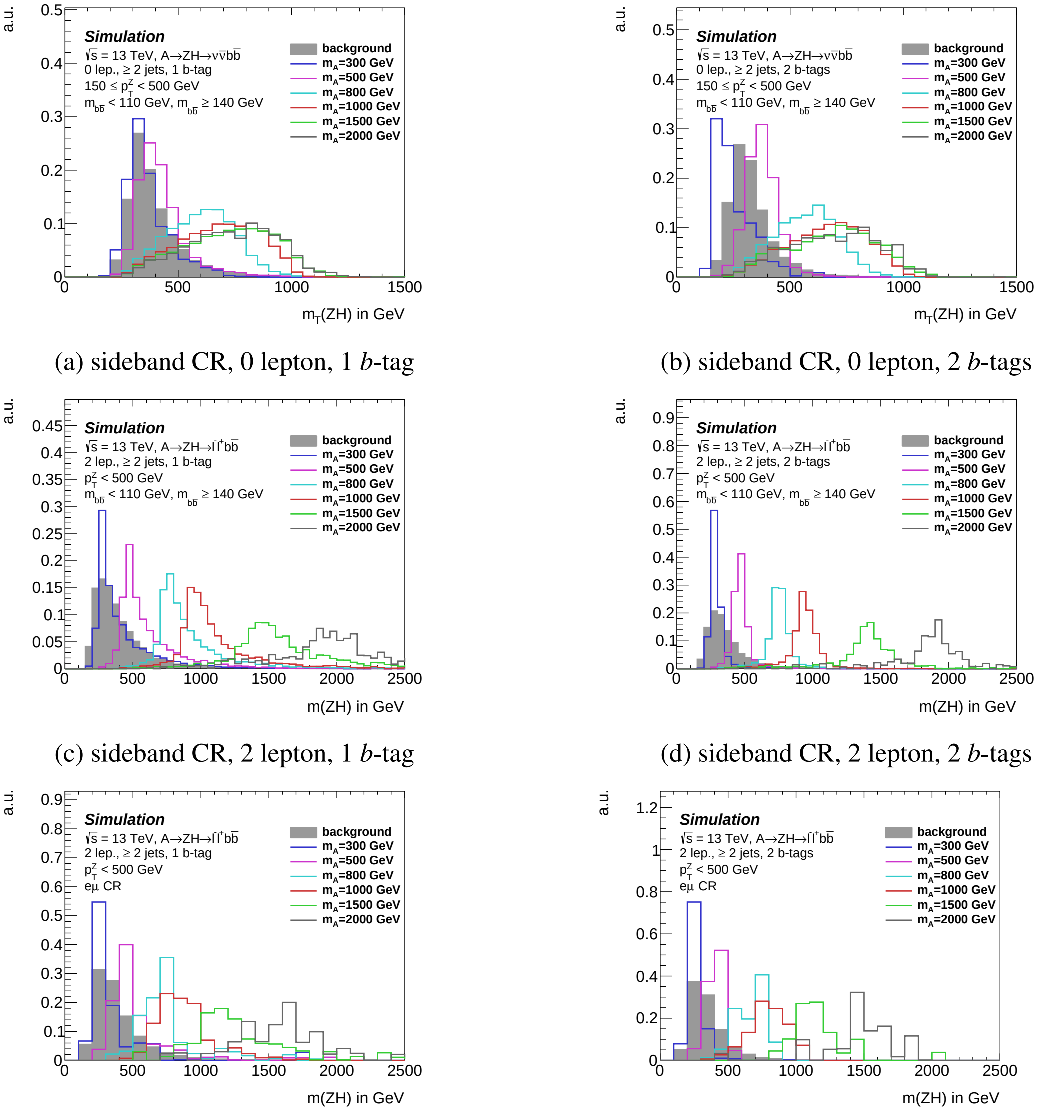
Appendix - $A \rightarrow ZH$ Search

D.1 $m(ZH)$ Distributions in CRs

Comparisons of the $m(ZH)$ distributions for different signal mass hypotheses and the sum of all backgrounds in the CRs of the analysis are shown in figure D.1.

D.2 $A \rightarrow ZH$ Fit Model Studies

To enhance the sensitivity to certain background components, especially increase of the separation of V+jets and $t\bar{t}$ events, a split based on the jet multiplicities is tested. To quantify the sensitivity the expected limit of the 2 lepton channel, which has a very good $m(ZH)$ resolution, is calculated for different fit configurations. The default fit set-up utilises the analysis categories, as described in section 10.2, and a bin width of 40 GeV for the $m(ZH)$ distributions in the SRs and of 200 GeV in the CRs. Figure D.2a) shows the relative improvement of the expected limit with respect to this default fit set-up. It demonstrates that an improvement of up to 20% is achieved for $m_A = 800$ GeV with a split of each category in three categories: 2 or 3 jets, 4 jets and $\geq$ 5 jets. However, given the small amount of expected events in some analysis categories this agressive splitting scheme might introduce instabilities due to large statistical uncertainties in some bins of the $m(ZH)$ distributions. Thus an alternative strategy is explored, which decreases the bin size in the 2 lepton SRs. To avoid similar problems due to statistical uncertainties finer bins can only be used in more inclusive regions that provide a larger amount of expected events. The results of this study are shown in figure D.2a) as well. The final chosen fit configuration uses bins of 10 GeV width in the 2 lepton SRs and does not introduce additional splits based on the number of *signal* jets. This yields an improvement of 20% on the limit for $m_A = 300$ GeV and of 10% for the limit of $m_A = 300$ GeV since the shape information is optimally used. A second study is performed to determine if the fit model may be simplified by excluding categories. The results are quantified as improvements of the expected limit for the 0 and 2 lepton combined fit and are shown in figure D.2b). It shows that the $m_{b\bar{b}}$ sideband CRs and the $1b$-tag categories add 7% and an additional 5% improvement in the expected limit for $m_A = 300$ GeV. The improvement is smaller for $m_A = 800$ GeV: 6% in total. Since a non-negligible amount of the signal events are accepted in the $m_{b\bar{b}}$ sideband and $1b$-tag CRs these improvements are not surprising. The $e\mu$ CRs do not improve the expected limit but they are kept since they are important to extract the $t\bar{t}$ normalisation.



(a) sideband CR, 0 lepton, 1 *b*-tag

(b) sideband CR, 0 lepton, 2 *b*-tags

(c) sideband CR, 2 lepton, 1 *b*-tag

(d) sideband CR, 2 lepton, 2 *b*-tags

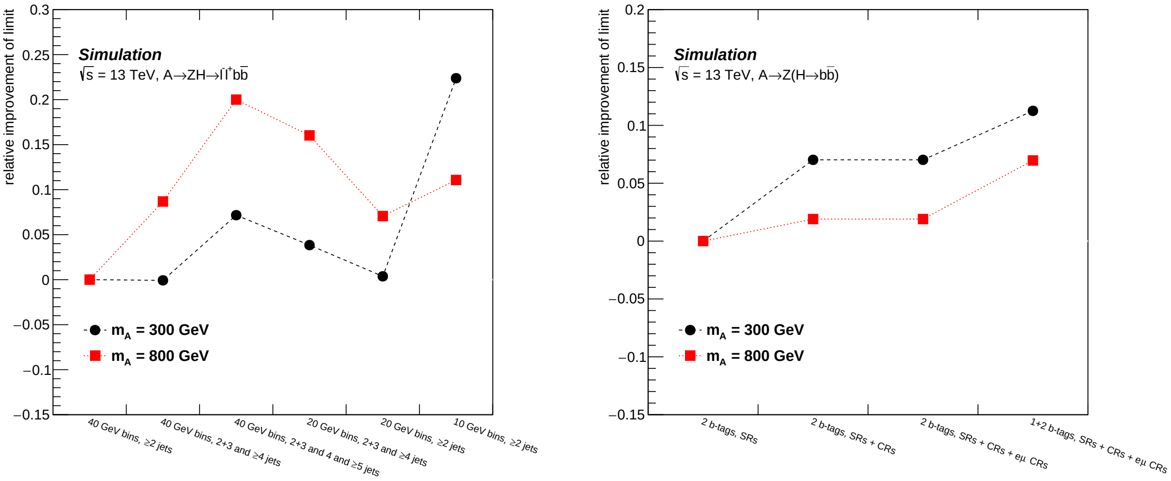
(e) $e\mu$ CR, 2 lepton, 1 *b*-tag

(f) $e\mu$ CR, 2 lepton, 2 *b*-tags

Figure D.1: The $m_T(ZH)$ and $m(ZH)$ distributions for different A boson mass hypotheses and the sum of all background processes in the CRs of the $A \to ZH$ analysis: a) sideband CR, 0 lepton, 1 *b*-tag, b) sideband CR, 0 lepton, 2 *b*-tags, c) sideband CR, 2 lepton, 1 *b*-tag, d) sideband CR, 2 lepton, 2 *b*-tags, e) $e\mu$ CR, 2 lepton, 1 *b*-tag, f) $e\mu$ CR, 2 lepton, 2 *b*-tags. All distributions are normalised to unity.

D.3 $m(ZH)$ 1 *b*-tag Postfit Distributions

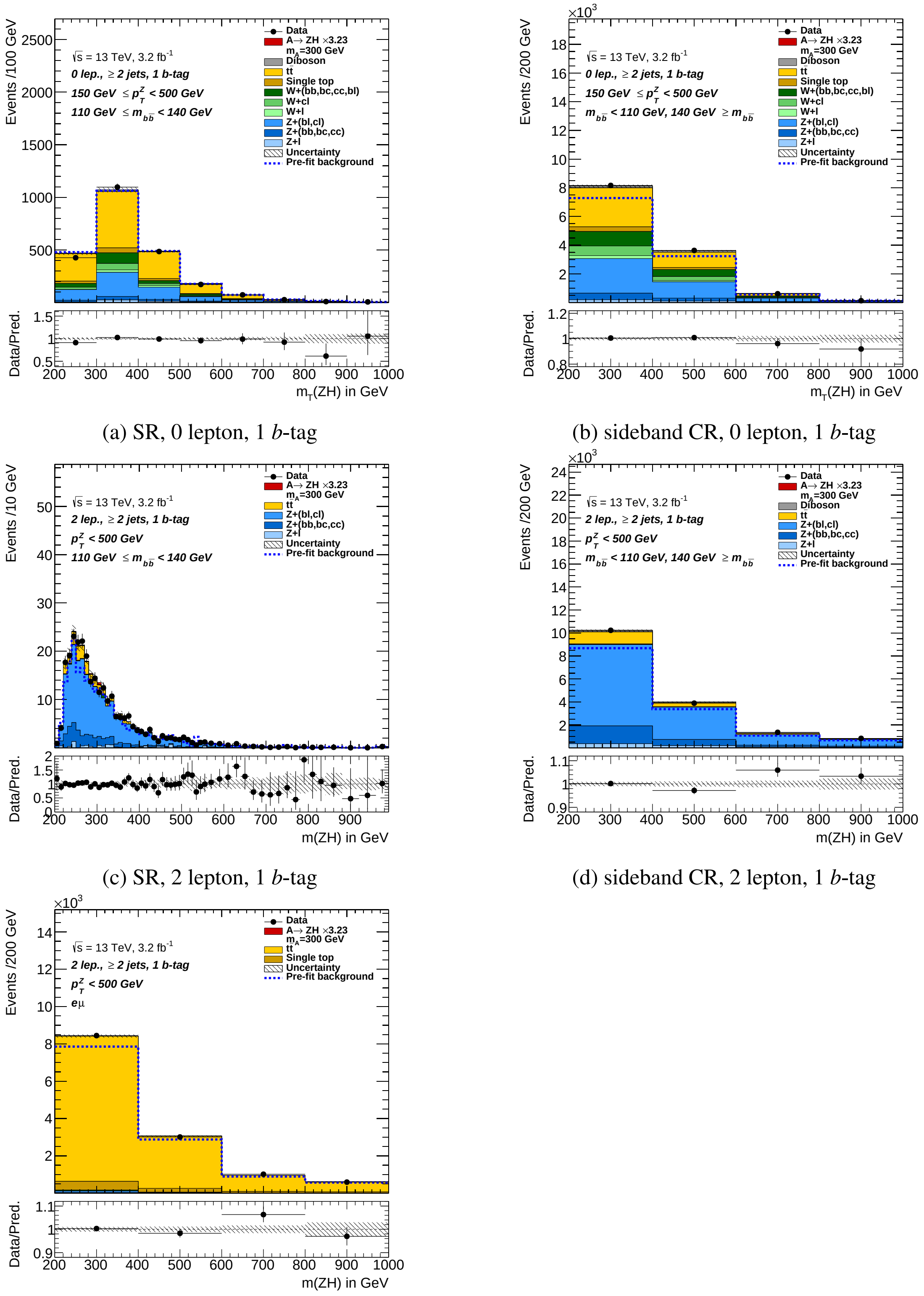
Figure D.3 shows the postfit $m_T(ZH)$ and $m(ZH)$ distributions in the 1 btag categories. Comparing data and simulated events. The simulated background events are scaled by their nuisance parameter values determined by the fit and the simulated signal event are scaled to the observed upper exclusion limit.



(a) jet multiplicity and binning studies

(b) analysis categories studies

Figure D.2: The evolution of the expected limits for A bosonmass hypothesis of 300 GeV and 800 GeV for different fit set-ups: a) testing different jet multiplicity splits and bin widths for the 2 lepton categories, b) studying the impact of CRs and the $1b$-tag categories. Shown are the relative improvements with respect to the first configuration of the test. The utilised set-up is the last one in both cases, i.e. they correspond to the same expected limit.



(a) SR, 0 lepton, 1 b-tag

(b) sideband CR, 0 lepton, 1 b-tag

(c) SR, 2 lepton, 1 b-tag

(d) sideband CR, 2 lepton, 1 b-tag

(e) $e\mu$ CR, 2 lepton, 1 b-tag

Figure D.3: The postfit $m_\mathrm{T}(ZH)$ and $m(ZH)$ distributions in the 1 btag categories; the simulated background events are scaled by their nuisance parameter values determined by the fit; the signal is scaled to its observed upper exclusion limit. a) SR, 0 lepton, 1 b-tag, b) sideband CR, 0 lepton, 1 b-tag, c) SR, 2 lepton, 1 b-tag, d) sideband CR, 2 lepton, 1 b-tag, e) $e\mu$ CR, 2 lepton, 1 b-tag

List of Figures

List of Tables

Glossary

***b*-hadron** A b-hadron is a hadron with at least one b-quark as a valence quark.

***c*-hadron** A c-hadron is a hadron with at least one c-quark as a valence quark and no b-quark as a valence quark.

beta function The beta function describes the size of a beam in the plane transverse to the beam. β^* refers to this size at the interaction point of the beam..

elastic scattering An elastic scattering event, in the context of pp collisions, is a collision event where the inner structure of the protons is not probed.

fermion A fermion is particle with a spin of $n \times \frac{1}{2}$ (n is an integer).

hard A hard object is an object that carries a significant amount of transverse momentum relative to a reference axis, e.g. objects with large momenta in the xy-plane which were created by the LHC proton-proton collisions taking place in the z-direction would be called hard objects..

hard scattering A hard scattering event, in the context of pp collisions, is a collision event where the inner structure of the protons is pierced and new particles with high transverse momenta with respect to the beam are produced.

inelastic scattering An inelastic scatterig event, in the context of pp collisions, is a collision event where the inner structure of the protons is partially pierced and mostly low energetic pions are produced.

ion An ion is an electrically charged atom or molecule.

jet energy regression Jet energy regression is a multivariate correction for the energy of b-jets. It is optimised and tested for b-jets that were recorded by the ATLAS detector and reconstructed with the anti-k_T algorithm with a radius parameter of $R = 0.4$ and underwent the ATLAS jet calibration chain.

leading A leading object is the object with the highest transverse momentum in a group of objects, e.g. signal jets..

light jet A light jet is a jet that originated from either an up, down or strange quark or a gluon. Since no experimental discrimination between them is possible (as opposed to a jet that originates from a charm or bottom quark) they are all grouped together..

normalised transverse beam emittance The normalised transverse beam emittance represents the average spread of the beam in the momentum-position phase space transverse to the beam..

out-of-cone leakage Out-of-cone leakage refers to particles of a jet that are not contained in the reconstructed jet's cone and therefore their energy is missing from the total energy of the jet.

parton A parton is a constituent of a hadron which might be either a quark or gluon. Most commonly it is used to refer to the constituents of the proton.

pile-up Pile-up refer to additional activity in the detector that does not originate from the hard scatter event.

relativistic gamma factor The relativistic gamma factor γ_r describes how mass, time and length of an object change if it is moving with a velocity v. It is defined as $\gamma_r = (\sqrt{1-(v/c)^2})^{-1}$, with the speed of light c. If $v << c$ then $\gamma_r \approx 1$. If v becomes close to c γ_r becomes large..

run 1 Run 1 refers to the LHC data taking period from 2010 to 2012.

run 2 Run 2 refers to the LHC data taking period from 2015 to 2018 (expected).

sub-leading A sub-leading object is the object with the second highest transverse momentum in a group of objects, e.g. signal jets..

synchrotron radiation Synchrotron radiation refers to photons that are radiated off from accelerated charged particles that are bent.

true The true value of a quantity represents the value of a quantity before it is affected by detector effects, such as resolution or reconstruction effects.

truth A truth object is a object before it is affected by detector effects, such as resolution or reconstruction effects. Therefore its properties have their true value.

Acronyms

pp proton-proton.

BDTs Boosted Decision Trees.

BR branching ratio.

CERN Conseil Européen pour la Recherche Nucléaire.

CRs control regions.

DR diagram removal.

DS diagram subtraction.

ECAL electro-magnetic calorimeter.

EW electroweak.

HCAL hadron calorimeter.

HLT high level trigger.

L1 level 1 trigger.

LEP Large Electron Positron Collider.

LHC Large Hadron Collider.

LO leading order.

MC Monte Carlo.

ME matrix element.

MVA multivariate analysis.

NLO next-to-leading order.

NNLO next-to-next-to-leading order.

PDF parton distribution functions.

PS parton shower.

PV primary vertex.

QCD quantum chromo dynamics.

SCT semiconductor tracker.

SJC standard ATLAS jet calibration.

SM Standard Model.

SRs signal regions.

SV secondary vertex.

TRT transition radiation tracker.

UE underlying event.

Acknowledgements

First I would like to thank Norbert Wermes who supervised my PhD thesis and allowed me to carry out my work within the research of his group. He already knew me as an undergraduate student and always supported my path to become a researcher. I would also like to thank the other members of my PhD committee: Ian Brock, Bastian Kubis and Sigurd Höger. A special thank you to Eckhardt von Toerne who opened the door into the world of particle physics research for me and was one of my biggest supporters from the start. I would also like to thank the other post-docs in the group that accompanied my PhD studies and were always available for help, comments, discussions, opinions and proof-reading: Tatjana Lenz, Götz Gaycken and Vadim Kostyukin.

An extra big thank you goes to my family, who supported me throughout my life and were always interested in my work (no matter how abstract, specific or complex it is).

I would also like to thank all my friends. David, thank you for great conversations inside and outside the office and your generally positive attitude. David and Irina, thank you for starting the „write-a-thon" - without it my thesis would probably not have been finished in time. Ulli, Tim, Andy, Ben, Jerry and Fynn who I know since the first semester of university: thank you for your friendship and long-lasting support. Thank you to my friends from the Super Collider Around the Moon collaboration. Thank you to Xin for introducing me to the concept of „eternal friends". Last but not least: thank you, to a long list of people who were not only colleagues but shared coffees, drinks and great conversations with me.

I would also like to thank my ATLAS colleagues from all over the world with whom I worked closely during my PhD thesis.

I would like to thank the Studienstiftung des Deutschen Volkes and the Bonn-Cologne Graduate School of Physics and Astronomy for supporting my PhD work.