Search for a heavy charged Higgs boson decaying into a W boson and a 125 GeV Higgs boson, in final states with leptons and jets, with the ATLAS experiment at the LHC

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

vorgelegt von
Shubham Bansal
aus
Agra, Uttar Pradesh

Bonn, 2024

Angefertigt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Rheinischen Friedrich-Wilhelms-Universität Bonn

Gutachter/Betreuer:	Prof. Dr. Jochen Christian Dingfelder
Gutachter:	Prof. Dr. Florian Urs Bernlochner
Tag der Promotion:	10. 04. 2025
Erscheinungsjahr:	2025

Abstract

This thesis presents a first search for a heavy charged Higgs boson produced in association with a top- and a bottom-quark, and decaying into a W boson and a 125 GeV Higgs boson decaying subsequently into a pair of b-quarks. The search is performed in final states with leptons and jets using proton-proton collision data at center-of-mass energy, $\sqrt{s}$ = 13 TeV, recorded with the ATLAS detector during Run 2 of the LHC at CERN. This dataset corresponds to a total integrated luminosity of 140 fb^{-1}. Multivariate analysis techniques are used to reconstruct the charged Higgs boson candidate, and to improve the background rejection. The reconstructed invariant mass distribution of the charged Higgs boson candidate is scrutinised for an excess in data events in the mass range from 250 GeV to 3 TeV. No significant excess is observed, and 95% confidence level upper limits are set between 2.77 pb and 1.18 fb on the production cross section times branching ratio of the charged Higgs boson decays.

Contents

CHAPTER 1

Introduction

The Big Bang model [1] is the most popular cosmological model so far that explains the birth of the universe. It claims the existence of a singular state of high density and high temperature, known as the Big Bang, around 13.8 billion years ago. Following this state, the universe has been expanding in all directions until today. After a time lapse of around 10^{-32} s to 10^{-6} s after the Big Bang, the universe constituted a hot soup of the first elementary particles: the quarks and gluons, and the subsequent clumping of the quarks to form the so-called *nucleons*: protons and neutrons. This event in time corresponds to an energy scale ranging from 0.1 GeV to 100 GeV. Other elementary particles such as electrons, photons, neutrinos etc. were formed spontaneously. The physics presented in this thesis focuses on the *high-energy regime* of this *particle era* (stage of the universe when particles begin to form). One possible way to study physics processes in this regime is to analyse data collected by the Large Hadron Collider (LHC) [2] - the largest particle accelerator in the world. At the LHC, protons or ions are brought into collisions at high centre-of-mass energy, recreating conditions similar to those of the early universe. The properties and the interactions of the elementary particles is described by the Standard Model (SM) of Particle Physics [3]. The SM describes three of the four fundamental interactions in nature: weak, strong, and electromagnetic, with the exception of gravity deduced by Einstein's general relativity [4]. However, the SM being an extremely successful model cannot provide explanations for certain observed phenomena which compells one to believe that the SM is incomplete. One such phenomenon is the presence of almost only matter in the universe, even if antimatter was believed to be originally produced with an equal amount - the so-called baryonic asymmetry. Other phenomena are the origin of dark matter and the non-zero mass of neutrinos.

After the discovery of Higgs boson in 2012 by the ATLAS and CMS experiments at the LHC [5, 6], the last missing piece of the SM was presumably found. This was the most recent development and paved the way for studying a wide range of interesting search channels involving the Higgs boson. Numerous studies have been performed to measure the Higgs boson properties [7, 8] to understand whether it is the Higgs boson predicted in the SM or one of the observed particles of an extended Higgs sector. Searches for an extended Higgs sector are important as large number of models of physics beyond the SM require additional scalar Higgs states. One such common example is the presence of two Higgs doublets as realised in the so-called Two-Higgs-Doublet Model (2HDM) [9]. A common example of the 2HDM is the minimal supersymmetric [10] extension of the SM (MSSM). An extended Higgs sector can help provide solutions to some of the aforementioned problems which SM fails to answer. It can help in modifying the *electroweak phase transition* [11] and facilitate

baryogenesis (creation of baryonic asymmetry in the universe), or provide a dark matter candidate in the context of MSSM.

Several extended Higgs models predicts the existence of at least one set of charged Higgs ($H^{\pm}$) bosons in addition to the discovered neutral one, such as the models that requires a second doublet, or one or more triplets. The main production and decay modes of $H^{\pm}$ bosons are strongly model dependent. For example, in the 2HDM, the dominant production mode, in case the $H^{\pm}$ boson mass is larger than the sum of the top- and bottom-quark masses, is expected to be in association with a top- and a bottom-quark. The dominant $H^{\pm}$ boson decay modes in this scenario are decays to fermions: $H^{\pm} \to tb$[1] and $H^{\pm} \to \tau^{\pm}\nu$. However, there are other models such as the Next-to-two-Higgs-Doublet Model (N2HDM) [12, 13], or the Georgi-Machacek (GM) model [14] in which other decay modes also become dominant. One such decay is the search for $H^{\pm}$ bosons decaying into a $W^{\pm}$ and a SM Higgs (h) boson: $H^{\pm} \to W^{\pm}h$. This decay mode is predicted to have significant branching ratios by various extended Higgs models [15–17]. The ATLAS and CMS collaborations have searched for $H^{\pm}$ bosons using proton-proton collisions at centre-of-mass energies of $\sqrt{s}$ = 7 TeV, 8 TeV, and 13 TeV. These include the direct searches for an $H^{\pm}$ boson decaying into fermions: $H^{\pm} \to \tau^{\pm}\nu$ [18], $H^{\pm} \to cb$ [19, 20], $H^{\pm} \to cs$ [21, 22], $H^{\pm} \to tb$ [23], and also for an $H^{\pm}$ boson decaying into bosons: $H^{\pm} \to W^{\pm}Z$ [24], $H^{\pm} \to W^{\pm}H(\tau^{+}\tau^{-})$ [25] with the mass of neutral Higgs boson (H) being 200 GeV. The search for $H^{\pm} \to W^{\pm}h$ remains uncovered so far at the LHC. This thesis deals with the search for this channel, with the h boson decaying into a pair of b-quarks, $pp \to tbH^{\pm} \to W^{\pm}h(\to b\bar{b})$. The search is performed using proton-proton collision data collected by the ATLAS experiment at $\sqrt{s}$ = 13 TeV. It covers two scenarios to ensure high sensitivity to both low- and high-mass resonances. At low resonance masses, when the final state particles have a relatively low Lorentz boost, the decay products of the h boson and hadronically decaying W boson are reconstructed as individual small-radius jets ("resolved topology"). At high resonance masses, when the final state particles have a relatively large Lorentz boost, the decay products of the h boson and hadronically decaying W boson are reconstructed as single large-radius jet ("merged topology").

The content of this thesis is organised as follows: Chapters 2 and 3 give an introduction to the theoretical and experimental basis of the work presented in this thesis. The analysis presented in this thesis was performed as a part of the *ATLAS Higgs, Multi-Boson, and SUSY searches Working Group* [26], and a publication [27] of this analysis is currently in Journal of High Energy Physics review. The author has been one of the leading contributors to the analysis, and the level of details in Chapters 4 to 7 reflects his specific contribution to the various aspects of the analysis. Chapters 4 and 5 are dedicated to the description of the analysis strategy used in the search for resolved $H^{\pm}$ boson decays. Chapter 6 details the systematic uncertainties and the statistical framework used to derive the sensitivity on the search for resolved $H^{\pm}$ boson decays. Chapter 7 summarises the calibration of a boosted Higgs tagger ($X \to b\bar{b}$ tagger) in $Z \to b\bar{b}$ + jet events [28], performed as a service task to attain authorship within the ATLAS collaboration. This chapter also briefly discusses the search for merged $H^{\pm}$ boson decays, which uses the $X \to b\bar{b}$ tagger as one of the analysis tools. Chapter 8 then presents the results obtained for the $H^{\pm} \to W^{\pm}h(\to b\bar{b})$ search. Finally, Chapter 9 concludes this thesis with a summary and an outlook for future work.

[1] An "overline" for an antiparticle is implied in the final state of the $H^{\pm}$ boson decays.

CHAPTER 2

Theoretical Framework

This chapter provides an introduction to the theoretical aspects of high-energy physics, and their relevance in the context of this thesis. It begins with Section 2.1 giving an overview of the SM, its particle content, and a description of the various forces between the particles constituting the SM. Section 2.2 introduces the Higgs mechanism responsible for the generation of particle masses, and its related physics. Following this, some of the open questions associated with the SM are detailed in Section 2.3. Finally Section 2.4 highlights various models in light of an extended Higgs sector as a potential solution to the presented concerns. The description of these models also accomodate the relevance of $H^{\pm}$ boson decay into a SM Higgs boson and a $W^{\pm}$ boson, which is the subject of the search presented in this thesis.

2.1 The Standard Model

The SM is a relativistic quantum field theory that describes all the known particles and their interactions, the electromagnetic, the weak, and the strong with the exception of gravitational force. However, the gravitational force between two interacting particles is extremely small in comparison to other forces, at energies where the SM is tested experimentally, and hence can be neglected. The particle content of the SM is illustrated in Figure 2.1 which presents all the elementary particles. They are categorised into two groups based on their spin, the half-integer-spin fermions and the integer-spin bosons. The interactions between the fermions are mediated by the bosons.

Fermions are often termed as the building blocks of matter, and are classified into two categories based on whether they interact strongly. The first class of particles known as quarks carry a colour charge and hence take part in the strong interactions. Colour charge can take three different values, commonly denoted as *red*, *blue*, and *green*. The second class of particles are known as leptons which do not interact strongly and therefore carry no colour charge. Each class is known to have their corresponding anti-particle with the same mass and spin but opposite quantum numbers. In total there are six quarks and six lepton types grouped in three generations. The first generation consists of the two lightest leptons and the two lightest quarks forming the ordinary matter. The remaining two generations are the replicas of the first but with particles of increasing mass.

Leptons are further separated into charged leptons and neutral leptons. Charged leptons carry an electric charge of -1: electrons (e^-), muons (μ^-), and tauons (τ^-) which interact via both weak and

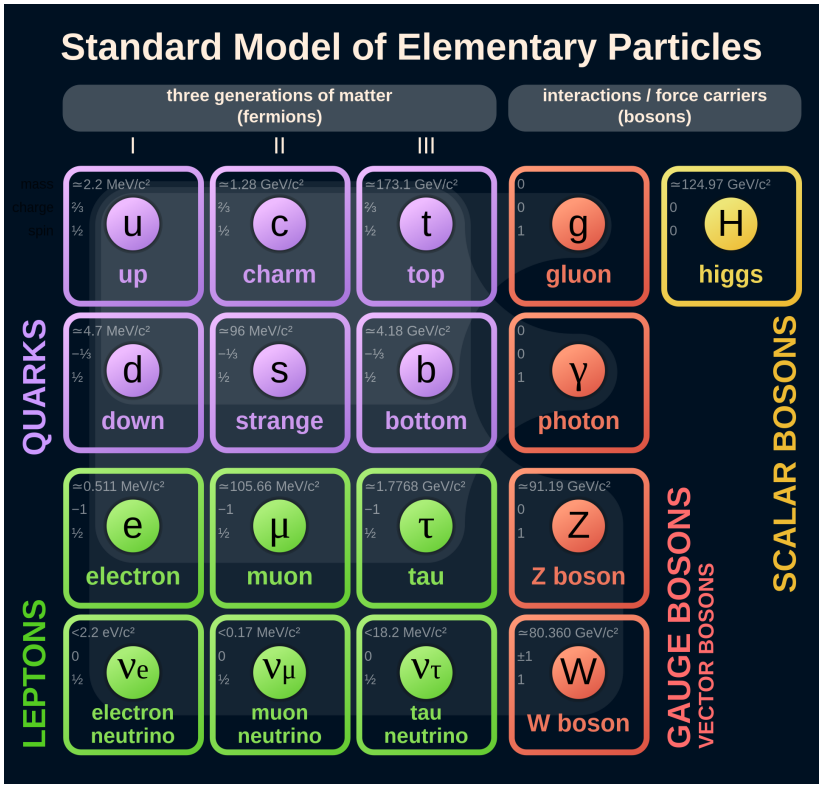


Figure 2.1: Schematic representation of the particles in the SM. The mass, spin, and charge are shown for each particle [29].

electromagnetic interactions. Neutral leptons or the neutrinos include an electron neutrino (ν_e), a muonic neutrino (ν_μ), and a tau neutrino (ν_τ) which interact only weakly. Each lepton generation comprises of a pair of a charged lepton and a neutrino: e^- - ν_e, μ^- - ν_μ, and τ^- - ν_τ. The quarks are also distributed in a similar way as the leptons, each generation contains a pair of particles with electric charge[1] of $+\frac{2}{3}$ and $-\frac{1}{3}$, respectively. The first generation is formed by the up (u) and down (d) quarks, while the remaining quarks including the charm (c) and strange (s) forming the second generation, and the top (t) and bottom (b) forming the third generation. Specifically, the b-quark having a mass of $4.18^{+0.03}_{-0.02}$ GeV [30] is the quark of interest for this thesis. The leptons can be observed freely in nature, but the quarks are only found in bound states e.g. the protons (uud) and neutrons (udd). They can be grouped together into states of quark and an antiquark called mesons or states of three quarks called baryons, collectively known as hadrons.

The second type of fundamental particles are the bosons. In the SM, there are in total twelve spin-1 bosons mediating the three fundamental interactions: one photon, γ, which mediates the electromagnetic interactions, three bosons W^+, W^-, and Z which mediates the weak interactions, and eight gluons which mediates the strong interactions. In addition to these, there exists a spin-0 particle or a scalar, the Higgs boson. This particle is responsible for all the SM particles, except neutrinos, attaining mass through the Higgs mechanism described further in Section 2.2.

[1] Throughout this thesis, the electric charges are expressed in units of the electron electric charge (e).

The symmetry of a physical system is its intrinsic feature that remains unchanged under some imposed transformation. The most natural ones which SM respects primarily are the *discrete symmetries*: the charge conjugation (C-symmetry), parity (P-symmetry), and time reversal (T-symmetry). Another important class of symmetries which play an important role in particle physics are the *gauge symmetries*. They are mathematically described using group theory [31] where one studies the transformations under which an object is symmetric. These symmetries facilitate a mathematically consistent and simple formulation of theories, and eventually predicts the particle that are exchanged - the gauge particles. The SM can be mathematically expressed within the Lagrangian[2] framework of the particle fields, and is based on the symmetry group defined as [32]:

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

where $SU(3)$ is the group governing the strong interaction, $SU(2)$ governs the weak interaction, and $U(1)$ describes the electromagnetic interaction. The SM Lagrangian is local gauge invariant under the symmetry group defined by Equation 2.1. The concept of local gauge invariance dictates that the dynamics of the system under consideration does not change under space-time-dependent transformations. Sections 2.1.1, 2.1.2, and 2.1.3 discuss how the theory of these fundamental interactions between the elementary particles arise as a result of local gauge invariance. The description and the mathematical formalism of these interactions is based on [32–34].

2.1.1 Quantum Electrodynamics

The Lagrangian of a free spin-$\frac{1}{2}$ particle is postulated on the basis of the Dirac [35] equation[3]:

$$\mathcal{L}_{\text{Dirac}} = \bar{\psi}(i\gamma^{\mu}\partial_{\mu} - m)\psi, \tag{2.2}$$

where m and ψ are the mass and field variable of the spin-$\frac{1}{2}$ particle, respectively, γ^{μ} are the Dirac matrices, and $\bar{\psi}$ is the field adjoint in the Dirac's representation. Under the $U(1)$ local gauge transformation, the Lagrangian, $\mathcal{L}_{\text{Dirac}}$ is found to be not invariant and contains an additional phase term $e\bar{\psi}\gamma^{\mu}\partial_{\mu}\chi(x)\psi$ where $\chi(x)$ and e represent the local phase and the elementary charge of the spin-$\frac{1}{2}$ particle, respectively.

$$\mathcal{L}_{\text{Dirac}} \rightarrow \mathcal{L}'_{\text{Dirac}} = \mathcal{L}_{\text{Dirac}} - e\bar{\psi}\gamma^{\mu}\partial_{\mu}\chi(x)\psi. \tag{2.3}$$

In order to restore local gauge invariance, one can replace the derivative ∂_{μ} with the so-called covariant derivative, D_{μ}:

$$\partial_{\mu} \rightarrow D_{\mu} = \partial_{\mu} + ieA_{\mu}, \tag{2.4}$$

where A_{μ} is a newly introduced field which transforms as:

$$A_{\mu} \rightarrow A'_{\mu} = A_{\mu} - \partial_{\mu}\chi. \tag{2.5}$$

This field A_{μ} provides the cancellation of the phase term, and is hence called the gauge field of

[2] A Lagrangian of a system is defined as the combination of its kinetic (T) and potential (V) energy : $L \equiv T - V$.

[3] Throughout this thesis, natural units are used, i.e. the speed of light and the reduced Planck's constant are set to unity ($c = \hbar = 1$). This chapter uses an Einstein's summation convention which sums over repeated indices in a mathematical expression as shown in Equation 2.2 where $\mu = 0, 1, 2, 3$ is the twice-repeated index, and ∂_{μ} is the four-gradient.

electromagnetic interactions, or more commonly the field of the mediator gauge particle - the photon. The associated Quantum Electrodynamics (QED) Lagrangian which describes the fields for the spin-$\frac{1}{2}$ particles, the photon and the interactions between them can therefore be written as:

$$\mathcal{L}_{\text{QED}} = \bar{\psi}(i\gamma^{\mu}D_{\mu} - m)\psi - \frac{1}{4}F_{\mu\nu}F^{\mu\nu}, \tag{2.6}$$

where $F_{\mu\nu} = \partial_{\mu}A_{\nu} - \partial_{\nu}A_{\mu}$, and $F_{\mu\nu}F^{\mu\nu}$ is the kinetic energy of the photon.

2.1.2 Electroweak Interactions

The aforementioned concept of gauge invariance can be generalised to cover both the electromagnetic and weak interactions collectively called the Electroweak (EW) interactions respecting the $SU(2)\otimes U(1)$ symmetry group. The unification of the two interactions was originally postulated by the *Glashow-Salam-Weinberg (GSW) model* [36, 37] which allows the mixing of the gauge boson states of the $SU(2)$ and $U(1)$ groups. In this scenario, the field A_{μ} is replaced by the four fields $W_{\mu}^{1,2,3}$ and B_{μ}. The covariant derivate defined previously, in the case of QED, now becomes:

$$D_{\mu} = \partial_{\mu} - ig\frac{\tau_a}{2}W_{\mu}^{a} - ig'\frac{Y}{2}B_{\mu},\ a = 1, 2, 3, \tag{2.7}$$

where g and τ_a are the coupling constant and group generators of the $SU(2)$ group, respectively, whereas g' and Y are the coupling constant and group generators of the $U(1)$ group[4], respectively. The EW Lagrangian is then modified as:

$$\mathcal{L}_{\text{EW}} = i\bar{\psi}\gamma^{\mu}D_{\mu}\psi - \frac{1}{4}W_{a\mu\nu}W_{a}^{\mu\nu} - \frac{1}{4}B_{\mu\nu}B^{\mu\nu}. \tag{2.8}$$

The four electroweak fields mix in the following way, resulting in the formation of the gauge particles: $W^{\pm}$, Z, and γ bosons.

$$W_{\mu}^{\pm} = \frac{1}{\sqrt{2}}(W_{\mu}^{1} \mp iW_{\mu}^{2}), \tag{2.9}$$

$$Z_{\mu} = \frac{gW_{\mu}^{3} - g'B_{\mu}}{\sqrt{g^2 + g'^2}}, \tag{2.10}$$

$$A_{\mu} = \frac{g'W_{\mu}^{3} + gB_{\mu}}{\sqrt{g^2 + g'^2}}. \tag{2.11}$$

In the GSW model, Z and γ bosons can be written as a linear combination of W_{μ}^{3} and B_{μ}.

$$\begin{pmatrix} Z_{\mu} \\ A_{\mu} \end{pmatrix} = \begin{pmatrix} \cos\theta_w & -\sin\theta_w \\ \sin\theta_w & \cos\theta_w \end{pmatrix} \begin{pmatrix} W_{\mu}^{3} \\ B_{\mu} \end{pmatrix}, \tag{2.12}$$

[4] In the context of GSW model, Y is the weak hypercharge similar to the electric charge in the case of QED.

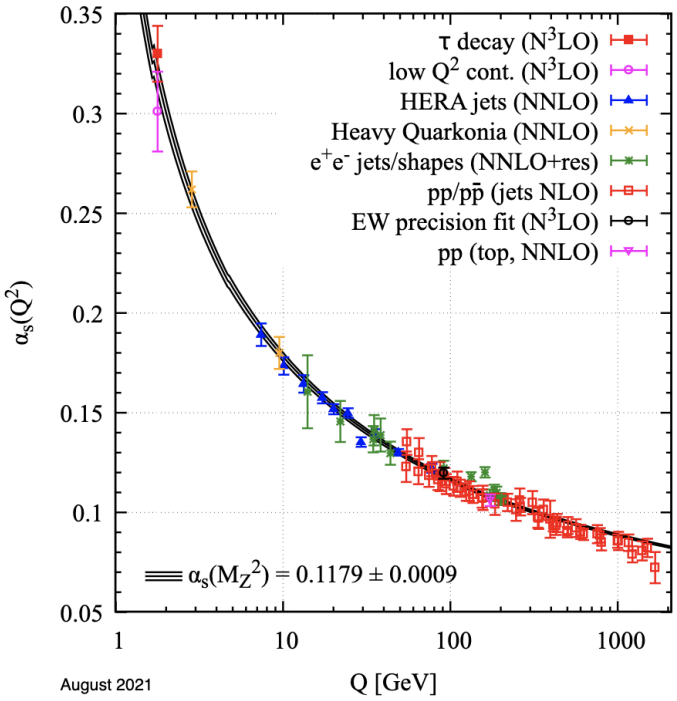


Figure 2.2: Strong coupling α_s as a function of momentum transfer, Q, summarised for various measurements [39].

where θ_w is the weak mixing angle defined as:

$$\cos\theta_w = \frac{g}{\sqrt{g^2 + g'^2}}. \tag{2.13}$$

2.1.3 Quantum Chromodynamics

The strong interactions are included in the SM through Quantum Chromodynamics (QCD) which are governed by the $SU(3)$ gauge group, and mediated by the gluon which carries a colour charge. Gluons lie in the adjoint representation of $SU(3)$ with eight[5] independent gluon fields, G_μ. The covariant derivative for QCD is expressed as shown in Equation 2.14 where λ_a are the generator *Gell-Mann* matrices [38], and α_s is the strong coupling constant.

$$D_\mu = \partial_\mu - i\alpha_s \frac{\lambda_a}{2} G^a_\mu, \; a = 1 \text{ to } 8. \tag{2.14}$$

The QCD Lagrangian can be expressed as:

$$\mathcal{L}_{\text{QCD}} = \bar{\psi_C}(i\gamma^\mu \partial_\mu - m)\psi_C - \frac{1}{4}\text{Tr}[G_{\mu\nu}G^{\mu\nu}], \tag{2.15}$$

where m is the the mass of the quark field ψ_C.

[5] For an $SU(N)$ group, the number of generators are N^2-1 such that there are eight gluon fields in QCD.

Given that the SM is a *renormalisable* theory, the QCD Lagrangian acquires a mass scale μ during the renormalisation procedure [40]. This mass scale, μ, is arbitarily chosen, and is the momentum at which any energy divergences for a strongly interacting process are subtracted. The strength of a strong interaction is parameterised by α_s, which is not a constant but varies with the momentum scale, Q. The dependence of α_s as a function of Q^2 is given as [34]:

$$\alpha_s(Q^2) = \frac{\alpha_s(\mu^2)}{1 + A\alpha_s(\mu^2)\ln\frac{Q^2}{\mu^2}}, \tag{2.16}$$

where A is defined as:

$$A = \frac{11N_C - 2N_F}{12\pi}. \tag{2.17}$$

In Equation 2.17, N_C is the number of colours and N_F is the number of quark flavours available at a scale Q^2. Figure 2.2 shows that α_s decreases with increasing momentum transfer leading to the *asymptotic freedom* of quarks. The gluons also self interact due to which there are both bosonic and fermionic loops which affect the development of α_s at higher energy scales. The closer the two quarks get, the less they will be influenced by each other via the strong force, but if the two quarks move away from each other, the potential between them increases linearly, leading to *colour confinement*. This process is important for the understanding of hadronisation i.e. when the energy in the gluon field is high enough, another pair of quark and anti-quark can be produced, and the observed bound states of these produced particles can only be colourless, known as hadrons.

2.2 The Higgs mechanism

The model presented in Section 2.1 explains the kinetic properties and the interactions of the fundamental fields, although it does not provide a way to generate boson and fermion masses. Constructing the SM on symmetries requires all its particles to be massless, as adding a mass term, for instance of the form $\frac{1}{2}m^2W^\mu W_\mu$ for gauge bosons would break the local gauge invariance. On the other hand, the electroweak gauge bosons are massive. The mass generation mechanism of the electroweak gauge bosons and of the fermions can be realised through the *Brout-Englert-Higgs mechanism* [41, 42], also commonly called the Higgs mechanism. This mechanism adds a scalar Higgs term to the SM Lagrangian as:

$$\mathcal{L}_{\text{Higgs}} = (D^\mu\phi)^\dagger(D_\mu\phi) - V(\phi), \tag{2.18}$$

where ϕ is a complex scalar field (a $SU(2)$ doublet), and D_μ is the covariant derivative of a complex scalar field as defined earlier in the context of electroweak interactions (see Equation 2.7). The first term of $\mathcal{L}_{\text{Higgs}}$ corresponds to the kinetic term, and describes the interaction of the scalar field with the gauge fields. The second term defines the Higgs field potential, $V(\phi)$, as shown in Equation 2.19, where $\lambda > 0$ and $\mu^2 < 0$.

$$V(\phi) = \mu^2\phi^\dagger\phi + \lambda(\phi^\dagger\phi)^2 \tag{2.19}$$

The Higgs field potential has a shape like a mexican hat as illustrated in Figure 2.3. It is observed from this figure that the vaccum expectation value (vev)[6], v, of the Higgs potential is not at $\phi^\dagger\phi = 0$,

[6] In quantum field theory, the vaccum expectation value of an operator is its average value in the vacuum.

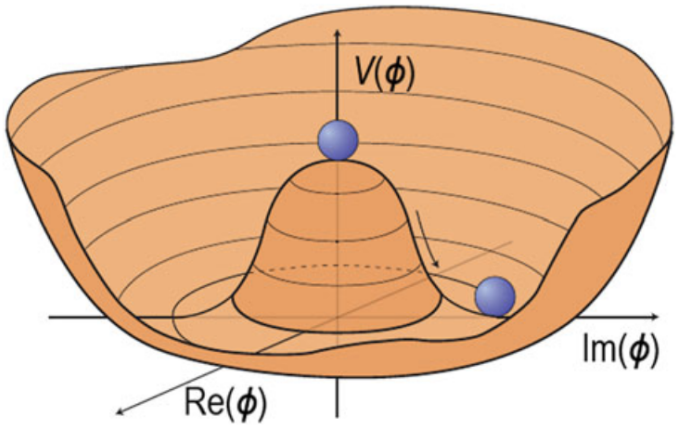


Figure 2.3: A schematic of the Higgs potential [43].

but is located at $\sqrt{-\mu^2/\lambda}$. Given the shape of the potential, there are an infinite number of degenerate states that have the minimum energy, provided one breaks the electroweak symmetry spontaneously, commonly known as the electroweak symmetry breaking (EWSB) [42].

As an initial value to mathematically formulate the mechanism, one can arbitarily choose the ground state of ϕ, ϕ_0 as:

$$\phi_0 = \frac{1}{\sqrt{2}}\begin{pmatrix}0\\v\end{pmatrix}. \tag{2.20}$$

The scalar field is then expanded around this ground state to obtain the Higgs doublet, $h(x)$, as following:

$$\phi(x) = \frac{1}{\sqrt{2}}\begin{pmatrix}0\\v+h(x)\end{pmatrix}. \tag{2.21}$$

On substituting $\phi(x)$ in the Higgs Lagrangian (Equation 2.18), the kinetic term is now modified to contain the mass terms for the gauge bosons:

$$(D^\mu\phi)^\dagger(D_\mu\phi) = \frac{1}{2}(\partial_\mu h)(\partial^\mu h) + \frac{g^2}{4}(v+h)^2 W^+_\mu W^{\mu-} + \frac{1}{8}(g^2+g'^2)(v+h)^2 Z_\mu Z^\mu. \tag{2.22}$$

The corresponding potential term describing the Higgs boson self-interaction becomes:

$$V(\phi) = \frac{\mu^2}{2}(v+h)^2 + \frac{\lambda}{4}(v+h)^4. \tag{2.23}$$

The masses of $W^\pm$ and Z bosons, m_W and m_Z, respectively, are hence directly derived from the kinetic term:

$$m_W = \frac{1}{2}gv, \tag{2.24}$$

$$m_Z = \frac{1}{2}(g^2+g'^2)v. \tag{2.25}$$

The $W^{\pm}$ and Z bosons, with a mass of 80.37 ± 0.012 GeV and 91.18 ± 0.002 GeV [30], respectively, are the heavy weights of the elementary particles. Since they are massive bosons, they have a very short lifetime, such that they are not directly observed but only their decay products can be measured. The $W^{\pm}$ boson can decay into a lepton and the corresponding anti-neutrino, or to a quark and an anti-quark pair of complementary types. On the other hand, the Z boson can decay into a fermion and an anti-fermion pair. Decays of both the $W^{\pm}$ and Z bosons to top-quark(s) is forbidden by the energy conservation.

The mass of the Higgs boson, m_h, is obtained by expanding the potential term:

$$m_h = \sqrt{2\lambda v^2}. \tag{2.26}$$

The Higgs field has a vev, $v \approx 246$ GeV, and allows one to generate the masses of the fermions in a gauge-invariant way. This is done by expressing the fermion masses, m_f, in a Yukawa Lagrangian defined as:

$$\mathcal{L}_y = -y_d \bar{Q}_L \phi d_R - y_u \bar{Q}_L \phi u_R - y_l \bar{L}_L \phi l_R + \text{h.c.}, \tag{2.27}$$

where $\bar{Q}_L = (\bar{u}_L, \bar{d}_L)$ and $\bar{L}_L = (\bar{\nu}_L, \bar{l}_L)$ are the quark and lepton left-handed doublets, respectively, u_R, d_R, and l_R are the corresponding right-handed fermion singlets, and h.c. stands for the hermitian conjugate of the scalar field. The masses of the fermions can hence be expressed as:

$$m_f = y_f \frac{v}{\sqrt{2}}, \tag{2.28}$$

where y_f is the Yukawa coupling which is the free parameter of the SM, and must be measured experimentally.

The current most precise mass measurement of the SM Higgs boson is 125.10 ± 0.14 GeV [30]. The main production modes at the LHC are shown in Figure 2.4. These are: gluon-gluon fusion (ggF) (a), in which the Higgs boson is created via a top-quark loop, vector boson fusion (VBF) (b), where the Higgs boson is created in association with two forward quark jets, vector boson associated production (c) which is relatively rare compared to the previous production modes, and the rarest top-quark associated production (d) mode which probes the coupling among the two heaviest known particles in the SM.

The SM Higgs boson can decay into both fermions and bosons. The decays into $b\bar{b}$ [44], $\tau\tau$ [45], $c\bar{c}$ [46], and into ZZ, $\gamma\gamma$, WW [47, 48] have been observed experimentally in ATLAS. Figure 2.5 shows the different branching ratios for a SM Higgs boson decaying into SM particles. The highest branching ratio of 58% is observed for the decay of Higgs boson into two b-quarks [44], which is also the decay mode of interest for this thesis.

2.3 Limitations of the SM

The SM has proven to be incredibly successful at describing the physics of subatomic particles. Some of the instances include the experimental measurement of the electron magnetic dipole moment with a precision of 1 part per 100 billion [50], another being the celebrated discovery of the Higgs boson at the LHC in 2012. While these and the other successes of the SM are a boon to the particle physics community, it is known that there exists observed phenomena that the SM can't explain. This compells

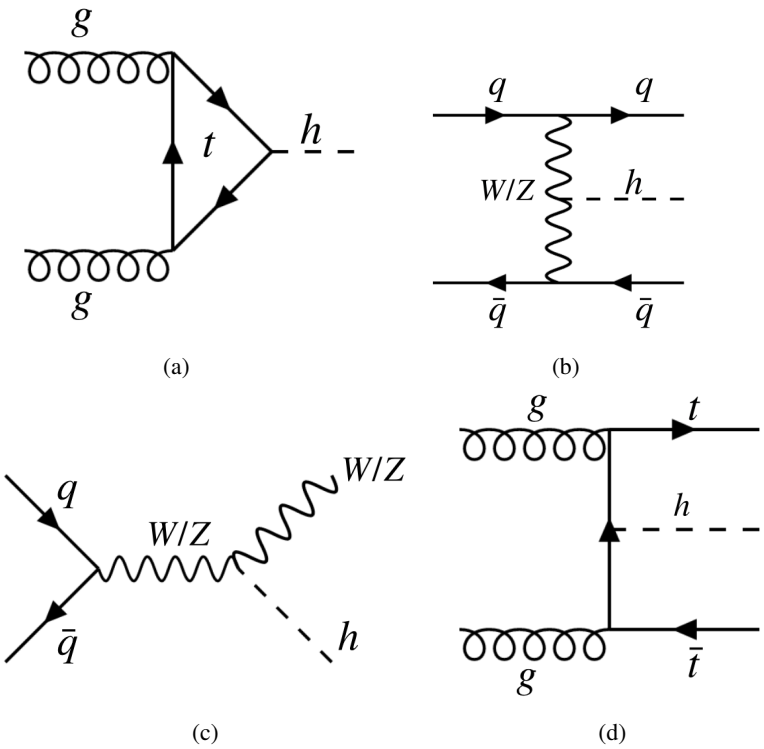


Figure 2.4: Leading order Feynman diagrams for SM Higgs boson production at the LHC. (a) gluon-gluon fusion, (b) vector boson fusion, (c) vector boson associated production, and (d) top-quark associated production [30].

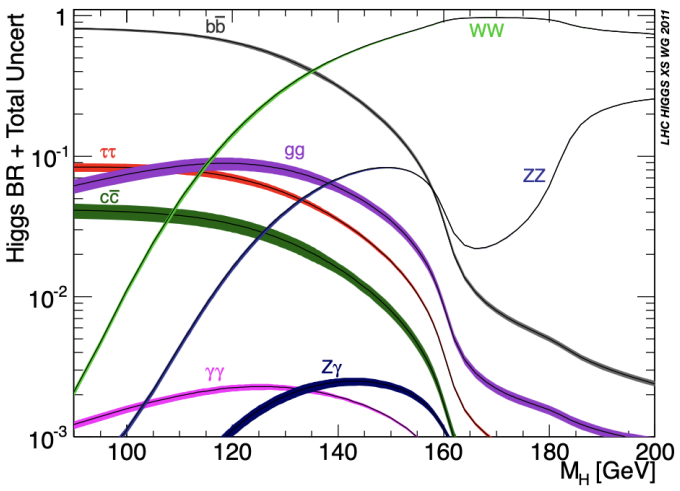


Figure 2.5: This SM Higgs boson branching ratios and the associated uncertainties [49].

us to believe that the SM is not the ultimate theory of fundamental particles and interactions. This section introduces some of these phenomena illustrating the limitations of the SM. These discussed limitations are not complete and represent only a selected set of problems.

2.3.1 Matter-antimatter asymmetry

An important physical observation that remains unexplained by the SM is the asymmetry between matter and anti-matter in the universe, the so-called baryonic asymmetry. In the Big Bang, equal amounts of matter and anti-matter should have been produced, although the present day universe comprises almost entirely of matter. The Sakharov conditions [51] postulate that a baryon-generating mechanism that produces the two matter states in different rates should violate the baryon number, the charge symmetry, and the CP symmetry, and should occur out of thermal equilibirium. However, the fact that CP violation in the SM which occur mainly in the quark sector [52], and to some extent in the neutrino sector is insufficient to account for the observed matter-antimatter asymmetry. Moreover, all existing experimental evidence so far indicates that baryon number is a conserved quantum number in fundamental particle interactions.

2.3.2 Neutrino oscillations

As mentioned in Section 2.1, neutrinos are neutral leptons categorised in three generations known as the three *flavours* of neutrinos: ν_e, ν_μ, and ν_τ. On the other hand, there exist three *mass states* of neutrinos: m_1, m_2, and m_3. The flavour and the mass states are not independent, for instance, ν_e is a superposition of the three mass states, referred to as *neutrino mixing*. The mass states of neutrinos propagate through space as waves with different frequencies. As the neutrinos propagates through space, the phase of the mass states advance at different rates due to the difference in the neutrino masses. This results in a changing mixture of the mass states and subsequently the change in the initial flavour state. For instance, a neutrino produced as a ν_e can *oscillate* to a ν_μ. However, since the quantum mechanical phase advances in a periodic fashion, after some distance ν_μ will oscillate back to ν_e. This phenomenon of *neutrino oscillations* was first observed in 1998 by the Super-Kamiokande experiment [53]. The neutrino oscillations proved that the neutrinos are massive, however small the mass may be. This in contrast to the SM which thought the neutrinos to be massless. Some of the remaining open questions in this context are the order of neutrino masses: $m_{\nu_e} < m_{\nu_\mu} < m_{\nu_\tau}$ and their respective values.

2.3.3 Dark matter

Various experimental and theoretical predictions [54, 55] point towards the existence of a type of matter that shows no signs of electromagnetic interactions, and is responsible for around 27% of the energy in the universe. One of the potential candidates of dark matter are neutrinos, although due to their low mass they do not account for the currently estimated dark matter component in the universe. Some models as mentioned in [56] hint towards the possibility of dark matter being purely from an astronomical origin. However, there are beyond SM theories which hypothesise the existence of additional neutral particles like *heavy sterile neutrinos* [57], *axions* [54], and *weakly interactive massive particles (WIMPs)* [58] as potential dark matter candidates.

2.3.4 The hierarchy problem

The calculation of the Higgs boson mass in the SM infers that it can receive contributions from all energy scales. The most naive choice of the highest energy scale at which the SM is valid is the Planck mass $\approx 10^{19}$ GeV. This difference in scale of O (10^{17}) GeV between the SM Higgs boson mass and the Planck mass give rise to the hierarchy problem [59, 60]. This scale difference leads to large corrections to the bare Higgs boson mass which would make the observed Higgs boson mass huge. To obtain the experimentally observed value of 125 GeV, one would require extreme tuning in the higher-scale theory (new physics theory) to cancel out the correction terms, leading to un-naturalness of the Higgs boson mass.

2.4 Extended Higgs Sector

Extended Higgs sectors are promising features of beyond SM (BSM) theories aiming to explain the aforementioned shortcomings of the SM. These models can contain new sources of CP violation which is one of the required ingredients to potentially explain the matter-antimatter asymmetry. This can occur in electroweak baryogenesis scenarios [61, 62]. Also, under an assumption of Supersymmetry (SUSY) [63], a supersymmetric partner of a particle would cancel the dominant particle's contribution to the Higgs boson mass correction discussed in the hierarchy problem. Supersymmetric models also predict an extended Higgs sector with at least three neutral Higgs bosons and two charged Higgs bosons. The extended Higgs sector can also be embedded in a larger theoretical scenarios such as dedicated neutrino models [64] and *Axion-like particles (ALPs)* [65]. This section describes the Two-Higgs-Doublet Model (2HDM) - the most widely used extension of the Higgs sector - and briefly summarises other exotic models in light of the direct $H^{\pm}$ boson searches with focus on the $H^{\pm} \rightarrow W^{\pm}h$ search presented in this thesis.

2.4.1 The Two-Higgs-Doublet Model

The 2HDM is the simplest way to extend the Higgs sector by adding a second Higgs doublet. The Higgs potential in 2HDM is given as [9]:

$$V(\phi_1,\phi_2) = m_{11}^2\phi_1^\dagger\phi_1 + m_{22}^2\phi_2^\dagger\phi_2 - [m_{12}^2\phi_1^\dagger\phi_2 + \text{h.c.}] + \frac{1}{2}\lambda_1(\phi_1^\dagger\phi_1)^2 + \frac{1}{2}\lambda_2(\phi_2^\dagger\phi_2)^2 + \\ + \lambda_3(\phi_1^\dagger\phi_1)(\phi_2^\dagger\phi_2) + \lambda_4(\phi_1^\dagger\phi_2)(\phi_2^\dagger\phi_1) + \frac{1}{2}[\lambda_5(\phi_1^\dagger\phi_2)^2 + \text{h.c.}] \quad (2.29)$$

where ϕ_1 and ϕ_2 are the scalar fields corresponding to the first and the second doublet, m_{11}^2, m_{22}^2, and m_{12}^2 are the quadratic parameters (mass parameters), and λ_1,....λ_5 are the five quartic parameters.

The scalar fields acquires non zero vevs, namely v_1 and v_2. Their expansion around the vacuum state give rise to five fields manifesting themselves as five physical particles ("five Higgs bosons"). These particles include two CP-even neutral Higgs bosons, h, H ($m_H > m_h$), one CP-odd Higgs boson, A, and two charged Higgs bosons, $H^{\pm}$. The 2HDM is described by six physical parameters:

- four Higgs boson masses (m_h, m_H, m_A, $m_{H^{\pm}}$),
- $\tan(\beta) = \frac{v_2}{v_1}$ (ratio of the two vevs),

Table 2.1: The four different 2HDM types defined based on the allowed couplings of ϕ_1 and ϕ_2 to the quarks and charged leptons. u_i and d_i are the up and down type quarks, respectively, l_i denotes the charged leptons. i refers to the generation index.

	Type I	Type II	Lepton-specific	Flipped
ϕ_1		d_i, l_i	l_i	d_i
ϕ_2	u_i, d_i, l_i	u_i	u_i, d_i	u_i, l_i

- mixing angle, α, which denotes the mixing between the two CP-even Higgs bosons.

The squared sum of the vevs are fixed to the SM value of 246 GeV.

$$v = \sqrt{v_1^2 + v_2^2} \tag{2.30}$$

From first principles, none of the 2 CP-even Higgs bosons in the 2HDM behaves as an SM-like Higgs boson (h_{SM}). In order to realise this, the model parameter space has to fall in the so-called alignment limit under which h or H transforms to an SM-like Higgs boson:

$$\lim_{\sin(\beta-\alpha)\to 1} h \to h_{\text{SM}} \tag{2.31}$$

$$\lim_{\cos(\beta-\alpha)\to 1} H \to h_{\text{SM}} \tag{2.32}$$

For a 2HDM without including *Flavour-Changing-Neutral-Currents (FCNCs)* which are the hypothetical interactions that change the flavour of a fermion without changing its electric charge, there are four possible 2HDM types distinguished based on whether the two doublets couple to the u-quark, d-quark, and leptons. These are namely the Type-I, Type-II, Lepton-specific, and Flipped scenarios as described in Table 2.1 with a model assumption that the second doublet always couple to the u-quark. The type-I 2HDM scenario is known as *fermiophobic* where the charged fermions couple only to the second doublet. In the type-II 2HDM scenario, the u- and d-quarks couple to separate doublets. The type-II 2HDM scenario corresponds to the Higgs sector of the *Minimal Supersymmetric Standard Model (MSSM)* [66, 67], and is a relevant mode in the context of the search presented in thesis.

The production and decay modes of the $H^\pm$ boson depend on its mass ($m_{H^\pm}$), $\tan(\beta)$, and mixing angle, α. For $H^\pm$ boson masses below the top-quark mass ($m_{H^\pm} < m_t$), the leading production mode is realised through the decay of a top-quark ($t \to bH^\pm$) in a double-resonant top-quark production as shown in Figure 2.6 (a). In this mass regime, the decay process $H^\pm \to \tau\nu$, $H^\pm \to cb$, and $H^\pm \to cs$ shows sizeable branching ratios[7]. For $H^\pm$ boson masses above the top-quark mass ($m_{H^\pm} > m_t$), the main production mode is $pp \to tbH^\pm$ through a single-resonant top-quark production as shown in Figure 2.6 (b). For this production mode, the dominant decay is $H^\pm \to tb$ and $H^\pm \to \tau\nu$ in a region close to the alignment limit ($\sin(\beta-\alpha) \to 1$) and $H^\pm \to W^\pm h$ becomes sizeable in the region excluded by the alignment limit ($\cos(\beta-\alpha) \to 1$). In the intermediate-mass regime ($m_{H^\pm} \approx m_t$), the non-resonant top-quark production (as shown in Figure 2.6 (c)) seem to be theoretically viable, as described in [68].

[7] $H^\pm \to cb$ and $H^\pm \to cs$ processes are relatively favoured at low $\tan(\beta)$.

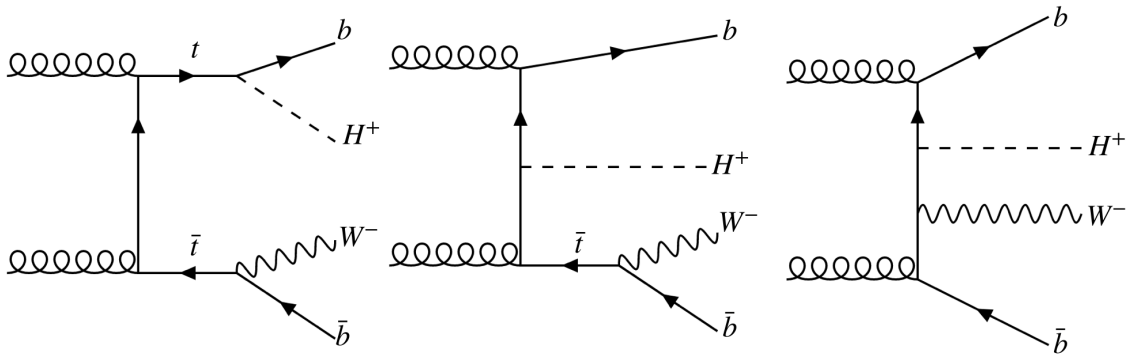


Figure 2.6: The leading-order Feynman diagrams for the production of a singly charged Higgs boson ($H^{\pm}$) in pp collision: (a) double-resonant top-quark production dominant at low $m_{H^{\pm}}$, (b) single-resonant top-quark production dominant at large $m_{H^{\pm}}$, and (c) non-resonant top-quark production preferred at intermediate $m_{H^{\pm}}$.

Figure 2.7 shows the variation of the branching ratio of the $H^{\pm} \to W^{\pm}h$ decay [69], for different $H^{\pm}$ boson masses, as a function of $\cos(\beta - \alpha)$ and for different $\tan\beta$ values, in the type-II 2HDM scenario. The branching ratio reaches a value above 50% for $\cos(\beta - \alpha) \geq 0.05$ (in the region away from the alignment limit) and for $\tan\beta > 25$. It can also be observed that the branching ratio increases with the $H^{\pm}$ boson mass. However, there are other exotic models where the $H^{\pm} \to W^{\pm}h$ decay is predicted to be more relevant than the 2HDM. These models are described in the following section.

CP violation in the 2HDM arises from the scalar interactions through bosonic charge and parity symmetry breaking [70], which is in contrast to CP violation in the SM where both charge and parity symmetry breaking is associated with the fermions. The 2HDM is classified as CP-conserving or CP-violating based on the nature of the quadratic and quartic parameters described in Equation 2.29. For a CP-conserving 2HDM, all the quadratic and quartic parameters are real, whereas for a CP-violating 2HDM, all parameters, except m_{12}^2 and λ_5, are real. Moreover, in a CP-violating case the limit (Equation 2.31) can never hold true since h would not be a pure CP-even state. However, since the allowed CP violating mixings are small, it is still valid to define an alignment limit for a CP-violating 2HDM, at least to a good approximation [71].

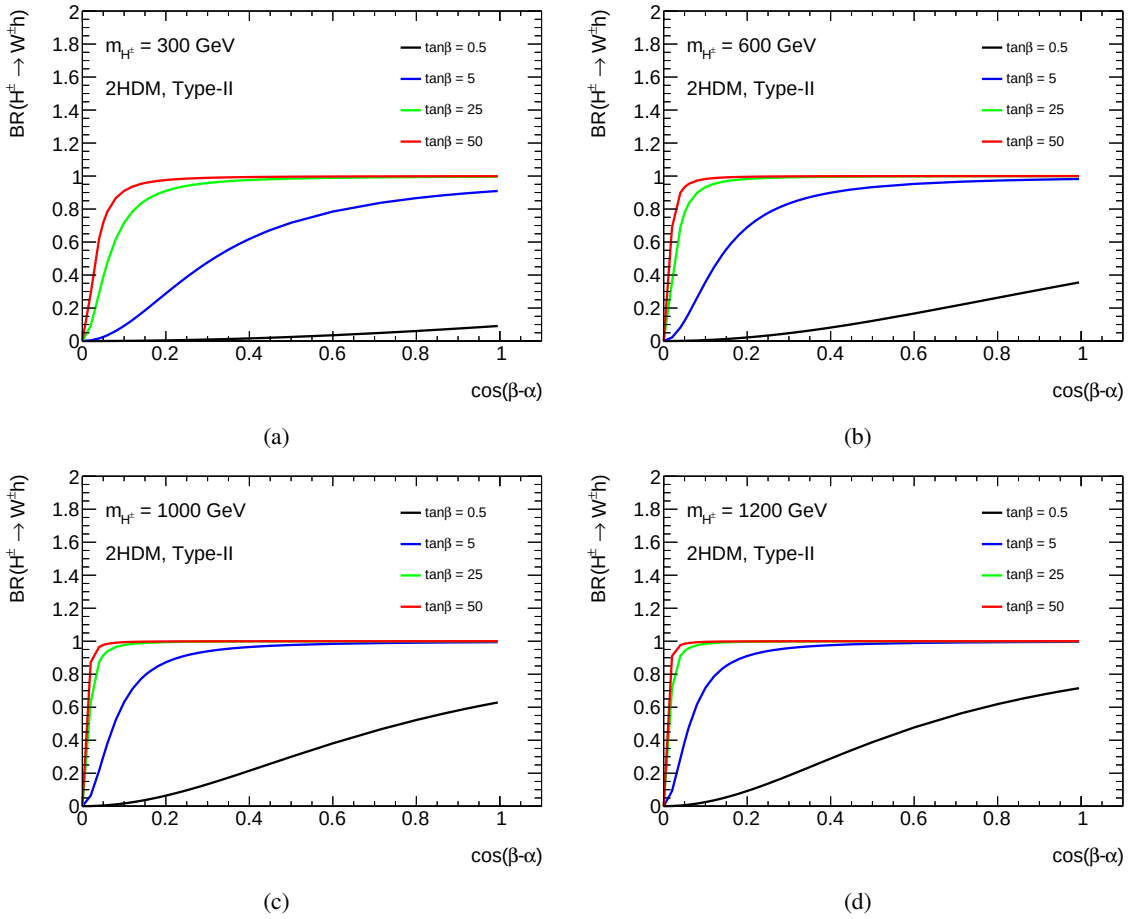


Figure 2.7: (a)-(d) shows the variation of the branching ratio of $H^{\pm} \to W^{\pm}h$, for different $H^{\pm}$ boson mass values, with $\cos(\beta - \alpha)$ and for different values of $\tan\beta$ in the type-II 2HDM scenario.

2.4.2 Other exotic models

The Next-to-Two-Higgs-Doublet Model (N2HDM)

The N2HDM extends the CP-conserving 2HDM by a real scalar singlet field. The scalar potential for the N2HDM is given as [12]:

$$\begin{aligned} V_{\text{N2HDM}} = {} & m_{11}^2\phi_1^\dagger\phi_1 + m_{22}^2\phi_2^\dagger\phi_2 - [m_{12}^2\phi_1^\dagger\phi_2 + \text{h.c.}] + \frac{1}{2}\lambda_1(\phi_1^\dagger\phi_1)^2 + \frac{1}{2}\lambda_2(\phi_2^\dagger\phi_2)^2 + \\ & + \lambda_3(\phi_1^\dagger\phi_1)(\phi_2^\dagger\phi_2) + \lambda_4(\phi_1^\dagger\phi_2)(\phi_2^\dagger\phi_1) + \frac{1}{2}[\lambda_5(\phi_1^\dagger\phi_2)^2 + \text{h.c.}] + \\ & \frac{1}{2}m_s^2\phi_s^2 + \frac{\lambda_6}{8}\phi_s^4 + \frac{\lambda_7}{2}(\phi_1^\dagger\phi_1)\phi_s^2 + \frac{\lambda_8}{2}(\phi_2^\dagger\phi_2)\phi_s^2 \end{aligned} \quad (2.33)$$

The terms in the first two lines of the given potential describes the 2HDM part of the N2HDM, while the remaining terms describes the contribution of the singlet field ϕ_s.

After EWSB, similar to the 2HDM, the two doublet fields acquire real vevs, v_1 and v_2, and the singlet field acquires a real vev, v_s. The N2HDM consists of 3 CP-even Higgs bosons, h_1, h_2, and h_3, 1 CP-odd Higgs boson, A, and 2 charged Higgs bosons, $H^\pm$. Conventionally, the CP-even Higgs bosons are ordered in ascending order of their masses:

$$m_{h_1} < m_{h_2} < m_{h_3} \tag{2.34}$$

A and $H^\pm$ are assumed to be heavier than the other two CP-even Higgs bosons. In total, the N2HDM is described by 12 independent parameters:

- three mixing angles, α_1, α_2, and α_3 which denote the mixing in the CP-even sector,
- $\tan(\beta) = \frac{v_2}{v_1}$ (ratio of the two vevs),
- SM vev, $v = \sqrt{v_1^2 + v_2^2}$,
- the vev of the singlet field, v_s,
- five Higgs masses (m_{h_1}, m_{h_2}, m_{h_3}, m_A, $m_{H^\pm}$),
- m_{12}^2.

The CMS collaboration reported an excess of $\approx 3\sigma$ in the search for light Higgs bosons at a mass of 96 GeV [72] in the $H \to \gamma\gamma$ decay mode, which coincides with a $\approx 2.3\sigma$ excess in the $e^+e^- \to ZH(\to bb)$ processes studied at the Large electron-positron collider (LEP) [73] at the same Higgs mass [74]. These excesses are seen to be accommodated in the N2HDM [15] where h_1 or h_2 is identified at this potential mass state of 96 GeV. The mixing of such a light Higgs boson with the SM Higgs boson leads to a scenario which deviates from the alignment limit of the 2HDM such that it leads to the occurence of sizeable branching ratios for $H^\pm \to W^\pm h_{1,2}$ decays with $m_{H^\pm} <$ 650 GeV [75]. This fact renders the N2HDM as a potential model in light of the search for heavy charged Higgs bosons.

Another model of interest that can be explained in the same phenomenological framework as the N2HDM is the *Three Higgs Doublet Model (3HDM)* [76] which is based on type-Z_3 models[8], and constitutes 9 physical states, namely 3 CP-even Higgs bosons, h_1, h_2, and h_3, 2 CP-odd Higgs boson, A_1 and A_2, and 4 charged Higgs bosons, $H_1^\pm$ and $H_2^\pm$.

The Georgi-Machacek Model

The Higgs Triplet Model (HTM) [77] is yet another way to extend the SM Higgs sector by providing a natural setting for the *see-saw mechanism* of neutrino mass generation [78]. The see-saw mechanism is used to understand the relative size of the observed neutrino masses when compared to those of quarks, and charged leptons which are a million times heavier. However, the HTM is unable to fulfil an experimental constraint on the tree-level value of ρ parameter: $\rho = \frac{M_W^2}{M_Z^2 \cos\theta_w} \approx 1$, needed to preserve

[8] Z_3 models accomodates scenarios in which the charged leptons, u-, and d-quarks couples to a different doublet individually.

the custodial $SU(2)$ symmetry[9] [79]. This problem can be solved by the introduction of a variant model by Georgi and Machacek [14] which consists of two scalar triplets, a real and a complex, in addition to the SM doublet and preserves this symmetry. The triplet vev induces a mixing between the GM and the SM sector leading to the existence of a rich particle phenomenology of $SU(2)$ bosons.

After the EWSB, the doublet and the triplets can be expanded around their vaccum expectation values, v_d and v_t, respectively. The SM vev, v, in terms of v_d and v_t can be represented as:

$$v_d^2 + 8v_t^2 = v^2 = (246\,\text{GeV})^2 \tag{2.35}$$

The GM model predicts following $SU(2)$ bosons:

- Custodial singlets, h, H (CP-even scalars),
- Custodial triplets, $H_3^{\pm}$, H_3^0,
- Custodial quintets, $H_5^{++/--}$, $H_5^{\pm}$, H_5^0.

Due to the preservation of the custodial $SU(2)$ symmetry, the quintets and the triplets neither mix with each other nor with the singlets. However, the singlet states mix with each other with the mixing angle α (analogous to the 2HDM), and also couple to the fermions and the weak gauge bosons. Given the fact that the neutral triplet state, H_3^0, couples only to the fermions and not to the weak gauge bosons, and all the quintet states couple only to the weak gauge bosons and not to the fermions, one of the two singlet states, h/H, can be identified as the SM Higgs boson. The charged triplet state $H_3^{\pm}$ couples to both fermions and weak gauge bosons, where the decay mode $H_3^{\pm} \to W^{\pm}(h/H)$ is predicted to have a sizeable branching ratio in the context of this model [16].

[9] A residual global $SU(2)$ symmetry that remains after EWSB and prevents higher-order corrections to degrade the property of the theory.

CHAPTER 3

Experimental Framework

This chapter gives a description of the experimental framework used in this thesis. It is structured as follows: Section 3.1 outlines the world's largest and most powerful accelerator, the LHC. This section introduces the LHC machine and its various features. The particle collisions are brought about at the four interaction points around which the four main detectors - ATLAS [80], CMS [81], ALICE [82], and LHCb [83] - are situated. Section 3.2 describes in detail the ATLAS detector, its coordinate system, the various detector components, and substantiates the trigger and data acquisition system. Following this, Section 3.3 summarises the generation of simulated physics processes which are needed in order to perform physics analyses, and test their agreement with collision data. Section 3.4 details the reconstruction of various physics objects, and their related physics in light of this thesis. Various requirements on these physics objects are applied in both data and simulation, to target a specific region needed for physics analyses. Multivariate algorithms, which are used for tasks such as event reconstruction and signal-to-background discrimination in the context of this thesis, are briefly discussed in Section 3.5. Finally, this chapter concludes with a short description of the statistical methods used in this thesis in Section 3.5.

3.1 Overview of the collider physics at the LHC

The LHC is a superconducting hadron accelerator and collider installed in a 27 km tunnel on the French-Swiss border at CERN, Geneva, where the LEP collider was previously operating. The main goal of the accelerator is to collide protons with a centre-of-mass energy, $\sqrt{s}$, up to 13.6 TeV, in order to enhance our understanding of particle physics. The first physics run, referred to as Run 1, was completed between 2010 and 2012 with $\sqrt{s}$ up to 8 TeV. This physics run saw the discovery of the Higgs boson in 2012 by the ATLAS and CMS experiments [5, 6]. In the year 2015, LHC started its second phase called Run 2, following a two year upgrade period, with an increased $\sqrt{s}$ = 13 TeV which lasted until 2018. The increase in luminosity helped to more precisely measure the properties of the Higgs boson, and search for physics beyond the Standard Model. Currently, the LHC is in mid of its third phase of the physics run, Run 3, which commenced in July 2022 at an increased $\sqrt{s}$ = 13.6 TeV.

The search presented in this thesis is performed using proton-proton collision data collected by the ATLAS detector at $\sqrt{s}$ = 13 TeV during the Run 2 data taking period. This dataset corresponds to a total integrated luminosity of 140 fb^{-1}.

3.1.1 Introduction to the machine

The tunnel of the collider houses two adjacent parallel beam pipes each containing a beam, which travels in opposite directions around the ring before the two beams collide. 1232 superconducting niobium-titanium dipole magnets bend the path of the proton beams travelling through the ring and keeping them in a circle. These dipole magnets are cooled down to a temperature of 1.9 K using 96 tonnes of superfluid helium[1], and provides a magnetic field of around 8 T. The beams are further stabilised and focused using additional 392 quadrupole magnets complementing the dipole system.

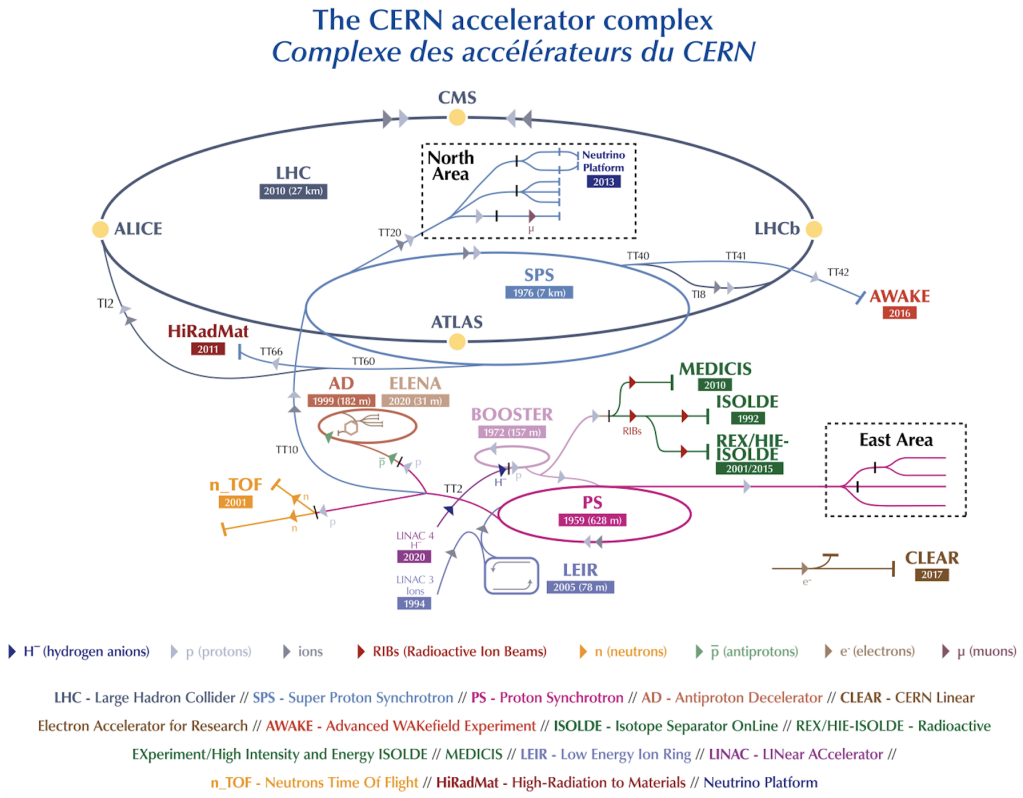


Figure 3.1: A schematic of the CERN accelerator complex [84].

The acceleration process is carried out in a step-wise manner using a large fraction of the CERN accelerator complex as shown in Figure 3.1, to help achieve the desired design energy of the LHC. The primary source of protons is a hydrogen gas. The protons after being ionised enter the *Linac 4* [85], a linear accelerator, which accelerates the protons up to 160 MeV and squeezes them into "bunches" which constitutes the beam. The beam is guided into the smallest circular machine of the chain, called the *Proton Synchrotron Booster (PSB)* [86]. The PSB consists of four superimposed synchrotron rings which accelerates the protons to an energy of 2 GeV. This higher energy allows for more protons to be injected into the next step of the accelerator chain. This leads to a higher final luminosity in comparison to the point where the injection was done directly from *Linac 4*. The beam is then further injected into the *Proton Synchrotron (PS)* [87] leading to an accelerated proton energy

[1] This fact makes the LHC the largest cryogenic facility in the world.

of around 26 GeV. From the PS, the beam is then sent into the *Super Proton Synchrotron (SPS)* [88] where the protons are accelerated to an energy of 450 GeV. The protons are finally transferred to the two beam pipes of the LHC where they circulate in opposite directions, and are accelerated up to an energy of 6.5 TeV (during Run 2). These are then brought into collisions at the interaction points, and the data is collected by the different detectors around the interaction points.

As mentioned earlier, the protons in the LHC ring travel in bunches with each beam being formed by a train of a maximum of 2808 bunches with approximately 10^{11} protons per bunch. The separation between these bunches is called the bunch spacing. The bunch spacing was set to 25 ns during Run 2.

3.1.2 Luminosity

Another important feature besides the centre-of-mass energy is the the number of interactions per bunch crossing generated in the LHC collisions. The luminosity (L) or commonly called the *instantaneous luminosity* is directly connected to the beam structure. For two colliding bunches with a transverse profile distributed according to a Gaussian-distribution, the luminosity L is given by the following expression [89]:

$$L = \frac{N_1 N_2 f N_b}{4\pi\sigma_x\sigma_y} S \tag{3.1}$$

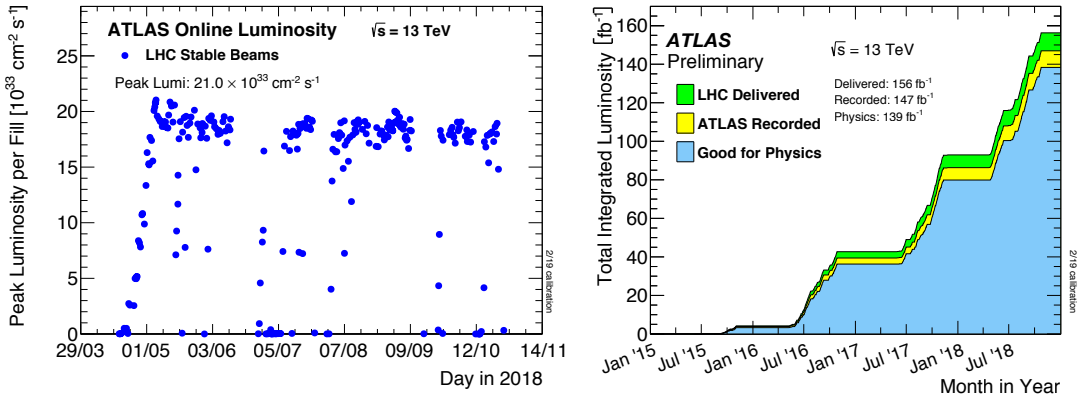
where N_1 and N_2 are the numbers of particles in the two colliding bunches, respectively, N_b is the number of bunches per beam, f denotes the revolution frequency of the bunches, and σ_x and σ_y correspond to the transverse beam sizes at the collision point. The factor S is a geometrical factor taking into account the reduction of the luminosity under a crossing-angle scheme[2].

A precise knowledge of the luminosity is needed for performing physics analyses. In order to achieve this, the ATLAS detector uses mutliple dedicated detectors called luminometers. These luminometers have complementary capabilities and different systematic uncertainties. The usage of multiple luminometers with different capabilities ensures a mechanism that offers redundancy and robustness against individual measurement biases. During Run 2, the primary bunch-by-bunch luminosity measurement was provided by the `LUCID 2` Cherenkov detector [90], which is placed approximately at ± 17 m from the interaction point. This was complemented by the measurements from the *beam conditions monitor (BCM)* [91] diamond detectors at the bunch-crossing level, and from the offline measurements of the multiplicity of reconstructed charged particles in randomly selected colliding bunch crossings. The measurements of the two detectors are related to the luminosity via an absolute calibration procedure. This procedure is performed in dedicated sessions during special LHC fills in each data taking year, called *van der Meer (vdM) scans* [92]. The vdM procedure scans the overlap of the two beams in the transverse plane ($x - y$ plane) to the beam axis, and subsequently measures the relative interaction rates as a function of the transverse beam separation. The two transverse planes are scanned independently such that the size and shape of the interaction region can be measured completely. Figure 3.2(a) shows the evolution of the instantaneous luminosity as a function of time in the year 2018 during Run 2. It can be observed already from this figure that the LHC showed an outstanding performance throughout its operation during Run 2, and successfully surpassed its design value of 10^{34} cm^{-2}s^{-1}.

Another quantity of interest with regard to the physics potential of the experiments at the LHC is

[2] At the LHC, the two beams collide at a crossing angle of ≈ 300 µrad in order to decrease the number of additional interactions.

the integrated luminosity which is the luminosity integrated over the entire data taking period. Figure 3.2(b) shows the integrated luminosity as a function of time for the ATLAS detector during Run 2. It is observed from this figure that an integrated luminosity of 156 fb^{-1} was delivered which is six times more than the one delivered during Run 1. The ATLAS detector recorded 147 fb^{-1} of data, corresponding to an impressive data-taking efficiency of almost 95%. Only a subset of this recorded data which satisfies a strict quality criteria discarding any faulty inputs arising from the hardware or software failures, is used in the physics analyses. In Run 2, 140 fb^{-1} of data suitable for physics was recorded.



(a) (b)

Figure 3.2: (a) Instantaneous luminosity delivered to ATLAS as a function of time for the 2018 data taking during Run 2. (b) Total integrated luminosity as a function of time delivered to ATLAS (green), recorded by ATLAS (yellow), and certified to be good physics data (blue) during stable beams for Run 2 proton-proton collisions [93].

3.1.3 Pile-up

An increased instantaneous luminosity at the LHC is beneficial as a large amount of data can be collected over a shorter duration of time. However, this leads to an increase in the number of proton-proton collisions per bunch crossing. This feature is an evident drawback as the LHC experiments are only interested in a single interaction - the so-called hard-scatter interaction. This interaction of interest can be contaminated by multiple inelastic interactions, per bunch crossing, which are collectively known as pile-up, and can deteriorate the performance of the detector. The number of pile-up interactions per bunch crossing is denoted by μ and is expressed as [94]:

$$\mu = \frac{L \times \sigma_{\text{inel}}}{N_b f} \tag{3.2}$$

where σ_{inel} is the proton-proton inelastic cross-section. Pile-up can be in-time which originates from the additional interactions in the same bunch crossing, or out-of-time arising from signals from the previous bunch crossings. It is hard to mitigate the in-time pile-up, and is best taken care by the high granularity of the detectors which can distinguish among different interaction vertices and their associated particles. The out-of-time pile-up becomes relevant in cases where the detector

signal-processing time is long compared to the time between the crossings. This is seen for instance in the ATLAS Liquid Argon calorimeters described further in Section 3.2.2. Figure 3.3 shows the distribution of average pile-up, $\langle\mu\rangle$, during Run 2 proton-proton collisions.
The average pile-up for the entire run was seen to be around 34 interactions per bunch crossing. These conditions poses an additional challenge in the calibration and performance of the detectors. Hence, a continuous evolution of techniques is needed to mitigate these undesired effects with each successive data taking periods.

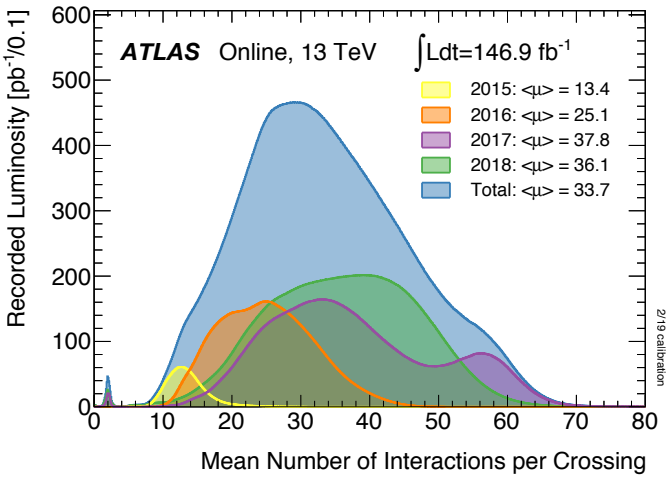


Figure 3.3: Distribution of $\langle\mu\rangle$ per bunch crossing for each year of Run 2 proton-proton collision data recorded by the ATLAS detector [93].

3.2 ATLAS

As mentioned earlier, ATLAS [80] (A Toroidal LHC ApparatuS) is one of the main experiments taking data at the LHC. It is one of the general-purpose detectors used to study a wide range of physics from the search for the Higgs boson to that of particles that could make up the dark matter or many other new interesting particles. It is situated in a hall about 100 m underground in a beam interaction point of the LHC ring. The detector is split into a barrel part and two end-cap parts. The detector is designed in such a way that a large number of different particles can be measured within a full 4π solid angle. It has a cylindrical shape around the beam pipe, with a height of 25 m and a length of 44 m, and weighs about 7,000 tonnes. The scattered particles resulting from the proton-proton interactions between the beam particles are measured by a series of sub-detectors which are placed around the beam pipe forming the barrel part. Figure 3.4 shows a schematic of the detector setup. The detector magnet system consists of a solenoid and a set of toroid magnets. The magnet system is needed to bend the charged particles, helping in their separation, and for their charge and momentum estimation. Close to the beam pipe, there is an innermost system called the Inner Detector (ID) [95] which is surrounded by the solenoid. Surrounding the ID, are the two calorimeters, namely the electromagnetic calorimeter (ECAL) [96], and the hadronic calorimeter (HCAL) [97]. The outermost detector is the Muon Spectrometer (MS) [98]. The toroids are situated outside the calorimeters, and within the MS.

The two end-cap parts of the ATLAS detctor are placed in a direction perpendicular to the beam, situated at the end of the detector. These end-caps improve the detector coverage in the forward region. The following section describes the detector coordinate system, and the different detector components in detail.

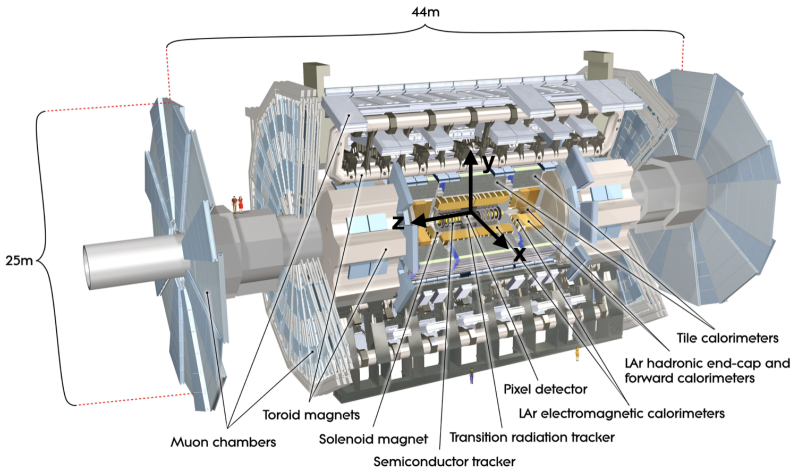


Figure 3.4: A computer generated image of the ATLAS detector. The coordinate system is marked in black [84].

3.2.1 Coordinate system

The coordinate system of the ATLAS detector and the commonly used kinematic terms used for the particles emerging from the proton-proton collisions are summarised in this section. The interaction point is chosen to be the origin of the coordinate system. The z-axis is in the direction of the beam axis which is counter-clockwise to the LHC ring. The x-axis points towards the center of the LHC ring, and the y-axis points upwards towards the surface of the earth. This cartesian coordinate system is marked in Figure 3.4. The four-momentum of a particle in the cartesian coordinate system can be expressed as:

$$\boldsymbol{p} = (E, p_x, p_y, p_z) \tag{3.3}$$

where E is the energy of the particle, and p_x, p_y, and p_z are the particle's momentum in the x, y, and z directions, respectively. The kinematic properties of the particles are often described in the transverse plane to the beam axis. The most common one is the transverse momentum, p_{T}, of a particle, defined as:

$$p_{\mathrm{T}} = \sqrt{p_x^2 + p_y^2} \tag{3.4}$$

In proton-proton collisions, the overall momentum along the beam axis is unknown. This is because when two protons collide, it is actually the partons (guarks and gluons forming a proton) that take part in the interactions. Each of these partons carries an unknown fraction of the proton's momentum. However, the overall momentum perpendicular to the beam axis is known to be approximately zero before the collisions. Thus, the p_{T} of a particle is considered as an important quantity in collider physics.

The transverse plane is often described using a spherical coordinate system ($r - \theta - \phi$). The radial

distance, $r = \sqrt{x^2 + y^2}$, is the distance of the particle from the beam axis. The polar angle θ is the angle from the beam axis, and the azimuthal angle ϕ is measured in the transverse plane around the beam axis. The θ coordinate is most often replaced by the pseudorapidity, η:

$$\eta = -\ln \tan\left(\frac{\theta}{2}\right) \tag{3.5}$$

The pseudorapidity is preferred over the θ coordinate as the pseudorapidity differences are Lorentz-invariant under boost along the beam axis. η in the massless particle limit is equivalent to rapidity, $y = \frac{1}{2}\ln\left(\frac{E+p_z}{E-p_z}\right)$. A particle with $\eta = 0$ passes through the detector in a direction perpendicular to the beam axis ($\theta = 90^\circ$), and if the particle is travelling in a direction parallel to the beam axis ($\theta = 0^\circ$ or $\theta = 180^\circ$), the pseudorapidity of the particle tends to be infinite ($|\eta| \to \infty$).

The angular separation (ΔR) between two objects is another useful quantity which is defined in the $\eta - \phi$ plane. It is expressed as:

$$\Delta R = \sqrt{(\eta_i - \eta_j)^2 + (\phi_i - \phi_j)^2} \tag{3.6}$$

where i and j denote two objects in the detector. ΔR is often used in the context of jet reconstruction as described further in Section 3.4.3, and for evaluating the spatial overlap between two physics objects.

3.2.2 Detector components

A brief description of the ATLAS detector components [80] used to identify and measure the properties of charged and neutral particles is given in this section.

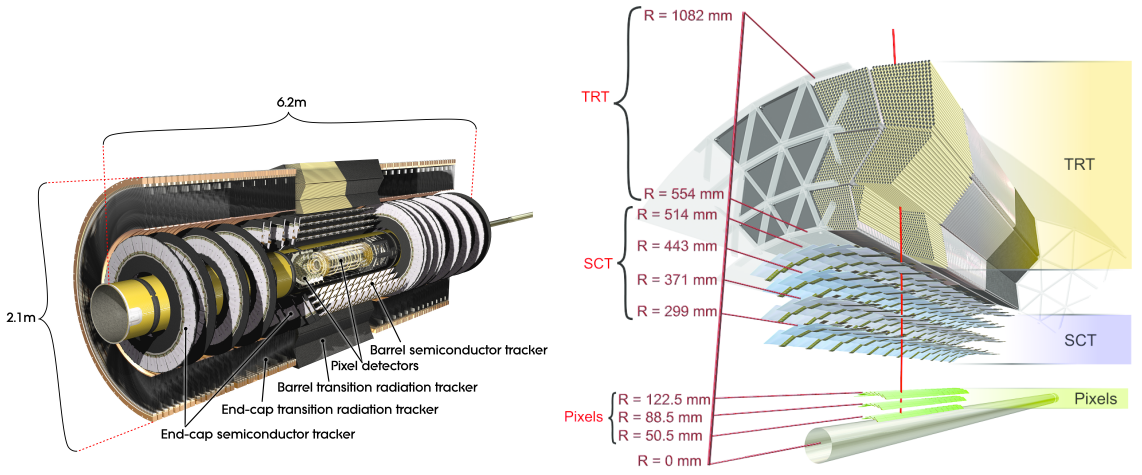
Inner Detector

The ID is the innermost detector closest to the beam pipe. It extends to a radius of 1.2 m, and is 6.2 m in length along the beam pipe and covers up to $|\eta| < 2.5$. The main function of the ID is to reconstruct the trajectory of the charged particles, measure their charges and momenta, and locate their interaction points. It operates inside a 2 T magnetic field provided by the solenoid, which is used to curve the charged particles. This feature makes it possible to measure a particle momenta via their curvature. The momentum resolution, $\frac{\sigma_{p_T}}{p_T}$, deteriorates with increasing p_T. It is expressed as [99]:

$$\frac{\sigma_{p_T}}{p_T} = 0.05\% p_T \oplus 1\% \tag{3.7}$$

Being nearest to the collision point, the ID is designed to resolve the large track density. For this purpose, high granularity sub-detectors are used which are the Pixel detector [100], the Semi-Conductor Tracker (SCT) [101], and the Transition Radiation Tracker (TRT) [102]. The ID is divided into three separate regions: the barrel and the two end-caps. In the barrel region, the sub-detectors are arranged in concentric cylinders around the beam axis. On the other hand, in the end-cap regions, the sub-detector are placed on disks perpendicular to the beam axis. Figure 3.5 shows a layout of the inner detector and its different sub-detector components.

An Insertable B-Layer (IBL) [105], which forms a part of the Pixel detector, was installed in 2014,



(a) (b)

Figure 3.5: (a) A computer generated image of the inner detector of ATLAS [103] (b) A schematic depicting the Pixel detector, SCT, and TRT [104].

during the long shut-down of the LHC between Run 1 and Run 2. It is named so because it is located close to the region where the long-lived b-hadrons decay. The main purpose of its installation was to ensure high-quality tracking despite the large radiation damage caused due to the increased luminosity. The addition of the IBL improved the resolution of the track impact parameters[3] by about 40 % for tracks with $p_T < 1$ GeV. In addition, the Pixel detector consists of three pixel modules arranged in concentric layers in the barrel region covering $|\eta| < 1.5$. In the end-cap regions, there are three disks of silicon modules covering the region $1.5 < |\eta| < 2.5$. The pixels allow for a high spatial resolution of about 10 µm in the transverse direction and 115 µm in the longitudinal direction.

The next part of the inner detector is the SCT. It has a similar function as the Pixel detector but consists of long, narrow strips rather than small pixels, enabling a large area coverage. It consists of four cylindrical layers in the barrel region ($|\eta| < 1.4$), and two sets of nine planar disks in the end-cap region ($1.4 < |\eta| < 2.5$). The strips provide a good spatial resolution of 17 µm in the transverse direction and 580 µm in the longitudinal direction. The silicon strips are preferred over the silicon pixels due to the lower expected particle density in the region. This choice maintains an optimal spatial resolution while containing the large number of readout channels. The last layer of the ID is the TRT. It surrounds the rest of the ID and provides a coverage of up to $|\eta| = 2.0$. As suggested by its name, the detector is a combination of the transition radiation detector and the straw trackers which are drift tubes. There are about 50,000 straws in the barrel region, and around 125,000 straws in the end-cap region. Each straw is filled with xenon gas that gets ionised when a charged particle traverses through the medium. The straw wall kept at high negative voltage creates an electric field which drives the ionised particles towards the central anode and produces an electric signal. The TRT has a reduced spatial resolution of 0.17 mm, but it can be used for particle identfication. This is brought about by the usage of materials with different refractive indices between the straws, which causes ultra-relativistic charged particles to produce transition radiation. This radiation heavily depends on the mass of the travsersing particles, and hence the intensity of the radiated photons can be used to distinguish between particles.

[3] The impact parameters are the track extrapolated position in the interaction plane, described further in Section 3.4.1.

Calorimeters

Calorimeters are situated outside of the solenoid magnet that surrounds the ID. They have a high coverage up to $|\eta| < 4.9$. Their main task is to measure the energy of the particles interacting with the detector material. The deposition of the energy in the calorimeters take place in the form of particle showers. These showers are cascades of secondary particles produced by the parent particle. There are mainly two kinds of showers: electromagnetic and hadronic. The electromagnetic showers are initiated by, and contains only electrons and photons. The hadronic showers are initiated by hadrons, and are propagated by strong interactions between the particles and the detector material. The calorimeter system consists of two sections: electromagnetic calorimeter (ECAL) which is responsible for the measurement of electromagnetic showers, and the hardonic calorimeter (HCAL) whose aim is to provide measurements of the hadronic showers. Figure 3.6 shows a schematic of the ATLAS calorimetric system. The calorimeters consists of a number of *sampling detectors* which have alternating layers of material that produces particle showers (*absorber*), and the material that measures the deposited energy (*active material*). This strategy allows for a precise reconstruction of shower evolution in all dimensions.

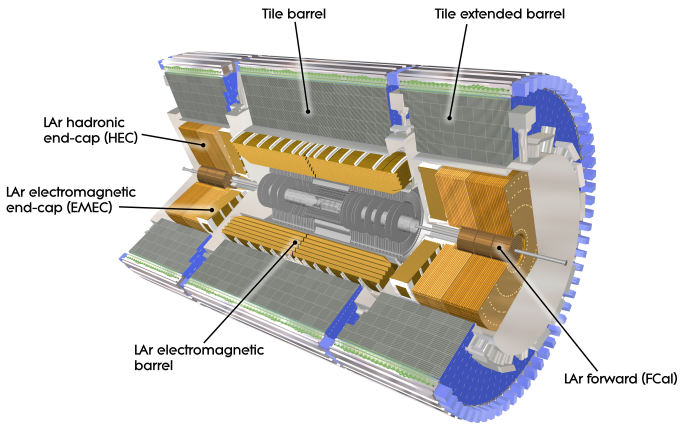


Figure 3.6: A computer-generated image of the calorimeter system of ATLAS showing the electromagnetic and hadronic calorimeters [106].

The ECAL is a sampling calorimeter made up of lead absorbers casted in liquid argon (LAr). Particles interacting with the ECAL ionise the liquid argon, and this allows for their energy measurement. The ECAL is divided into three parts, namely the barrel part covering the range $|\eta| < 1.47$, and the two end-caps with a coverage of $1.375 < |\eta| < 3.2$. The electromagnetic shower is characterised by the material's *radiation length,* X_0, defined as the mean distance over which the particle loses all but $\frac{1}{e}$ of its energy by Bremsstrahlung[4]. The thickness of the ECAL is around 24 X_0 in both the barrel and end-cap regions.

The HCAL is also a sampling calorimeter. It is formed of three parts: The Tile Calorimeter (TileCal) [107], the Hadronic End-cap Calorimeter (HEC) [108], and the Forward Calorimeter (FCal)

[4] Bremsstrahlung is an electromagnetic radiation produced by the deceleration of a charged particle when deflected by another charged particle.

[109]. The hadronic showers are characterised by the material's *interaction length,* λ which is the mean distance travelled by a hadron until it loses $\frac{1}{e}$ of its energy. The HCAL has a total thickness of 11 λ. The TileCal surrounds the ECAL envelopes and is made up of steel plates acting as the absorber, and plastic scintillating tiles as the active material. It consists of three barrels with a coverage of $0.8 < |\eta| < 1.7$. The HEC has two wheels per end-cap and is placed behind the end-cap ECAL, and uses the copper plates and liquid argon as absorber and active material, respectively. It provides a pseudorapidity coverage of $1.5 < |\eta| < 3.2$. Finally, the FCal uses the copper and tungsten as absorber, and liquid argon as active material. It covers the region $3.1 < |\eta| < 5.0$.

The ATLAS calorimeter system provides high granularity and an optimal energy resolution. The target energy resolution provided by the different sub-detector components of the calorimeter system is summarised in Table 3.1.

Table 3.1: The energy resolutions of the different sub-detector components of the ATLAS calorimeter system.

Sub-detector	Energy resolution
ECAL	$\frac{\sigma_E}{E} \approx \frac{10\%}{\sqrt{E\,[\mathrm{GeV}]}} \oplus 0.7\%$
TileCal, HEC	$\frac{\sigma_E}{E} \approx \frac{50\%}{\sqrt{E\,[\mathrm{GeV}]}} \oplus 3\%$
FCal	$\frac{\sigma_E}{E} \approx \frac{100\%}{\sqrt{E\,[\mathrm{GeV}]}} \oplus 10\%$

Muon Spectrometer

Due to their large mass, muons emit less radiation than electrons, i.e. their energy loss through Bremsstrahlung is reduced, and they pass through the ATLAS calorimeters with depositing only little energy. Information on the momentum of the muons is already given by the ID. However, one wants to measure the momentum of highly energetic muons precisely. This is done by the outermost part of the detector, called the Muon Spectrometer (MS). Figure 3.7 shows a schematic of the MS.

The MS is an extremely large tracking system immersed inside a toroid providing a magnetic field of up to 1 TeV enabling a precise measurement of the muon p_T. In order to efficiently cover a large area, the MS is built as a *gaseous ionisation detector* [110]. When a charged particle traverses the active area, it ionises the gasesous medium. An applied electric field then guides the electrons and ions to be collected at the anode. The MS detector extends up to $|\eta| < 2.7$.

The MS consists of four sub-detectors, two in the central barrel, and two in the end-cap regions. In each region, one detector system is needed for a precise trajectory measurement, and the other is used to trigger on muons. The trajectory measurement is provided by the Monitored Drift Tubes (MDTs) [111] in the barrel region and the Cathode Strip Chambers (CSCs) [112] in the end-cap regions. The MDTs are aluminium drift tubes with a tungsten wire filled with an argon gas. They provide a coverage of $-2.7 < |\eta| < 2.7$. The CSCs are multiwire proportional gas chambers with cathodes segmented into strips, and provide a coverage between $-2.0 < |\eta| < 2.7$.

The muon trigger system consists of Resistive Plate Chambers (RPCs) [113] in the barrel region, and Thin Gap Chambers (TGCs) [114] in the end-cap regions. They provide fast trigger signals with timing between 15 to 25 ns. RPCs consists of two parallel electrode plates with the gap in between filled with a $C_2H_2F_4$-based gas mixture with small contents of SF_6. They cover a pseudorapidity

region of $|\eta| < 1.05$. The TGC modules consist of a wire plane maintained at high positive voltage, between two resistive grounded cathode planes covering the range $1.05 < |\eta| < 2.4$.

Overall, the MS offers a high p_T resolution for muons between 3% to 12% for p_T values between 10 GeV to 1 TeV. In addition, it provides an excellent charge identification in the p_T range from 3 GeV to 3 TeV.

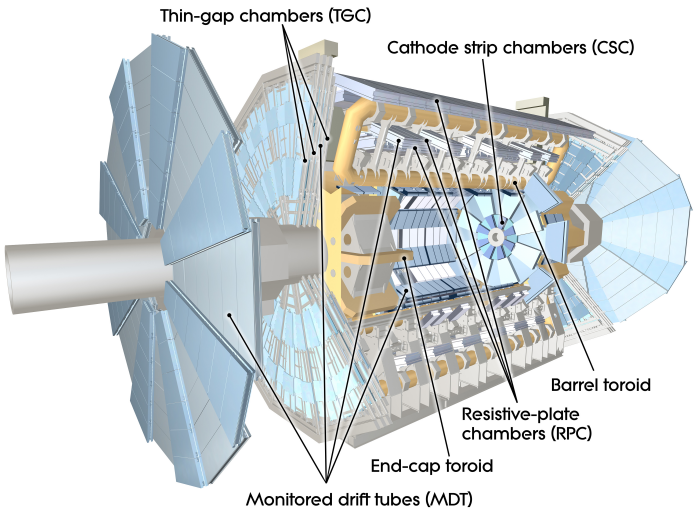


Figure 3.7: A computer-generated image of the ATLAS muon spectrometer [115].

3.2.3 Trigger and Data Acquisition

Proton-proton collisions at the LHC occur at a frequency of 40 MHz, i.e. every 25 ns. The available information from all the ATLAS sub-detectors lead to an overwhelming data collection rate of around 60 TB/s. In practise, only a fraction of events out of this information is used further for the physics analyses. The selection of the events is necessary due to limited resources for readout, storage, and offline processing of data. The ATLAS trigger system [116] is divided into two independent levels: a hardware-based first Level (L1) trigger and a software-based High Level Trigger (HLT). The layout of the ATLAS trigger and data acquisition system is shown in Figure 3.8.

The L1 trigger is based on custom-made electronics with a decision output rate of less than 40 MHz. It makes an initial selection based on the information from the calorimeters and the MS. The L1 trigger decision algorithm finds more than one Regions-of-Interest (RoIs) in coordinates of η and ϕ. In each of the RoIs, electron candidate objects, muon objects, and the missing transverse momentum are reconstructed (see Section 3.4), and they contain information on the type of feature identified and the criteria passed. The output of the L1 trigger system is a single-bit decision transmitted across the detector system to either read-out the full event or reject it. The event rate is reduced from 40 MHz to 100 kHz after the L1 trigger stage.

The accepted events are then passed on to the HLT which refines the decision based on additional selection criteria. The HLT is a farm of CPU cores, and can help in the execution of complex algorithms within the L1 RoIs to build objects in the detector such as tracks, electrons, muons, and jets. The HLT decision is based on a list of pre-defined trigger selection criteria, known as the *trigger*

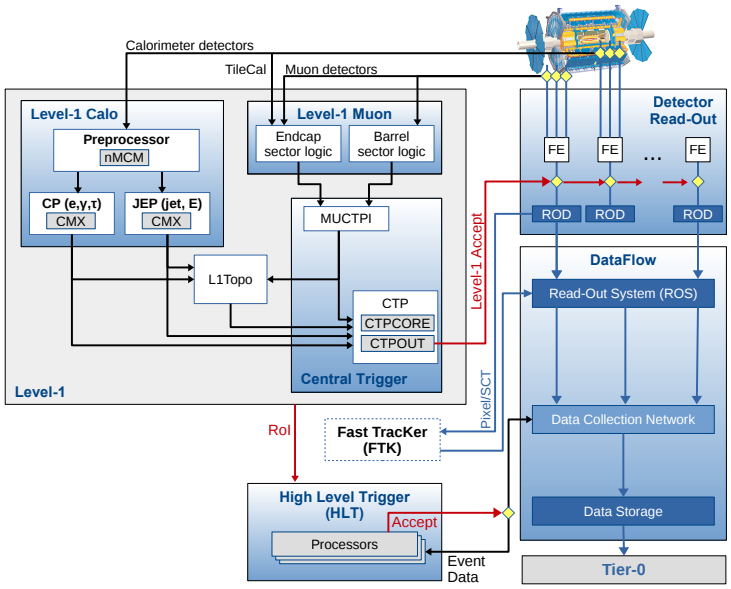


Figure 3.8: A schematic illustration of the ATLAS Run 2 trigger and data acquisition system [116].

menu [117]. The trigger menu includes various interesting event topologies for physics analyses, but also for calibration and performance measurements. The HLT reduces the event rate from 100 kHz to 1 kHz. Events that pass at least one of the trigger menu selections are recorded on disk, and can be analysed offline.

3.3 Simulation

The physics analyses depends on reliable simulations of signal and background processes. They are needed in order to optimise and validate analysis techniques, perform comparisons with collision data to test their level of agreement with the SM predictions, and calculate systematic uncertainties (see Section 3.6 and Chapter 6). These simulated processes are produced with state-of-art Monte Carlo (MC) event generators [118] such as PYTHIA, HERWIG etc. described later in this section. The simulation begins with a hard-scatter process in which the partons interact at high p_{T}, leading to the formation of quarks, leptons, bosons, or any new BSM particle. The quarks from this hard-scatter process radiate gluons as they evolve, which can further split into quarks in an iterative process forming the so-called parton showers. The particles constituting the parton showers hadronise to form colourless hadrons. As most of the produced hadrons are unstable, the next step of the simulation is that of a hadron decay. The constituent partons that do not take part in the hard-scatter process undergo multiple interactions forming the underlying event. Finally, the ATLAS detector response is

simulated using GEANT4 [119]. The aforementioned simulation steps (except for the simulation of the detector response) are shown in Figure 3.9. The various steps are described further in this section.

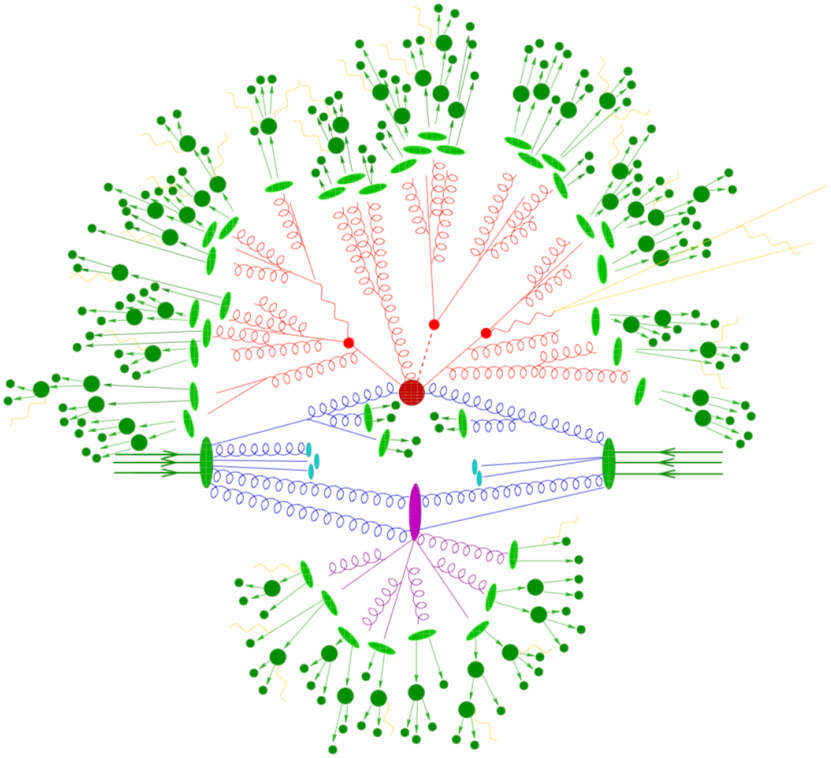


Figure 3.9: An overview of the various stages of event simulation. The points of hard-scatter interactions are shown in red. Parton showers are depicted in blue. Hadronisation and decay processes are shown in green. The underlying event is shown in purple [120].

- **Hard-scatter event simulation**: Most of the interesting physics processes at the LHC involve large momentum transfers, such as the production of high-p_{T} jets (see Section 3.4.3). These kinds of interactions can be described with the help of *perturbation theory* [121] because QCD quantas are asymptotically free. The simulation of such processes form the basis of any simulation of collider events with the help of MC generators. The cross-section for a scattering process of two hardons, h_1 and h_2, at hadron colliders is given using the *factorisation theorem*. The cross-section is expressed as:

$$\sigma_{h_1 h_2 \to n} = \sum_{a,b} \int_0^1 \int_0^1 dx_a dx_b \int d\Phi_n f_a^{h_1}(x_a, \mu_F) f_b^{h_2}(x_b, \mu_F) \times \frac{1}{2\hat{s}} |\mathcal{M}_{ab\to n}(\Phi_n; \mu_F, \mu_R)|^2 \tag{3.8}$$

  In the above equation, $f_a^{h_1}(x_a, \mu_F) f_b^{h_2}(x_b, \mu_F)$ is the parton distribution function[5] (PDF) which depends on the momentum fraction, x_a and x_b, of partons a and b with respect to its parent

[5] These functions are used to describe the probability of finding a parton with momentum x of the total momentum.

hadrons h_1 and h_2, respectively, and on the factorisation scale μ_F. $\mathcal{M}_{ab\to n}$ is the matrix-element, $d\Phi_n$ is the differential phase space element for the n-particle final state. $\hat{s} = 2x_a x_b s$ is the effective (paton-parton) centre-of-mass energy. μ_R denotes the renormalisation scale.

- **Parton showers**: Equation 3.8 gives a fixed-order framework for the calculation of a scattering process at a specific order in perturbation theory. The parton shower algorithms [122] extend the calculations of a hard-scatter process to higher orders. The algorithm recursively produces the parton splitting processes (e.g. $g \to q\bar{q}, q \to gq$) until a low-energy scale (O (1 GeV)) is achieved. There can be an incompatibility between the matrix-element and parton shower calculations for a full cross-section computation at order $n > 1$, as there is a potential overlap in the phase space of the extra partons that are considered for the matrix-element at order n with the ones considered in the splitting at order n-1. To handle this double-counting, there are different approaches, commonly known as matrix-element-to-paton-shower matching, which can be used. The most commonly used approach is to define a matching scale where the higher-energy region is covered only by the matrix-element simulation, while any additional parton with energy below this scale are vetoed and are covered by the parton shower algorithm.

- **Hadronisation**: The parton shower algorithms continues until a low energy scale is achieved. At this state, the quarks and gluons hadronise into bound (colour-singlet) final states. The hadronisation process is independent of the initial hard-scatter process, thus the hadronisation models [123, 124] can be used and tuned in some reference processes.

- **Decay**: After the hadronisation of the partons, unstable hadrons decay into stable particles. This is considered to be an important step of the event simulation process as the observed final state hadrons are produced as an outcome of this simulation stage.

- **Underlying event**: The underlying event simulation consists of all interactions that do not arise from the hard-scatter process. These include initial- and final-state radiations[6], beam-beam remnants, or multiple parton interactions [125, 126]. These events are represented by low energy scales. The simulation of the underlying event is overlaid with the simulation of pile-up events, which are explained in the next step.

- **Pile-up events**: As discussed in Section 3.1.3, there is a high probability of multiple inelastic interactions in the same or the neighbouring bunch crossing which can happen simultaneously at various collision points, known as the pile-up. The pile-up simulation is performed for signal events prior to the conversion of energy deposits to detector signals.

- **Detector simulation**: The last stage of the simulation process is to simulate the interactions of the final state particles created by the MC generators with the ATLAS detector. The widely used tool for this purpose is the GEANT4 package. The kind of simulation that uses the GEANT4-based simulation of the ATLAS detector geometry and response are referred to as Full Simulation (`Full Sim`). `Full Sim` samples are computationally intensive but provide the most precise result. Since more than 90% of the computation time is spent on the calorimeter simulations, faster alternatives are needed in practise. The `AtlFast II` (`AFII`) simulation is one such

[6] The showering process that is applied to the final products after the hard-scattering is referred to as final-state radiation (FSR), and the simulation of initial-state radiation (ISR) is simulated for the incoming partons.

algorithm which is used as an ATLAS Fast Calorimeter Simulation (FASTCALOSIM) for the calorimeters, and as Fast ATLAS Tracking Simulation (FATRAS) for the ID. The remaining detector sub-systems are simulated with GEANT4 which requires less compuation time, and provides a reasonable accuracy.

There are several kinds of MC event generators used in this thesis as described in the following. Some of them are multi-purpose generators, and can be used to simulate various parts of the event. Others are used to simulate only a certain part of the event, and can be used together with the multi-purpose ones depending on the need of an analysis.

- PYTHIA [127] is a multi-purpose generator which provides parton shower, underlying event, and matrix-element simulations. The matrix-element is simulated up to the leading order (LO), while both the parton shower and the underlying event models are tuned based on the existing measurements. This generator can also be used to simulate a single stage of an event simulation. For most of the simulated processes used in the analysis presented in this thesis, PYTHIA is used to provide a description of the parton showers, which shows a good agreement with data.

- HERWIG [128] is another multi-purpose generator used for providing a description of the parton showers. It is primarily used to evaluate the systematic effects due to parton shower algorithms by comparing its prediction with that of PYTHIA.

- The SHERPA [120] event generator covers the matrix-element calculation and parton showers description. It is capable of providing next-to-leading order (NLO) calculation for a variety of processes, and has been the preferred choice for processes with additional radiated jets. Unlike PYTHIA and HERWIG, SHERPA cannot be used to generate a single stage of an event simulation.

- The POWHEG BOX [129] generator can provide matrix-element simulations in NLO. It is best used with other generators that can simulate parton showers. For instance, it is used with PYTHIA to model $t\bar{t}$ events.

- The MADGRAPH [130] generator can provide matrix-element simulations in LO and NLO. In most of the cases, it is used to evaluate the systematic effects due to the matrix-element-to-parton-shower matching approach.

3.4 Object reconstruction with the ATLAS detector

Different kinds of particles interact in specific ways with the various components of the ATLAS detector which give rise to measurable signals, as shown in Figure 3.10. These signals can be realised in the form of various physics objects which are used in physics analyses. These physics objects are reconstructed using dedicated offline event reconstruction techniques. The different objects are most often reconstructed independently in ATLAS. However, there can be ambiguities between the final state particles, which are treated on a case-by-case basis at the analysis level. This section describes the reconstruction of the physics objects used in the analysis presented in this thesis, in both simulated events and collision data.

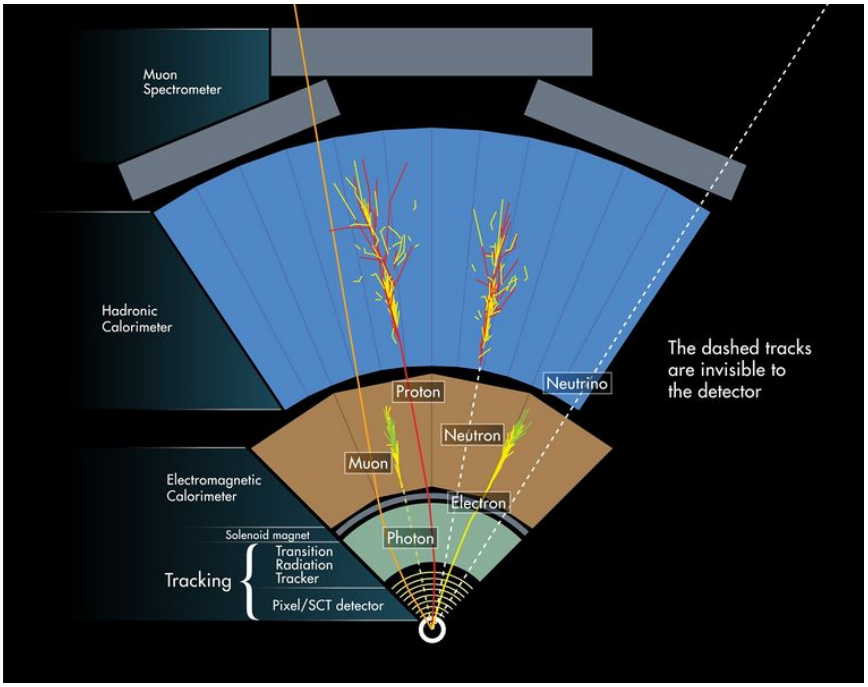


Figure 3.10: A schematic illustrating the interactions of various kinds of particles with the ATLAS detector [131].

3.4.1 Tracks and vertices

A track is a reconstructed-trajectory of a charged particle. Tracks are reconstructed from hits[7] of the charged particles in the ID by the ATLAS track reconstruction algorithm [132]. This algorithm follows a step-wise procedure. It begins with clustering hits in the different ID sub-detectors, followed by a combinatorial track candidate finding algorithm [133]. Once all possible track candidates are found, ambiguities are resolved based on a scoring system which ranks those track candidates higher which are more likely to correctly represent a charged particle track. A track fit is then performed for these track candidates using information from all ID sub-systems.

Several proton-proton collisions can appear per bunch crossing. Usually, one high-energy interaction is overlayed by several low-energy interactions. The task of the primary hard-scatter vertex reconstruction is a precise reconstruction of all interaction points in one bunch crossing. The reconstruction of the primary vertex (PV)[8] takes place in two steps [134, 135]. The first step associates the tracks with vertex candidates, where a set of tracks are used to determine possible vertex seed positions. The second step takes in the tracks and the seeds from the previous step as inputs, and fits the best vertex position. Once the vertex postion is determined, tracks that are not compatible with the vertex are removed from the list of the vertex-associated tracks, and can be used to find another vertex. The PV is taken as the one with the highest sum of the transverse momenta squared (p_{T}^2) of all the associated tracks.

The so-called impact parameters are among the quantities that parametrises a track, namely the transverse impact parameter, d_0 and longitudinal impact parameter, z_0. d_0 is defined as the shortest

[7] A hit is a measurement point for a track in the ID.

[8] For simplicity, the term "primary vertex" is henceforth used to denote a "primary hard-scatter vertex".

distance between a track and the beam axis in the transverse plane. z_0 is the distance along the z-axis between the PV and the track which is used to evaluate d_0.

3.4.2 Leptons

As mentioned earlier in Chapter 2, charged leptons are electrons (e), muons (μ), or tauons (τ). However in ATLAS analyses, the term "leptons" is exclusively used for electrons and muons. τ-leptons decay before they are detected. Neutrinos leaves no trace of a signal in the detector. The missing transverse momentum reconstruction is described separately in Section 3.4.5. Thus, the following sections describe the reconstruction procedures for electrons and muons. The description of the reconstruction procedure for τ-leptons is omitted as they are not relevant for the analysis described in this thesis.

Electrons

As shown in Figure 3.10, electrons are characterised by a track in the ID, and a matching compact energy deposition in the ECAL. Electron-track candidates are reconstructed in the same way as described in Section 3.4.1. The only variation is that the algorithm also accounts for the significant energy losses due to the Bremsstrahlung radiation. These electron-track candidates need to be compatible with an origin from the PV. The showers are localised energy deposits which are processed using an algorithm based on the clustering of calorimeter cells forming cluster seeds. Electron candidates are thereby built by matching an electron-track candidate to a calorimetric cluster seed. Electron candidates must satisfy the following requirements [136]:

- $p_{\mathrm{T}} > 27\,\mathrm{GeV}$,
- $|\eta| < 2.47$, with the transition region between the barrel and end-cap ECAL ,$1.37 < |\eta| < 1.52$, being excluded,
- $d_0/\sigma_{\mathrm{d_0}} < 5$, where $\sigma_{\mathrm{d_0}}$ is the d_0 uncertainty,
- $|z_0 \sin\theta| < 0.5\,\mathrm{mm}$.

However, not all objects which are built by the aforementioned electron reconstruction algorithm can be classified as prompt electrons. Most of the prompt electrons originate from $Z \rightarrow e^+e^-$, $W^{\pm} \rightarrow e^{\pm}\nu$, $J/\Psi \rightarrow e^+e^-$ decays. They can also include electrons from FSR or Bremsstrahlung processes. The non-prompt background objects include hadronic jets, electrons originating from photon conversion, as well as from semileptonic heavy-flavour hadron decays. To reject these backgrounds, a dedicated identification algorithm based on a multivariate technique is applied, widely known as the *likelihood-based (LH) method*. At the same time, one also needs to isolate the prompt electrons from the high activity in the area surrounding the candidate object. This kind of activity can arise from hadrons misidentified as leptons, or from the decay of heavy quarks, and can be quantified by multivariate techniques using the properties of tracks and calorimeter clusters. Three levels of identification and isolation working points are considered, referred to as Loose, Medium, and Tight in order of increasing background rejection [137]. For the analysis presented in this thesis, electrons with tight identification and isolation working points are used. Single-electron triggers [138] are used to retain events for further analysis, whose p_{T} thresholds range from 24 GeV to 26 GeV. The trigger-level electron is required to match a reconstructed electron with $p_{\mathrm{T}} > 27\,\mathrm{GeV}$ and within $\Delta R = 0.07$.

Muons

Muons are reconstructed based on the information from the ID, MS, and the calorimeter system. Tracks are reconstructed inside the ID using the same algorithm as described in Section 3.4.1. In the MS, tracks are reconstructed via a search for hit patterns in each muon chamber. Muon candidates are identified by matching MS tracks to ID tracks, and performing a global χ^2 fit based on the ID and MS hits, taking into account the energy loss in the calorimeters. The muon candidate is required to fulfil the following criteria [139]:

- $p_{\mathrm{T}} > 27\,\mathrm{GeV}$,
- $|\eta| < 2.5$,
- $d_0/\sigma_{d_0} < 3$,
- $|z_0 \sin\theta| < 0.5\,\mathrm{mm}$.

Prompt muons need to be distinguished from the ones originating from pion and kaon decays. It is relatively easy to separate muons originating from charged hardons in the ID, as they can be easily identified by the poor quality of the fit of the resulting combined track. However, to separate prompt muons from the ones originating from semileptonic decays in hadronic jets, additional cuts on d_0 and z_0 are required. Unlike the electron case, the muon identification is based on a cut-based analysis [139], and a tag-and-probe method is employed to measure the efficiency of the muon identification on a sample of $Z \to \mu^+\mu^-$ and $J/\Psi \to \mu^+\mu^-$ events. Similar to electrons, muons also need to be isolated from any additional activity in the vicinity of the candidate using track- and calorimeter-based isolation variables. For the analysis presented in this thesis, for muons with $p_{\mathrm{T}} < 300\,\mathrm{GeV}$, a medium identification working point is used, and for muons with $p_{\mathrm{T}} > 300\,\mathrm{GeV}$, a specialised high-$p_{\mathrm{T}}$ working point [140] is used which provides the best momentum resolution, and helps in the removal of poorly measured high-p_{T} tracks. For muon isolation, a tight working point is utilised. A single-muon trigger [141] with p_{T} threshold ranging from 20 GeV to 26 GeV is used, and a trigger-level muon should match a reconstructed muon with $p_{\mathrm{T}} > 27\,\mathrm{GeV}$ and within $\Delta R = 0.1$.

3.4.3 Jets and their tagging

As mentioned in Section 3.3, partons hadronise to form bound states called hadrons. Hadrons deposit energy in the calorimeters. Charged hadrons also leave a track in the ID. The resulting tracks and energy depositions lead to collimated bunches of particles in the detector, called jets. However, the kinematic properties of the initial partons are challenging to determine, as the only physical object observed in the detector is the hadronic jet. Hence, an efficient jet reconstruction algorithm is needed whose aim is to obtain a single four-vector from the energy depositions in the detector components, that are as closely as possible related to the four-vector of the bunch of partons after showering. The jet reconstruction algorithm which is commonly used in ATLAS is the anti-k_{T} algorithm [142].

Anti-k_{T} jet reconstruction algorithm

The anti-k_{T} algorithm is an iterative clustering algorithm which takes topological clusters as inputs. The algorithm is based on two distance parameters: d_{ij} and d_{iB}. d_{ij} is the distance between two

objects i and j called the *inter-particle distance*. d_{iB} is the distance between the object i and beam B called the *beam distance*. These two parameters are defined as following:

$$d_{ij} = \min\left(\frac{1}{k_{\mathrm{T},i}^2}, \frac{1}{k_{\mathrm{T},j}^2}\right)\frac{\Delta R_{ij}}{R^2}, \tag{3.9}$$

$$d_{iB} = \frac{1}{k_{\mathrm{T},i}^2} \tag{3.10}$$

where ΔR_{ij} is the angular separation between objects i and j, as described in Section 3.2.1, $k_{\mathrm{T},i}$ and $k_{\mathrm{T},j}$ are the transverse momentum of the object i and j, respectively. R denotes the radius of the output jet. For an object i and j in a list of objects, the algorithm is designed as follows:

- Calculate both the distances: d_{ij} and d_{iB}.
- Pick the smallest of the two distances:
  - If the smallest distance is d_{ij}, then the objects i and j are removed from the list, and are combined into a new single object.
  - If the smallest distance is d_{iB}, then the object i is classfied as a jet, and is removed from the list.
- The above process is repeated until all the objects in the list are exhausted.

At the end of the reconstruction procedure, the jet four-momentum is obtained as the sum of the cluster four-momenta. The biggest advantage of using this reconstruction algorithm is its soft-resilient nature. It clusters the hard objects first. A hard object will develop by successively accumulating soft particles around it until it has reached a distance R from the jet axis.

Jet candidates undergo a cleaning procedure to remove the calorimeter noise, any contribution from cosmic muons, and the beam-induced backgrounds [143]. They are also required to pass through a pile-up reduction procedure [144] to ensure their orgin from the hard-scatter interaction. At the hardware level, the signal pulse shape of the LAr calorimeter provides a good protection against out-of-time pile-up. At the reconstruction level, the first cleaning step is performed at the very beginning of the clustering algorithm, where only calorimeter cells above a given noise threshold are considered [145]. The energy ratio of the clusters in the ECAL and HCAL, and the tracking information from the ID are further used to suppress any kind of pile-up. The left-over pile-up contributions to the measured energy are taken care of during the calibration procedure of various kinds of jets as described later in this section.

Small-R jets

Jets reconstructed with a radius parameter, $R = 0.4$, using an anti-k_{T} algorithm, are termed as small-R jets. In this thesis, the jet reconstruction used for these small-R jets is based on a Particle Flow (PFlow) approach [146]. This approach combines the calorimeter deposits with the tracking information provided by the ID, and has several advantages. The ID information enhances the angular resolution of the particles in comparison to the calorimeters, and helps in better supressing the pile-up. In addition, for particles with $p_{\mathrm{T}} < 100\,\mathrm{GeV}$, the momentum resolution of the tracker is better than the energy resolution of the HCAL.

The first step of reconstructing PFlow jets is that of matching the good quality tracks[9] to clusters in the calorimeter. The energy deposits of these matched clusters are replaced by their corresponding track momentum. In the next step, the anti-k_T algorithm with $R = 0.4$ is employed where the clusters and the tracks matched to the PV are used as inputs.

The energy of the jet measured in the calorimeters might not be the true energy of the particles constituting them because of the inefficiencies in the reconstruction procedure. A dedicated jet calibration procedure, in the p_T range: 20 GeV < p_T < 1 500 GeV [147], is hence performed to match the jet energy scale (JES) [148] of the small-R jets to that of the jets reconstructed from simulated stable particles (particle-level). The JES calibration procedure is shown in the schematic in Figure 3.11. It proceeds in six consecutive steps, and each step corrects the full four-momentum scaling the jet p_T, energy, and mass. Initially, the four-momentum of jets is recalculated to make it point in the direction of the the PV in case it points to the center of the detector. This process is termed as the *origin correction*. Following that, two pile-up corrections are applied namely the *jet area-based pile-up correction*, and *residual pile-up correction*. These corrections subtract the average energy due to in-time pile-up, and other residual corrections, respectively, which might depend on the number of PVs and bunch crossings. The *absolute MC-based correction* corrects the four-momentum of the jets to the particle-level energy scale, and takes care of any biases in the jet η reconstruction. The penultimate step of *global sequential correction* further improves the reconstructed energy and its related uncertainties using calorimeter, MS, and track-based variables. Finally, a *residual in-situ calibration* is performed which accounts for the differences in data and simulation. These in-situ corrections are derived from analyses which focuses on γ/Z+jets or multijet processes. The reconstructed and calibrated small-R jets with p_T > 25 GeV and $|\eta|$ < 2.5 are used for the analysis presented in this thesis.

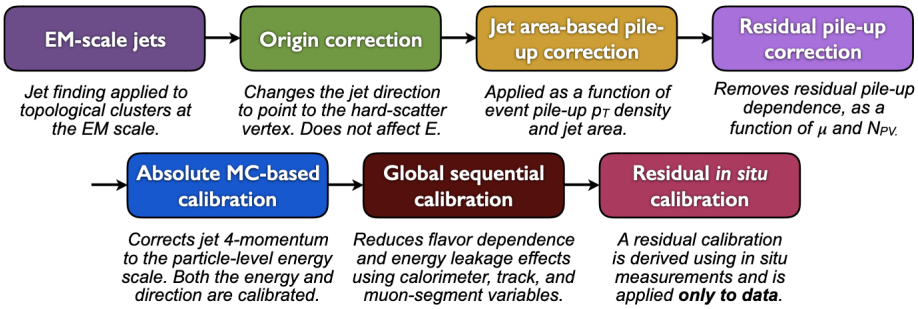


Figure 3.11: An overview of the small-R jet calibration scheme [148].

Large-R jets

The high centre-of-mass energy at the LHC is helpful in understanding the collider physics at high p_T. The large transverse momenta lead to large Lorentz boosts of the studied particles. For instance, the decay of the Higgs boson to two b-quarks at high p_T leads to a collimation of b-quarks to a point that the standard reconstruction techniques fail. The angular separation of the two decay products can be approximated as:

$$\Delta R(b_1, b_2) \approx \frac{2m}{p_T} \tag{3.11}$$

[9] High-energy tracks with p_T > 40 GeV are not used, since it is difficult to remove their associated calorimeter energy.

where b_1 and b_2 are the two b-quarks, m and p_{T} are the mass and the transverse momentum of the decaying Higgs boson (or any other particle of interest), respectively. The above expression guides us to identify a suitable radius of the Higgs jet. For a Higgs boson with $p_{\mathrm{T}} > 250\,\mathrm{GeV}$, the two b-quarks can be reconstructed as one large-R jet with radius parameter $R = 1.0$. The large-R jets can be suitably reconstructed using the anti-k_{T} algorithm. The soft-resilience nature of the anti-k_{T} algorithm is depicted in Figure 3.12 where the large-R jets have nearly perfect circular shape. Before any dedicated

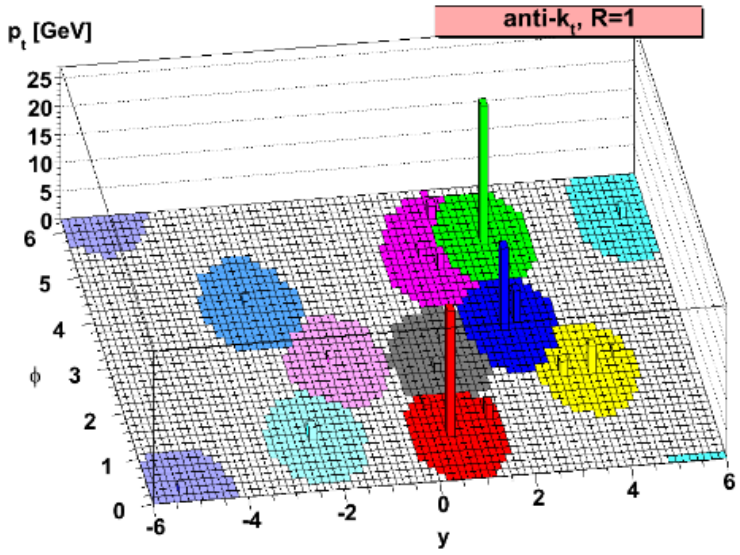


Figure 3.12: Jets reconstructed via anti-k_{T} algorithm with $R = 1.0$. The shaded regions shows the jet catchment area which depicts that the jets have a nearly perfect circular shape [149].

calibration is applied to these jets, they undergo a jet grooming procedure [150]. Jet grooming helps to remove any pile-up or soft particles unrelated to the hard-scatter event which can contaminate the jets. The grooming procedure used in this thesis is *jet trimming* [151]. This algorithm reclusters the constituents of an $R = 1.0$ jet (parent jet) using a k_{T} algorithm[10] into subjets with radius parameter, $R = 0.2$. Then, the contribution of each subjet i associated with the parent jet, J, is discarded if $p_{\mathrm{T}_i}/p_{\mathrm{T}_J} < f_{\mathrm{cut}}$, where f_{cut} is a fixed dimensionless parameter set to 0.05. The remaining constituents, after this procedure is applied, then forms the trimmed jet. The jet trimmimg technique is shown in Figure 3.13.

In principle, the mass of the large-R jet can be calculated from the four momentum of the calorimeter clusters, which is called the *calorimeter-based jet mass*, m^{calo}. Due to a worse particle resolution at high p_{T}, the angular separation of the two decay products falls below the calorimeter granularity threshold. Due to this reason, the tracking information is also taken into account which provides an

[10] The k_{T} algorithm is similar to the anti-k_{T} algorithm with modified distance parameters, $d_{ij} = \min(k^2_{\mathrm{T},i}, k^2_{\mathrm{T},j})\frac{\Delta R_{ij}}{R^2}$ and $d_{iB} = k^2_{\mathrm{T},i}$.

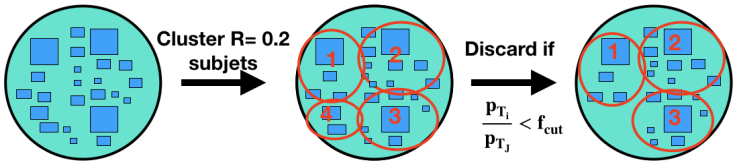


Figure 3.13: A schematic highlighting the large-R jet trimming procedure.

excellent directional resolution. The *track-assisted jet mass*, m^{TA} is expressed as:

$$m^{\mathrm{TA}} = \frac{p_{\mathrm{T}}^{\mathrm{calo}}}{p_{\mathrm{T}}^{\mathrm{track}}} \times m^{\mathrm{track}} \tag{3.12}$$

where $p_{\mathrm{T}}^{\mathrm{calo}}$ is the transverse momentum of the large-R jet, $p_{\mathrm{T}}^{\mathrm{track}}$ and m^{track} are the transverse momentum and mass components of the four momentum sum of the tracks associated with the large-R jet, respectively. The mass of the large-R jet (m^{comb}) is then evaluated as the weighted sum of m^{calo} and m^{TA}.

The JES and the jet mass scale (JMS) [152] of the large-R jets are corrected to the particle-level scale using the generator-level jet collections. These corrections are applied as a function of the reconstructed jet energy, mass, and η. Then, the in-situ correction factors are applied only to data for taking into account any residual mismodelling. These correction factors depend on the p_{T} imbalance in the relevant final states. Similar to the small-R jets, these corrections are derived in Z+jets, γ+jets, and multijet processes, and a statistical combination of these three measurements delivers an absolute calibration. A schematic depicting the large-R jet calibration procedure is shown in Figure 3.14. Large-R jets with $p_{\mathrm{T}} > 200\,\mathrm{GeV}$, $m^{\mathrm{comb}} > 30\,\mathrm{GeV}$, and $|\eta| < 2.0$ are used in this thesis.

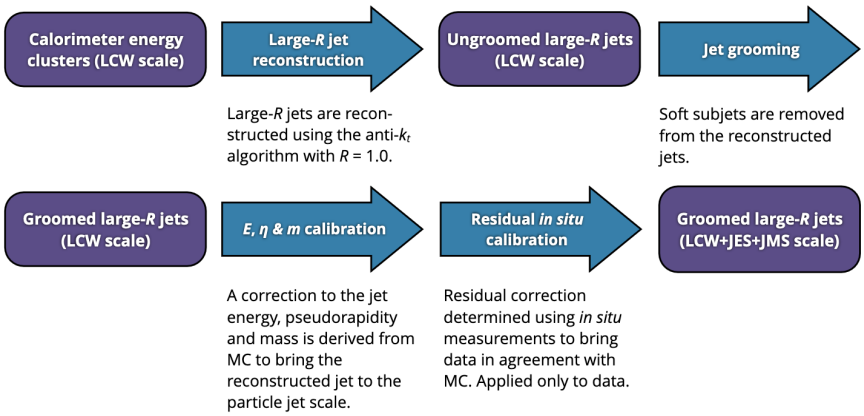


Figure 3.14: An overview of the large-R jet calibration scheme [153].

Track jets

Track jets are another useful category of jets which are reconstructed using only ID tracks originating from the hard-scatter PV. These tracks are fed as inputs to the anti-k_{T} algorithm, and must fulfil the

following requirements:

- $p_{\mathrm{T}} > 0.5\,\mathrm{GeV}$,
- $|\eta| < 2.5$,
- presence of at least seven hits in the pixel and SCT detectors. A maximum of one (two) of the expected hit(s) may be missing from the pixel (SCT) detector, and no more than one hit may be shared with other tracks,
- $|z_0 \sin\theta| < 3\,\mathrm{mm}$.

This analysis uses variable-radius (VR) track jets [154] with their radius parameter R varying with the $\frac{\rho}{p_{\mathrm{T}}}$ scale. The ρ parameter determines how fast the radius parameter decreases with the jet p_{T}. The VR track jets also take into account two additional parameters, namely R_{min} and R_{max} that defines the range of radius parameter of the jets. Outside these radius limits, the jet radius remains fixed to avoid events with jets having arbitarily small or large radii. For the presented analysis, VR track jets with $\rho = 30\,\mathrm{GeV}$, $R_{\mathrm{min}} = 0.02$ and $R_{\mathrm{max}} = 0.4$ are used [155].

In the context of this thesis, VR track-jets are used to tag the two b-hadrons inside the large-R jet in the identification of boosted $h \to b\bar{b}$ decays (see Chapter 7). Due to the dense hadronic environment, there can be some ambiguity in matching track jets to calorimeter jets. In order to resolve this, a robust matching technique, known by the name of ghost-association is used [156]. The "ghosts" are the track jets which are added to the large-R jet as an infinitesimally soft particle such that the jet reconstruction axis is not changed. The jet reconstruction anti-k_{T} algorithm is then re-run with both the calorimeter clusters and the track jets. A track jet is considered to be ghost-associated with the large-R jet if its ghost version remains contained in the large-R jet after the clustering. VR track jets which are ghost-associated with the large-R jet are required to have $p_{\mathrm{T}} > 7\,\mathrm{GeV}$, $|\eta| < 2.5$, and at least two track constituents.

Flavour-tagging

Flavour-tagging of jets refers to the identification of jets containing b-hadrons (b-jets), jets containing c-hadrons but no b-hadrons (c-jets) or jets containing neither b- nor c-hadrons (light-flavoured jets). The b-jets are of primary interest for this thesis and are characterised by the presence of a secondary vertex (SV)[11] reconstructed from displaced tracks with a large impact parameter separated from the PV, as shown in Figure 3.15. The properties of c-jets are similar to b-jets which make the distinction between the two kinds of jets a challenging task. At the same time, several other processes like photon conversions, or low-quality tracks can also give rise to displaced vertices and can make the jet classification process difficult. Hence, a suitable algorithm is needed for the b-tagging of jets.

The algorithm used for b-tagging of jets in this thesis is known as DL1r [157] which is a neural-network-based algorithm. It uses various observables related to the kinematic properties of small-R jets, properties of the SV, variables related to exclusively identify a c-jet, and the flavour-tagging probabilities of jets given by more advanced networks [158].The network produces three output nodes corresponding to the probabilities for identifying a jet as a b-, c-, or a light-flavoured jet. The training

[11] Given a long lifetime ($\approx 1.5\,\mathrm{ps}$), measurable b-hadrons have a significant flight length, leading to at least one vertex displaced from the hard-scatter interaction point - the secondary vertex.

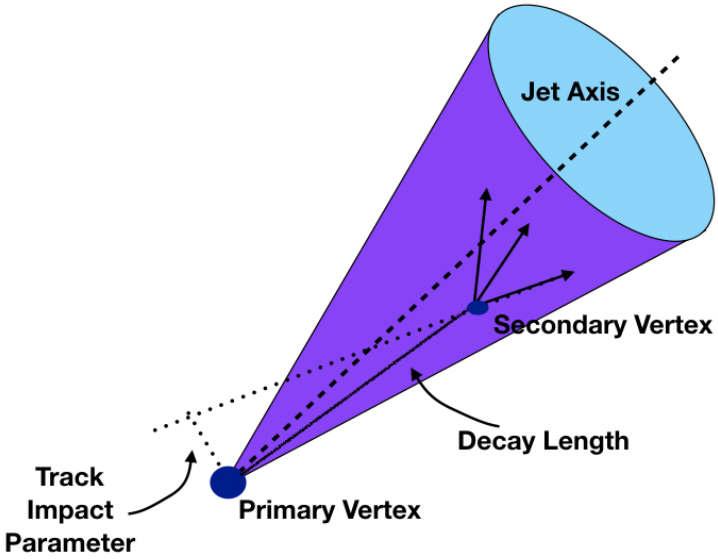


Figure 3.15: An illustration of a jet with a secondary vertex reconstructed from displaced tracks with a large impact parameter, and with a significant decay length indicating the decay of a b-hardon.

of the DL1r tagger is carried out on hybrid $t\bar{t}$, and Z' samples where jets with $p_{\mathrm{T}} < 250\,\mathrm{GeV}$ are taken exclusively from $t\bar{t}$ events, and jets with $p_T > 250\,\mathrm{GeV}$ are taken from $Z' \rightarrow$ dijet events. All the jet flavours are treated equally during the training, and hence this algorithm can be used for both b- and c-tagging. These probabilities are then combined in a final discriminant, D_{DL1r}, which is expressed as:

$$D_{\mathrm{DL1r}} = \ln\left(\frac{p_{\mathrm{b}}}{f_{\mathrm{c}} \cdot p_{\mathrm{c}} + (1 - f_{\mathrm{c}}) \cdot p_{\mathrm{light}}}\right) \tag{3.13}$$

where p_{b}, p_{c}, and p_{light} are the probabilities for tagging a jet as a b-, c-, and a light-flavoured jet, respectively. f_{c} denotes the effective c-jet fraction of the background training sample, and has been set to 0.018 according to ATLAS recommendations [158].

The performance of the DL1r tagger is evaluated in terms of the tagging efficiency of b-jets, and the background rejection of c- and light-flavoured jets. The background rejection factors are evaluated as the reciprocal value of the mistagging efficiency of the c-and light-flavoured jet. Different working points (WPs) are defined based on the cut on D_{DL1r} corresponding to different b-tagging efficiencies. The WPs vary from very tight to loose with efficiencies of 60%, 70%, 77%, and 85%. The b-tagging efficiency for a 77% WP, as a function of jet p_{T} for a DL1r tagger is shown in Figure 3.16. Subsequently, Figure 3.17 shows a substantial improvement in the c- and light-flavoured jet rejection by the DL1r tagger for 77% b-tagging efficiency WP, in comparison to their predecessors, the MV2 [159], and DL1 tagger [158].

For the presented analysis, jets tagged with a 77% WP are used, which provides a good balance between the number of accepted b-jets, and misidentified c- and light-flavoured jets for a given sample size. However, in order to further refine the identification of b-jets, a requirement combining all the available WPs can be used. One such useful observable is the pseudo-continuous b-tagging (PCBT) score which is a discretised observable consisting of five intervals. Each interval represents whether a

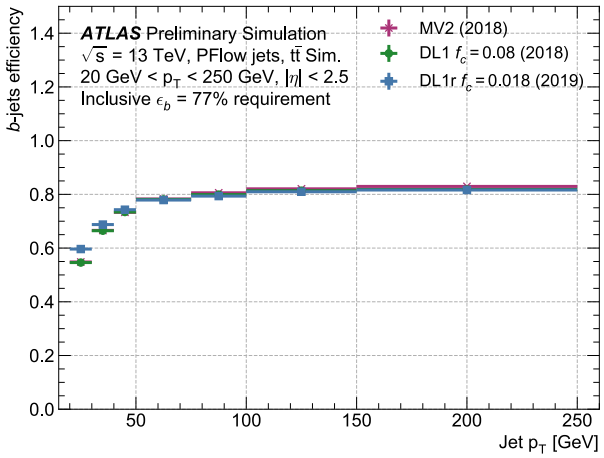
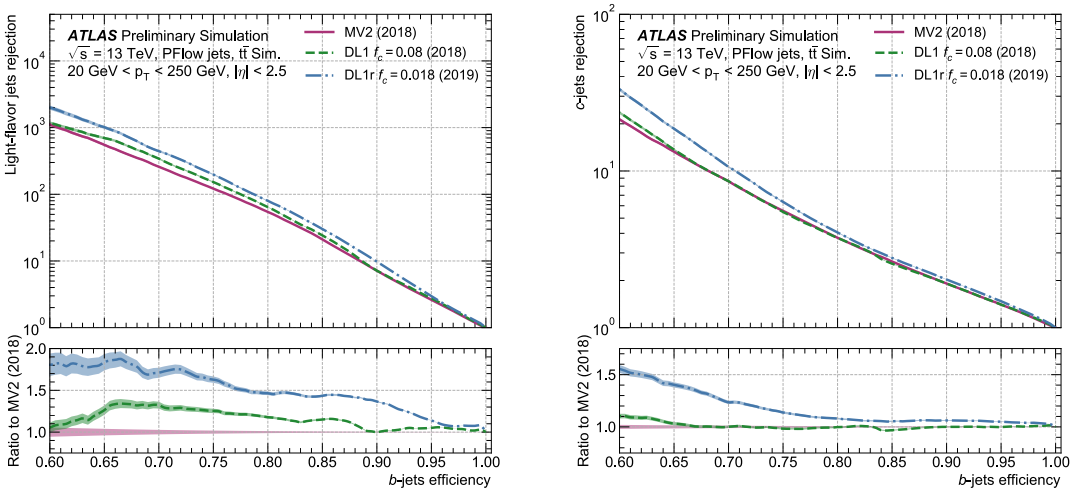


Figure 3.16: The b-jet tagging efficiency (denoted as "b-jets efficiency") as a function of jet p_{T} for various flavour-tagging algorithms [158].



(a) (b)

Figure 3.17: (a) Light-flavoured jet rejection rate and (b) c-jet rejection rate as a function of b-jet tagging efficiency (denoted as "b-jets efficiency") at the 77% efficiency WP for various flavour-tagging algorithms [158].

jet passes or fails a specific efficiency WP. The first interval correspnds to jets failing the 85% WP, the second interval corresponds to jets passing the 85% WP but failing the 77% WP, the third interval corresponds to jets passing the 77% WP but failing the 70% WP, the fourth interval corresponds to jets passing the 70% WP but failing the 60% WP, and the last interval corresponds to jets passing the 60% WP. In the context of this thesis, this observable is used in correctly identifying b-jets that can originate from a Higgs boson, described later in Chapter 5.

3.4.4 Overlap removal

The reconstruction and identification of electrons, muons, and jets is performed independently. This can lead to an ambiguity in identifying these objects correctly as they can be spatially close to each other. To get rid of this ambiguity, an overlap removal procedure is applied. This overlap removal procedure is based on the angular separation, ΔR, in the η-ϕ plane between the two reconstructed objects. Once all the objects are fully reconstructed and calibrated, the overlap removal algorithm is implemented in a sequential manner. The algorithm used in the context of this thesis is as follows [27]:

- Electron candidates that lie within $\Delta R = 0.01$ of a muon candidate, are removed to supress the contributions from the muon bremsstrahlung processes,
- The closest small-R jet within a cone of size $\Delta R = 0.2$ around an electron is removed,
- A small-R jet with fewer than three associated tracks is removed if the jet is within a cone of $\Delta R = 0.2$ around a selected muon,
- Electrons and muons are discarded if they are within a cone of size $\Delta R = \min(0.4, 0.04 + 10\ \text{GeV}/p_{\text{T}}^{\ell})$ around the axis of any surviving jet, where p_{T}^{ℓ} is the transverse momentum of the electron or muon. This requirement reduces the background contribution from semileptonic decays of heavy-flavour hadrons,
- Any large-R jet found within $\Delta R < 0.1$ of an electron is rejected.

3.4.5 Missing Transverse Momentum

According to the conservation of momentum, the vectorial sum of the transverse momentum of all particles produced in an event must be zero. A sizeable deviation from a null value can primarily indicate the presence of neutrinos which are solely weakly interacting and escape the direct detection in the detector. One can define the missing transverse momentum, $\vec{E}_{\text{T}}^{\text{miss}}$, in this regard as the negative of the vectorial momentum sum of all the reconstructed objects in the detector. The $\vec{E}_{\text{T}}^{\text{miss}}$ reconstruction is split into two contributions, namely the hard term and the soft term. The hard term includes all objects which are fully reconstructed and calibrated: electron, muons, and small-R jets, whereas the soft term includes all tracks that originate from the PV, but are not associated with any object from the hard term. The $\vec{E}_{\text{T}}^{\text{miss}}$ components, E_x^{miss} and E_y^{miss} can be expressed as [160, 161]:

$$E_{x(y)}^{\text{miss}} = -\sum_{i \in \text{hard objects}} p_{x(y),i} - \sum_{j \in \text{soft signals}} p_{x(y),j} \tag{3.14}$$

where p_x and p_y are the transverse momentum components of the objects in the transverse plane. The vector $\vec{E}_{\text{T}}^{\text{miss}}$ provides the amount of missing transverse momentum via its magnitude $E_{\text{T}}^{\text{miss}}$ and its direction in the transverse plane in terms of the azimuthal angle ϕ^{miss}, defined as $\tan^{-1}(\frac{E_y^{\text{miss}}}{E_x^{\text{miss}}})$.

3.5 Multivariate techniques

Multivariate techniques are used in the evaluation of data to explain relationships and patterns between different observables associated with this data. Within the context of particle physics, they can be used

in various analysis tasks, for example in particle identfication, energy and momentum measurements, signal to background discrimination, function fitting[12], and many more. One of the most common tasks is the identification of events (signal) that are rare, and are mimicked by wide variety of physics processes (backgrounds) or signal to background discrimination. The conventional approach of selecting events by using selection requirements on individual kinematic variables is not optimal for this task, and has been greatly surpassed by the usage of multivariate techniques. Mathematically, the underlying task to carry out this discrimination is the so-called *functional approximation* [162]. It is a technique of estimating an unknown underlying function (*mapping function*) that can map the several correlated observables or feature variables to a well defined output. These feature variables usually describe an object or an event, and can be represented by a vector $\vec{x} = (x_1, x_2, ..., x_d)$ in a d-dimensional space[13]. Given $\vec{x}$, the goal is to construct a function $y = f(\vec{x})$ such that a map of the given form can be extracted:

$$f : R^d \rightarrow R^N \tag{3.15}$$

where $N << d$, and R^m is a generic real vector space of dimension m. In practice, one tries to approximate the desired function with $\tilde{y} = f(\vec{x}, \vec{w})$, where $\vec{w}$ denotes some adjustable parameters.

The most commonly used multivariate technique using computer algorithms is machine learning, which has its origin in the field of *Artificial Intelligence* [163]. In machine learning, a functional approximation is derived automatically from the given data without having prior information about the function. One such strong approach to obtain the approximation, $f(\vec{x}, \vec{w})$, is known as supervised learning. Within the supervised learning, a training (learning) dataset comprising of feature vector or commonly called inputs, and the corresponding targets is used. This training dataset, $(y, \vec{x})$ where y are the targets from the true function $f(\vec{x})$, encapsulates the information between the input-output relation which is meant to be learnt. The training dataset contains both signal and background events, and is taken from MC simulations in the context of this thesis. The goal of training is to correctly look for parameters $\vec{w}$ of the model under consideration, that is a functional approximation of the desired input-output map.

In a problem involving functional approximation, the incured information loss has to be minimised, which is evaluated by a loss function $L(y, f(\vec{x}, \vec{w}))$. The minimisation is considered to be more efficient if one minimises the loss function averaged over the training dataset, commonly called the risk function $R(\vec{w})$. This risk function provides information about the cost of mistakes made in the predictions, and looks for the best parameters $\vec{w}$. On the other hand, if there is a possible constraint $Q(\vec{w})$ that can impact the training procedure, it is further added to the risk function leading to a *cost function* (defined in Equation 3.16) which needs to be minimised.

$$C(\vec{w}) = R(\vec{w}) + \lambda Q(\vec{w}) \tag{3.16}$$

In the above Equation 3.16, λ is an adjustable parameter which controls the model complexity and helps avoid overfitting[14]. The performance of a machine learning classifier is generally evaluated using a test dataset which is independent of the training dataset. Currently the most widely used classifiers, based on supervised learning, in high-energy physics are Boosted Decision Trees and Neural Networks.

[12] Function fitting, in this context, refers to the derivation of continuous functions of variables.

[13] When sizeable correlations exist between variables, the dimensionality of a problem is smaller than d.

[14] Overfitting is an unwanted behavior that occurs when a model gives accurate predictions for the training dataset, but not for any new data.

Boosted decision trees have been used primarily in the event reconstruction technique employed in the search for resolved $H^{\pm}$ boson decays (see Chapter 5). Neural networks are used to tag boosted Higgs bosons, and also to provide a signal to background discrimination in the search for merged $H^{\pm}$ boson decays (see Chapter 7). Thus, keeping within the scope of the author's involvement with these algorithms, the following Sections 3.5.1 and 3.5.2 describe these two classifiers with varying level of detail.

Other important types of machine learning are unsupervised and reinforcement learning. In the unsupervised approach, no targets are provided and the algorithm is expected to find connections among the input vectors. In contrast, in the reinforcement approach, correct outputs are rewarded and the incorrect ones are penalised. These two methods are not used this thesis and are hence not discussed further.

3.5.1 Boosted Decision Trees

Decision Trees (DTs) [164] are based on an algorithm that uses a sequential set of cuts such as $x_1 > z_1$, $x_2 > z_2$... where $(z_1, z_2....z_d)$ forms a cut-point in the d-dimensional feature space. This sequential process begins at a *root node*, where an optimal cut is searched for one feature variable at a time which gives the best separation between signal and background. The data is then split using this best cut thereby forming two branch nodes. A variable used in a previous node may be used again at several other nodes, while the other variables might not be used at all. The splitting process is continued recursively until a terminating condition is met. The feature space is split this way into multiple regions that are eventually classified as signal or background events, depending on the majority of the training events that end up in the final leaf node. A schematic view of a decision tree is shown in Figure 3.18.

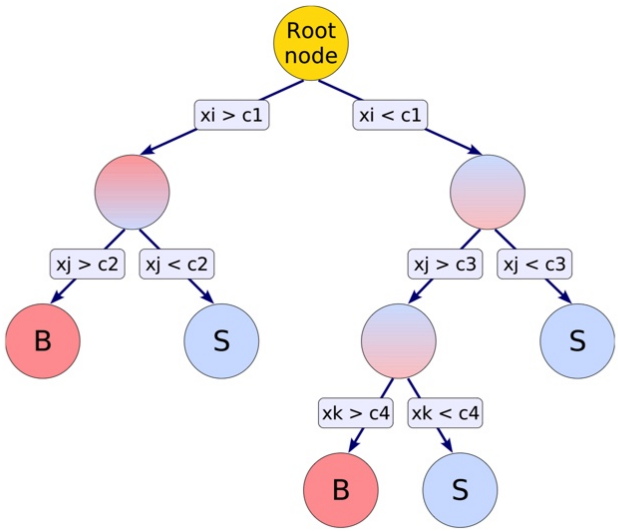


Figure 3.18: A schematic view of a decision tree [164].

`GiniIndex` is commonly used to evaluate the separation criteria (*Separation Index*) for node

splitting. It is defined as:

$$\texttt{GiniIndex} = \frac{s}{s+b} \tag{3.17}$$

where s and b are the signal and background yields, respectively, at any step of the splitting process. The splitting at a branch node is stopped if the `GiniIndex` does not change any further, or in other words when the background contamination is not further reduced. At this state, an output response of $\frac{s}{s+b}$ is assigned to the leaf node.

In general, the decision trees provide transparency to the classification procedure, and are easy to interpret. However, they can be prone to instabilities with respect to statistical fluctuations in the training dataset from which the tree structure is formed. The recursive splitting in a single decision tree leads to increasingly fewer events of the training dataset in each step, and can lead to a compromised classifier response. Also, if the two input variables show similar separation power, a fluctuation in the training dataset can favour the split on one variable over the other. In such a case the whole tree structure is altered below that node which can further lead to a different classifier response.

To solve the aforementioned problems, one tends to create an ensemble of classifiers that can collectively "boost" the performance, and can make the classification procedure less prone to fluctuations in the training dataset. These kinds of classifers are known by the name of Boosted Decision Trees (BDTs). For an ensemble of M classifiers, the predictions of the combined classifier is given as:

$$f(\vec{x}, \vec{w}) = \Sigma_{m=1}^{M} \alpha_m y_m(\vec{x}, \vec{w_m}) \tag{3.18}$$

where $\vec{w_m}$ are the parameters of the m^{th} classifier. The *weighing coefficients* α_m are determined based on the chosen boosting algorithm. For this thesis, a GradientBoost algorithm[15] is used, which tries to adjust α_m by minimising a loss function $L(y, f(\vec{x}, \vec{w}))$ of the following form:

$$L(y, f(\vec{x}, \vec{w})) = \ln(1 + e^{-2f(\vec{x}, \vec{w})y}) \tag{3.19}$$

The GradientBoost algorithm works best on classifiers having a small number of constituting nodes (tree depth). Having such small sized trees are much less susceptible to overfitting compared to simple decision trees. Their robustness can be further enhanced by reducing the learning rate of the algorithm, evaluated by the `Shrinkage` parameter which controls the weight of the individual trees. A small value of `Shrinkage` (0.1 to 0.3) would require more trees to be used in the algorithm but can help improve the accuracy of the classifier response in complex settings. Other common training parameters include `nCuts` and `MinNodeSize`. `nCuts` is the number of optimisation steps used in the splitting of decision tree nodes. `MinNodeSize` denotes the minimum number of events per leaf node, and takes care of overtraining[16]. For this thesis, BDTs are used as a binary classifier, but they can also be used efficiently in the discrimination of signal against multiple classes of background, which adds a further complexity in node-splitting.

3.5.2 Neural Networks

Another useful technique for data classification is that of a feed-forward neural network (NN) called a *multilayer percepton (MLP)* [165]. It consists of an inter-connected group of nodes or *neurons*

[15] Another commonly used algorithm is AdaBoost, which has an exponential loss function and works as efficiently as GradBoost.

[16] Overtraining refers to a scenario when a model is trained on a single set of data for too long leading to overfitting.

which are arranged in layers. A schematic of an MLP is shown in Figure 3.19. Each node processes information received by it using a transformation function, also commonly called the *activation function* [166], and then passes the result on to the next layer of nodes. The first layer is called the input layer which receives the feature variables. This layer is followed by one or more hidden layers of nodes. Each of the interconnections in the network is characterised by a weight, and each of the processing nodes may have a definite bias or a threshold. These weights and thresholds are the network parameters whose values are learnt during the training phase. The last layer of the network gives an output of the final response of the network.

There are several features that are needed for an efficient functioning of the neural networks. Since the hidden nodes are crucial in extracting relevant information from the feature variables, their number depends on the density of the underlying data. To avoid any overfitting that might arise from the number of nodes, a regularisation procedure is commonly used. Regularisation penalises the complexity of the network by adding a penalty term to the risk function. The activation functions are best chosen as non-linear functions that allow for a flexible modelling of the network. A network is trained by cycling through the training data about O (10^2 to 10^3) times, and the performance of the network is periodically tested on separate sub-sets of data.

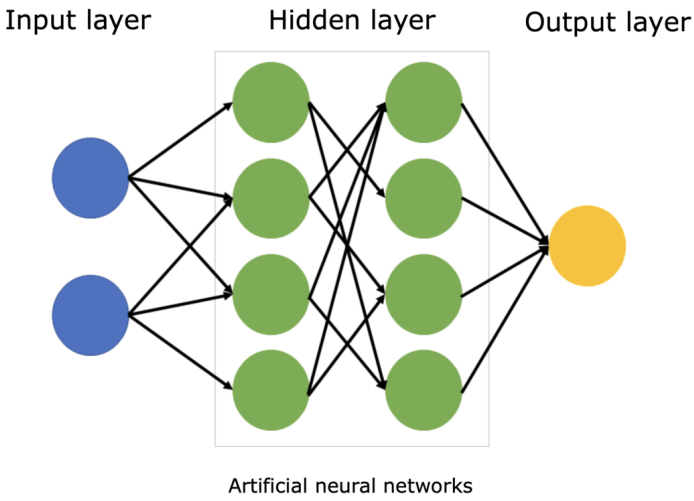


Figure 3.19: A schematic view of a feed-forward neural network [164].

3.6 Statistical methods

Data analysis techniques play an important role for the discovery reach of any particle collider experiment, and lie at the heart of *statistical inference* - i.e. deriving knowledge from data. This section focuses on the statistical methods of data analysis used in the context of this thesis. These statistical methods provide a mathematical framework to evaluate how well the collision data agree with the predictions. The presented statistical methods include the description of a likelihood model, and the role of hypothesis testing in searches for new particles.

3.6.1 The Likelihood

In particle physics, one often tries to evaluate whether some collected data is consistent with the SM predictions, known as the hypothesis H_0 (null hypothesis). Alternatively, one tries to check whether the collected data favours a new physics process in addition to the SM backgrounds - known as the hypothesis H_1 (alternative hypothesis). A hypothesis under consideration is usually modelled in terms of a Probability Density Function (PDF). The probability of observing n events under an assumed model that predicts ν events is given by a Poisson distribution [167]:

$$P(n;\nu) = \frac{\nu^n}{n!}e^{-\nu} \tag{3.20}$$

The two aforementioned hypotheses can be discriminated by constructing a function of x independent meaurements called the test statistic. The Neyman-Pearson lemma [168] states that the test statistic with the maximum separating power between the two hypotheses is a likelihood ratio. This likelihood ratio is defined as the ratio of the PDFs of the two hypotheses, $f(x|H_0)$ and $f(x|H_1)$:

$$t(x) = \frac{f(x|H_0)}{f(x|H_1)} \tag{3.21}$$

The PDFs $f(x|H_0)$, and $f(x|H_1)$ are characterised by a set of unknown parameters, θ. The deduction of these unknown parameters that best agree with the data is an important step of a statistical test, and can be addressed in a maximum-likelihood (ML) method [169]. This method is used for estimating the parameter values of a model of interest that best fit a finite sample of events, and is based on the maximisation of the likelihood function:

$$L(x;\theta) = \prod_{i=1}^{N} f(x_i,\theta) \tag{3.22}$$

where $(x_1$,...., $x_N)$ are N statistically independent measurements, and $f(x_i,\theta)$ is a hypothesised PDF of the model of interest. Generally, the negative logarithm of the likelihood (NLL) is considered which is defined as:

$$-\ln(L(x;\theta)) = -\sum_{i=1}^{N} \ln(f(x_i,\theta)) \tag{3.23}$$

The negative logarithmic transformation is easier to handle as one tends to search for a minimum rather than a maximum of $f(x_i,\theta)$ which can be rapidly fluctuating. The best-fit values of the unknown parameters can then be derived by solving:

$$-\frac{\partial \ln(L(x;\hat{\theta}))}{\partial\theta_i} = 0 \tag{3.24}$$

where $\hat{\theta}$ are the best-fit values for which the Likelihood reaches a minimum.

The unknown parameters of the hypothesis under test can be divided into two categories, namely the parameters of interest and the nuisance parameters. The former, as the name suggests, are the parameters which govern the interpretation of a statistical result. For the search presented in this thesis, the signal strength, μ, is taken as the parameter of interest (POI) which is defined as

$\sigma(pp \to tbH^{\pm}) \times BR(H^{\pm} \to W^{\pm}h(\to b\bar{b}))$, where $\sigma(pp \to tbH^{\pm})$ is the production cross section of the $tbH^{\pm}$ process and $BR(H^{\pm} \to W^{\pm}h(\to b\bar{b}))$ is the branching ratio of $H^{\pm} \to W^{\pm}h(\to b\bar{b})$ decay. The scenarios where $\mu = 0$ refers to a SM-only hypothesis, whereas the cases where $\mu = 1$ corresponds to an existence of an $H^{\pm}$ boson with a production cross section of 1 pb with a 100% branching ratio of the $H^{\pm} \to W^{\pm}h(\to b\bar{b})$ decay. The nuisance parameters (NPs), $\vec{\alpha}$, form the second category, and are the statistical and systematic uncertainties (see Chapter 6) associated with the simulated processes that are needed to define the model. They provide additional degrees of freedom and are incorporated in the statistical model with a specific form of probability distribution i.e. they are said to be constrained in the model in a specific way. To test a hypothesised value of μ, and in parallel infer the best-fit values of $\vec{\alpha}$ for the said value of μ, one considers a *profile-likelihood ratio (PLR)*:

$$\lambda(\mu) = \frac{L(\mu, \hat{\hat{\vec{\alpha}}})}{L(\hat{\mu}, \hat{\hat{\vec{\alpha}}})} \tag{3.25}$$

where $\hat{\hat{\vec{\alpha}}}$ denotes the value of $\vec{\alpha}$ that maximises L for a specified value of μ. The numerator, $L(\mu, \hat{\hat{\vec{\alpha}}})$, hence is a *conditional likelihood function* of $\vec{\alpha}$ as a function of μ. The denominator, $L(\hat{\mu}, \hat{\hat{\vec{\alpha}}})$, on the other hand corresponds to an *unconditional likelihood function* where $\hat{\mu}$, and $\hat{\hat{\vec{\alpha}}}$ are the values obtained from maximising the likelihood.

A suitable approach of using PLR is by defining a test statistic, $t(\mu) = -2\ln(\lambda(\mu))$. The test statistic is modified depending on different ranges of $\hat{\mu}$:

$$t(\mu) = \begin{cases} -2\ln \frac{L(\mu, \hat{\hat{\alpha}}(\mu))}{L(0, \hat{\hat{\alpha}}(0))}, & \hat{\mu} < 0 \\ -2\ln \frac{L(\mu, \hat{\hat{\vec{\alpha}}})}{L(\hat{\mu}, \hat{\hat{\vec{\alpha}}})}, & 0 \leq \hat{\mu} \leq \mu \\ 0, & \hat{\mu} > \mu \end{cases} \tag{3.26}$$

If $\hat{\mu} < 0$, the test statistic $t(\mu)$ is modified in a way such that the parameters in the denominator optimise the likelihood for $\mu = 0$. If $\hat{\mu} > \mu$, the test statistic is set to zero, as a presence of signal below an observed measurement is in complete agreement with the hypothesis. A higher value of $t(\mu)$ corresponds to an increasing incompatibility between data and hypothesis. In order to begin with a statistical test, the distribution of $t(\mu)$ need to be determined. For $t(\mu)$ being Gaussian-distributed, it can be approximated with an *asymptotic formula*. Within an asymptotic approximation, the estimation of the spread of the distribution is based on the generation of an artificial dataset called an Asimov dataset [170]. The Asimov dataset is defined in a way that when it is used to evaluate the best-fit values of the model parameters, the true parameter values are obtained. These datasets are conveniently used in estimating the test statistic distribution for large numbers of events, and wherever a Gaussian approximation is valid. However, a more accurate approach of estimating $t(\mu)$ can be achieved by the usage of pseudo-datasets generated directly from the PDF. This method samples the true distribution of the test statistic commonly known as *toy Monte Carlo sampling*. This method is most powerful for determining the test statistics for a small number of events[17] and is limited by the size of generated pseudo-datasets which is a computationally intensive task.

[17] Currently, the ATLAS recommendation is to use an asymptotic approximation for the cases where all the bins of the studied kinematic distribution contain at least five background events. Failing this requirement, a toy-based method is employed.

3.6.2 Hypothesis Testing

The two aforementioned hypotheses, H_0 and H_1, can be placed readily in the context of a classical example of search for a new particle. In a generic new physics search, the null hypothesis is usually defined as a scenario where no new particle is found - i.e. the existence of only previously known SM background processes. Resultantly, H_0 can be termed as the background-only hypothesis. On the other hand, the alternative hypothesis, H_1 used for setting limits on the POI, is defined as the signal-plus-background hypothesis. This hypothesis indicates the presence of a new particle in addition to the known SM background processes. The level of agreement between the observed data and a given hypothesis H is determined with the help of a test statistic as defined in Equation 3.26, and can be quantified by a p-value. This p-value gives the probability under the assumption of H being either H_0 or H_1, of observing a result of equal or greater incompatibility with the predictions of H. Often the p-value is replaced by an equivalent Gaussian significance, Z, which is defined as the quantile[18] of 1 - p:

$$Z = \Phi^{-1}(1 - p) \tag{3.27}$$

In the above equation, a Gaussian-distributed variable that is around Z standard deviations above its mean has an upper-tail probability equal to p. In high-energy particle physics, a signal discovery is claimed with a significance of at least $Z = 5$, also known as the 5σ significance level, which corresponds to $p = 2.87 \times 10^{-7}$. An evidence for new physics is announced at a 3σ significance level.

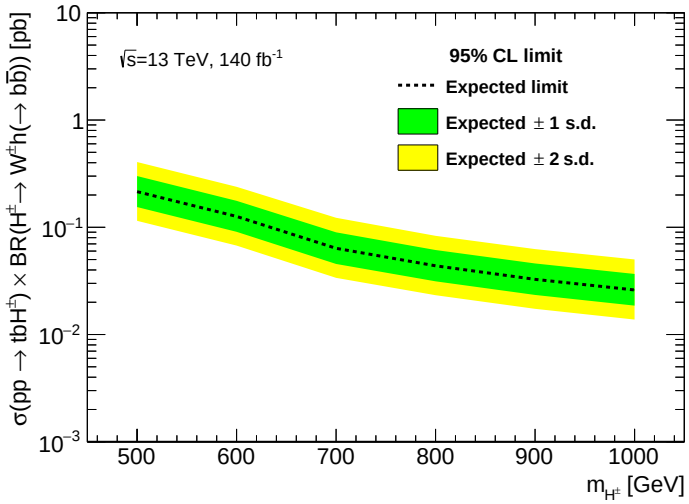


Figure 3.20: Representative 95% CL expected upper limits on $\sigma(tbH^{\pm}) \times BR(H^{\pm} \rightarrow W^{\pm}h(\rightarrow b\bar{b}))$ as a function of $m_{H^{\pm}}$.

When trying to exclude a hypothesis, usually in absence of strong evidence of new physics, a search can be used to constrain the parameters of H_1, the signal-plus-background hypothesis. In the context of the search presented in this thesis, the signal strength, μ, of the $H^{\pm}$ boson process is constrained with respect to H_0, the background-only hypothesis. Upper limits on μ are set at a fixed confidence level, CL = 1 - p. This confidence level is pre-defined, and is generally set to 95% for most

[18] A quantile is the inverse of the cumulative distribution of a unit Gaussian.

of the particle physics experiments, which corresponds to a Z value of 1.64 (p-value of 0.05). The confidence level method (CL_s method) [171] protects the statistical interpretation of any search results from pathologies which might result from a downward fluctuation of the background leading to a false signal evidence. CL_s is calculated as:

$$CL_s = \frac{CL_{s+b}}{1 - CL_b} \tag{3.28}$$

where CL_{s+b} is the confidence level computed for a signal-plus-background hypothesis, and CL_b corresponds to the confidence level computed for a background-only hypothesis. Following Equation 3.28, all $\mu > \mu_{CL_s=0.05}$ are excluded at a confidence level of 95%. To determine whether the observed data is in agreement or not with the background-only hypothesis, the observed limits are overlayed with the expected limits. These expected limits are the median limits calculated based on the Asimov datasets generated for the signal-plus-background hypothesis against the background-only hypothesis. These limits are generally shown in a so-called "brazillian format" with the green and yellow shaded bands indicating the limit range where the 68% (1σ) and 95% (2σ) of the results are expected if the limit determination is repeated fairly multiple times with an assumption that the background-only hypothesis is true. This brazillian format is depicted in Figure 3.20 which shows representative 95% CL expected upper limits on $\sigma(tbH^{\pm}) \times BR(H^{\pm} \rightarrow W^{\pm}h(\rightarrow b\bar{b}))$ as a function of $H^{\pm}$ boson mass ($m_{H^{\pm}}$).

CHAPTER 4

Resolved Topology: Background modelling and event classification

4.1 Introduction

The search for charged Higgs bosons[1] produced in association with a top- and a bottom-quark ($pp \to tbH^+$), and decaying via $H^+ \to Wh(\to b\bar{b})$ targets events with multiple jets, and leptons from the top-quark or W boson decays i.e. $l\nu q\bar{q}b\bar{b}b\bar{b}$ final states. The search covers two topologies: the resolved one, where the H^+ boson's hadronic decay products are reconstructed individually as small-R jets, and the merged one, where ΔR between the hadronic decay products of the h/W boson is small due to the large p_{T} of the H^+ boson, such that the decay products are reconstructed inside a large-R jet. This and the next chapter detail the development of the analysis strategy used in the search for resolved $H^+ \to Wh(\to b\bar{b})$ decays. A brief discussion of the search for merged $H^+ \to Wh(\to b\bar{b})$ decays is given in Chapter 7.

The main background in this search is the top-quark pair production in association with jets ($t\bar{t}$ + jets) [172]. The sub-leading background includes single top-quark and weak boson processes [173, 174]. The $t\bar{t}$ + jets background is known to be mismodelled due to missing higher-order QCD and EW corrections [175], and is corrected using a dedicated reweighting procedure. The remaining backgrounds are derived solely from simulations. The search targets two H^+ boson decay channels based on the decay products of the W boson from the H^+ boson decay: $H^+ \to Wh \to \ell\nu b\bar{b}$ decays ($\ell\nu b\bar{b}$ channel) and $H^+ \to Wh \to q\bar{q}b\bar{b}$ decays ($q\bar{q}b\bar{b}$ channel). The leading-order Feynman diagrams representing these channels are shown in Figure 4.1. A selection requirement on the leptonic top-quark[2] candidate mass ($m_{\mathrm{top}}^{\mathrm{lep}}$) is used to classify events into two channels. BDTs are trained to identify the correct pairing of jets to reconstruct the h boson, W boson, and H^+ boson candidates. Control and signal regions are then defined based on the BDT score. The reconstructed charged Higgs boson mass (m_{Wh}) is used as the main discriminant between signal and backgrounds, and is used for signal extraction.

[1] A notation of H^+ is used henceforth, with the charged-conjugate H^- always implied. Also, a generic notation for particles produced in association with charged Higgs bosons and in their decays (W, ℓ, q etc.) are used henceforth for simplicity.

[2] In principle, W boson from top-quark ($t \to Wb$) decays leptonically and the top-quark itself decays semileptonically. However, the terms "leptonic top-quark" or "leptonically decaying top-quark" are used in this thesis for simplicity and are meant to imply the same meaning.

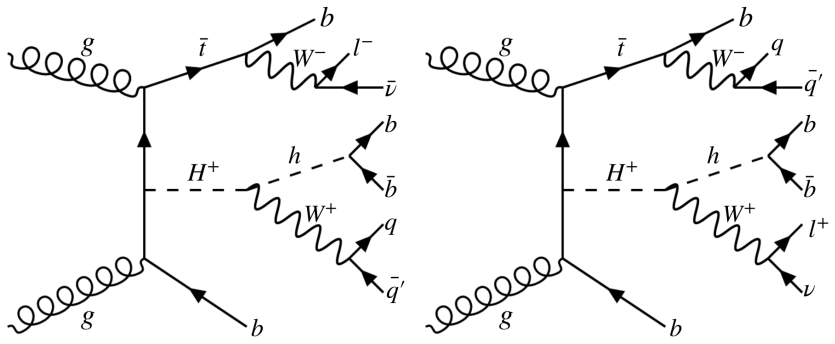


Figure 4.1: Leading-order Feynman diagrams for (a) the $q\bar{q}b\bar{b}$ channel and (b) the $\ell\nu b\bar{b}$ channel.

This chapter is structured as follows: Section 4.2 summarises the various kinds of simulated processes used in this analysis. Section 4.3 gives an overview of the event pre-selection requirements. Section 4.4 then discusses in detail the treatment of the dominant background in this search: $t\bar{t}$ + jets. Finally, Section 4.5 explains the reconstruction and performance of the $m_{\text{top}}^{\text{lep}}$ to classify events into $\ell\nu b\bar{b}$ and $q\bar{q}b\bar{b}$ channels. The remaining aspects of the analysis strategy covering the reconstruction BDTs, and the optimisation of signal and control regions are discussed in Chapter 5.

4.2 Simulated processes

This section summarises the simulated signal and background processes used in the analysis, and gives an overview of the used event generators. A complete description of the generation parameters of the simulated processes can be found in [176].

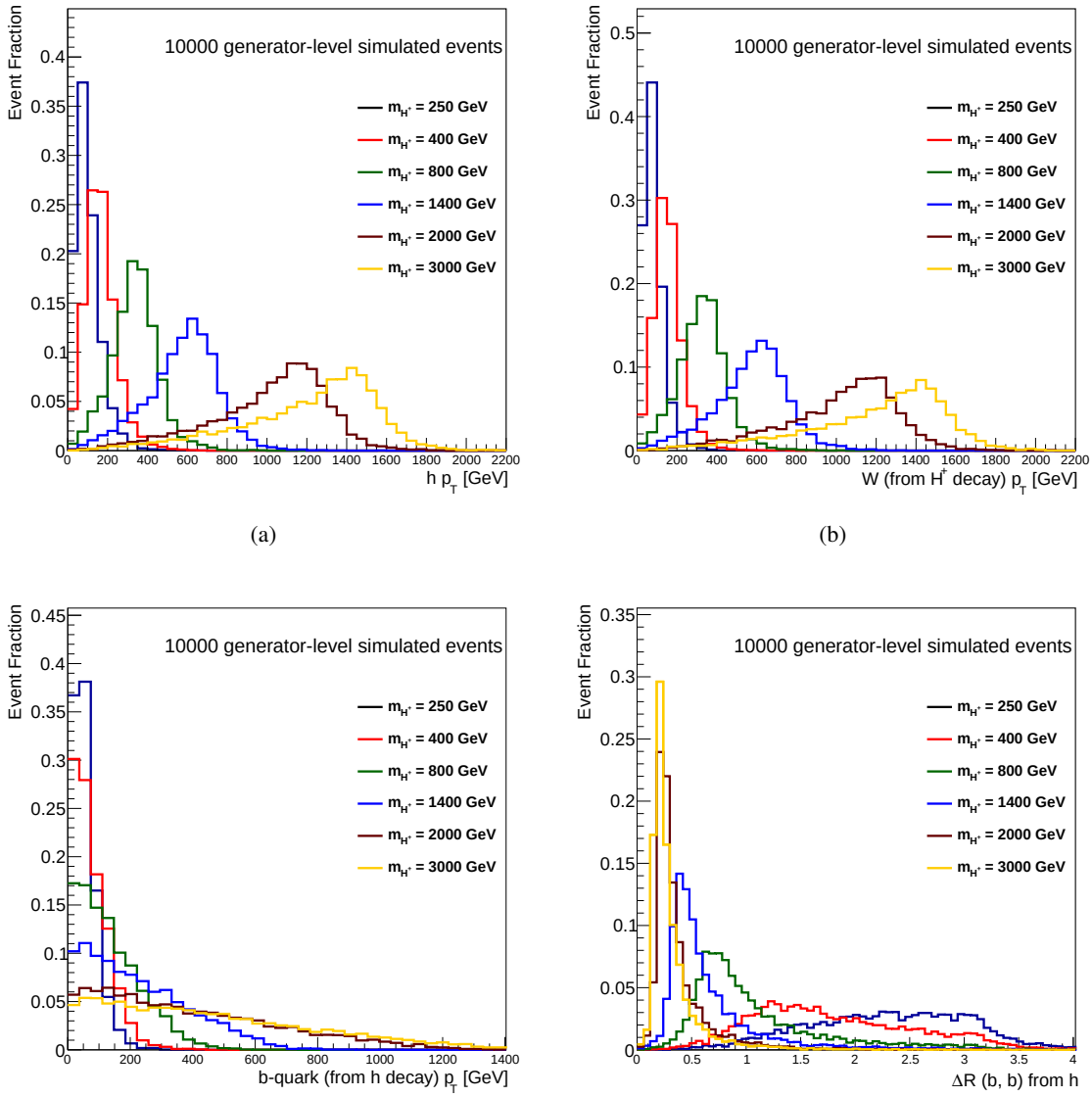
4.2.1 Signal

The signal process was simulated using MadGraph 5 [177], at NLO precision in QCD, for modelling the matrix-element (ME) . In addition, Pythia 8.244 [178] was used to model the parton showers (PS) and hadronisation. A narrow-width approximation [179] was used for the simulation of the tbH^+ process. This assumption has a negligible impact on the analysis, as the experimental resolution is much larger than the H^+ boson natural width. Seventeen signal mass hypotheses covering a range from 250 GeV to 3 TeV were simulated. W boson was allowed to decay both hadronically and leptonically. However, only Higgs boson decays to a pair of b-quarks were considered assuming a SM Higgs boson of mass m_h = 125 GeV. Simulated events of different signal mass hypotheses are normalised to a reference cross section of 1 pb for the purpose of analysis optimisation.

Figure 4.2 shows the generator-level[3] event distributions for p_{T} of the h boson, W boson, b-quarks from the h boson, and ΔR between the two b-quarks from the h boson, overlayed for several signal mass hypotheses. From these distributions, one can observe that the shapes of the p_{T} distribution for the h and W bosons get increasingly broader for higher signal mass (m_{H^+}) values, with the mean of the distribution peaking approximately at $\frac{m_{H^+}}{2}$ implying an equal split of p_{T} between the two decay

[3] Throughout this thesis, generator-level events correspond to events without detector simulation.

candidates of the H^+ boson in an ideal case. The p_{T} distribution of the two b-quarks from the h boson also gets smeared out with higher m_{H^+} due to the large Lorentz boost of the h boson. The same feature also explains why the shape of the $\Delta R(b, b)$ distribution takes increasingly lower values for higher m_{H^+} values due to the collimation of two b-quarks from the h boson.



(a) (b)

(c) (d)

Figure 4.2: The distributions of (a) p_{T} of the h boson, (b) p_{T} of the W boson, (c) p_{T} of the two b-quarks originating from the h boson, and (d) ΔR between the two b-quarks originating from the h boson at generator-level.

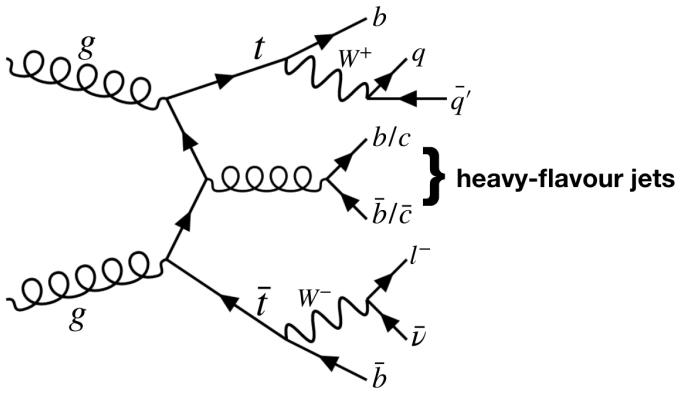


Figure 4.3: Leading-order Feynman diagram for a representative $t\bar{t}$ + HF process.

4.2.2 Backgrounds

$t\bar{t}$ + jets

The production of events containing top-quark pairs ($t\bar{t}$) were performed using the Powheg Box v2 generator [180] for modelling the ME at NLO precision in QCD. The PS, hardonisation, and underlying event were modelled using Pythia 8.230 [181]. These events are normalised to the predicted cross sections calculated with `Top++ 2.0` [182] at next-to-next-to-leading order (NNLO) in QCD including the resummation of next-to-next-to-leading logarithmic (NNLL) soft gluon terms.

The background events are categorised depending on the flavour of the jets produced in association with the $t\bar{t}$ system. The categorisation scheme is based on the generator-level information, where jets are reconstructed from stable particles with a mean lifetime $\tau > 3 \times 10^{-11}$ s using the anti-k_{T} algorithm with radius parameter $R = 0.4$. These jets are required to have $p_{\mathrm{T}} > 15$ GeV and $\eta < 2.5$, and their flavour is determined by counting the number of b- or c-hadrons with $p_{\mathrm{T}} > 5$ GeV located within a cone of $\Delta R = 0.4$ around the jet axis. Jets containing at least one b-hadron are labelled as b-jets, whereas jets that do not contain any b-hadron, but contains at least one c-hadron are labelled as c-jets. Events that contain at least one b- or c-jet, but do not contain any additional jets including b- or c-hadrons from the top-quark, W, Z, or h boson decays are termed as $t\bar{t}$ + HF where HF denotes the heavy-flavour jets. Events that do not contain any heavy-flavour jets, except from those arising from the top-quark, or W boson decays are termed as $t\bar{t}$ + LF where LF denotes the light-flavour jets. $t\bar{t}$ + HF and $t\bar{t}$ + LF are two separate background categories in this analysis. The leading-order Feynman diagram for a representative $t\bar{t}$ + HF process is shown in Figure 4.3.

Single Top

The associated production of top-quarks with W bosons (Wt), and the s- and t-channel production of single top-quarks were modelled using the Powheg Box v2 generator with NLO precision in QCD, interfaced to Pythia 8.230 to simulate the PS. The *diagram removal (DR)* scheme [183] was used to remove the interference and overlap with the $t\bar{t}$ events that can occur at NLO precision in QCD. Other processes including top quarks such as the tWZ and tZq processes are also considered and included in the single top category of backgrounds. These backgrounds were generated using MadGraph 5 at NLO precision in QCD, interfaced to Pythia 8.210 to simulate the PS.

Diboson (VV) and V + jets

The production of a $V = W$ or Z boson in association with jets was modelled using the Sherpa v2.2.11 generator [120] for both ME and PS. The ME and PS matching and merging at NLO is performed using the MEPS@NLO formalism [184].

Diboson production (VV with $V = W$ or Z) with V decays into semileptonic final states were simulated using Sherpa 2.2.1 [120], whereas events with decays into fully leptonic final states were simulated using Sherpa 2.2.2 [120]. Both these kinds of simulated processes include off-shell effects and Higgs boson contributions wherever appropriate.

$t\bar{t}$ + X (W/Z/h)

The production of $t\bar{t}$ events in association with a h boson ($t\bar{t}+h$) was carried out using the Powheg Box v2 generator at NLO precision in QCD for modelling the ME. Pythia 8.230 [181] was used in addition to simulate the PS. The $t\bar{t}+W$ and $t\bar{t}+Z$ events were modelled using MadGraph 5 at NLO precision in QCD for generating the ME. Pythia 8.244 was used in addition to simulate the PS.

Other rare top-quark processes

Rare top-quark processes such as $thjb$ and $t\bar{t}t\bar{t}$ production were also considered as a separate background category. These rare events were simulated using MadGraph 5 at NLO precision in QCD to generate the ME. In addition, they were interfaced to Pythia 8.230 to simulate the PS.

All simulated signal and background processes include the effect of pile-up by overlaying simulated miniumum-bias events on each generated signal or background event. The minimum-bias events[4] were simulated with the single-, double-, and non-diffractive proton-proton processes of Pythia 8.186 [185]. The generated events were processed using either the Geant4-based full detector simulation (`Full Sim`), or using a fast simulation (`AtlFast II`) (see Section 3.3) where the calorimeter response is replaced by a detailed parametrisation of shower shapes. The simulated events were reconstructed using the same algorithms as used for the collision data events.

4.3 Event pre-selection

A common set of pre-selection requirements on the reconstructed physics objects (see Section 3.4) are applied to both simulated events and data. The selection requirements are defined as follows:

- Exactly one prompt $e^{\pm}$ or $\mu^{\pm}$,
- $E_{\mathrm{T}}^{\mathrm{miss}} > 30\,\mathrm{GeV}$,
- $\geq$ five small-R jets,
- $\geq$ two b-tagged small-R jets passing the 77% working point of the DL1r tagger.

[4] Minimum-bias events are inelastic events selected by a loose trigger with as little bias as possible. They are a factor of two less dense than the underlying event.

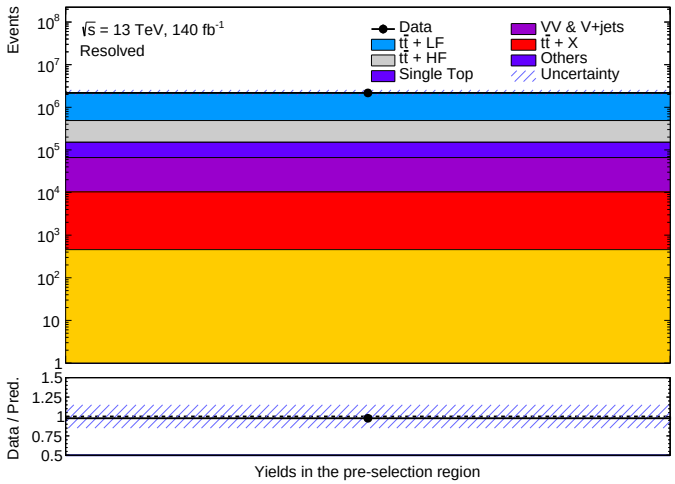


Figure 4.4: The background composition of the events passing the pre-selection requirements. The background uncertainty band shows the statistical and systematic uncertainties added in quadrature.

4.4 Background estimation

$t\bar{t}$ + jets ($t\bar{t}$ + LF and $t\bar{t}$ + HF) is the dominant background in this search as shown in Figure 4.4, thus it is important to model this background well. The modelling of the $t\bar{t}$ + jets background by the Powheg Box generator is limited to NLO precision in QCD [186] for the ME. The missing higher-order QCD and EW corrections in simulation are visible in particular at high top-quark momenta and for large jet multiplicities (n_{jets}). $H_{\text{T}}^{\text{all}}$, defined as the scalar sum of the transverse momentum of all visible objects in the detector, is an important observable in this regard. It is defined as:

$$H_{\text{T}}^{\text{all}} = \Sigma_{i=1}^{n_{\text{jets}}} p_{\text{T}}^{\text{jet}} + p_{\text{T}}^{\text{lep}} \tag{4.1}$$

Figure 4.5 presents the $H_{\text{T}}^{\text{all}}$ and n_{jets} distributions in events passing the pre-selection requirements, and shows the potential mismodelling due to the $t\bar{t}$ + jets background. The $H_{\text{T}}^{\text{all}}$ observable encodes the hardness of an event, thus serves as a suitable candidate to account for the $t\bar{t}$ + jets mismodelling. To handle the mismodelling, the $t\bar{t}$ + jets background is corrected using a reweighting procedure in order to match data in the relevant kinematic regions. This reweighting procedure corrects the shape of the $t\bar{t}$ + jets background. There are residual normalisation effects which are visible after the corrections are applied, as discussed in the following section. The remaining background processes (15% of the total background yield in the pre-selection) are derived purely from simulations.

4.4.1 Reweighting technique for the $t\bar{t}$ + jets background

The correction strategy presented in this section is inspired from other ATLAS analyses with $t\bar{t}$ + jets events as the main background [20, 23]. The mismodelling of the $t\bar{t}$ + jets background is assumed to be due to additional radiations in the PS, which are independent of the flavour of the associated jets. Signal events are expected to have a higher number of b-tagged jets, thus the correction factors are derived in a signal-depleted region by requiring exactly two b-tagged jets. These correction factors

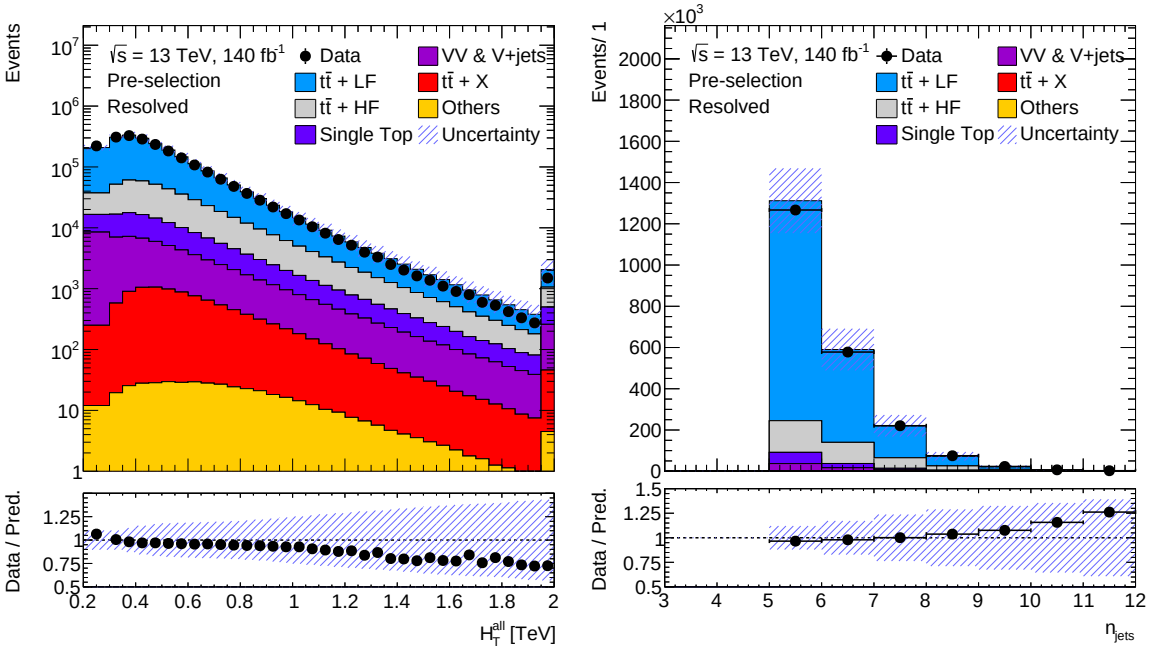


Figure 4.5: $H_{\mathrm{T}}^{\mathrm{all}}$ and n_{jets} distributions in events passing the pre-selection requirements. The ratio panels show the ratio of the observed data and the SM background prediction. The background uncertainty band shows the statistical and systematic uncertainties added in quadrature.

are subsequently applied in events containing three or more b-tagged jets. The correction factors are derived as a function of $H_{\mathrm{T}}^{\mathrm{all}}$ for different jet multiplicities n_{jets}, of five, six, seven, and at least eight. A correction factor, $C(H_{\mathrm{T}}^{\mathrm{all}})$, is defined as:

$$C(H_{\mathrm{T}}^{\mathrm{all}}) = \frac{N_{\mathrm{data}}(H_{\mathrm{T}}^{\mathrm{all}}) - N_{\mathrm{non-t\bar{t}}}(H_{\mathrm{T}}^{\mathrm{all}})}{N_{\mathrm{t\bar{t}}}(H_{\mathrm{T}}^{\mathrm{all}})} \tag{4.2}$$

where $N_{\mathrm{data}}(H_{\mathrm{T}}^{\mathrm{all}})$, $N_{\mathrm{t\bar{t}}}(H_{\mathrm{T}}^{\mathrm{all}})$, and $N_{\mathrm{non-t\bar{t}}}(H_{\mathrm{T}}^{\mathrm{all}})$ denote the yields observed in data, and the predictions for the $t\bar{t}$ + jets, and non-$t\bar{t}$ backgrounds, respectively.

Different functional forms are tested to model the shape of $H_{\mathrm{T}}^{\mathrm{all}}$, namely hyperbola-plus-sigmoid, exponential-plus-sigmoid, and a combination of second-order polynomial and a first-order exponentiated polynomial. The definitions of these functional forms are given in Equations 4.3, 4.4 and 4.5, respectively.

$$f(H_{\mathrm{T}}^{\mathrm{all}}) = a + \frac{b}{(H_{\mathrm{T}}^{\mathrm{all}})^{c}} - \frac{d}{1 + \exp(e - f \cdot H_{\mathrm{T}}^{\mathrm{all}})} \tag{4.3}$$

$$f(H_{\mathrm{T}}^{\mathrm{all}}) = a + b \cdot \exp(c \cdot H_{\mathrm{T}}^{\mathrm{all}}) - \frac{d}{1 + \exp(e - f \cdot H_{\mathrm{T}}^{\mathrm{all}})} \tag{4.4}$$

$$f(H_{\mathrm{T}}^{\mathrm{all}}) = a + b \cdot H_{\mathrm{T}}^{\mathrm{all}} + c \cdot (H_{\mathrm{T}}^{\mathrm{all}})^{2} + d \cdot \exp(e \cdot H_{\mathrm{T}}^{\mathrm{all}}) \tag{4.5}$$

Out of the three tested functional forms, the exponential-plus-sigmoid functional form is chosen as

the most suitable functional form. This choice is based on the goodness of the analytical fit describing the data points, and the non-divergent behaviour of the error variations associated with each functional form. These error variations are derived by varying each fit parameter, one at a time, within their statistical uncertainties and re-evaluating the functional form fits. Analytical fits to the correction factor for different jet multiplicity regions, using an exponential-plus-sigmoid functional form are shown in Figure 4.6. The corresponding analytical fits for the other two functional forms are shown in Figures A.1 and A.2 in Appendix A.3.

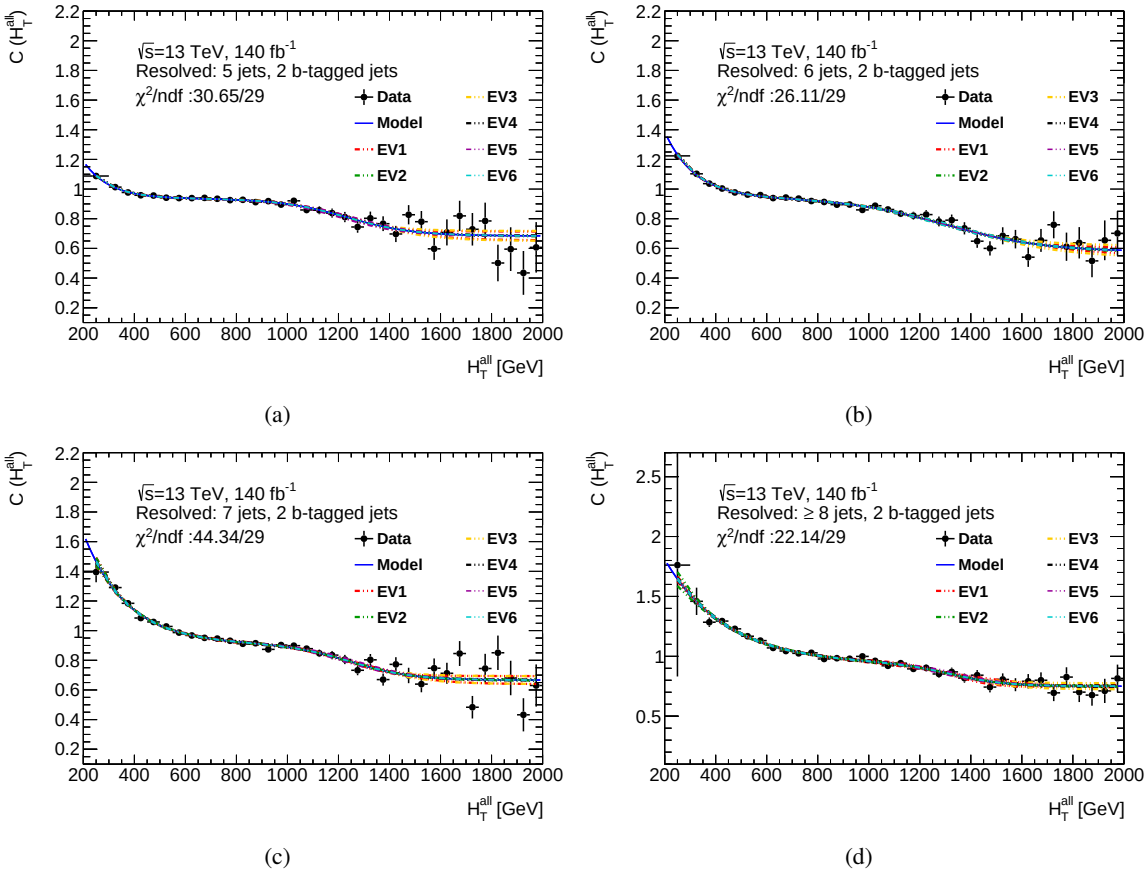
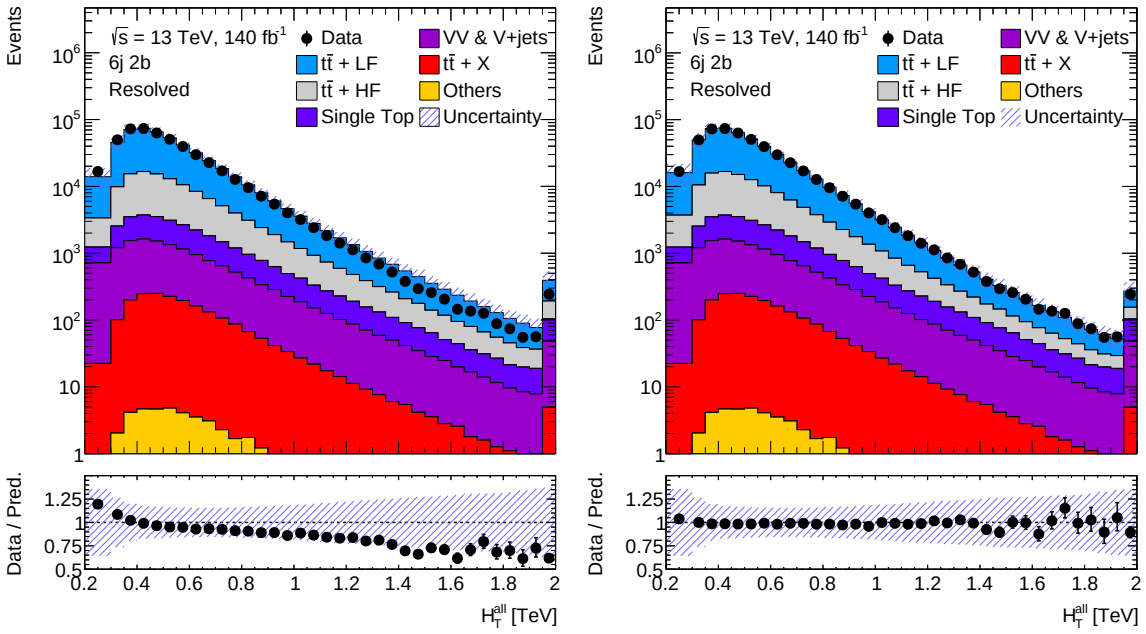


Figure 4.6: (a) to (d) shows the analytical fits to the correction factor, C($H_{\mathrm{T}}^{\mathrm{all}}$), in bins of $H_{\mathrm{T}}^{\mathrm{all}}$ modelled by an exponential-plus-sigmoid functional form for different jet multiplicity regions (a: five jets, b: six jets, c: seven jets, d: at least eight jets) for events containing two b-tagged jets. "ndf" here refers to the number of degrees of freedom which is the difference between the number of fitted bins and the number of functional form parameters. "EV" denotes the error variation (uncertainty) associated with each fit parameter.

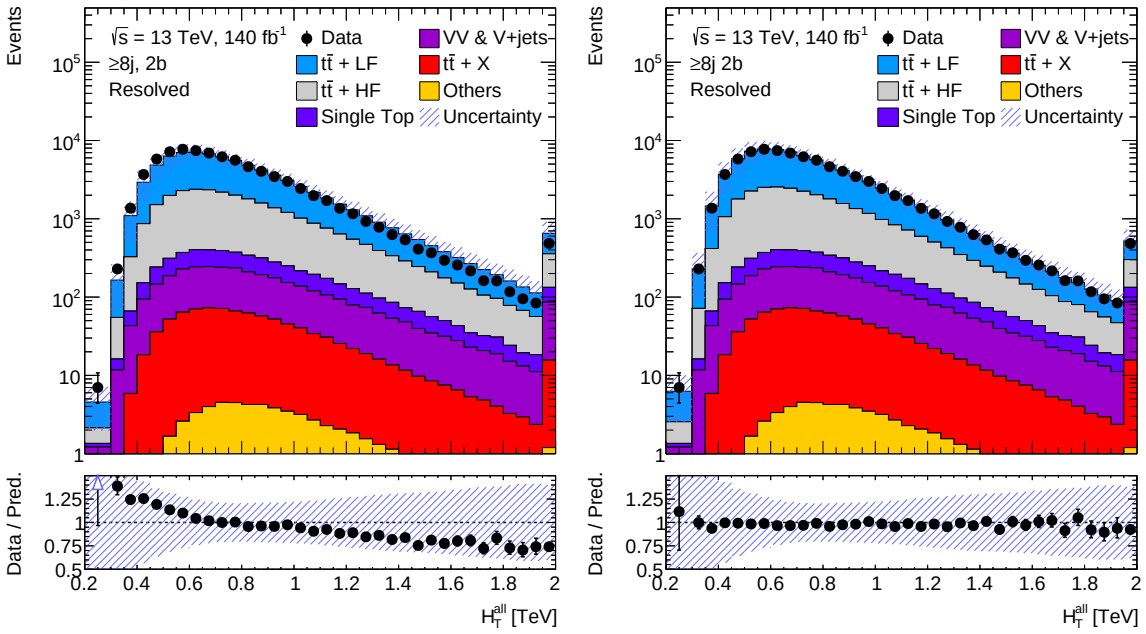
4.4.2 Validation of the reweighting technique

The $H_{\mathrm{T}}^{\mathrm{all}}$ distributions before and after applying corrections to the $t\bar{t}$ + jets background in events containing two b-tagged jets are shown in Figure 4.7 for two different jet multiplicity regions. It can be observed from Figures 4.7(b) and 4.7(d) that the reweighting procedure improves the modelling of the $t\bar{t}$ + jets background, in regions where the corrections are derived, with a data-to-prediction ratio consistent with unity. As discussed earlier, the corrections are subsequently applied in events containing three or more b-tagged jets. The $H_{\mathrm{T}}^{\mathrm{all}}$ distributions, before and after applying the corrections, in events containing three b-tagged and at least four b-tagged jets are shown in Figures 4.8 and 4.9, respectively. The reweighting procedure corrects the shape of the $t\bar{t}$ + jets background as observed in Figures 4.8(b) and 4.8(d). The offset in the ratio panel is due to a residual normalisation effect from the $t\bar{t}$ + HF background. The error variations resulting from the analytical fit constrains the normalisation of the $t\bar{t}$ + LF background within 6% [187] which is the cross-section uncertainty on the $t\bar{t}$ + jets background. However, these error variations do not constrain the normalisation of the $t\bar{t}$ + HF background, which is reflected as a normalisation effect after the corrections are applied. This feature is taken care of in the statistical model (see Chapter 6) where the normalisation factor of the $t\bar{t}$ + HF background is allowed to float independently, and is derived from a fit to data.

The reweighting procedure is further validated for other kinematic observables, namely $E_{\mathrm{T}}^{\mathrm{miss}}$, electron p_{T}, leading jet p_{T}, and muon p_{T}. The corresponding distributions for the first two observables, before and after the corrections are applied, in events containing at least five jets and two b-tagged jets are shown in Figure 4.10. A good agreement between data and simulation is observed within the uncertainties. The distributions for the remaining observables can be found in Figure A.6 in the Appendix.

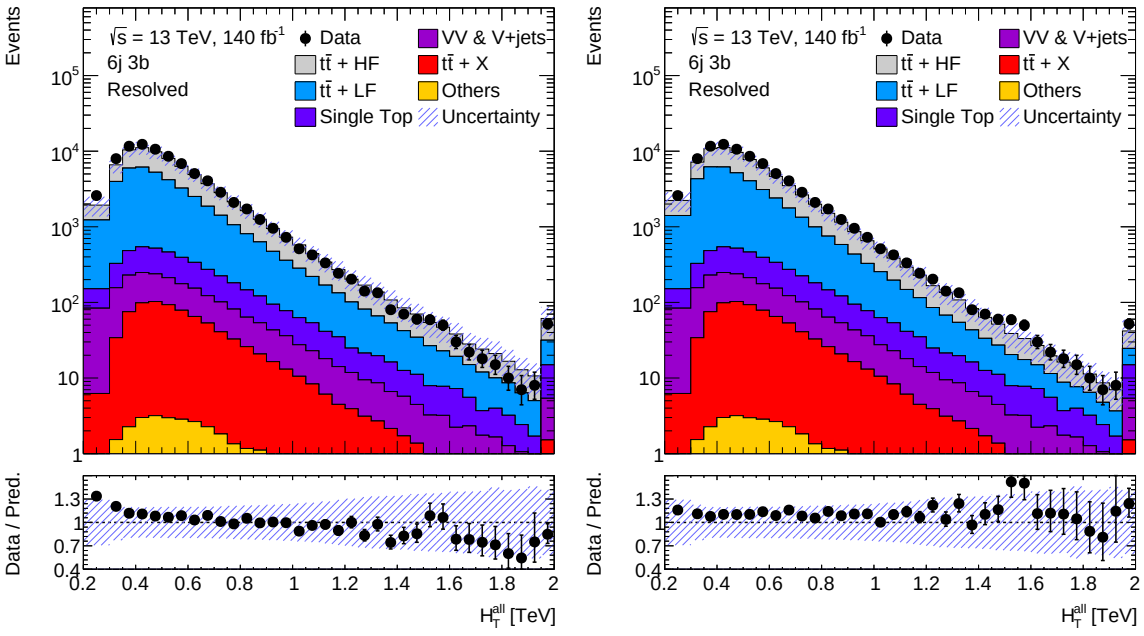


(a) (b)

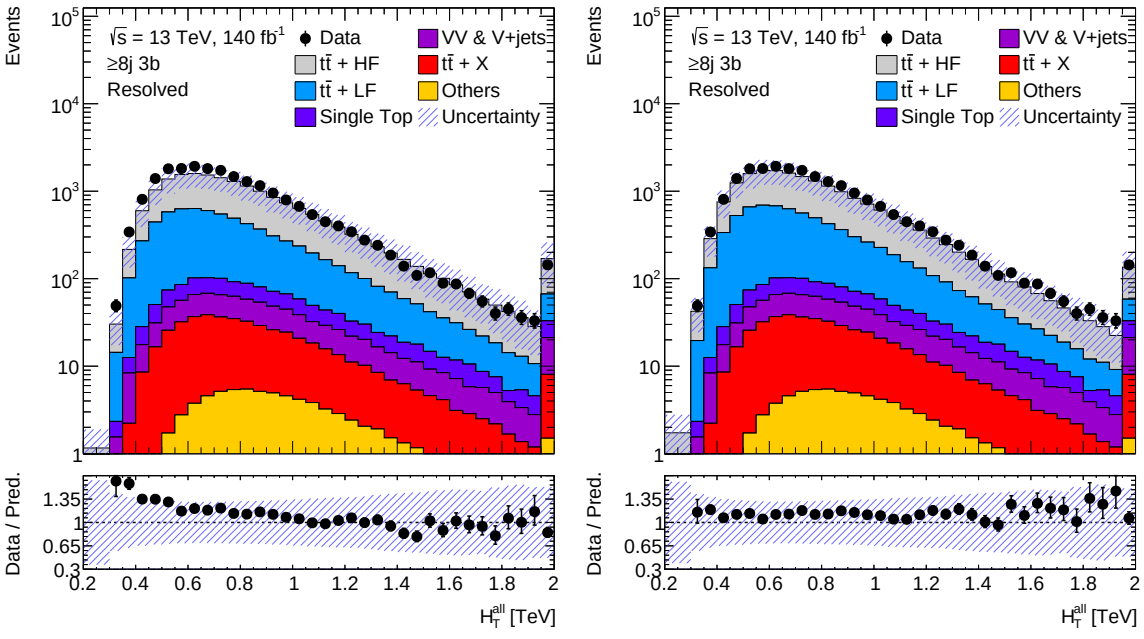


(c) (d)

Figure 4.7: $H_{\mathrm{T}}^{\mathrm{all}}$ distributions, for $n_{\mathrm{jets}} = 6$ and ≥ 8, before (left) and after (right) the reweighting is applied in events containing two b-tagged jets. The ratio panels show the ratio of the observed to the SM background prediction. The background uncertainty band shows the statistical and systematic uncertainties added in quadrature. The notation "$6j$ $2b$" refers to the jet (j) and b-jet multiplicities (b).

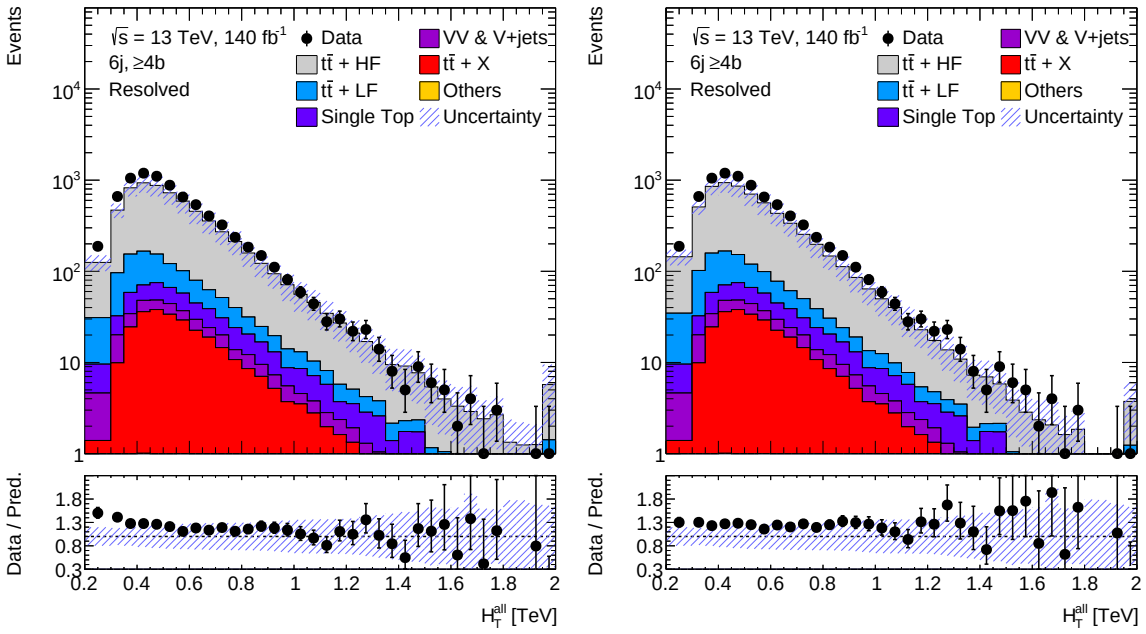


(a) (b)



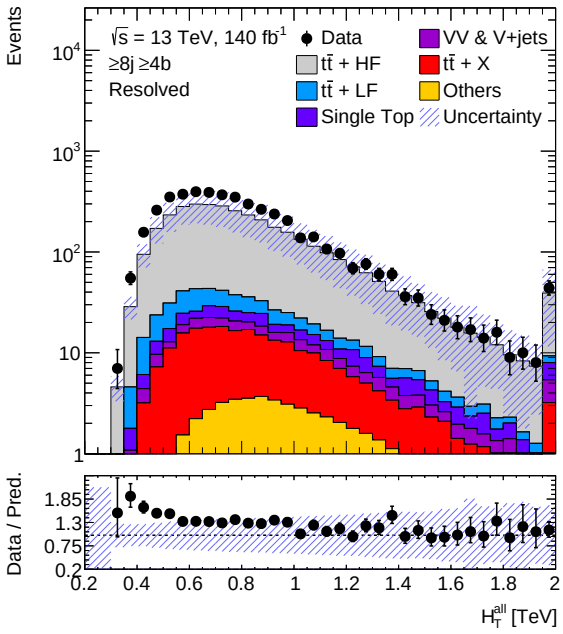
(c) (d)

Figure 4.8: $H_{\mathrm{T}}^{\mathrm{all}}$ distributions, for $n_{\mathrm{jets}} = 6$ and ≥ 8, before (left) and after (right) the reweighting is applied in events containing three b-tagged jets. The ratio panels show the ratio of the observed to the SM background prediction. The background uncertainty band shows the statistical and systematic uncertainties added in quadrature. The notation "6j 3b" refers to the jet (j) and b-jet multiplicities (b).

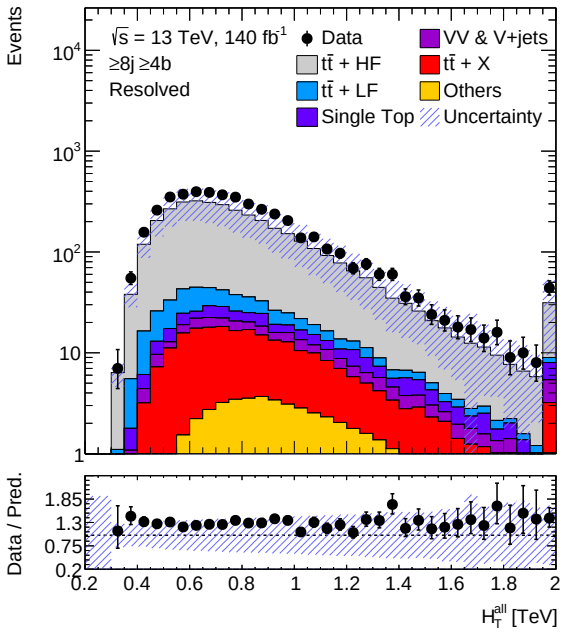


(a)

(b)



(c)



(d)

Figure 4.9: $H_{\mathrm{T}}^{\mathrm{all}}$ distributions, for n_{jets} = 6 and ≥ 8,, before (left) and after (right) the reweighting is applied in events containing at least four b-tagged jets. The ratio panels show the ratio of the observed to the SM background prediction. The background uncertainty band shows the statistical and systematic uncertainties added in quadrature. The notation "6j ≥4b" refers to the jet (j) and b-jet multiplicities (b).

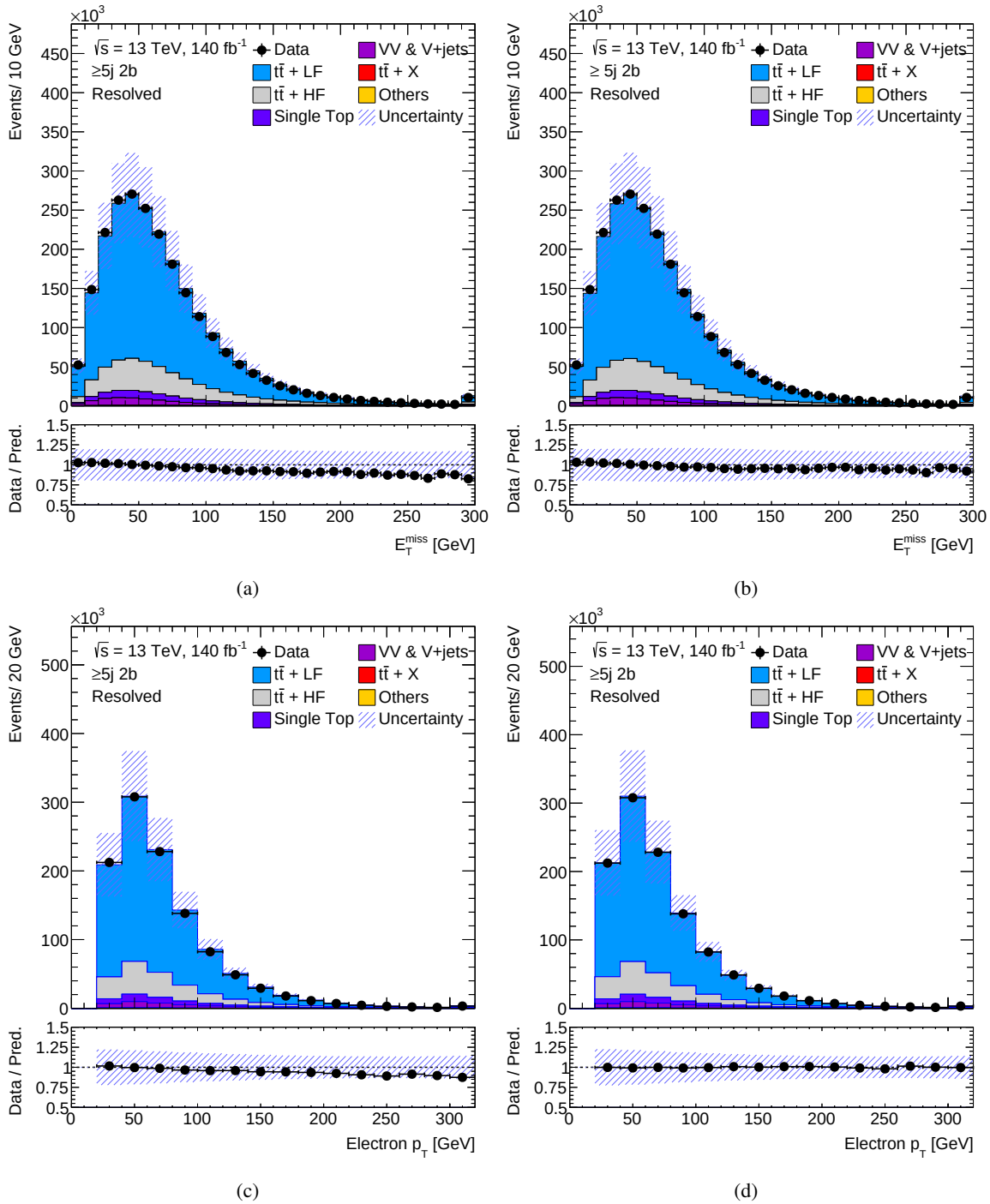


Figure 4.10: The distributions for $E_{\mathrm{T}}^{\mathrm{miss}}$ and electron p_{T}, before (left) and after (right) the reweighting is applied in events containing at least five jets and two b-tagged jets. The ratio panels show the ratio of the observed to the SM background prediction. The background uncertainty band shows the statistical and systematic uncertainties added in quadrature. The notation "$\geq 5j\ 2b$" refers to the jet (j) and b-jet multiplicities (b).

4.5 Event classification in $\ell\nu b\bar{b}$ and $q\bar{q}b\bar{b}$ channels

As discussed earlier in Section 4.1, the final state studied in this analysis can be realised in two different decay channels, namely $\ell\nu b\bar{b}$ and $q\bar{q}b\bar{b}$. To differentiate between the two channels, events are classified based on a requirement on the reconstructed leptonic top-quark candidate mass, $m_{\text{top}}^{\text{lep}}$. $m_{\text{top}}^{\text{lep}}$ is a suitable observable for classifying events in the two channels, as the $q\bar{q}b\bar{b}$ channel contains a leptonically decaying top-quark, whereas the $\ell\nu b\bar{b}$ channel does not (see Figure 4.1). Moreover, the reconstruction of $m_{\text{top}}^{\text{lep}}$ is simpler in contrast to reconstructing a hadronically decaying top-quark present in the $\ell\nu b\bar{b}$ channel, which would require selecting a combination of three jets to reconstruct the top-quark candiate from a complex final state. The reconstruction procedure, the optimisation of the selection requirement on the $m_{\text{top}}^{\text{lep}}$ to classify events in the two channels, and the $m_{\text{top}}^{\text{lep}}$ classification performance are discussed in this section.

4.5.1 $m_{\text{top}}^{\text{lep}}$ reconstruction

The $m_{\text{top}}^{\text{lep}}$ is reconstructed as the four-vector sum of a b-tagged jet of interest and a leptonically decaying W boson (W_{lep}). W_{lep} candidate is reconstructed using a charged lepton and a neutrino candidate. The x and the y components of the neutrino candidate are directly obtained from E_x^{miss} and E_y^{miss} (see Section 3.4.5). The longitudinal component of the neutrino momentum is determined by applying an on-shell W boson mass constraint [188] to the $E_{\text{T}}^{\text{miss}}$ + charged lepton system. This procedure leads to a quadratic equation which provides either two, one, or zero real solutions. If it does not have a real solution, $\vec{E}_{\text{T}}^{\text{miss}}$ is rotated until a real solution is found. If this approach leads to ambiguities, the rotation which provides the minimum change in $E_{\text{T}}^{\text{miss}}$ is chosen. If two real solutions are found, then the solution with the smallest $|p_z|$ is chosen. The b-tagged jet of interest, to reconstruct the $m_{\text{top}}^{\text{lep}}$ is chosen as the one which returns the minimum value for the following expression [189]:

$$\frac{|(m(b\text{-jet} + W_{\text{lep}}) - 172.73)|}{m(b\text{-jet} + W_{\text{lep}})} \tag{4.6}$$

where 172.73 (in units of GeV) is the top-quark pole mass.

The reconstructed $m_{\text{top}}^{\text{lep}}$ distribution is presented for $m_{H^+} = 800\,\text{GeV}$ for the $q\bar{q}b\bar{b}$ and $\ell\nu b\bar{b}$ channels, and the sum of backgrounds in Figure 4.11. One can observe that $m_{\text{top}}^{\text{lep}}$ tends to be distributed around the top-quark mass for events containing a leptonically decaying top-quark which is expected for background events and signal events in the $q\bar{q}b\bar{b}$ channel. For signal events in the $\ell\nu b\bar{b}$ channel, this observable depends on the energy scale of the event and shows broad tails towards high $m_{\text{top}}^{\text{lep}}$ values. The $m_{\text{top}}^{\text{lep}}$ distribution for the background overlaps with the corresponding signal distribution in the $q\bar{q}b\bar{b}$ channel, whereas a good separation is observed between background, and signal events in the $\ell\nu b\bar{b}$ channel. Thus, the $\ell\nu b\bar{b}$ channel has a much lower background contamination compared to the $q\bar{q}b\bar{b}$ channel.

In order to optimise the selection requirement on the $m_{\text{top}}^{\text{lep}}$ to distinguish between two channels, a $q\bar{q}b\bar{b}$ signal hypothesis is assumed as a benchmark[5]. Three selection requirements, namely $m_{\text{top}}^{\text{lep}} \leq$

[5] The choice of this benchmark is primarily due to analysis development reasons, as the $q\bar{q}b\bar{b}$ channel was studied prior to the $\ell\nu b\bar{b}$ channel.

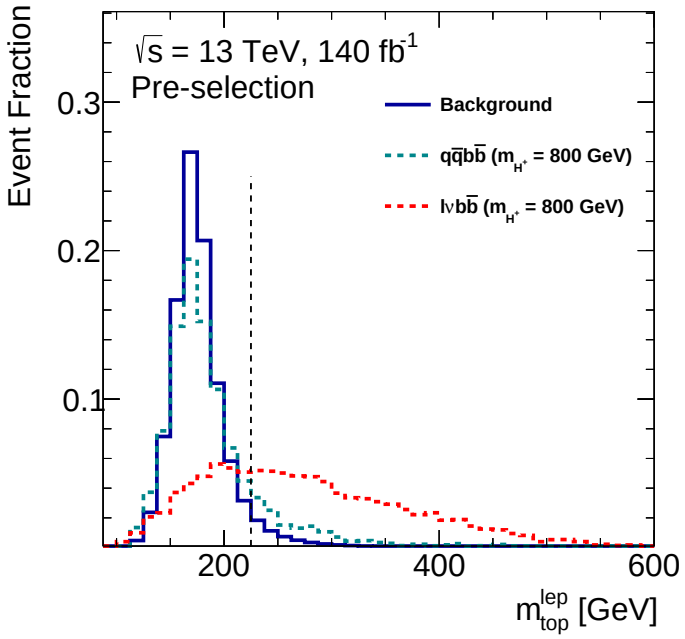


Figure 4.11: $m_{\text{top}}^{\text{lep}}$ distribution for two different mass hypotheses (m_{H^+} = 500 GeV and 1.4 TeV) for the $q\bar{q}b\bar{b}$ and $\ell\nu b\bar{b}$ channels, overlayed with the corresponding distribution for the total background.

200 GeV, $m_{\text{top}}^{\text{lep}} \leq 225$ GeV, and $m_{\text{top}}^{\text{lep}} \leq 250$ GeV are tested to find the best selection requirement. The best selection requirement is chosen based on the most stringent 95% CL expected upper limits on the signal strength, $\sigma(pp \rightarrow tbH^+) \times BR(H^+ \rightarrow Wh(\rightarrow b\bar{b}))$, obtained for the tested selection requirements. The upper limits are calculated using an Asimov fit to the m_{Wh} distribution[6], in a region containing at least five jets and at least three b-tagged jets[7], under a signal-plus-background hypothesis considering statistical uncertainties on the simulated events only. A comparison of the limits for the three tested selection requirements is shown in Figure 4.12. It can be observed from this figure that for $m_{\text{top}}^{\text{lep}} \leq 225$ GeV and $m_{\text{top}}^{\text{lep}} \leq 250$ GeV, one tends to obtain relatively better (especially at high H^+ boson masses) limits than those obtained for $m_{\text{top}}^{\text{lep}} \leq 200$ GeV. However, a looser selection requirement on $m_{\text{top}}^{\text{lep}}$ leads to a higher miscategorisation of events in the $q\bar{q}b\bar{b}$ channel from the $\ell\nu b\bar{b}$ channel, in the low-mass tails of the reconstructed m_{Wh} distribution. This is shown in Figure 4.13 which presents the reconstructed m_{Wh} distribution for m_{H^+} = 600 GeV and m_{H^+} = 1.2 TeV for the two selection requirements. Hence, the best cut to classify $\ell\nu b\bar{b}$ ($q\bar{q}b\bar{b}$) channels is chosen as $m_{\text{top}}^{\text{lep}} >$ 225 GeV ($m_{\text{top}}^{\text{lep}} \leq 225$ GeV) which serves as a good trade-off between the expected sensitivity and miscategorisation of events.

[6] The reconstruction of m_{Wh} is described in Chapter 5, and is independent of the $m_{\text{top}}^{\text{lep}}$ reconstruction.

[7] The optimisation of the selection requirement is performed using events containing at least three b-tagged jets which are used to define the signal and control regions (see Chapter 5).

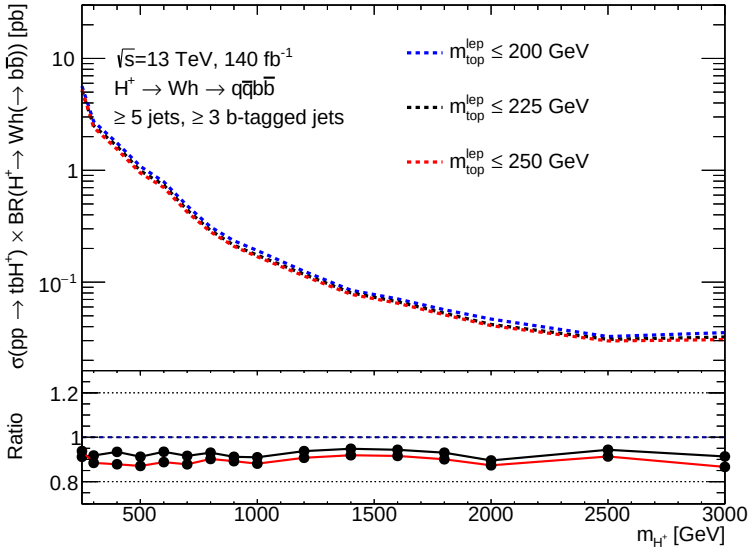


Figure 4.12: 95 % CL expected upper limits on $\sigma(pp \to tbH^+) \times BR(H^+ \to Wh(\to b\bar{b}))$ for different $m_{\text{top}}^{\text{lep}}$ requirements in events containing at least five jets and at least b-tagged jets. The ratio panel depicts the ratio of the expected upper limits w.r.t the limits obtained for $m_{\text{top}}^{\text{lep}} \leq 200\,\text{GeV}$ selection requirement.

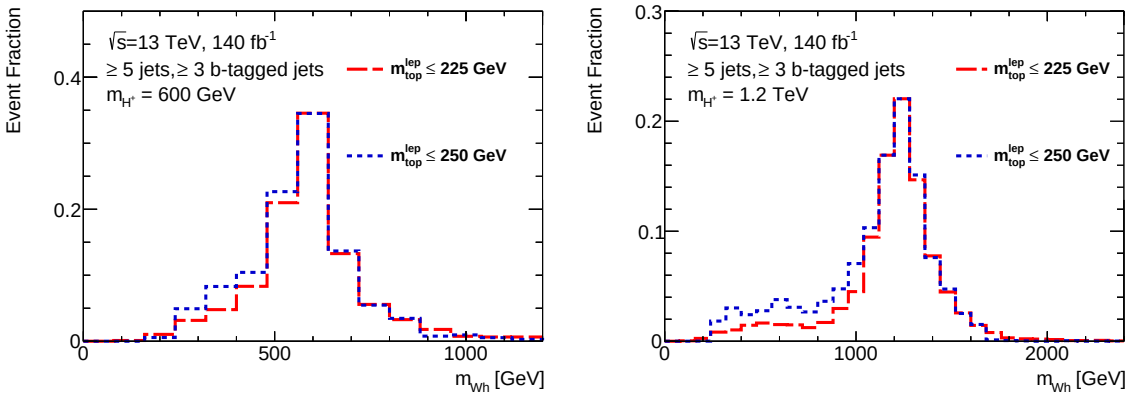
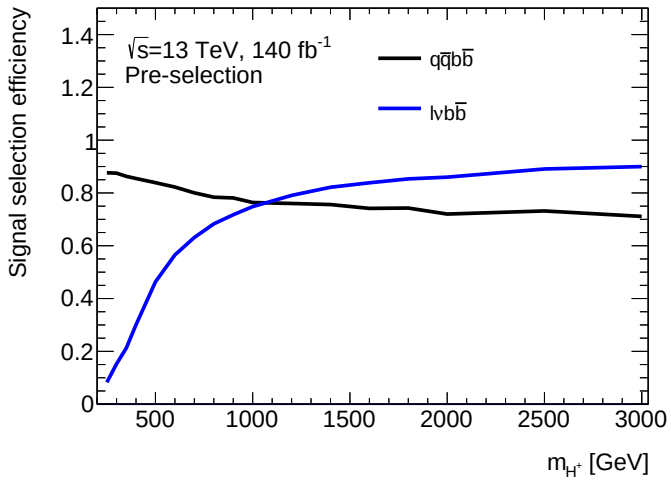


Figure 4.13: The reconstructed m_{Wh} distribution for $m_{H^+} = 600\,\text{GeV}$ and $m_{H^+} = 1.2\,\text{TeV}$ in the $q\bar{q}b\bar{b}$ channel for the $m_{\text{top}}^{\text{lep}} \leq 225\,\text{GeV}$ and $m_{\text{top}}^{\text{lep}} \leq 250\,\text{GeV}$ selection requirements.



(a)

Figure 4.14: Signal selection efficiency in events passing the pre-selection requirements.

4.5.2 Event classification performance

The event classification performance is evaluated in terms of the signal selection efficiency. The signal selection efficiency (ϵ_{sig}) is defined as the ratio of the number of signal events at generator-level passing the selection requirement on $m_{\text{top}}^{\text{lep}}$, to be classified as $q\bar{q}b\bar{b}$ or $\ell\nu b\bar{b}$ events, to the total number of signal events at generator-level. It is expressed as:

$$\epsilon_{\text{sig}} = \frac{N_{\text{gen}}^{\text{total}} | m_{\text{top}}^{\text{lep}}}{N_{\text{gen}}^{\text{total}}} \tag{4.7}$$

where $N_{\text{gen}}^{\text{total}}$ denotes the total number of signal events at generator-level, for the two channels, passing the pre-selection requirements. Figure 4.14 shows the signal selection efficiency as a function of m_{H^+} for the two channels. The efficiency varies from 10% (88%) for the lowest considered m_{H^+} to 95% (76%) for the highest considered m_{H^+} in the $\ell\nu b\bar{b}$ ($q\bar{q}b\bar{b}$) channels. The low signal selection efficiency in the $\ell\nu b\bar{b}$ channel for $m_{H^+} < 700$ GeV is due to the miscategorisation of $\ell\nu b\bar{b}$ events as $q\bar{q}b\bar{b}$ events. At low masses, the H^+ boson decay products in the $\ell\nu b\bar{b}$ channel have low transverse momentum, and can pass the jet and lepton p_{T} acceptance requirements leading to low $m_{\text{top}}^{\text{lep}}$ values which in turn populate the $q\bar{q}b\bar{b}$ channel. At high H^+ boson masses, the reconstruction procedure is seen to work comparably well for both channels.

CHAPTER 5

Resolved Topology: Event Reconstruction BDTs

This chapter covers the remaining aspects of the analysis strategy used to perform the search for resolved $H^+ \to Wh(\to b\bar{b})$ decays. After the classification of events into $q\bar{q}b\bar{b}$ and $\ell\nu b\bar{b}$ channels, BDTs are trained in both the channels to select the correct combination of jets to reconstruct the W and h boson candidates originating from the H^+ boson. Section 5.1 discusses in detail the BDT training, application, and performance in the $q\bar{q}b\bar{b}$ channel, and gives a brief overview of the BDT training and application in the $\ell\nu b\bar{b}$ channel. Dedicated control and signal regions are defined in different intervals of the BDT score for the two channels. The optimisation of signal and control regions is described in Section 5.2.

5.1 Event reconstruction BDTs

The $tbH^+ \to tbWh(\to b\bar{b})$ final state contains at least five jets, and hence the main challenge in the event reconstruction is to correctly assign jets to the original particles, the W and h bosons. To tackle this challenge two sets of BDTs are trained separately in the two channels to perform the event reconstruction. They are trained using the simulated events of all seventeen signal mass hypotheses in the analysis pre-selection region. Up to nine jets arranged in decreasing order of their transverse momentum are used to perform the training. The first set of BDTs targeting the $H^+ \to Wh \to q\bar{q}b\bar{b}$ decays ($q\bar{q}b\bar{b}$ channel) are trained on input variables built from the four-vectors of the four jets used to construct the W and h boson candidates. The second set of BDTs targeting the $H^+ \to Wh \to \ell\nu b\bar{b}$ decays ($\ell\nu b\bar{b}$ channel) are trained on input variables built from the four-vectors of the charged lepton and the neutrino[1] candidates to construct the W boson candidate, and from the four-vectors of the two jets to construct the h boson candidate. The BDTs are trained to distinguish between the correct pairings of the final state objects, i.e. leptons and jets, labelled as signal, and the incorrect pairings labelled as background (combinatorial background). The correct pairing is found using a spatial matching between the generator-level particles and the reconstructed objects. A jet pair is considered to be correctly matched to a generator-level W/h boson, if its angular distance (ΔR) to the generator-level

[1] The p_z component of the neutrino is reconstructed using the W boson mass constraint method as described in Section 4.5.1.

W/h boson is less than 0.3 (ΔR (gen-level, reco-level) < 0.3)[2]. The matching between the charged lepton + neutrino candidate and the corresponding generator-level W boson, in the $\ell\nu b\bar{b}$ channel, is based on the same criteria.

5.1.1 $q\bar{q}b\bar{b}$ channel

BDT Training

The training of the decision trees is carried out using the *Toolkit for MultiVariate data Analyses (TMVA)* framework [164]. A general overview on the BDTs, and the common training parameters known as the hyperparameters was given in Section 3.5.1. To protect against any potential biases due to overtraining, and to ensure that the training is not tested (evaluated) on events which were used for the training, a two-fold cross-validation method is employed. The input dataset is split into two independent datasets based on the event number. One dataset is built from the even numbered events (`DATASET 1`), and the other dataset is built from the odd numbered events (`DATASET 2`). Resultantly, two trainings are performed: The first training is performed on `DATASET 1` and is tested on `DATASET 2`, while the second training is performed on `DATASET 2` and is tested on `DATASET 1`.

In the $q\bar{q}b\bar{b}$ channel, the training is based on ten input variables. These variables are built from the four-vectors of the four jets denoted by j_{W_1}, j_{W_2} and j_{h_1}, j_{h_2} to construct the hadronically decaying W boson (W_{had}) and h boson candidates, respectively. The list of variables include the invariant masses of the two boson candidates and the pseudo-continuous b-tagging (PCBT) (see Section 3.4.3) scores[3] of the four jets, which are expected to show distinct distributions between correct object pairings (signal) and incorrect object pairings (combinatorial background). A dependence on the H^+ boson mass is reduced by taking the ratio of the transverse momentum of the reconstructed objects with respect to the reconstructed charged Higgs boson mass. In addition, variables associated with the angular separation between the reconstructed objects are also taken into account which can vary between signal and combinatorial background. The input variables are defined as following:

- The invariant masses of the W_{had} and h boson candidates: m_W ($m_{j_{W_1} j_{W_2}}$) and m_h ($m_{j_{h_1} j_{h_2}}$), respectively,
- the PCBT scores of the four jets: b-tag j_{h_1}, b-tag j_{h_2}, b-tag j_{W_1}, b-tag j_{W_2},
- the ratio of the W_{had} transverse momentum to the mass of the reconstructed charged Higgs boson, m_{Wh}: p_T^W/m_{Wh},
- the ratio of the h boson transverse momentum to the mass of the reconstructed charged Higgs boson, m_{Wh}: p_T^h/m_{Wh},
- the pseudorapidity difference between the h boson and W_{had} candidates: $|\Delta\eta(h, W)|$,
- the azimuthal angle difference between the h boson and W_{had} candidates: $\Delta\Phi(h, W)$.

[2] If more than two jet pairs satisfy the matching criteria, the pair with the smaller ΔR to the generator-level W/h boson is considered as the correct pairing.

[3] Following the definition of the PCBT score in Section 3.4.3, the correct jet pairings used to construct the h boson candidate populates the higher PCBT score bins whereas the correct jet pairings used to construct the W boson candidate populates the lower PCBT score bins.

Table 5.1: Hyperparameters used for the training of decision trees.

Parameter	Value
Boosting type	Gradient
Separation type	`GiniIndex`
`Shrinkage`	0.1
N^{trees}	10, 50, 100, 200, **400**, 600, 1000
`MinNodeSize`	2%
`nCuts`	10, 20, **40**, 60, 80
Tree depth	3, **5**, 7, 9

Table 5.2: Variable ranking of BDT input variables in the $q\bar{q}b\bar{b}$ channel.

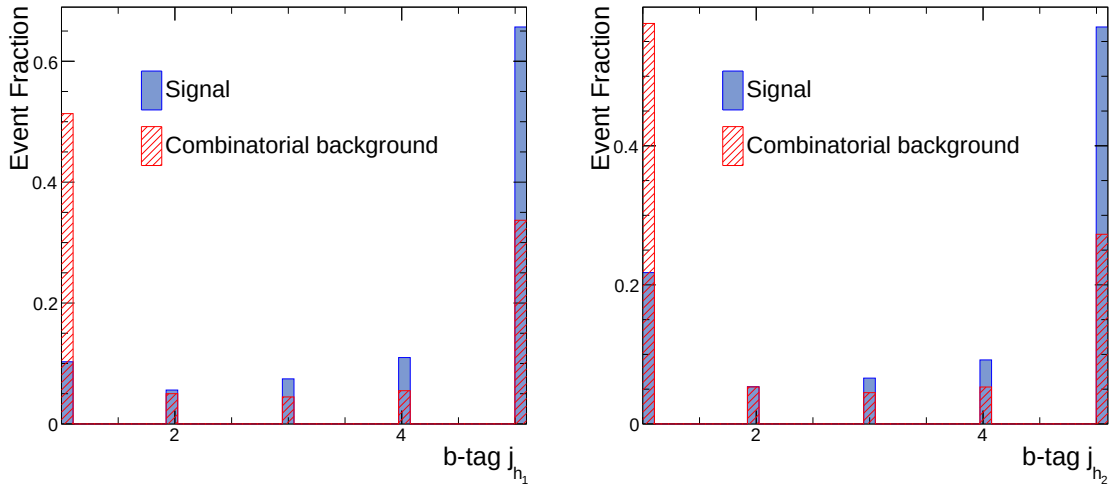
Rank	Variable name
1	$p_{\text{T}}^{\text{h}}/m_{\text{Wh}}$
2	$\lvert\Delta\eta(h,W)\rvert$
3	$p_{\text{T}}^{\text{W}}/m_{\text{Wh}}$
4	m_{h}
5	m_{W}
6	$\Delta\Phi(h,W)$
7	b-tag j_{h_1}
8	b-tag j_{h_2}
9	b-tag j_{W_1}
10	b-tag j_{W_2}

The distributions of training input variables overlayed for signal and combinatorial background are shown in Figures 5.1, 5.2, and 5.3. A reasonable separation between signal and combinatorial background is observed for all the ten input variables.

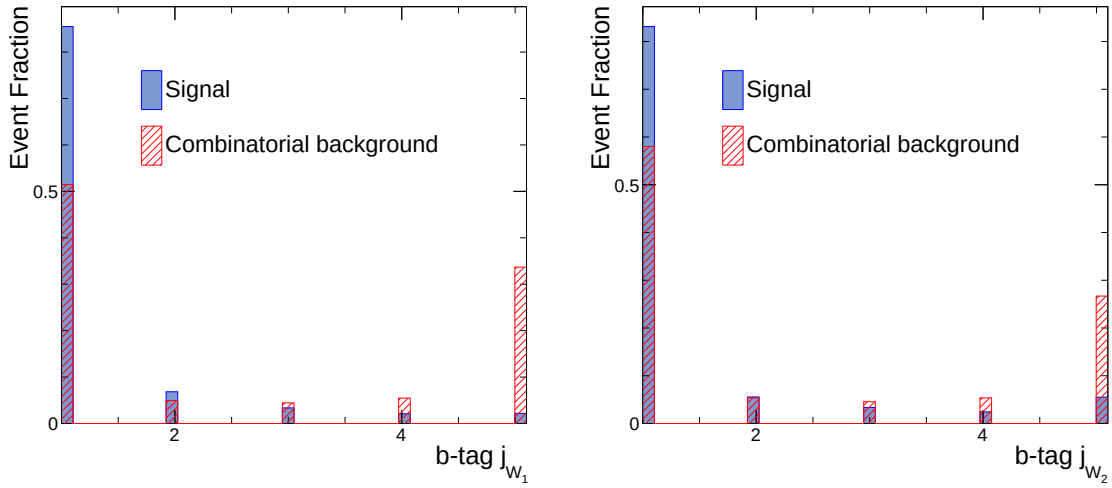
Table 5.1 summarises the hyperparameters and their corresponding values used for the training. For some of the considered hyperparameters, different values were studied and the value highlighted in bold was chosen for the training. The choice of an optimal value out of the many studied ones is based on attaining the best area under the *Receiver Operator Characteristic (ROC)* curve[4], or yielding no further improvement in this area on further increasing the hyperparameter value. The ranking of the input variables resulting from the BDT training is shown in Table 5.2 with $p_{\text{T}}^{\text{h}}/m_{\text{Wh}}$, $|\Delta\eta(h,W)|$[5], $p_{\text{T}}^{\text{W}}/m_{\text{Wh}}$, and m_{h} being the most important variables. The response distribution (BDT score) resulting from the training is presented in Figure 5.4 which shows a minimum overlap between the signal and background distributions. One can also observe that the training and test datasets for signal and background distributions do not display any significant differences, confirming that there are no signs of over-training.

[4] A ROC curve evaluates the performance of the classifier at various classification threshold values [190].

[5] The $|\Delta\eta(h,W)|$ variable shown in Figure 5.2(a) shows a smaller degree of separation between signal and background in comparison to other variables, but is among the highly ranked variables due to its moderate correlations with $p_{\text{T}}^{\text{W}}/m_{\text{Wh}}$ and $p_{\text{T}}^{\text{h}}/m_{\text{Wh}}$ (see Figure A.9 in Appendix A).



(a) (b)



(c) (d)

Figure 5.1: The signal and combinatorial background discrimination for the input variables used for the BDT training in the $q\bar{q}b\bar{b}$ channel. The variables are shown in the following order: (a) b-tag j_{h_1}, (b) b-tag j_{h_2}, (c) b-tag j_{W_1}, and (d) b-tag j_{W_2}.

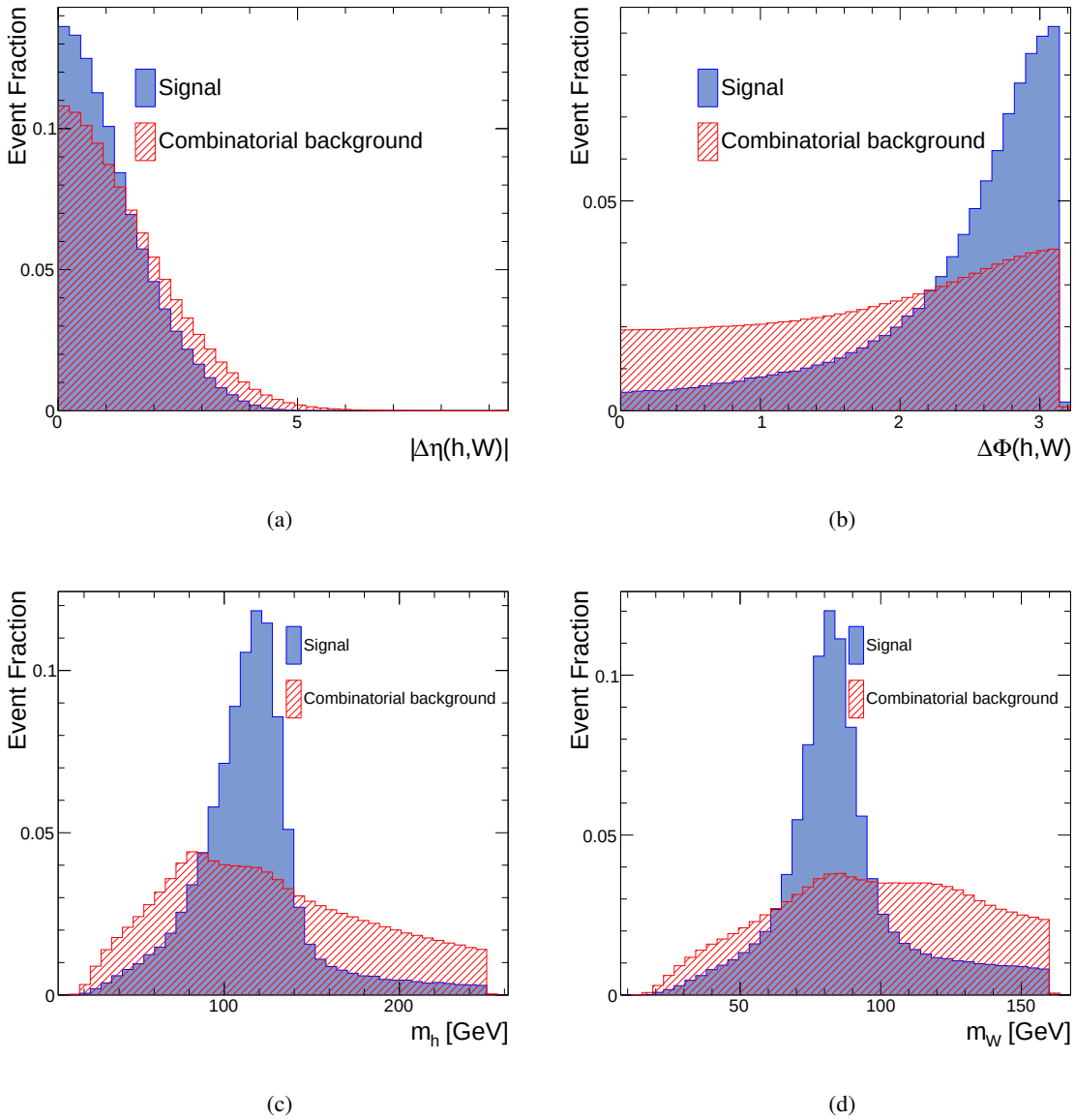
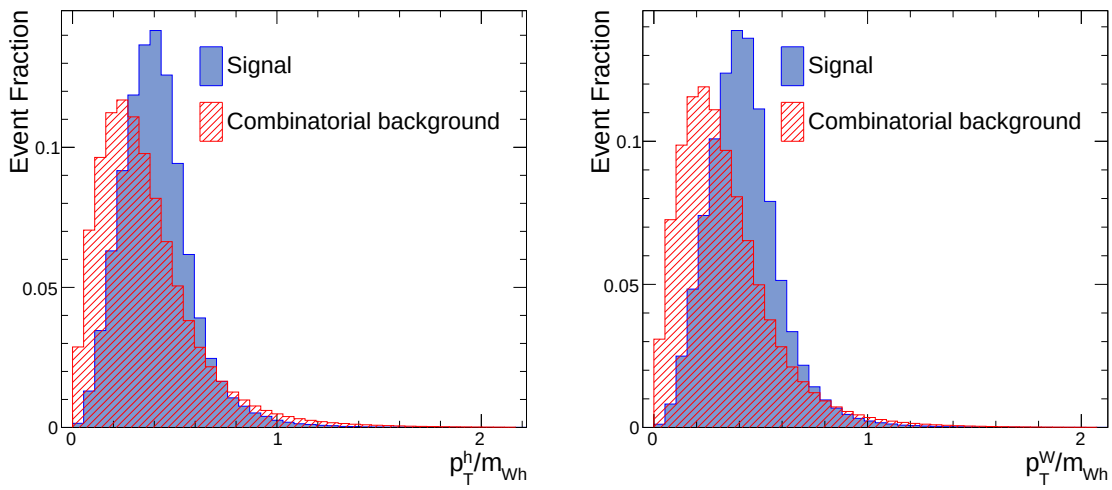


Figure 5.2: The signal and combinatorial background discrimination for the input variables used for the BDT training in the $q\bar{q}b\bar{b}$ channel. The variables are shown in the following order: (a) $|\Delta\eta(h,W)|$, (b) $\Delta\Phi(h,W)$, (c) m_{h}, and (d) m_W



(a) (b)

Figure 5.3: The signal and combinatorial background discrimination for the input variables used for the BDT training in the $q\bar{q}b\bar{b}$ channel. The variables are shown in the following order: (a) $p_{\mathrm{T}}^{\mathrm{h}}/m_{\mathrm{Wh}}$ and (b) $p_{\mathrm{T}}^{\mathrm{W}}/m_{\mathrm{Wh}}$

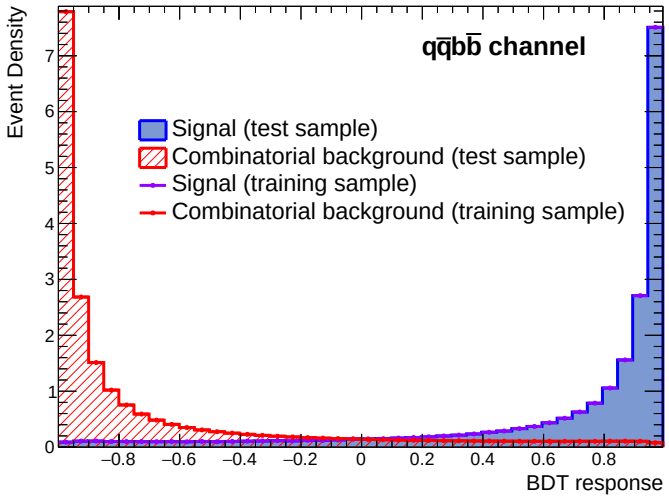


Figure 5.4: The BDT response distribution resulting from the training in the $q\bar{q}b\bar{b}$ channel.

BDT Application

BDTs trained in the $q\bar{q}b\bar{b}$ channel are applied to all possible four-jet pairings in an event. The pairing corresponding to the largest BDT score ($w_{\text{BDT}}^{\text{max}}$) for an event is used to reconstruct the h and W boson candidates, and subsequently the invariant mass of the charged Higgs boson, m_{Wh}. Data-to-simulation comparisons for the highly ranked BDT input variables in events containing at least three b-tagged jets[6] are shown in Figure 5.5. For all the presented data-to-simulation comparisons in this chapter, the $t\bar{t}$ + LF and $t\bar{t}$ + HF backgrounds are corrected via the $H_{\text{T}}^{\text{all}}$ reweighting procedure (see Section 4.4.1). In addition, the $t\bar{t}$ + HF background is scaled to its normalisation factor obtained from a background-only fit to data discussed further in Chapter 8. This scaling mitigates any residual normalisation effects for the $t\bar{t}$ + HF background after the reweighting is applied. A good agreement between data and simulated backgrounds is observed within the uncertainties. Data-to-simulation comparisons for the remaining BDT input variables are shown in Figures A.10 and A.11, in Appendix A.

A comparison of $w_{\text{BDT}}^{\text{max}}$ for events passing the pre-selection requirements for two signal mass hypotheses (m_{H^+} = 400 GeV and m_{H^+} = 800 GeV) and the sum of backgrounds is shown in Figure 5.6(a). One can observe that the background distribution tends to peak at high $w_{\text{BDT}}^{\text{max}}$ values. The dominant $t\bar{t}$ + jets background can potentially mimick the signal, particularly in the case of jets originating from hadronically decaying W bosons which can populate the correct jet pairings, leading to high $w_{\text{BDT}}^{\text{max}}$ values. A comparison of m_{Wh} for events passing the pre-selection requirements for a range of signal mass hypotheses is presented in Figure 5.6(b). The m_{Wh} distributions for different signal mass hypotheses are centered around their pole masses. Data-to-simulation comparsions for $w_{\text{BDT}}^{\text{max}}$ and m_{Wh} in events containing at least three b-tagged jets are shown in Figure 5.7. A good agreement between data and simulated backgrounds is observed within the uncertainties. The m_{Wh} distributions for the backgrounds peak at lower mass values, in contrast to the m_{Wh} distributions for various signal mass hypotheses (cf. Figure 5.6(b)) which peak at their respective pole masses.

[6] The comparison is shown for events containing at least three b-tagged jets which are used to define signal and control regions (see Section 5.2).

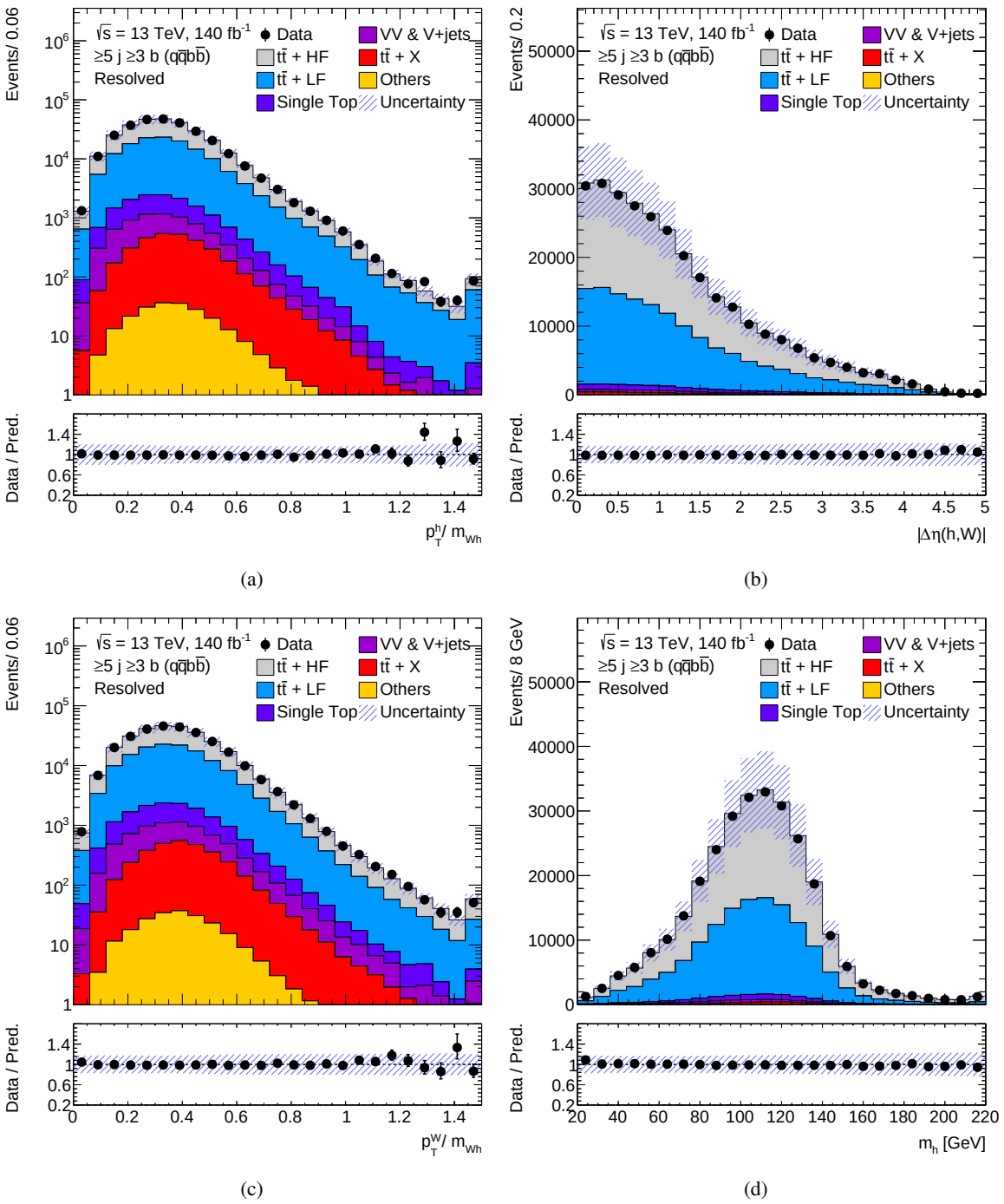
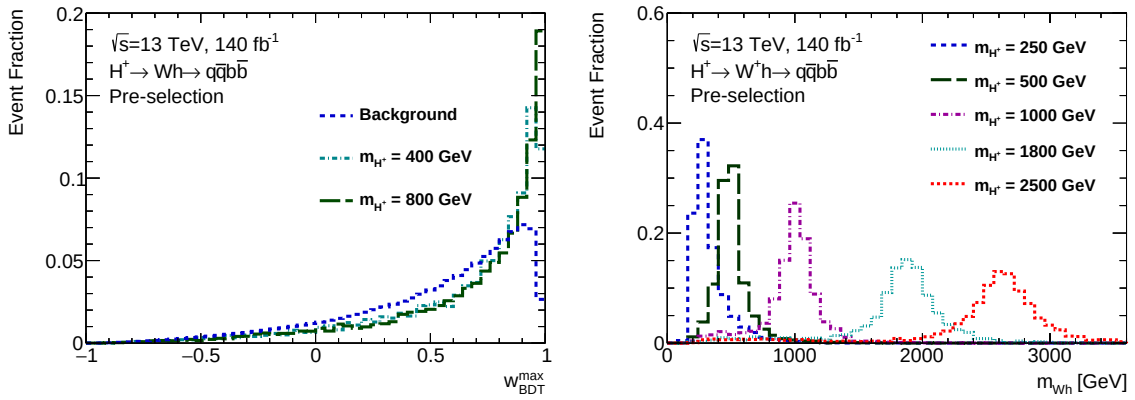


Figure 5.5: Data-to-simulation comparisons for selected BDT input variables in events containing at least three b-tagged jets in the $q\bar{q}b\bar{b}$ channel. (a) $p_{\mathrm{T}}^{\mathrm{h}}/m_{\mathrm{Wh}}$, (b) $|\Delta\eta(h,W)|$, (c) $p_{\mathrm{T}}^{\mathrm{W}}/m_{\mathrm{Wh}}$, and (d) m_{h}. The background uncertainty band shows the statistical and systematic uncertainties added in quadrature. The notation "$\geq 5j \geq 3b$" refers to the jet (j) and b-jet multiplicities (b).



(a) (b)

Figure 5.6: (a) Comparison of $w_{\mathrm{BDT}}^{\mathrm{max}}$ for two signal mass hypotheses and the sum of backgrounds. (b) Comparison of m_{Wh} for a range of signal mass hypotheses.

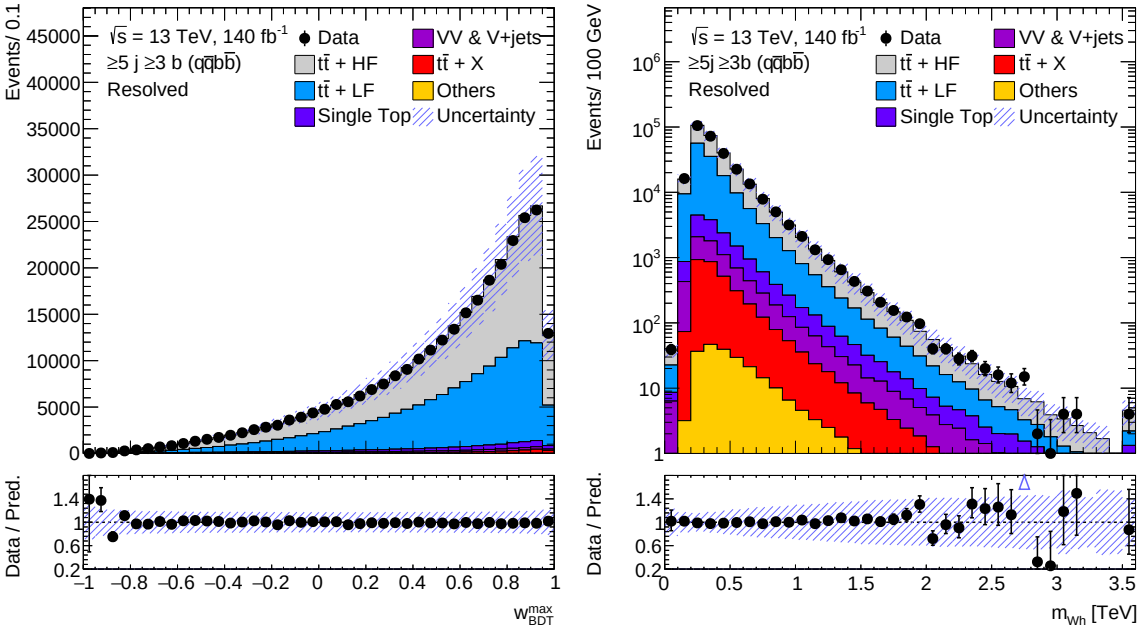


Figure 5.7: Data-to-simulation comparisons for (a) $w_{\mathrm{BDT}}^{\mathrm{max}}$ and (b) m_{Wh} in events containing at least three b-tagged jets in the $q\bar{q}b\bar{b}$ channel. The background uncertainty band shows the statistical and systematic uncertainties added in quadrature. The notation "$\geq 5j$ $\geq 3b$" refers to the jet (j) and b-jet multiplicities (b).

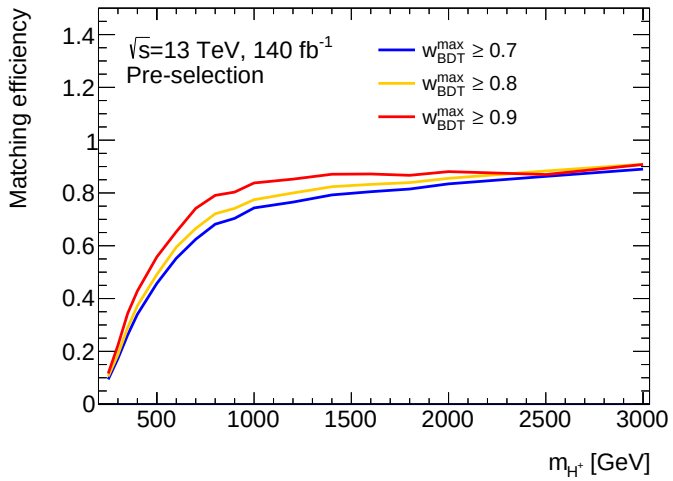


Figure 5.8: Matching efficiency for different selection requirements on $w^{\text{max}}_{\text{BDT}}$ in the $q\bar{q}b\bar{b}$ channel.

BDT Performance

The performance of the BDTs is studied by evaluating how often the reconstructed h and W boson candidates are chosen correctly in the $q\bar{q}b\bar{b}$ channel. This is evaluated in terms of a metric which encodes the efficiency to match the reconstructed objects to the corresponding generator-level particles[7], after applying the BDTs. Furthermore, the relative m_{Wh} resolution is also studied as an additional metric to test the BDT performance. This section explains these two different metrics in detail.

The matching efficiency is defined as:

$$\varepsilon^{\text{match}} = \frac{N^{\text{match}}_{\text{q}\bar{\text{q}}\text{b}\bar{\text{b}}}}{N^{\text{total}}_{\text{q}\bar{\text{q}}\text{b}\bar{\text{b}}}} \tag{5.1}$$

where $N^{\text{match}}_{\text{q}\bar{\text{q}}\text{b}\bar{\text{b}}}$ denotes the number of signal events for which the reconstructed h and W boson candidates are matched to the corresponding generator-level particles, and $N^{\text{total}}_{\text{q}\bar{\text{q}}\text{b}\bar{\text{b}}}$ denotes the total number of signal events in the $q\bar{q}b\bar{b}$ channel. The matching efficiency is evaluated for three different selection requirements on $w^{\text{max}}_{\text{BDT}}$ ($w^{\text{max}}_{\text{BDT}} \geq 0.7$, $w^{\text{max}}_{\text{BDT}} \geq 0.8$, and $w^{\text{max}}_{\text{BDT}} \geq 0.9$) as a function of the H^+ boson mass (m_{H^+}), and is presented in Figure 5.8. It can be observed from this figure that the matching efficiency ranges from 20% for the lowest considered m_{H^+} to 90% for $m_{H^+} > 1$ TeV. For low H^+ boson masses, the reconstructed h/W candidates are more likely to originate from jets with low transverse momentum, such that the reconstructed candidates do not fulfil the matching criteria to the generator-level particles, leading to lower efficiencies with respect to those obtained at higher H^+ boson masses.

The matching efficiency was also used in order to study two different BDT training scenarios. The first scenario corresponds to a training including simulated events of all seventeen signal mass hypotheses, which is the default training used for event reconstruction. The second scenario

[7] The generator-level and the reconstruction-level objects are considered to be matched if they fulfil ΔR (gen-level, reco-level) < 0.3 as described earlier in Section 5.1.

Table 5.3: The matching efficiency evaluated for different m_{H^+} hypotheses for a unified training compared to the training carried out on individual H^+ boson masses.

m_{H^+}	Default training	Individual mass training
400 GeV	0.48 ± 0.08	0.49 ± 0.09
800 GeV	0.69 ± 0.12	0.68 ± 0.15
1.6 TeV	0.74 ± 0.13	0.73 ± 0.17

corresponds to a training performed exclusively using simulated events of an individual signal mass hypothesis. The efficiencies were evaluated for three representative H^+ boson masses, namely 400 GeV, 800 GeV, and 1.6 TeV for the two training scenarios. The comparison of efficiencies for these two scenarios is shown in Table 5.3, with consistent results obtained for both cases. This study proves that there is no gain observed from training for individual signal mass hypothesis, and the default training is sufficient for event reconstruction.

Another relevant quantity to test the reconstruction performance is the relative mass resolution defined as:

$$\frac{m_{\mathrm{Wh}}^{\mathrm{reco}} - m_{\mathrm{Wh}}^{\mathrm{pole}}}{m_{\mathrm{Wh}}^{\mathrm{pole}}} \tag{5.2}$$

where $m_{\mathrm{Wh}}^{\mathrm{reco}}$ refers to the reconstructed charged Higgs boson mass and $m_{\mathrm{Wh}}^{\mathrm{pole}}$ refers to the charged Higgs boson pole mass. The relative mass resolution is modelled by fitting a Bukin's function [191] to the distribution for each signal mass hypothesis, and is able to fit asymmetric distributions. The Bukin's function is a convolution of a Gaussian function and an exponential function:

$$f\left(m|m_p, h, \lambda\right) = A \exp\left[-\frac{1}{2\sigma^2} \ln^2 \frac{m_s - m}{m_s - m_p}\right], \int_{\frac{m_s - m}{m_s - m_p}} f\,dx = 1 \tag{5.3}$$

where m_p is the peak position (function maximum), h is the full width at half maximum (FWHM), λ is an asymmetry parameter, A is the normalisation factor, and σ corresponding to $\frac{\mathrm{FWHM}}{2.35}$ is taken as the mass resolution. Definitions of various variables used in this function are decribed in Equations 5.4, 5.5, and 5.6.

$$m_s = m_p - \frac{\left(z + \frac{1}{z}\right)}{2\lambda} \tag{5.4}$$

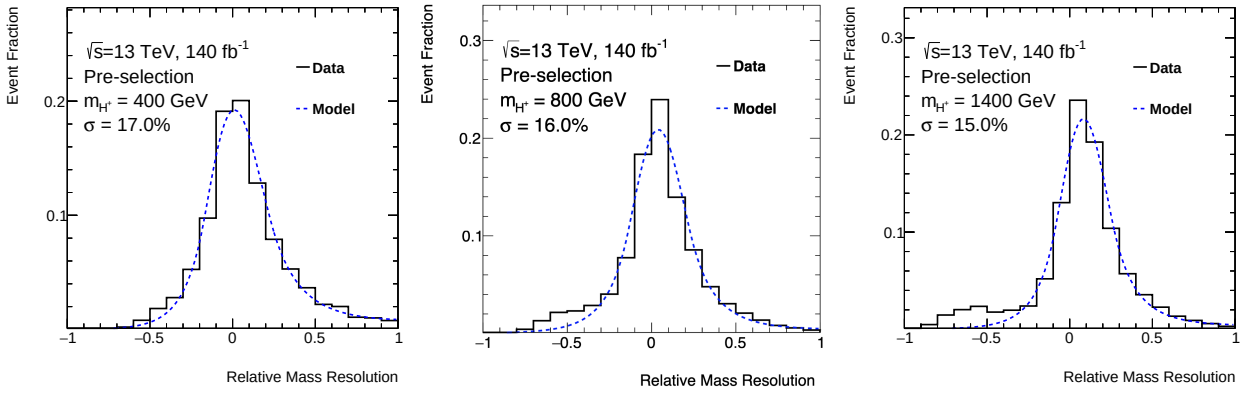
$$\sigma = \frac{1}{\sqrt{2\ln 2}} \ln z, A = \sqrt{\frac{2}{\pi}} \frac{|\lambda|}{(z + \frac{1}{z})\sigma} \exp\left[-\sigma^2/2\right] \tag{5.5}$$

$$z = \sqrt{h|\lambda| + \sqrt{h^2\lambda^2 + 1}} \tag{5.6}$$

The fits to the relative mass resolution distribution for three representative signal mass hypotheses corresponding to $m_{H^+} = 400\,\mathrm{GeV}$, $m_{H^+} = 800\,\mathrm{GeV}$, and $m_{H^+} = 1.4\,\mathrm{TeV}$[8] performed using events

[8] The left sideband of the distribution (in the range from -1 to -0.5) for $m_{H^+} = 1.4\,\mathrm{TeV}$ has slightly more events due to a

passing the pre-selection requirements in the $q\bar{q}b\bar{b}$ channel are shown in Figure 5.9. The distributions in these figures are centered around zero with moderate asymmetrc tails, which demonstrates that the reconstruction BDTs perform well in identifying the correct jet pairing from the several jet combinations.



(a) (b) (c)

Figure 5.9: Relative mass resolution distributions modelled by a Bukin's function in the $q\bar{q}b\bar{b}$ channel for (a) $m_{H^+} = 400\,\text{GeV}$, (b) $m_{H^+} = 800\,\text{GeV}$, and (c) $m_{H^+} = 1.4\,\text{TeV}$.

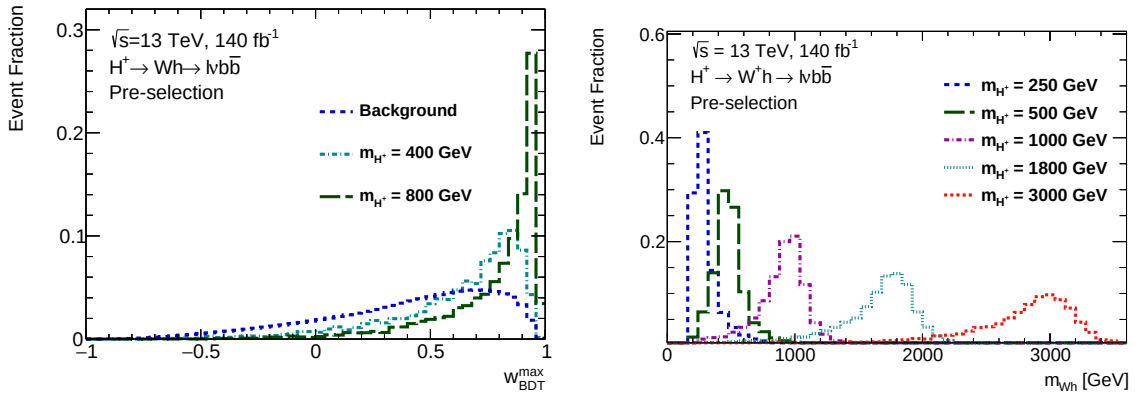
5.1.2 $\ell\nu b\bar{b}$ channel

The training in the $\ell\nu b\bar{b}$ channel follows the same configuration as in the $q\bar{q}b\bar{b}$ channel, and is based on seven input variables. These variables are built from the four-vectors of the charged lepton and neutrino candidates to construct the leptonically decaying W boson (W_{lep}) candidate, and from the four-vectors of the two jets denoted by j_{h_1} and j_{h_2} to construct the h boson candidate. The input variables are defined as following:

- The invariant mass of the h boson candidate: m_{h} ($m_{j_{\text{h}_1} j_{\text{h}_2}}$),
- the PCBT score of the two jets: b-tag j_{h_1}, b-tag j_{h_2},
- the ratio of the W_{lep} transverse momentum to the mass of the reconstructed charged Higgs boson, m_{Wh}: $p_{\text{T}}^{\text{W}}/m_{\text{Wh}}$,
- the ratio of the h boson transverse momentum to the mass of the reconstructed charged Higgs boson, m_{Wh}: $p_{\text{T}}^{\text{h}}/m_{\text{Wh}}$,
- the pseudorapidity difference between the h boson and W_{lep} candidates: $|\Delta\eta(h, W)|$,
- the azimuthal angle difference between the h boson and W_{lep} candidates: $\Delta\Phi(h, W)$.

Similar to the $q\bar{q}b\bar{b}$ channel, the charged lepton + neutrino candidate and the jet pairing corresponding to the largest BDT score ($w_{\text{BDT}}^{\text{max}}$) for an event is used to reconstruct the W and h boson candidates, and subsequently the m_{Wh} observable. A comparison of $w_{\text{BDT}}^{\text{max}}$ for events passing the pre-selection

higher miscategorisation of events from the $\ell\nu b\bar{b}$ channel.



(a) (b)

Figure 5.10: (a) Comparison of $w_{\text{BDT}}^{\text{max}}$ for two signal mass hypotheses and the sum of backgrounds. (b) Comparison of m_{Wh} for a range of signal mass hypotheses.

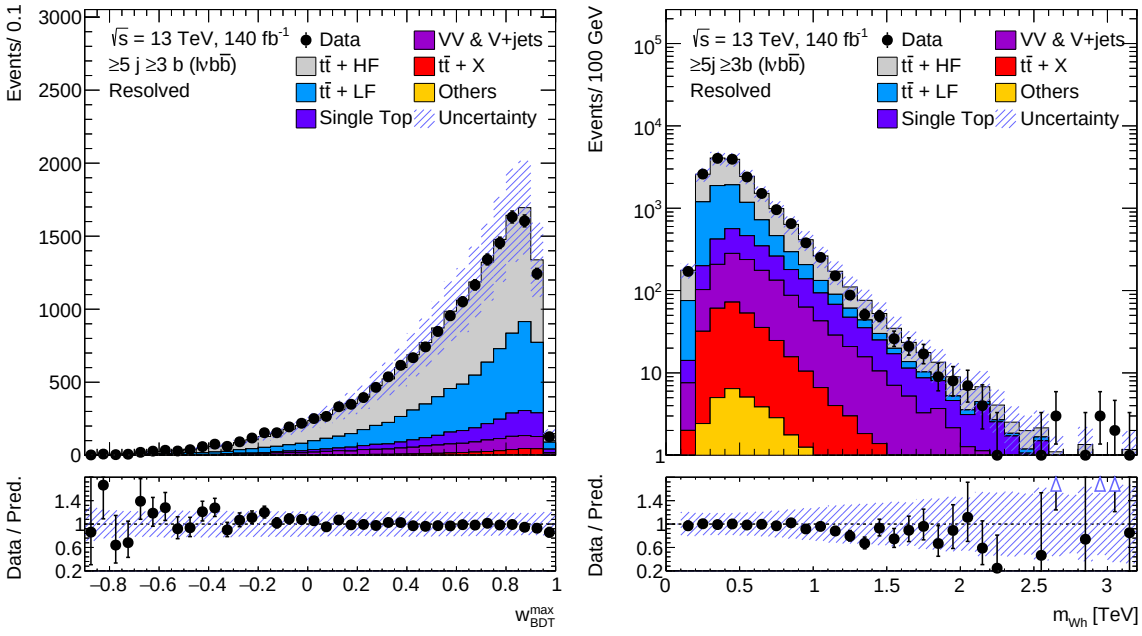
requirements for two signal mass hypotheses (m_{H^+} = 400 GeV and m_{H^+} = 800 GeV) and the sum of backgrounds is shown in Figure 5.10(a). One can observe a reasonable separation between the signal and background distributions. A comparison of m_{Wh} for events passing the pre-selection requirements for a range of signal mass hypotheses is presented in Figure 5.10(b). The m_{Wh} distributions for different signal mass hypotheses are centered around their pole masses. Data-to-simulation comparsions for $w_{\text{BDT}}^{\text{max}}$ and m_{Wh} in events containing at least three b-tagged jets are shown in Figure 5.11. A good agreement between data and simulated backgrounds is observed within the uncertainties. Data-to-simulation comparsions for BDT input variables in events containing at least three b-tagged jets are shown in Figures A.12 and A.13 in the Appendix. A detailed description of the BDT training and its performance in the $\ell\nu b\bar{b}$ channel can be found in [176].

5.2 Signal and control regions

The definition of signal and control regions (SRs and CRs) used in the search for resolved $H^+ \to Wh$ decays is based on $w_{\text{BDT}}^{\text{max}}$ allocated to an event. This is motivated by the fact that the backgrounds tend to have lower $w_{\text{BDT}}^{\text{max}}$ values than the signal process. Events are categorised according to the jet (j) and b-tagged jet (b) multiplicities. In this context, four exclusive categories are defined: $5j\ 3b$, $5j\ \geq 4b$, $\geq 6j\ 3b$, and $\geq 6j\ \geq 4b$. A dedicated signal region and a control region are defined for each event category. The m_{Wh} observable is used as the final discriminant between signal and backgrounds to determine the sensitivity for $H^+ \to Wh(\to b\bar{b})$.

5.2.1 $q\bar{q}b\bar{b}$ channel

Three selection requirements, namely $w_{\text{BDT}}^{\text{max}} \geq 0.7$, $w_{\text{BDT}}^{\text{max}} \geq 0.8$, and $w_{\text{BDT}}^{\text{max}} \geq 0.9$, are tested to define the signal region. The selection requirements are set around high $w_{\text{BDT}}^{\text{max}}$ values due to high signal



(a) (b)

Figure 5.11: Data-to-simulations comparison for (a) $w^{\text{max}}_{\text{BDT}}$ and (b) m_{Wh} in events containing at least three b-tagged jets in the $\ell\nu b\bar{b}$ channel. The background uncertainty band shows the statistical and systematic uncertainties added in quadrature. The notation "$\geq 5j$ $\geq 3b$" refers to the jet (j) and b-jet multiplicities (b).

significance ($S/\sqrt{B}$)[9] in these $w^{\text{max}}_{\text{BDT}}$ intervals. A signal region is chosen based on the most stringent 95% CL expected upper limits on the signal strength, $\sigma(pp \to tbH^+) \times BR(H^+ \to Wh(\to b\bar{b}))$, obtained for the tested selection requirements on $w^{\text{max}}_{\text{BDT}}$. The upper limits are calculated using an Asimov fit to the m_{Wh} distribution, in a region containing at least five jets and at least three b-tagged jets, under a signal-plus-background hypothesis considering statistical uncertainties on simulated events only. A comparison of limits for the three tested selection requirements is shown in Figure 5.12. It can be observed that the requirement $w^{\text{max}}_{\text{BDT}} \geq 0.9$ leads to better limits in comparison to other selection requirements. The order of the limit curves reverses for higher H^+ boson masses ($m_{H^+} = 2.5$ TeV and $m_{H^+} = 3$ TeV) for increasingly tighter selection requirements on $w^{\text{max}}_{\text{BDT}}$. The signal selection efficiency drops significantly for tighter selection requirements owing to the shape of the $w^{\text{max}}_{\text{BDT}}$ distribution (see Figure 5.6(a)). The search for merged $H^{\pm} \to Wh$ decays (see Chapter 7) is expected to be more sensitive at high H^+ boson masses, therefore the decision on the selection requirement is taken based on the performance at low H^+ boson masses. Hence, setting a requirement of 0.9 seems to be a suitable choice to define a signal region.

Control regions are defined by scanning $S/\sqrt{B}$ for all signal mass hypotheses for different intervals of $w^{\text{max}}_{\text{BDT}}$ in events containing three b-tagged jets and at least four b-tagged jets[10] as shown in Figures

[9] A $S/\sqrt{B}$ significance metric is used where S is the number of signal events, and B is the number of background events in the region under consideration.

[10] Unlike signal regions, dedicated control regions are optimised for regions split in different b-tagged jet multiplicities to constraint the background model better in regions of kinematically different phase space.

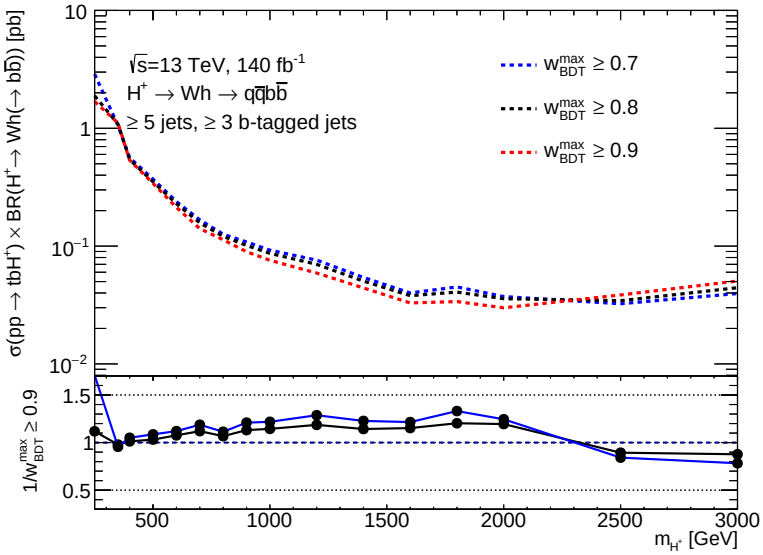


Figure 5.12: 95% CL expected upper limits on $\sigma(tbH^+) \times BR(H^+ \rightarrow Wh(\rightarrow b\bar{b}))$ for different $w_{\mathrm{BDT}}^{\mathrm{max}}$ requirements in the $q\bar{q}b\bar{b}$ channel.

5.13 and 5.14, respectively. The $w_{\mathrm{BDT}}^{\mathrm{max}}$ interval with low signal significance and large number of background events is chosen as a control region. Based on this criteria, control regions are defined in the range -0.5 $\leq w_{\mathrm{BDT}}^{\mathrm{max}} <$ 0.0 for the $5j$ $3b$ and $\geq 6j$ $3b$ event categories, and in the range -0.5 $\leq w_{\mathrm{BDT}}^{\mathrm{max}} <$ 0.6 for the $5j$ $\geq 4b$ and $\geq 6j$ $\geq 4b$ event categories.

Additional signal regions are defined in the $q\bar{q}b\bar{b}$ channel to gain any signal sensitivity which can arise from within the large-gap region in the $w_{\mathrm{BDT}}^{\mathrm{max}}$ distribution between the signal and control regions. These regions are defined in the range $0.0 \leq w_{\mathrm{BDT}}^{\mathrm{max}} < 0.9$ for the $5j3b$ and $\geq 6j3b$ event categories, and in the range $0.6 \leq w_{\mathrm{BDT}}^{\mathrm{max}} < 0.9$ for the $5j$ $\geq 4b$ and $\geq 6j$ $\geq 4b$ event categories. The inclusion of these regions improves the signal sensitivity by 60% for low H^+ boson masses. These regions are termed as "low-purity" signal regions, and the signal regions described earlier are thereby termed as "high-purity" signal regions. The gain in the signal sensitivity by including the low-purity signal regions is shown in Figure 5.15. A schematic representing the split of the $w_{\mathrm{BDT}}^{\mathrm{max}}$ spectrum in signal and control regions for each event category of the $q\bar{q}b\bar{b}$ channel is shown in Figure 5.16.

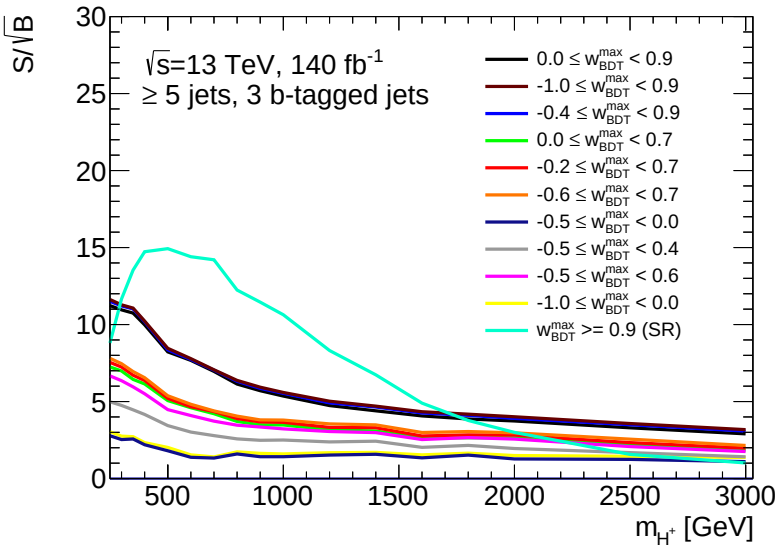


Figure 5.13: Significance scans for a range of $w_{\text{BDT}}^{\text{max}}$ intervals for events containing exactly 3 b-tagged jets in the $q\bar{q}b\bar{b}$ channel.

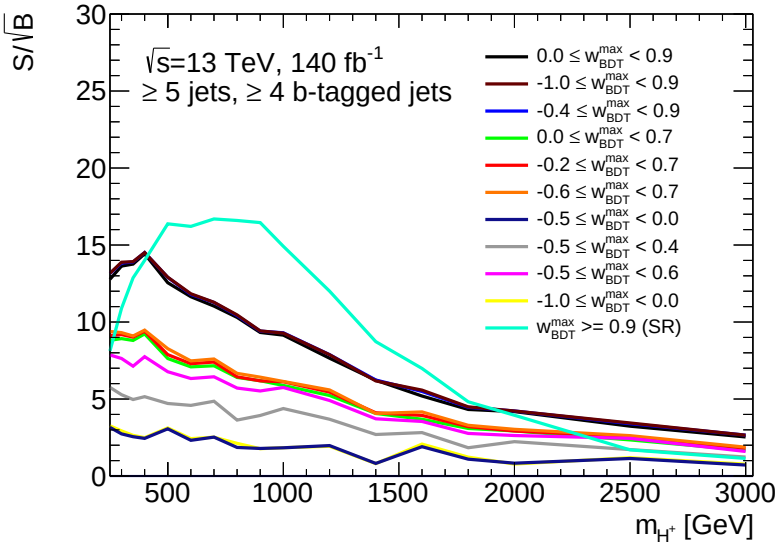


Figure 5.14: Significance scans for a range of $w_{\text{BDT}}^{\text{max}}$ intervals for events containing ≥ 4 b-tagged jets in the $q\bar{q}b\bar{b}$ channel.

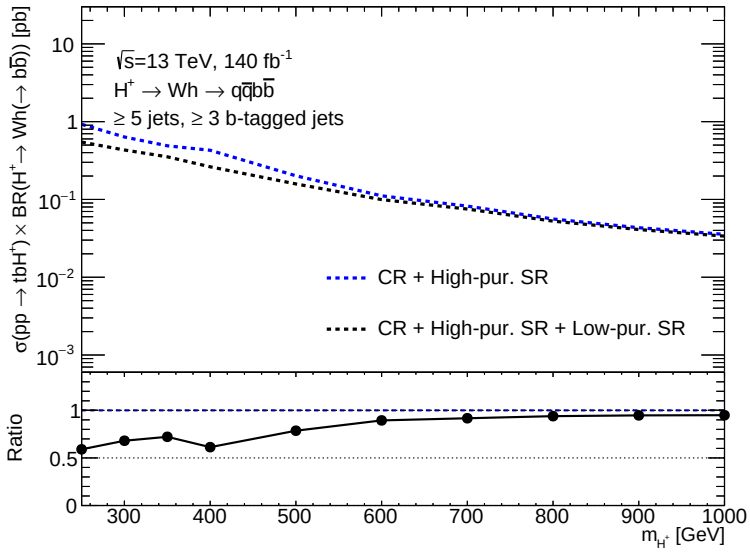


Figure 5.15: 95% CL expected upper limits on $\sigma(pp \to tbH^+) \times BR(H^+ \to Wh(\to b\bar{b}))$ in the $q\bar{q}b\bar{b}$ channel with and without including the low-purity signal regions. The upper limits are estimated using an Asimov fit to the m_{Wh} distribution under a signal-plus-background hypothesis. The ratio panel indicates the ratio of limits obtained from the control and low-purity signal regions to the limits obtained from the control, low-purity, and high-purity signal regions.

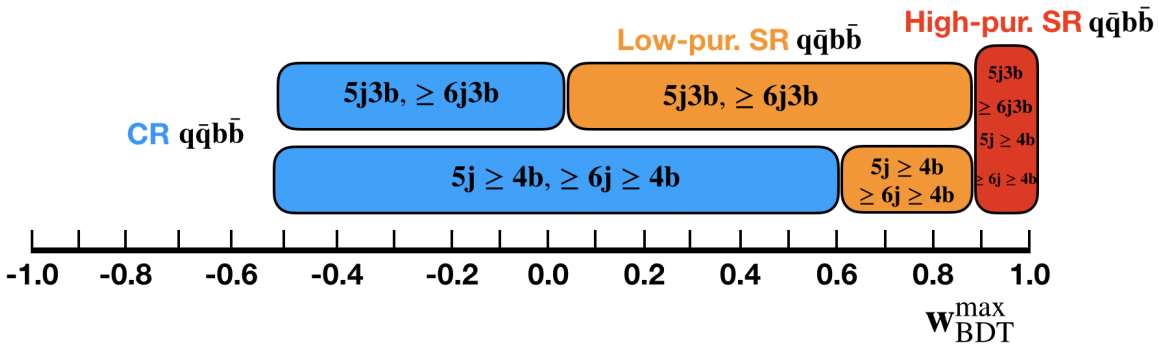


Figure 5.16: A schematic representing the definitions of control and signal regions in the $q\bar{q}b\bar{b}$ channel.

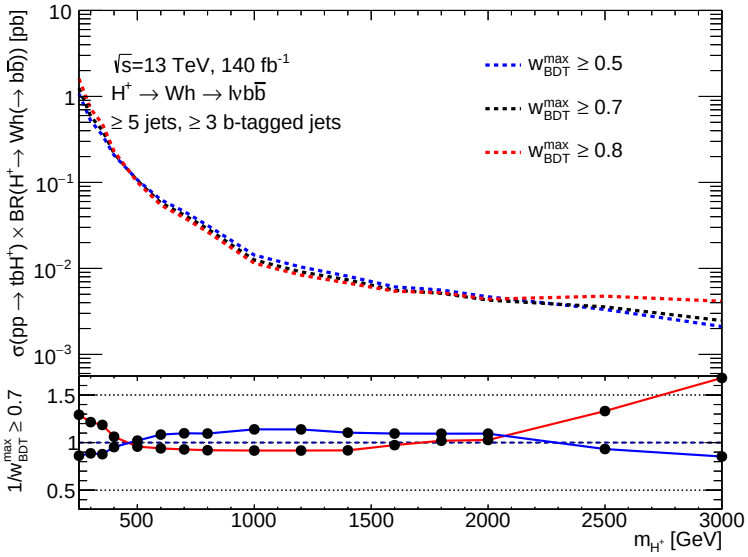


Figure 5.17: 95% CL expected upper limits on $\sigma(pp \rightarrow tbH^+) \times BR(H^+ \rightarrow Wh(\rightarrow b\bar{b}))$ for different $w_{\text{BDT}}^{\text{max}}$ selection requirements in the $\ell\nu b\bar{b}$ channel. The upper limits are estimated using an Asimov fit to the m_{Wh} distribution under a signal-plus-background hypothesis.

5.2.2 $\ell\nu b\bar{b}$ channel

A similar procedure as in the $q\bar{q}b\bar{b}$ channel is used to define a signal region in the $\ell\nu b\bar{b}$ channel. 95% CL expected upper limits on $\sigma(pp \rightarrow tbH^+) \times BR(H^+ \rightarrow Wh(\rightarrow b\bar{b}))$ are estimated for events containing at least three b-tagged jets for different selection requirements on $w_{\text{BDT}}^{\text{max}}$, namely $w_{\text{BDT}}^{\text{max}} \geq 0.5$, $w_{\text{BDT}}^{\text{max}} \geq 0.7$, and $w_{\text{BDT}}^{\text{max}} \geq 0.8$, as shown in Figure 5.17. The selection requirements corresponding to $w_{\text{BDT}}^{\text{max}} \geq 0.7$ and $w_{\text{BDT}}^{\text{max}} \geq 0.8$ give most stringent limits for a large range of H^+ boson masses. However, $w_{\text{BDT}}^{\text{max}} \geq 0.7$ performs better than $w_{\text{BDT}}^{\text{max}} \geq 0.8$ with an improvement of 20% to 30% for the first three mass hypotheses. The performance for both these selection requirements are comparable for intermediate masses. At high H^+ boson masses, $w_{\text{BDT}}^{\text{max}} \geq 0.7$ again outperforms $w_{\text{BDT}}^{\text{max}} \geq 0.8$ due to higher signal yields in the former region. Hence, $w_{\text{BDT}}^{\text{max}} \geq 0.7$ is chosen as a signal region which should deliver a good trade-off between the expected sensitivity and a sufficient number of signal events in the $\ell\nu b\bar{b}$ channel.

Control regions in the $\ell\nu b\bar{b}$ channel are defined in the range -0.5 $\leq w_{\text{BDT}}^{\text{max}} <$ 0.0 for all four event categories. The optimisation procedure of these regions is detailed in [176]. A schematic representing the split of the $w_{\text{BDT}}^{\text{max}}$ spectrum into signal and control regions for all event categories of the $\ell\nu b\bar{b}$ channel is shown in Figure 5.18.

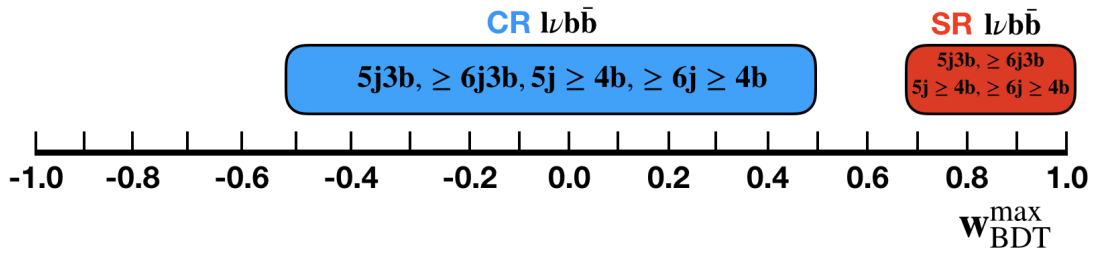
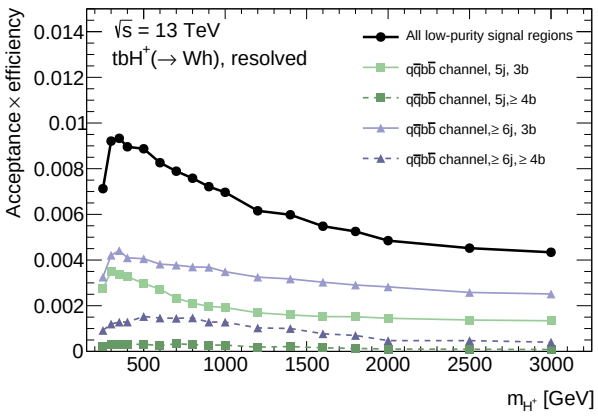


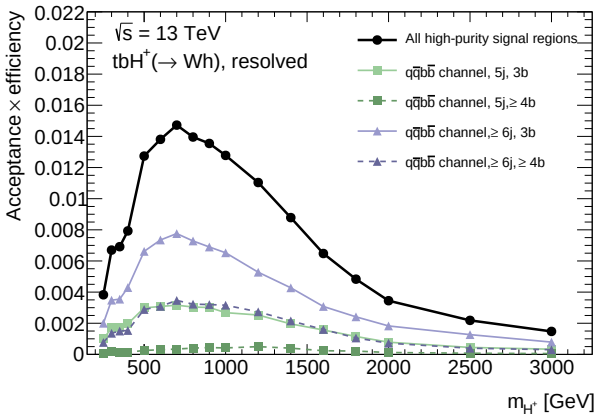
Figure 5.18: A schematic representing the definitions of the control, signal, and low-purity signal regions in the $\ell\nu b\bar{b}$ channel.

5.2.3 Acceptance times efficiency for the signal regions

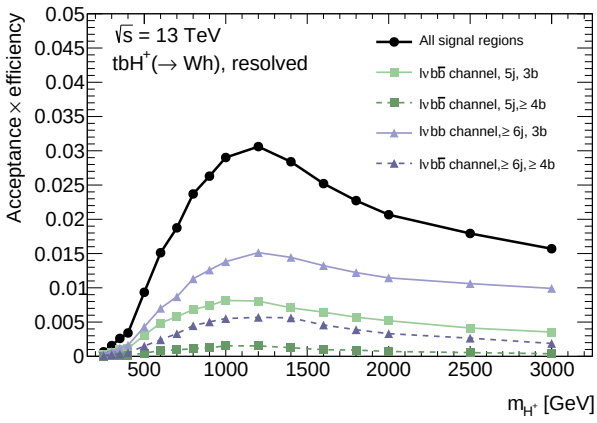
The product of kinematic acceptance and reconstruction efficiency for all signal regions in the $\ell\nu b\bar{b}$ and $q\bar{q}b\bar{b}$ channels is presented in Figure 5.19. In this context, the acceptance is defined as the fraction of simulated signal events for which the expected final state particles pass all object definition requirements (see Section 4.3). The reconstruction efficiency is defined as the fraction of simulated signal events passing the selection criteria of a given signal region. The product of acceptance and efficiency ranges from 0.1% to 3%, and is seen to be highest for the signal regions in the $\ell\nu b\bar{b}$ channel which is expected as these regions cover a wider signal-enriched $w^{\text{max}}_{\text{BDT}}$ spectrum. The $\geq 6j\ 3b$ event category contributes the most among all the signal regions. For the low-purity signal regions in the $q\bar{q}b\bar{b}$ channel, the maximum is observed at $m_{H^+} = 350\,\text{GeV}$ as these regions improve the signal sensitivity for low H^+ boson masses (see Section 5.2.1). For the high-purity signal regions in the $q\bar{q}b\bar{b}$ channel, the maximum is observed at $m_{H^+} = 700\,\text{GeV}$, and at $m_{H^+} = 1.2\,\text{TeV}$ for signal regions in the $\ell\nu b\bar{b}$ channel. The drop at increasingly higher m_{H^+} is due to a lower number of small-R jets passing the selection requirements. This happens due to the large Lorentz boost of the H^+ boson at high masses resulting in the collimation of the h and W boson decay products, and their possible reconstruction as single large-R jets which are not included in the selection of the search for resolved H^+ boson decays.



(a) Resolved low-purity SRs in the $q\bar{q}b\bar{b}$ channel



(b) Resolved high-purity SRs in the $q\bar{q}b\bar{b}$ channel



(c) Resolved SRs in the $\ell\nu b\bar{b}$ channel

Figure 5.19: Product of acceptance and efficiency as a function of the H^+ boson mass for (a) resolved low-purity SRs in the $q\bar{q}b\bar{b}$ channel, (b) resolved high-purity SRs in the $q\bar{q}b\bar{b}$ channel, and (c) resolved SRs in the $\ell\nu b\bar{b}$ channel.

CHAPTER 6

Resolved Topology: Statistical Framework

This chapter focuses on the statistical model used to scrutinise the reconstructed charged Higgs boson mass distribution for the presence of an excess in data events over the SM background prediction. Section 6.1 introduces the model used in the search for resolved $H^+ \to Wh(\to b\bar{b})$ decays. Section 6.2 discusses the systematic uncertainties and their treatment in the statistical model. Finally, Section 6.3 presents the various studies performed in order to validate the statistical model.

6.1 Statistical Model

The signal-plus-background hypothesis for the presented search is tested with a binned maximum-likelihood fit performed in the `RooStats` framework based on the `RooFit` toolkit [192]. As discussed in Section 3.6.1, the likelihood function L enters the *profile-likelihood ratio* as a function of the signal strength, μ, and a set of nuisance parameters (NPs), $\vec{\alpha}$. The nuisance parameters are either left unconstrained, i.e. they are kept as independently floating in the fit, or they are left constrained where a Gaussian probability distribution is chosen for the systematic uncertainties, and a Poisson probability distribution is chosen for the statistical uncertainty of the simulated processes in each bin of the m_{Wh} distribution. The floating nuisance parameters are generally the normalisation factors of the different background processes. In the search for resolved $H^+ \to Wh(\to b\bar{b})$ decays, the overall normalisation factor of the $t\bar{t}$ + HF background (k_{HF}) is left independently floating in the fit under any tested hypothesis.

The explicit form of the likelihood function is a double-product of the Poissonian probabilities in the regions r and bins b multiplied by the product of $P(\vec{\alpha}_j)$ denoting the constraint on the j^{th} nuisance parameter:

$$L(\mu, \vec{\alpha}) = \prod_r \prod_b \mathrm{Poiss}\left(n_{rb}^{\mathrm{obs}} \mid n_{rb}^{\mathrm{sig}}(\mu, \vec{\alpha}) + n_{rb}^{\mathrm{bkg}}(\vec{\alpha})\right) \prod_{\alpha_j} P(\vec{\alpha}_j) \tag{6.1}$$

where the indices r and b denote the b^{th} bin of the m_{Wh} distribution in the r^{th} region of the analysis. n_{rb}^{obs}, n_{rb}^{sig}, and n_{rb}^{bkg} are the numbers of observed data, signal, and background events, respectively. As described in Section 5.2, the search for resolved $H^+ \to Wh(\to b\bar{b})$ decays is divided into several control and signal regions in both the $q\bar{q}b\bar{b}$ and $\ell\nu b\bar{b}$ channels. In total, these amount to 20 regions split in different jet and b-jet multiplicity regions that enter the maximum-likelihood fit as summarised in Table 6.1.

Binning of the m_{Wh} distribution The m_{Wh} distribution in the control regions is divided into 6 bins covering a mass range from 0 to 3.6 TeV with an exception of the $5j \geq 4b$ and $\geq 6j \geq 4b$ $\ell\nu b\bar{b}$ channel control regions which are defined as single-bin regions due to fewer background events. The first 5 bins are placed equidistantly with the 6^{th} bin covering the remaining events of the m_{Wh} distribution. Each bin contain at least five background events[1]. The binning in the signal regions is optimised using the following algorithm based on the m_{Wh} resolution and the number of background events. The m_{Wh} distribution in the signal regions is intially divided uniformly into 200 bins, and the merging of the bins is initiated from the right end of the distribution to the left end until the following conditions are met:

- A minimum of five background events in each bin.
- A minimum bin width of 200 GeV. The choice of this bin width is motivated from the m_{Wh} resolution. The average m_{Wh} resolution in the probed H^+ boson mass range is 10%. For a benchmark H^+ boson mass of 1 TeV, a minimum bin width of 200 GeV - twice the value of the average mass resolution times 1 TeV - is a justified initial value.
- An exclusive condition of merging the leftmost bin of the distribution to the adjacent bin is applied, in case the aforementioned conditions are not satisfied for the leftmost bin.

The m_{Wh} distributions in the control and signal regions are shown in Figures 6.1 to 6.5.

Table 6.1: A summary of the selection requirements used to define the signal and control regions in the search for resolved $H^{\pm} \to Wh(b\bar{b})$ decays. The regions are defined for the $5j\ 3b$, $5j \geq 4b$, $\geq 6j\ 3b$, and $\geq 6j \geq 4b$ event categories. All the requirements are applied to all event categories, unless explicitly stated.

Region	$\boldsymbol{w}_{\mathrm{BDT}}^{\mathrm{max}}$ $(\ell\nu b\bar{b})$	$\boldsymbol{w}_{\mathrm{BDT}}^{\mathrm{max}}$ $(q\bar{q}b\bar{b})$
(High-purity) signal regions	$w_{\mathrm{BDT}}^{\mathrm{max}} \geq 0.7$	$w_{\mathrm{BDT}}^{\mathrm{max}} \geq 0.9$
Low-purity signal regions	-	$0.0 \leq w_{\mathrm{BDT}}^{\mathrm{max}} < 0.9$ (for $5j\ 3b$ and $\geq 6j\ 3b$ categories) $0.6 \leq w_{\mathrm{BDT}}^{\mathrm{max}} < 0.9$ (for $5j \geq 4b$ and $\geq 6j \geq 4b$ categories)
Control regions	$-0.5 \leq w_{\mathrm{BDT}}^{\mathrm{max}} < 0.5$	$-0.5 \leq w_{\mathrm{BDT}}^{\mathrm{max}} < 0.0$ (for $5j\ 3b$ and $\geq 6j\ 3b$ categories) $-0.5 \leq w_{\mathrm{BDT}}^{\mathrm{max}} < 0.6$ (for $5j \geq 4b$ and $\geq 6j \geq 4b$ categories)

[1] This requirement ensures that the distribution of the test-statistic $t(\mu)$ (see Section 3.6.1) can be determined using an asymptotic approximation.

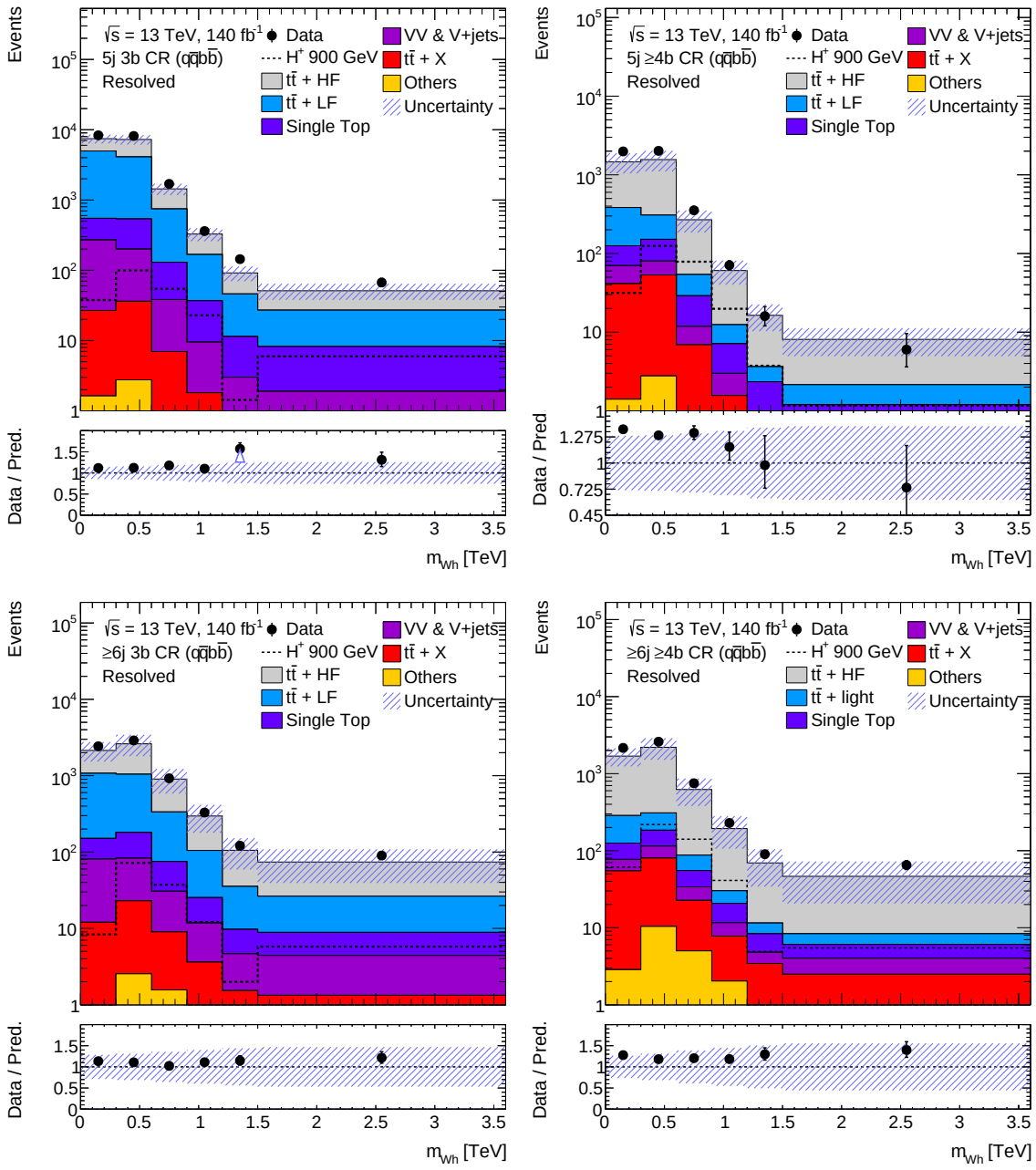


Figure 6.1: Data-to-simulation comparisons for the m_{Wh} in the control regions of the resolved $q\bar{q}b\bar{b}$ channel. The corresponding distribution for $m_{H^+} = 900\,\mathrm{GeV}$ (normalised to a cross section of 1 pb) is overlaid. The uncertainty band shows the statistical and systematic uncertainties for all the background processes added in quadrature.

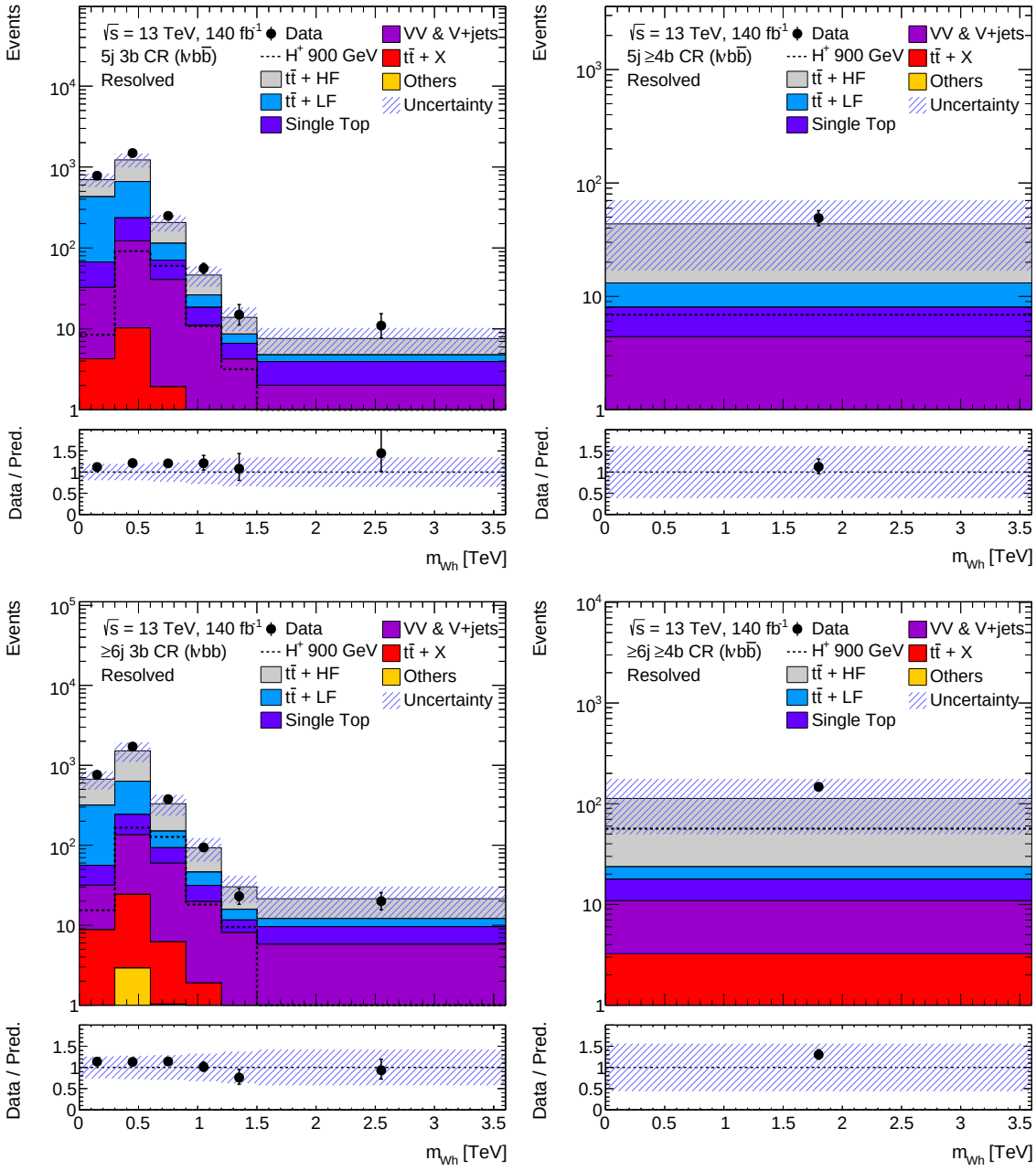


Figure 6.2: Data-to-simulation comparisons for the m_{Wh} in the control regions of the resolved $\ell\nu b\bar{b}$ channel. The corresponding distribution for m_{H^+} = 900 GeV (normalised to a cross section of 1 pb) is overlaid.The uncertainty band shows the statistical and systematic uncertainties for all the background processes added in quadrature.

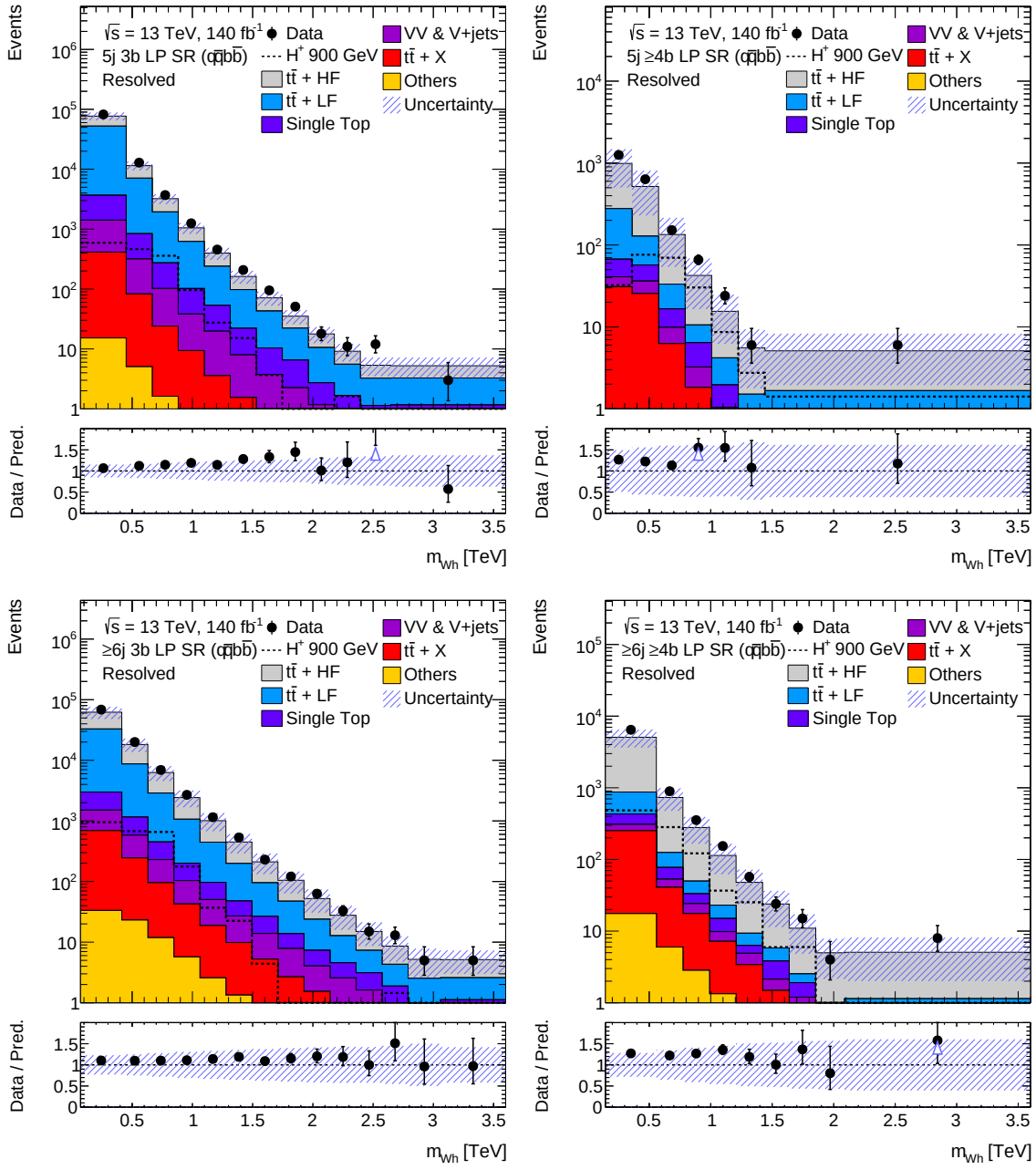


Figure 6.3: Data-to-simulation comparisons for the m_{Wh} in the low-purity (LP) signal regions of the resolved $q\bar{q}b\bar{b}$ channel. The corresponding distribution for $m_{H^+} = 900\,\mathrm{GeV}$ (normalised to a cross section of 1 pb) is overlaid. The uncertainty band shows the statistical and systematic uncertainties for all the background processes added in quadrature.

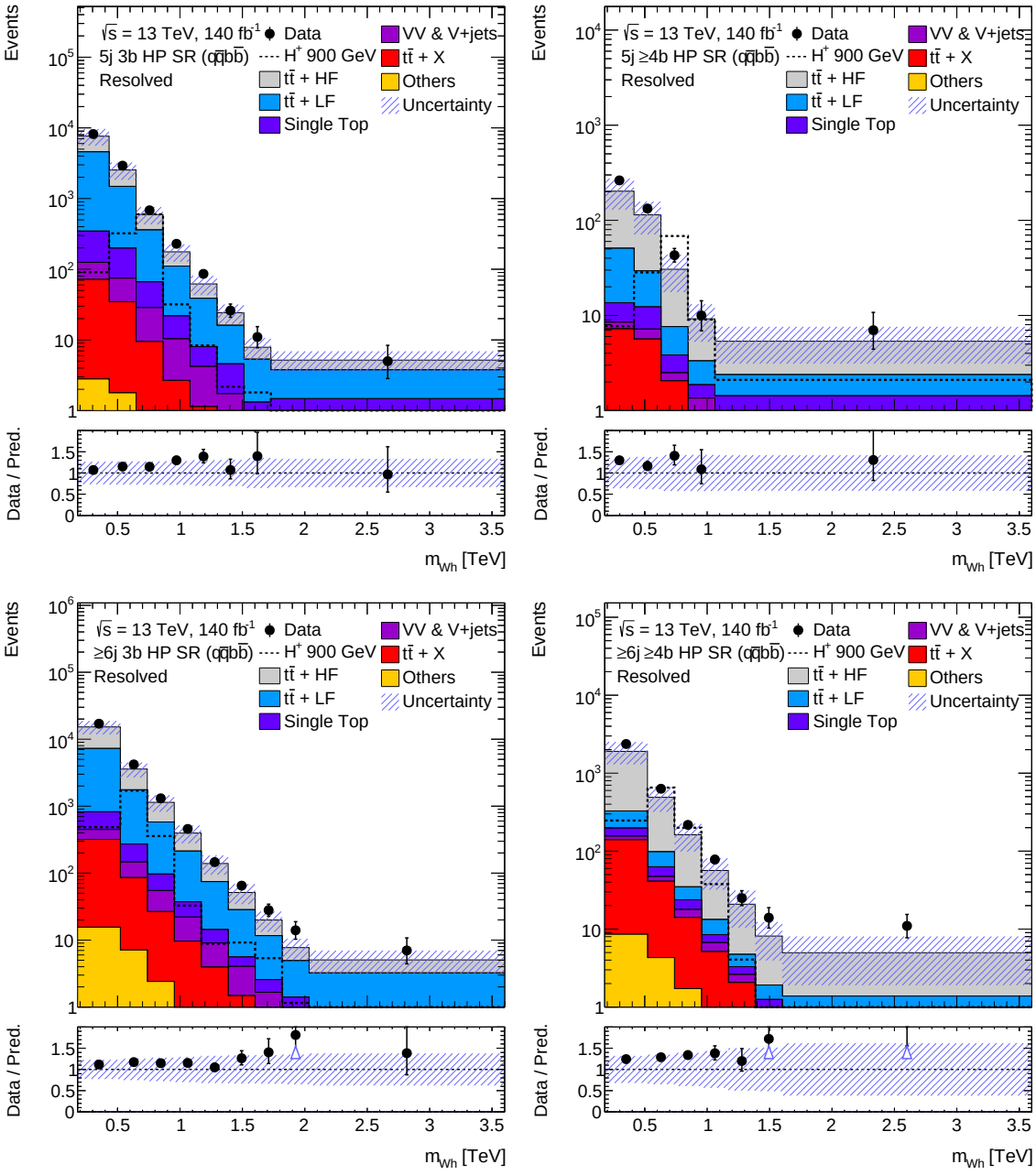


Figure 6.4: Data-to-simulation comparisons for the m_{Wh} in the high-purity (HP) signal regions of the resolved $q\bar{q}b\bar{b}$ channel. The corresponding distribution for $m_{H^+} = 900\,\mathrm{GeV}$ (normalised to a cross section of 1 pb) is overlaid. The uncertainty band shows the statistical and systematic uncertainties for all the background processes added in quadrature.

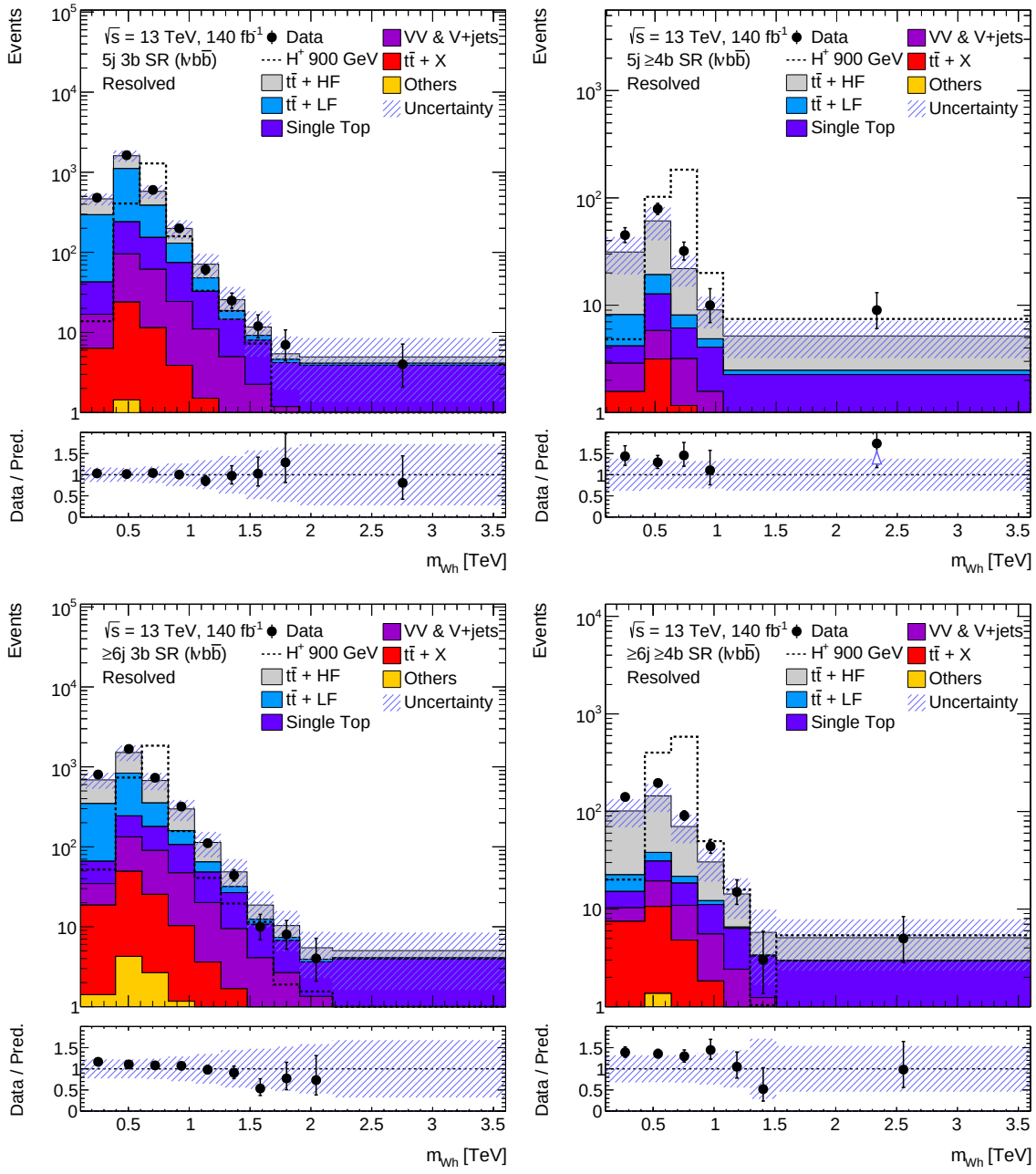


Figure 6.5: Data-to-simulation comparisons for the m_{Wh} in the signal regions of the resolved $\ell\nu b\bar{b}$ channel. The corresponding distribution for $m_{H^+} = 900\,\mathrm{GeV}$ (normalised to a cross-section of 1 pb) is overlaid. The uncertainty band shows the statistical and systematic uncertainties for all the background processes added in quadrature.

Table 6.2: List of experimental systematic uncertainties used in the search for resolved $H^+ \to Wh(\to b\bar{b})$ decays. This table presents the uncertainty type, specifies the uncertainty impact (shape (S), normalisation (N), or both S and N), and the number of associated nuisance parameter components that enter the likelihood.

Type	**Impact**	**Components**
Luminosity	N	1
Pileup modelling	SN	1
Physics objects		
Electrons	SN	7
Muons	SN	15
Jet energy scale	SN	31
Jet energy resolution	SN	9
Jet vertex tagger	SN	1
$E_{\mathrm{T}}^{\mathrm{miss}}$	SN	3
Flavour-tagging of jets		
b-tagging efficiency	SN	45
c-tagging mistag rate	SN	20
light-flavour tagging mistag rate	SN	20
high p_{T} extrapolation	SN	3

6.2 Systematic Uncertainties

Systematic uncertainties enter the likelihood as nuisance parameters. The uncertainties are classified into two categories, namely experimental uncertainties and modelling uncertainties, and are taken into account for all simulated processes. Experimental uncertainties are associated with the reconstruction of various physics objects. Modelling uncertainties are associated with the normalisation of simulated processes and also take into account the shape differences in the m_{Wh} distribution as observed from simulations generated using different event generators (see Section 3.3). The experimental and modelling uncertainties can either impact the overall normalisation (denoted by N), the shape (denoted by S), or both the shape and normalisation of the m_{Wh} distribution. The uncertainties are evaluated by comparing the nominal m_{Wh} distribution to the m_{Wh} distribution produced by varying the source of uncertainty by $\pm 1\sigma$. This section describes in detail the various sources of experimental and modelling uncertainties used in the search for resolved $H^+ \to Wh(\to b\bar{b})$ decays. A list of the systematic uncertainties with the number of associated NP components that enter the likelihood is shown in Tables 6.2 and 6.3.

Table 6.3: List of theoretical systematic uncertainties used in the search for resolved $H^+ \to Wh(\to b\bar{b})$ decays. This table presents the uncertainty type, specifies the uncertainty impact (shape (S), normalisation (N), or both S and N), and the number of associated nuisance parameter components that enter the likelihood.

Type	**Impact**	**Components**
Signal and background modelling		
Signal		
PDF	N	1
μ_F, μ_R scale	SN	2
Parton Shower	N	1
$t\bar{t}$ + *jets background*		
PDF	SN	30
$t\bar{t}$ + jets reweighting	SN	24
$t\bar{t}$ + HF normalisation	N (*free floating*)	1
$t\bar{t}$ + LF modelling	SN	21
$t\bar{t}$ + HF modelling	S	57
Cross-region extrapolation	S	5
Heavy-flavour composition	S	1
Remaining backgrounds		
$t\bar{t} + W$ cross-section	N	2
$t\bar{t} + Z$ cross-section	N	2
$t\bar{t} + W$ modelling	SN	1
$t\bar{t} + Z$ modelling	SN	1
$t\bar{t} + h$ cross-section	N	1
$t\bar{t} + h$ modelling	SN	2
Single Top cross-section	N	6
Single Top modelling	SN	12
W + jets normalisation	N	1
Z + jets normalisation	N	1
VV normalisation	N	1
Other backgrounds cross-section	N	2
All simulated processes		
Non-closure	SN	1

6.2.1 Experimental Uncertainties

This section describes the various kinds of experimental uncertainties. These uncertainties are correlated across all analysis regions and among all simulated processes, and can impact both the normalisation and the shape of the m_{Wh} distribution.

Luminosity and pile-up modelling

The uncertainty on the integrated luminosity for the full ATLAS Run 2 dataset is considered as 0.83% [193]. It is the only experimental uncertainty that impacts the overall normalisation of all simulated processes. The measurement of the luminosity is described in Section 3.1.2.

An uncertainty related to the variation in the pile-up simulation is also taken into account. This uncertainty covers the inaccuracies in the ratio of the predicted and measured inelastic cross-sections in the fiducial[2] volume defined for $M_X > 13$ GeV, where M_X is the mass of the hadronic system [194].

Uncertainties on the reconstructed objects

Charged leptons: Uncertainties associated with the charged leptons arise from the trigger selection, object reconstruction, identification, and isolation criteria (see Section 3.4.2), as well as from the lepton momentum scale and resolution.

The trigger, reconstruction, identification, and isolation efficiencies of electrons and muons differ between data and simulation which is compensated for by the usage of dedicated efficiency scale factors. These efficiency scale factors are derived using tag-and-probe techniques in $Z \to l^+l^-$ events [195, 196]. The effect of these scale factors as well as their uncertainties are propagated as corrections to all simulated events.

Other sources of uncertainty include those arising from the corrections applied to simulated events to adjust the lepton momentum scale and resolution in order to reach an optimal agreement in data. These uncertainties are calculated using the reconstructed mass of the two leptons from $Z \to l^+l^-$ and $J/\psi \to l^+l^-$ decays, as well as using the E/p ratio measured in $W \to e\nu$ events, where E and p are the electron energy and momentum measured by the calorimeter and tracker, respectively [196, 197].

Jets: There are multiple sources of jet uncertainties, namely the ones arising from the efficiency of the pile-up rejection by the jet vertex tagging (JVT), the jet energy scale (JES), the jet energy resolution (JER), and the flavour-tagging of jets.

Dedicated scale factors are applied to correct for the differences between data and simulation for the JVT efficiencies. These scale factors are estimated with $Z \to \mu^+\mu^-$ events using tag-and-probe techniques [198]. The effect of these scale factors as well as their associated uncertainties are propagated as corrections to all simulated events. The uncertainties related to the JES are derived by combining the information from data and simulation [199]. These uncertainties include the ones related to the jet flavour which assumes a conservative default value of ±50% (relative uncertainty) on the quark-gluon fraction for the simulation of jets with different flavours. Other uncertainties include the ones from the pile-up corrections, jet kinematics, and the differences between the full and fast calorimeter simulation (see Section 3.3). The JER is measured in simulation and data, as a function of

[2] A fiducial region is a region defined at the parton level that corresponds closely to a region experimentally accessible by the detector.

jet p_{T} and η using dijet events [200]. The combined uncertainty is propagated by smearing the jet p_{T} in simulation.

The b-tagging of jets is crucial for the presented search, and so are its associated uncertainties. Efficiencies to tag jets containing b-hadrons, or mistag jets containing c-hadrons and jets containing neither b- nor c-hadrons (light-flavoured jets) are corrected in simulation to match their efficiencies in data. The efficiency scale factors are derived as a function of jet p_{T} in dedicated calibration analyses. b-jet tagging efficiencies are measured using $t\bar{t}$ events in a dilepton topology making use of the pure selection of b-jets arising from the decay of the top-quarks [201]. c-jet mistag rates are calculated using $t\bar{t}$ events in a single-lepton topology using c-jets from hadronically decaying W bosons [202]. On the other hand, light-flavoured jet mistag rates are derived using the so-called *negative-tag method*[3] [203] in Z + jets events. Additional uncertainty components which extrapolate the aforementioned uncertainties to p_{T} regions not covered by the calibration analyses are also considered [204]. These uncertainties are calculated from simulated events containing the decay of Z' bosons to quark-antiquark pairs [205] by considering variations of the quantities affecting the b-tagging performance such as the impact parameter resolution, percentage of tracks from random combinations of measurements in the ID, description of the detector material, and track multiplicity per jet.

Missing transverse momentum: As mentioned in Section 3.4.5, $E_{\mathrm{T}}^{\mathrm{miss}}$ is calculated from the reconstructed physics objects (hard-scatter objects) and a soft term not associated with any reconstructed objects. Therefore, the $E_{\mathrm{T}}^{\mathrm{miss}}$ reconstruction is affected by the uncertainties associated with leptons, JES, and JER. Additional uncertainties in the scale and resolution of the soft term are also considered. These uncertainties account for the disagreement between data and simulation for the p_{T} balance between the hard and soft components [160, 161].

6.2.2 Modelling Uncertainties

Unlike the experimental uncertainties, the modelling uncertainties are not correlated across all background and signal processes. However, they are correlated across analysis regions with few exceptions related to modelling uncertainties of the $t\bar{t}$ + jets and single-top backgrounds, as discussed in this section. The modelling uncertainties can impact both the shape and the normalisation of the m_{Wh} distribution.

Signal modelling

Three kinds of signal modelling uncertainties are considered for the search presented in this thesis. Firstly, the uncertainties associated with the PDF set are evaluated by replacing the nominal NNPDF3.0NNLO PDF set [206] by the alternative CT14 [207] and MMHT2014 [208] PDF sets. The PDF uncertainty is evaluated by comparing the maximum difference between the alternative and the nominal PDF sets with the difference of the alternative PDF set from the root-mean-squared spread of the nominal sets. The larger of the two differences is taken as the uncertainty. Secondly, the uncertainty associated with the parton shower (PS) is evaluated by comparing the prediction of the

[3] Negative-tag method relies on a modified tagger with reduced b-jet tagging and c-jet mistagging efficiencies but a similar light-flavoured jet mistagging efficiency w.r.t the nominal tagger.

nominal event generator, MadGraph 5[4] + Pythia 8, with an alternative[5] event generator, MadGraph 5 + Herwig 7. Lastly, the QCD scale uncertainties are derived by varying μ_F and μ_R (the factorisation and renormalisation scales) by a factor of 0.5 (2.0) for the $+1\sigma$ (-1σ) variation. No additional signal modelling uncertainty is considered, and the most relevant signal acceptance effects are covered in the aforementioned ones.

Background modelling

***$t\bar{t}$* + jets** As discussed in Section 4.2, the $t\bar{t}$ + jets background is split into two categories depending on the flavour of the additional jets in an event, namely $t\bar{t}$ + HF (heavy-flavour) and $t\bar{t}$ + LF (light-flavour). The $t\bar{t}$ + jets background is known to be mismodelled due to missing higher-order QCD and EW corrections, and is corrected using a reweighting procedure (see Section 4.4). The error variations (uncertainty components) resulting from the variation of the reweighting corrections within their statistical uncertainties are constrained in the statistical model, and impact both $t\bar{t}$ + LF and $t\bar{t}$ + HF backgrounds. These uncertainty components change the normalisation of the $t\bar{t}$ + LF background within 6% which is the cross-section uncertainty on the $t\bar{t}$ + jets background, and are thus sufficient to estimate the $t\bar{t}$ + LF background from simulation. The normalisation of the $t\bar{t}$ + HF background is not corrected by the variation of the uncertainty components, and hence its overall normalisation factor (k_{HF}) is allowed to float independently in the fit under a tested hypothesis.

The PDF uncertainties are derived by replacing the nominal NNPDF3.0nnlo set by the symmetrised Hessian set PDF4LHC15_nlo_30 following the PDF4LHC recommendations for LHC Run 2 [209]. The uncertainty associated with the PS is evaluated by comparing the predictions of the nominal event generator, Powheg Box + Pythia 8, to the predictions of an alternative event generator, Powheg + Herwig 7. The uncertainty associated with the matching of the matrix-element to the parton shower (ME-PS matching) is obtained by comparing the predictions of the Powheg Box + Pythia 8 event generator with that of MadGraph 5 + Pythia 8. The uncertainties associated with the initial- and final-state radiation (ISR and FSR) are derived by varying the strong coupling constant, α_s, independently at the matrix-element and the parton shower generation stage. In the matrix-element, the parameter α_s is increased (decreased) to 0.140 (0.115) instead of the nominal value of 0.127, while in the parton shower α_S is increased (decreased) to 0.142 (0.115) instead of the nominal value [210]. The QCD scale uncertainties are derived by varying μ_F and μ_R by a factor of 0.5 (2.0) for the $+1\sigma$ (-1σ) variation. All the aforementioned uncertainties are applicable for both $t\bar{t}$ + HF and $t\bar{t}$ + LF backgrounds. An additional uncertainty (4FS vs. 5FS) in the associated production of the top-quark pairs and b-quarks is further considered. This uncertainty is estimated by comparing the predictions of the `Powheg + Pythia 8` event generator in which the extra b-quarks arise from the PS - *five-flavour scheme* (5FS) - to the predictions of the same generator which uses the *four-flavour scheme* (4FS). In the latter, additional b-quarks are produced directly in the matrix-element calculation. Since this uncertainty is only related to the b-quark associated with the $t\bar{t}$ system, it is applicable only for $t\bar{t}$ + HF background.

All the aforementioned modelling uncertainties contribute as two individual uncertainty components, each for $t\bar{t}$ + LF and $t\bar{t}$ + HF backgrounds, correlated across all analysis regions. The pre-fit impact of these uncertainties on the m_{Wh} distribution for few representative analysis regions is shown in Figures

[4] Throughout this thesis, the MadGraph 5 event generator is used with NLO precision in QCD unless explicitly stated.

[5] In general, the perturbative precision and the PDF sets used in these alternative generator configurations match those of the nominal generators unless explicitly stated.

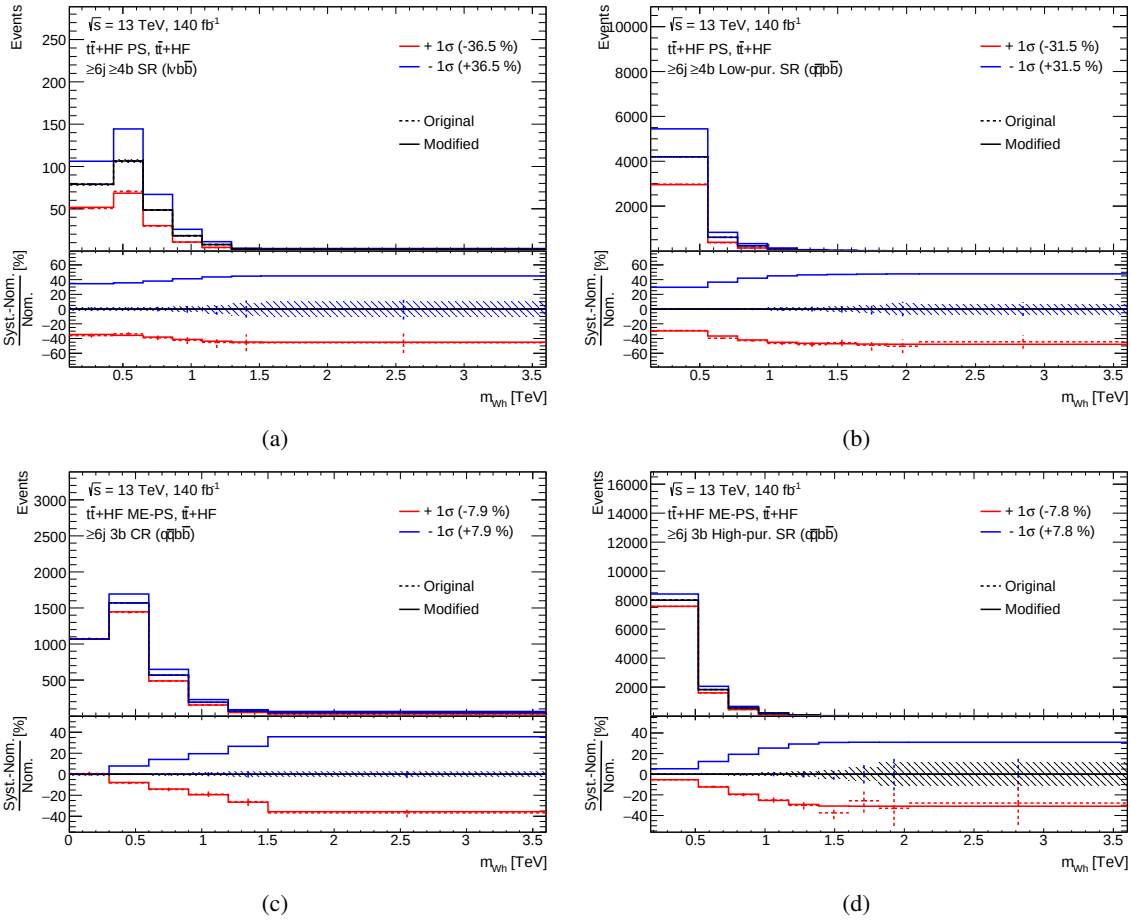


Figure 6.6: Figures (a) and (b) show the pre-fit size of the $t\bar{t}$ + HF PS uncertainty in the $\geq 6j \geq 4b$ SR in the $\ell\nu b\bar{b}$ channel and the $\geq 6j \geq 4b$ low-purity SR in the $q\bar{q}b\bar{b}$ channel, respectively. Figures (c) and (d) show the pre-fit size of the $t\bar{t}$ + HF ME-PS uncertainty in the $\geq 6j$ $3b$ CR and the $\geq 6j$ $3b$ high-purity SR in the $q\bar{q}b\bar{b}$ channel, respectively.

6.6 and 6.7. The size of the uncertainty ranges from 7% to 37%. A large uncertainty for several analysis regions leads to over-constrained[6] NPs from a signal-plus-background fit using Asimov data. A step-wise decorrelation of uncertainties in analysis regions is hence performed in order to avoid an over-estimation of the uncertainties. The scheme is discussed in Appendix B.2. The uncertainties which are decorrelated in various analysis regions are the PS uncertainties for both $t\bar{t}$ + LF and $t\bar{t}$ + HF backgrounds, the ME-PS matching uncertainties for both $t\bar{t}$ + LF and $t\bar{t}$ + HF backgrounds, the 4FS vs. 5FS uncertainty for the $t\bar{t}$ + HF background, and the FSR uncertainty for the $t\bar{t}$ + LF background.

[6] Here the term "constraint" of an NP should not be confused with the Gaussian or Poisson constraints of NPs in the likelihood model as mentioned in Section 3.6.1. The constraint here refers to an NP leading to a shrinkage of the parabola shape in the log-likelihood ratio. Further explanations are given in Section 6.3 which describes the pull distributions of the NPs.

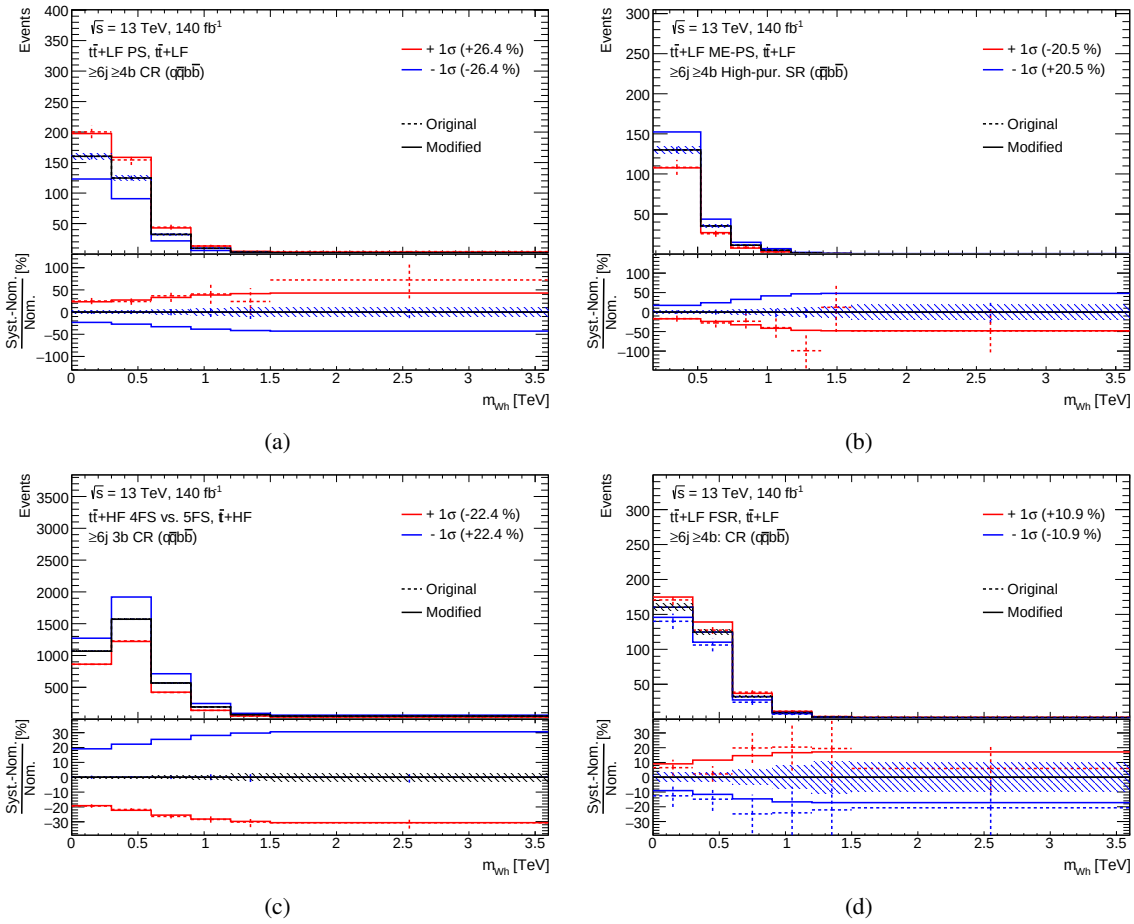


Figure 6.7: Figures (a) and (b) show the pre-fit size of the $t\bar{t}$ + LF PS and ME-PS uncertainty in the $6j \geq 4b$ CR and the $\geq 6j \geq 4b$ high-purity SR in the $q\bar{q}b\bar{b}$ channel, respectively. Figure (c) shows the pre-fit size of the $t\bar{t}$ + HF 4FS vs. 5FS uncertainty in the $\geq 6j$ $3b$ CR in the $q\bar{q}b\bar{b}$ channel. Figure (d) show the pre-fit size of the $t\bar{t}$ + LF FSR uncertainty in the $\geq 6j \geq 4b$ CR in the $q\bar{q}b\bar{b}$ channel.

Other modelling uncertainties exclusively affecting the $t\bar{t}$ + HF background are the relative acceptance uncertainties. Relative acceptance uncertainties account for the relative normalisation differences between regions and simulated processes with a common floating normalisation factor. These uncertainties include the cross-region extrapolation uncertainties and an uncertainty controlling the heavy-flavour composition of the $t\bar{t}$ + HF background. The cross-region extrapolation uncertainties help to describe the extrapolation of the floating normalisation factor from a region A into a region B. The cross-region extrapolation factor between two regions A and B is defined as:

$$R_{\mathrm{A}\to\mathrm{B}} = \frac{n_{\mathrm{A}}}{n_{\mathrm{B}}}, \tag{6.2}$$

where n_{A} and n_{B} are the yields of the $t\bar{t}$ + HF background in the regions A and B, respectively. The associated cross-region extrapolation uncertainty between the two regions is then calculated as:

$$\sigma_{R_{\mathrm{A}\to\mathrm{B}}} = \sqrt{\sum_{i}^{M} \left(\frac{R_{\mathrm{A}\to\mathrm{B}}^{\mathrm{var},i} - R_{\mathrm{A}\to\mathrm{B}}^{\mathrm{nominal}}}{R_{\mathrm{A}\to\mathrm{B}}^{\mathrm{nominal}}} \right)^2}, \tag{6.3}$$

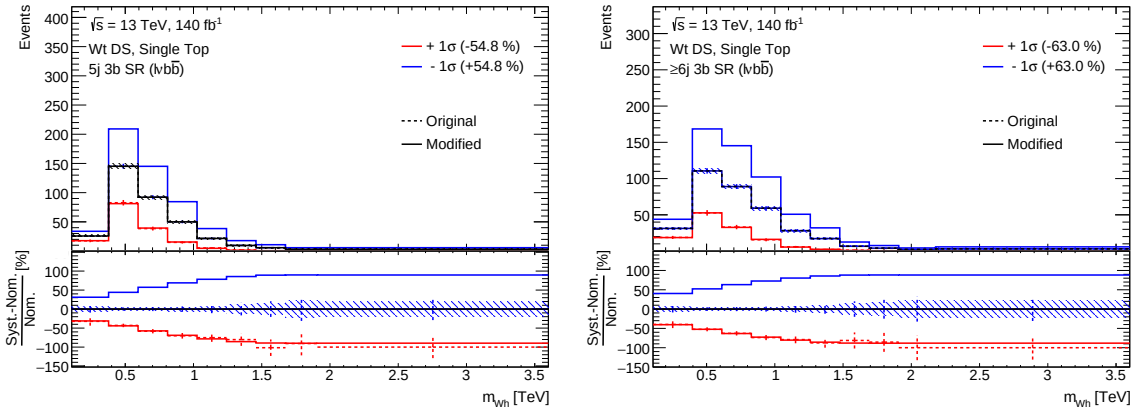
where $R_{\mathrm{A}\to\mathrm{B}}^{\mathrm{nominal}}$ is the cross-region extrapolation factor corresponding to the $t\bar{t}$ + HF background process generated with the nominal event generator, Powheg Box + Pythia 8, and $R_{\mathrm{A}\to\mathrm{B}}^{\mathrm{var,i}}$ are the cross-region extrapolation factors corresponding to the i^{th} alternative generator used to simulate the $t\bar{t}$ + HF background.

The cross-region extrapolation uncertainties are calculated between regions in $q\bar{q}b\bar{b}$ and $\ell\nu b\bar{b}$ channels ($\ell\nu b\bar{b} \leftrightarrow q\bar{q}b\bar{b}$[7]), between regions with five and at least six jets ($5j \leftrightarrow \geq 6j$), between signal and control regions ((high-purity) SRs $\leftrightarrow$ CRs), between high-purity and low-purity signal regions (SRs $\leftrightarrow$ low-purity SRs) in the $q\bar{q}b\bar{b}$ channel, and between low-purity signal regions and control regions (low-purity SRs $\leftrightarrow$ CRs) in the $q\bar{q}b\bar{b}$ channel. The second kind of relative acceptance uncertainty is the one that handles the ratio of the heavy-flavour composition ($t\bar{t}$ + $\geq 1b$ and $t\bar{t}$ + $\geq 1c$ components) of the $t\bar{t}$ + HF background. It is calculated in the same way as the cross-region extrapolation uncertainties. The extrapolation effect in this case is considered between the $t\bar{t}$ + $\geq 1b$ and $t\bar{t}$ + $\geq 1c$ background components instead of two regions.

Single Top uncertainties An uncertainty of 5% is considered on the cross-sections of the s-channel, t-channel, and Wt production modes [211–215]. The tZ background is assigned a cross-section uncertainty of 7.9% for assessing the variations in μ_F and μ_R, and an uncertainty of 0.9% for assessing the PDF variations. A 50% cross-section uncertainty is applied to the tWZ background [216]. Uncertainties associated with the PS and ME-PS matching are evaluated by comparing the predictions of the nominal event generator, Powheg Box + Pythia 8, with that of alternative event generators, Powheg Box + Herwig 7 and MadGraph 5 + Herwig 7, respectively for each single-top background process. An additional modelling uncertainty associated with the Wt background modelling is considered which reflects the interference between Wt and $t\bar{t}$ + jets production at NLO [217]. This uncertainty is evaluated by comparing the predictions of the `Powheg + Pythia 8` event generator produced using the *diagram removal (DR)* scheme with the predictions of the same generator produced using the *diagram subtraction (DS)* [183] scheme. The pre-fit impact of this uncertainty (denoted as Wt DS) on the m_{Wh} distribution in two representative analysis regions is shown in Figure 6.8 which illustrates that the size of the uncertainty ranges up to 63%. A large uncertainty for several analysis regions over-constraints the associated NP resulting from a signal-plus-background fit using Asimov data. Hence, this uncertainty is decorrelated in analysis regions following the same technique as used for the $t\bar{t}$ + jets background modelling uncertainties (see Appendix B.2).

***V* + jets and *VV* uncertainties** A 40% normalisation uncertainty is considered for the W + jets background, and a 35% normalisation uncertainty is considered for the Z + jets background. These normalisation uncertainties account for the uncertainties from assessing the variations in μ_F, μ_R, PDF, and any potential uncertainty in the extraction of the correction factor from data for the heavy-flavor component of the V + jets processes [206, 218]. A 50% normalisation uncertainty is assigned to

[7] A double-sided arrow is used to denote this uncertainty, as the extrapolation factors are calculated from a region with higher $t\bar{t}$+HF background yield to a region with lower $t\bar{t}$+HF background yield. The associated uncertainties and then applied to the latter region.



(a) (b)

Figure 6.8: Pre-fit size of the Wt DS uncertainty in the (a) $5j\ 3b$ SR and (b) $\geq 6j\ 3b$ SR in the $\ell\nu b\bar{b}$ channel.

the VV background, which includes the uncertainties on the cross-section and any additional jet production [219]. No additional modelling uncertainty is assumed on these backgrounds as their relative yield in comparison to the $t\bar{t}$ + jets background is significantly smaller.

***$t\bar{t}+X$* uncertainties** The $t\bar{t}+h$ background is assigned a cross-section uncertainty of 12% [220–225]. Predictions of the nominal event generator, POWHEG BOX + PYTHIA 8, are compared with that of an alternative event generator, POWHEG BOX + HERWIG 7, to evaluate the uncertainty due to PS. Predictions of the POWHEG BOX + PYTHIA 8 event generator are compared with that of MADGRAPH 5 + PYTHIA 8 to evaluate the uncertainty due to ME-PS matching. The $t\bar{t}+W$ background is assigned a cross-section uncertainty of 12.9% for assessing the variations in μ_F and μ_R, and an uncertainty of 3.4% for assessing the PDF variations. The $t\bar{t}+Z$ background is assigned a cross-section uncertainty of 11.3% for assessing the variations in μ_F and μ_R, and an uncertainty of 4% for assessing the PDF variations [226, 227]. An additional modelling uncertainty on the $t\bar{t}+W/Z$ background related to the ME-PS matching is also considered which is evaluated by comparing the predictions of the nominal event generator, MADGRAPH 5 + PYTHIA 8, with that of an alternative event generator, SHERPA 2.2.1.

Uncertainties on other rare top-quark backgrounds A 50% normalisation uncertainty is considered for the $t\bar{t}t\bar{t}$ background covering effects from varying μ_{F}, μ_{R}, PDF, and α_{s} [216, 225]. The $thjb$ background is assigned a cross-section uncertainty of 14.9% for assessing the variations in μ_F and μ_R, and an uncertainty of 3.7% for assessing the PDF variations [216].

Non-closure uncertainty

The transverse momentum of the W boson (from the H^+ boson decay) in the $\ell\nu b\bar{b}$ channel is found to be mismodelled in events containing exactly two b-tagged jets as shown in Figure 6.9. This residual[8] mismodelling is prominent in the tails of the p_{T} distribution and is taken to into account by including a

[8] The mismodelling is termed as "residual" as it is observed after applying the $t\bar{t}$ + jets reweighting procedure (see Section 4.4).

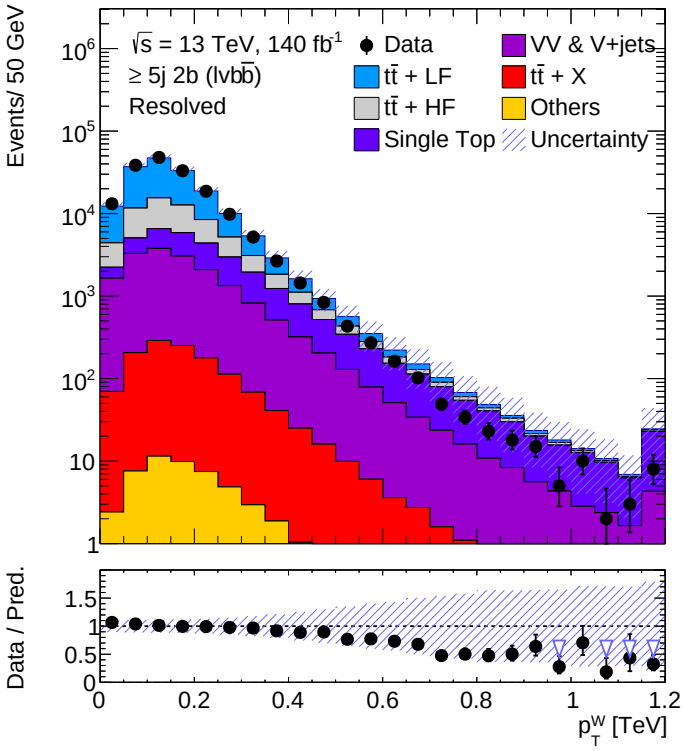
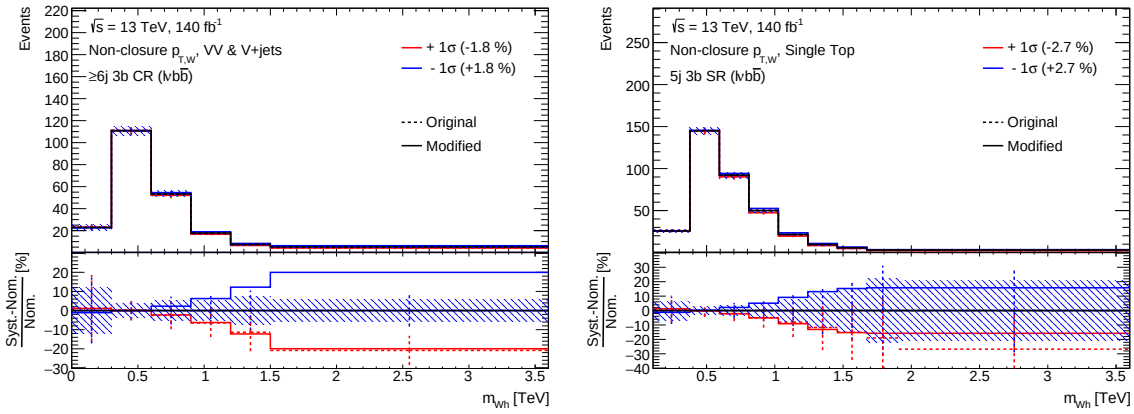


Figure 6.9: Data-to-simulation comparison for the p_{T} of the W boson in the $\ell\nu b\bar{b}$ channel in events containing at least five jets and exactly two b-tagged jets.

non-closure uncertainty in the statistical model. This non-closure uncertainty is evaluated by deriving a correction function on the data-to-simulation ratio of this observable in the presented region. This correction is then propagated as an uncertainty $+1\sigma$ variation to the m_{Wh} distribution in all analysis regions. This uncertainty is applied to all simulated processes in the $\ell\nu b\bar{b}$ channel. The pre-fit impact of this uncertainty on the m_{Wh} distribution for the single-top and VV & V+jets backgrounds in two representative analysis regions is shown in Figure 6.10.

A summary of the contribution of the systematic modelling uncertainties (at the pre-fit stage) is given in Table 6.4.



(a) (b)

Figure 6.10: Pre-fit size of the non-closure uncertainty in the (a) $6j\ 3b$ CR and (b) $\geq 5j\ 3b$ SR for VV & V+jets and Single Top backgrounds, respectively in the $\ell\nu b\bar{b}$ channel.

Table 6.4: The contribution of the systematic modelling uncertainties (at the pre-fit stage). '$A \leftrightarrow B$' indicates cross-region extrapolation uncertainties between regions A and B, and 'Norm.' is the product of the cross section and acceptance variations. A value of 'float' indicates that the parameter is not constrained in the fit. A range of values means that the size of the uncertainties varies between the regions included in a fit.

	$t\bar{t}$ + HF	$t\bar{t}$ + LF	$t\bar{t}$ + X	Single Top	VV & V + jets	Others	Signal
Norm.	float	fixed	0.8%–12%	1.1%–4.7%	2.0%–34%	2.9%–47%	float
PDF	S	0.2%–0.5%, S	–	–	–	–	0.5%–6.0%, S
ISR	S	0.7%–5.0%, S	–	–	–	–	–
FSR	S	1.3%–19%, S	–	–	–	–	–
ME–PS Matching	S	2.0%–40%, S	0.6%–7.0%	0.6%–53%, S	–	–	–
Parton Shower	S	3.3%–22%, S	1.7%–13%	2.2%–61%, S	–	–	1.0%–50%, S
4FS vs. 5FS	S	–	–	–	–	–	–
Heavy-flavour composition	S	–	–	–	–	–	–
DS vs. DR scheme (Wt)	–	–	–	3.3%–68%, S	–	–	–
Renormalisation/Factorisation Scales	S	0.5%–3.0%, S	–	–	–	–	1.0%–13%, S
H_{T} reweighting	S	0.5%–6.0%, S	–	–	–	–	–
$p_{\mathrm{T},W}$ non-closure	S	0.5%–0.9%, S	0.7%–1.6%, S	2.7%–4.0%, S	1.8%–3.3%, S	1.1%–2.0%, S	3.0%–30%, S
$5j \leftrightarrow \geq 6j$	17%–34%	–	–	–	–	–	–
(High-purity) SRs $\leftrightarrow$ CRs	11%–33%	–	–	–	–	–	–
High-purity SRs $\leftrightarrow$ Low-purity SRs ($q\bar{q}b\bar{b}$)	6.8%–17%	–	–	–	–	–	–
Low-purity SRs $\leftrightarrow$ CRs ($q\bar{q}b\bar{b}$)	5.2%–12%	–	–	–	–	–	–
$\ell\nu b\bar{b} \leftrightarrow q\bar{q}b\bar{b}$	5.1%–17%	–	–	–	–	–	–

6.2.3 Treatment of uncertainties in the statistical model

Three kinds of procedures are applied to the systematic uncertainties before they enter the likelihood fit, namely symmetrisation, smoothing, and pruning which are defined in this section.

Symmetrisation An uncertainty symmetrisation procedure is performed for all systematic uncertainties[9] to avoid asymmetric nuisance parameters in a fit which can impact the fit stability. The uncertainties are either one-sided where only a single uncertainty variation exists by construction, for instance an uncertainty $+1\sigma$ variation (up variation) or -1σ variation (down variation). On the other hand, the uncertainties can be two-sided where both up and down variations exist by construction. For one-sided uncertainties, the other variation is taken as a mirrored version of the given variation. For two-sided uncertainties, a secondary variation is calculated bin by bin defined as half of the difference between the up and down variations: $\frac{|\sigma_{\text{up}} - \sigma_{\text{down}}|}{2}$. This secondary variation is then taken as the primary up variation which is mirrored as the down variation.

Smoothing and pruning A smoothing procedure is used to mitigate statistical fluctuations that can arise in the systematic variations due to few signal and background events. The smoothing algorithm used in this thesis is based on an iterative rebinning of the systematic variations as described in [228]. The smoothing is applied on all systematic uncertainties that impact the shape of the m_{Wh} distribution, except the ones where an explicit event weight is used to form a systematic variation.

Several systematic uncertainties entering the likelihood show only a negligible impact on the m_{Wh} distribution in various analysis regions. The statistical model is therefore simplified by removing uncertainties according to the following pruning criteria:

- Uncertainties whose systematic variation is smaller than the statistical uncertainties in all bins of the m_{Wh} distribution, and for all backgrounds and signal processes are neglected,
- normalisation uncertainties are neglected if they lead to a systematic variation lower than 0.5%,
- shape uncertainties are neglected if only one bin has a systematic variation higher than 0.5% after the overall normalisation component has been removed.

[9] The symmetrisation procedure is applied only for the uncertainties described by a Gaussian probability distribution, hence the statistical uncertainties are neglected as they are described by a Poisson probability distribution.

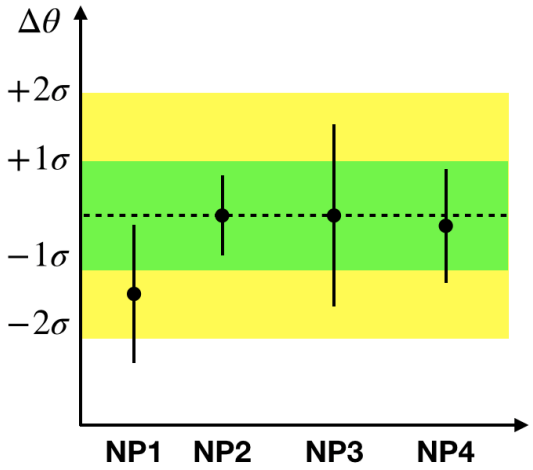


Figure 6.11: A pictorial representation of the pulls for the various NPs.

6.3 Validation of the statistical model

The statistical model described in Section 6.1 is a complex model with 20 analysis regions, and a total of 140 bins in the m_{Wh} distribution. Hence, it is important to study the performance of this model with the help of fit diagnostic tests which are discussed in the following section.

6.3.1 Pull distributions of nuisance parameters

Studying the NP pull distributions resulting from a fit under a tested hypothesis is a necessity while validating a statistical model. This distribution indicate the best-fit values of the NPs and the associated $1\sigma/2\sigma$ uncertainty variations around the miniumum of the negative log likelihood (NLL). An NP pull is mathematically expressed as:

$$\mathcal{P} = \frac{\hat{\theta}_i - \theta_{i,0}}{\Delta\theta_i} \tag{6.4}$$

where $\hat{\theta}_i$ is the best-fit value of an NP (θ_i) resulting from a likelihood fit under a tested hypothesis, $\theta_{i,0}$ is the value of θ_i at the pre-fit level, and $\Delta\theta_i$ is the pre-fit uncertainty of θ_i. A pictorial representation of the pulls of the various NPs is shown in Figure 6.11 which depicts four different kinds of scenarios. NP1 is termed as a pulled NP as its best-fit value deviates outside the 1σ uncertainty band of the pull distribution. This can occur from disagreement between data and simulation or any model specific feature related to the derivation of this NP which can cause such a pull. NP2 is termed as an over-constrained NP[10] where the pull uncertainty is seen to be considerably smaller than the 1σ uncertainty band. One common cause of such a scenario is the large size of the pre-fit uncertainty for this particular NP or a fit convergence issue. NP3, on the other hand, is termed as an under-constrained NP which can occur from the double counting of NPs in a fit or a fit covergence issue. NP4 rather

[10] For this thesis, an NP is considered over-constrained if its pull uncertainty is less than 0.5.

exhibits a balanced behaviour and is well constrained where the NP pull matches the 1σ uncertainty band.

Figures 6.12 and 6.13 show the comparison of the NP pulls obtained from fits to collision data (black) and Asimov data (see Section 3.6.1) (red), under a background-only hypothesis, in all analysis regions. The names of the NPs in these figures denote the various experimental and theoretical uncertainties as described in Section 6.2, and follow a naming convention for the following uncertainties:

- As discussed in Section 6.2.2, the Wt DS and $t\bar{t}$ + jets background modelling uncertainties are decorrelated across analysis regions due to their over-constrained behaviour. The naming of the NPs associated with these uncertainties follow the given order: uncertainty name, type of region (CR/ SR/ jet, b-jet multiplicity region), type of analysis channel ($q\bar{q}b\bar{b}$ / $\ell\nu b\bar{b}$), and in certain cases a symbol L/H is used, where L denotes the low-mass region of the m_{Wh} distribution, and H denotes the high-mass region of the m_{Wh} distribution. If an uncertainty is not decorrelated into any region of a channel, then the second category in the naming convention for the associated NP is skipped,
- NP names for the uncertainties associated with the reweighting of the $t\bar{t}$ + jets background are given as the jet multiplicity followed by the number of error variation ("EV") component,
- NP names for the different components of the jet energy scale and jet energy resolution uncertainties are given as the pre-fix "JES" and "JER" followed by the component name,
- NP names for the flavour tagging uncertainties for the b-, c-, and light-flavoured jets follow the given order: pre-fix "FTAG, b-/c-/l-jets" followed by the component number.

Overall, the pulls and constrains of all the NPs agree well between the fits using collision data, and Asimov data indicating a stable behaviour of the statistical model. The pulls resulting from the collision data fit are seen to be within the 1σ uncertainty band. The over-constrained behaviour of the following NPs associated with $t\bar{t}$ + HF PS uncertainty, arise from the first two to three bins of the m_{Wh} distribution where the uncertainty size is up to 40% (see Figure 6.6):

- $t\bar{t}$ + HF PS $\geq 6j \geq 4b$ CR $q\bar{q}b\bar{b}$ L,
- $t\bar{t}$ + HF PS $\geq 6j \geq 4b$ Low-purity SR $q\bar{q}b\bar{b}$ L,
- $t\bar{t}$ + HF PS $\geq 6j \geq 4b$ SR $q\bar{q}b\bar{b}$ L,
- $t\bar{t}$ + HF PS $\geq 6j \geq 4b$ SR $\ell\nu b\bar{b}$ L.

Another NP which is over-constrained is the HF composition ($t\bar{t}$ + $\geq 1b \leftrightarrow t\bar{t}$ + $\geq 1c$). This over-constraint arises from its large uncertainty of up to 30% in some analysis regions (see Table B.6).

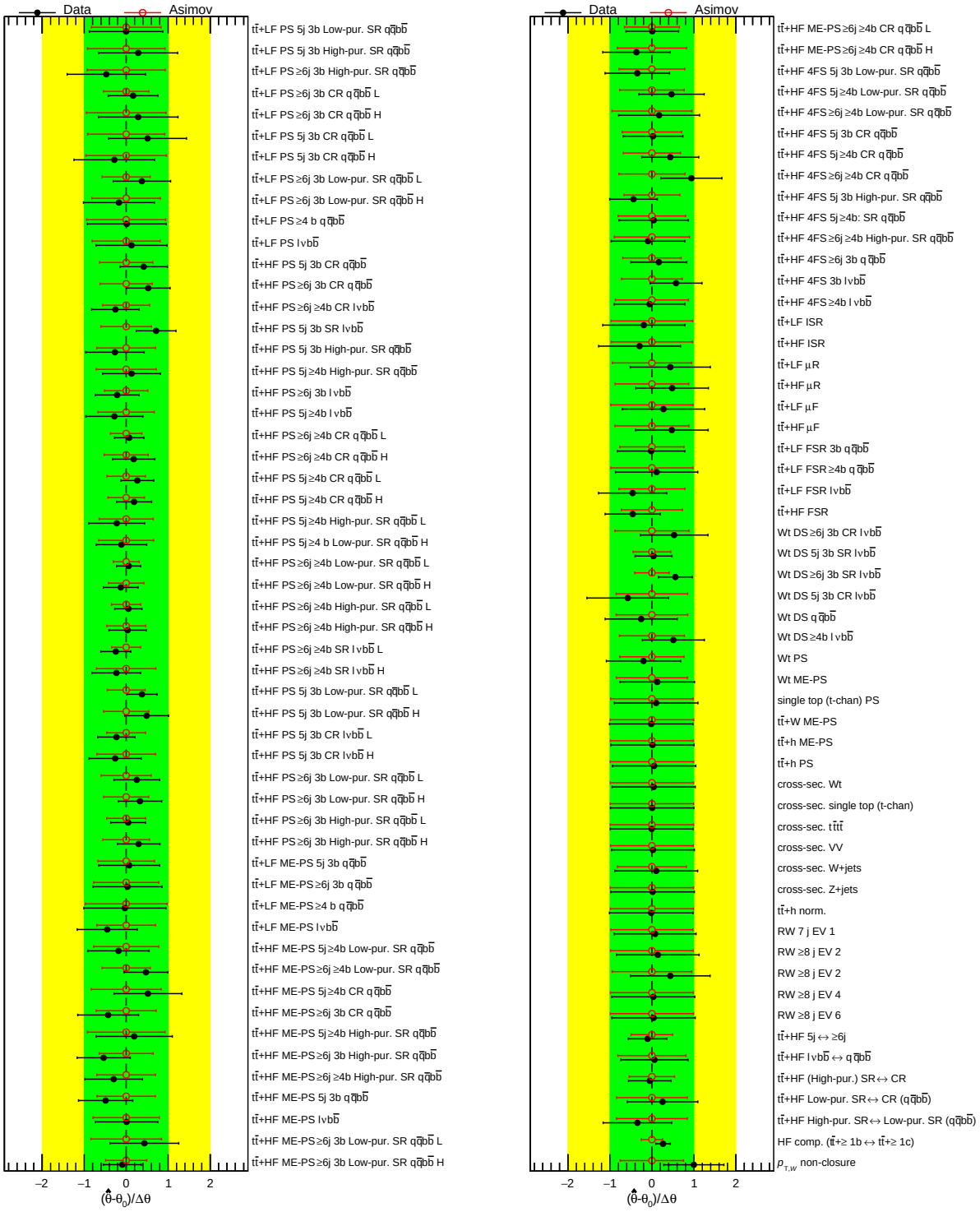


Figure 6.12: Comparison of the pull distributions from fits to collision data (black) and Asimov data (red), under a background-only hypothesis, in all analysis regions.

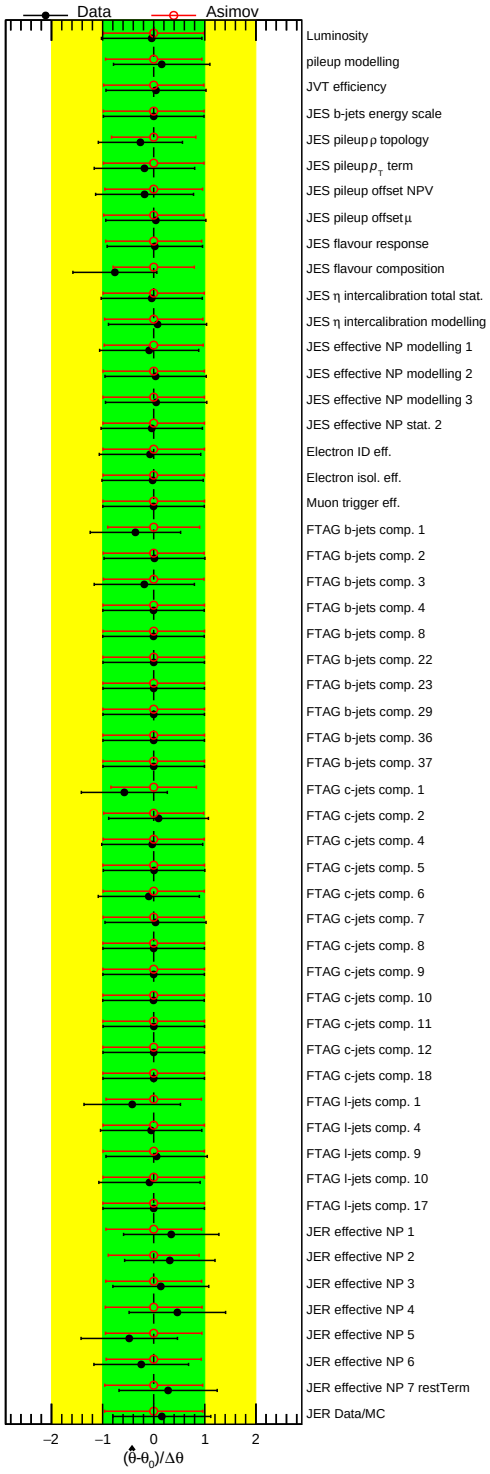


Figure 6.13: Comparison of the pull distributions from fits to collision data (black) and Asimov data (red), under a background-only hypothesis, in all analysis regions.

6.3.2 Nuisance parameters ranking

Another widely used method to understand the behaviour of the likelihood fit is to determine the degree of influence of the systematic uncertainties on the signal strength, μ. The impact of an individual NP on μ is determined by scanning the NP within $\pm 1\sigma$ around its nominal value during the likelihood fit while all the other NPs are allowed to vary. The best-fit μ value, $\hat{\mu}$, is re-calculated for each value of the NP under study. The NP which returns the maximum $\Delta\mu$ (difference of $\hat{\mu}$ from the nominal μ value) is placed the highest in the impact plot or the so-called *ranking plot*. Figures 6.14 to 6.17 show the NP ranking resulting from fits to Asimov data, under a signal-plus-background hypothesis, in all analysis regions for various H^+ boson mass hypotheses: 250 GeV, 600 GeV, 1.2 TeV, and 2 TeV. A high post-fit impact of the $t\bar{t}$ + jets, and single-top background modelling systematic uncertainties on μ is observed for all the presented mass hypotheses, which is expected as they are the dominant backgrounds in the analysis. A high impact of statistical uncertainties[11] for the 1.2 TeV and 2 TeV signal mass hypotheses is also expected due to fewer events in the high mass tails of the m_{Wh} distribution where the bulk of the signal distribution is expected to lie. For all the presented ranking plots, the NPs show a symmetric $\pm 1\sigma$ post-fit impact on $\Delta\mu$. However, NPs associated with the single-top (Wt) background modelling uncertainties: Wt DS $5j$ $3b$ SR $\ell\nu b\bar{b}$, Wt DS $\geq 6j$ $3b$ SR $\ell\nu b\bar{b}$, and Wt DS $\geq 4b$ $\ell\nu b\bar{b}$ show asymmetric and one-sided post-fit impacts on μ for some of the H^+ boson mass hypotheses. This one-sidedness and asymmetry seems to arise from the non-linear correlations of these NPs with other single top background NPs [176], and do not indicate any instable behaviour of these NPs in the statistical model.

[11] Statistical uncertainties, shown in the figures, are the combined uncertainties for all simulated processes, and are represented by γ for each bin of an analysis region.

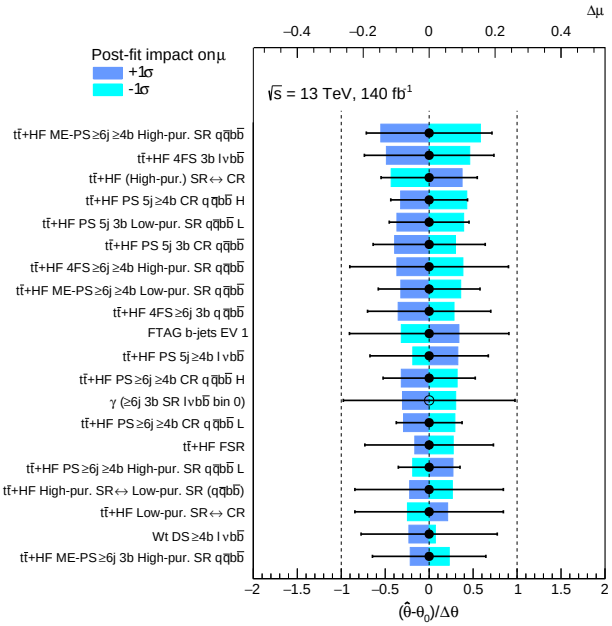


Figure 6.14: NP ranking from a fit under a signal-plus-background hypothesis using Asimov data for m_{H^+} = 250 GeV. The nomenclature of some specific NPs is discussed in Section 6.3.1, and the impact of uncertainties on the m_{Wh} distribution was listed in Tables 6.2 and 6.3.

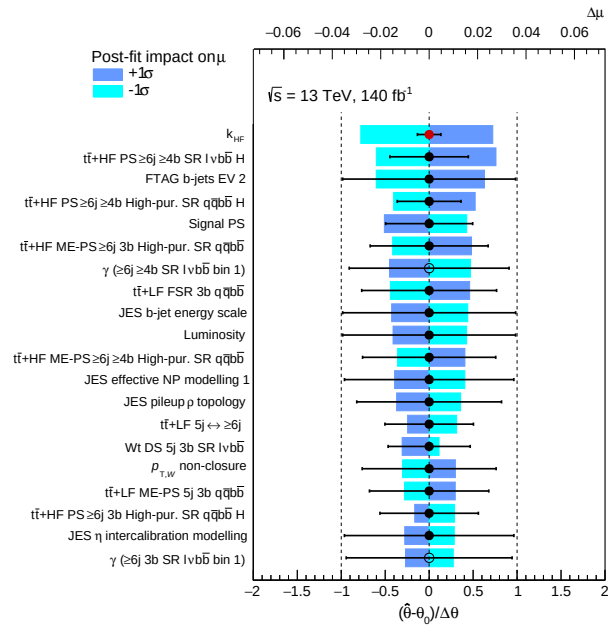


Figure 6.15: NP ranking from a fit under a signal-plus-background hypothesis using Asimov data for m_{H^+} = 600 GeV. The nomenclature of some specific NPs is discussed in Section 6.3.1, and the impact of uncertainties on the m_{Wh} distribution was listed in Tables 6.2 and 6.3.

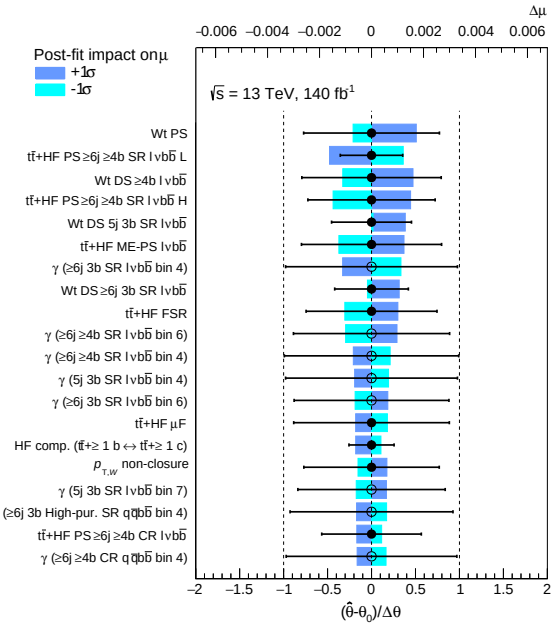


Figure 6.16: NP ranking from a fit under a signal-plus-background hypothesis using Asimov data for m_{H^+} = 1.2 TeV. The nomenclature of some specific NPs is discussed in Section 6.3.1, and the impact of uncertainties on the m_{Wh} distribution was listed in Tables 6.2 and 6.3.

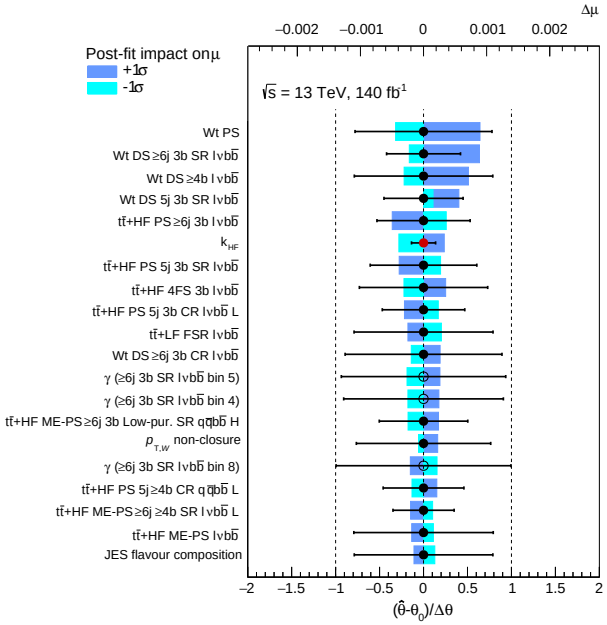
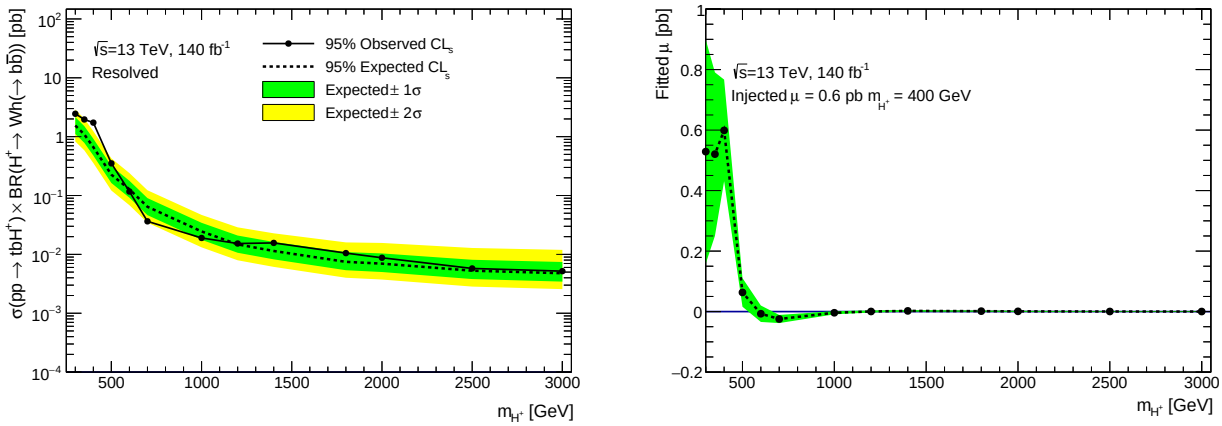


Figure 6.17: NP ranking from a fit under a signal-plus-background hypothesis using Asimov data for m_{H^+} = 2 TeV. The nomenclature of some specific NPs is discussed in Section 6.3.1, and the impact of uncertainties on the m_{Wh} distribution was listed in Tables 6.2 and 6.3.



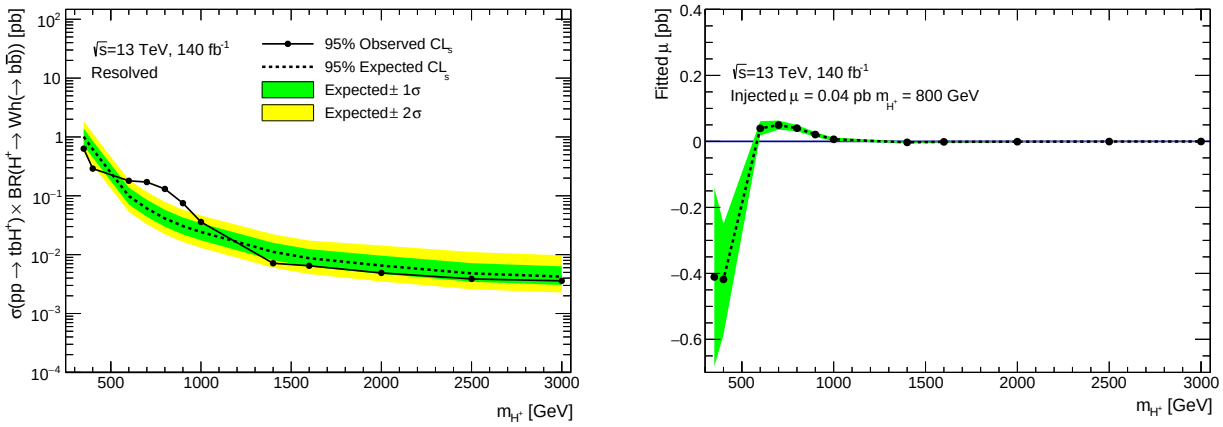
(a) (b)

Figure 6.18: 95% CL upper limits on $\sigma(tbH^+) \times BR(H^+ \rightarrow Wh(\rightarrow b\bar{b}))$, and the corresponding best-fit μ values as a function of m_{H^+} for an injected $\sigma = 0.6$ pb for $m_{H^+} = 400$ GeV.

6.3.3 Signal injection test

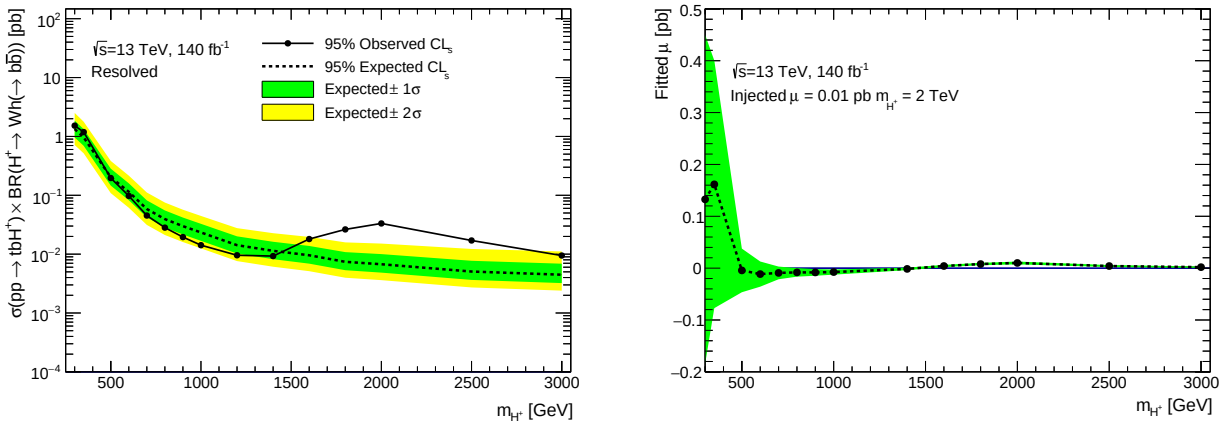
A final study in the series of statistical model validations is a signal injection test. A signal injection test is performed using a fit to a pseudo-dataset, under a signal-plus-background hypothesis, in all analysis regions. This pseudo-dataset is constructed from the simulated events of all background processes plus the 'injected' simulated signal events of a tested H^+ boson mass hypothesis normalised to a specific cross-section[12]. The aim of this test is to check if the fitted μ value is equal to the injected signal strength for the tested mass hypothesis, and at the same time cross-check the model behaviour in terms of observing a *"signal bump"* created artificially by the pseudo-dataset. Figures 6.18 to 6.20 present the 95% CL upper limits on μ ($\sigma(pp \rightarrow tbH^+) \times BR(H^+ \rightarrow Wh(\rightarrow b\bar{b}))$), and its corresponding best-fit values as a function of m_{H^+} for three different fit scenarios. Figure 6.18 shows the fit results for an injected signal strength of 0.6 pb for $m_{H^+} = 400$ GeV, Figure 6.19 shows the fit results for an injected signal strength of 0.04 pb for $m_{H^+} = 800$ GeV, and Figure 6.20 shows the fit results for an injected signal strength of 0.01 pb for $m_{H^+} = 2$ TeV. In all three presented cases, a broad excess depicting a maximum expected-observed discrepancy around the tested H^+ boson mass hypothesis is observed. This broad excess can potentially arise due to the mass resolution of the reconstructed H^+ boson being larger than the mass difference used to simulate the various signal hypotheses. One can also observe that the best-fit μ values for $m_{H^+} = 400$ GeV, $m_{H^+} = 800$ GeV, and $m_{H^+} = 2$ TeV are equal to the injected signal strengths which validates the statistical model.

[12] The injected signal strength should be equal to the injected signal cross-section value, given that $BR(H^+ \rightarrow Wh(\rightarrow b\bar{b}))$ = 1.0 for the simulated signal.



(a) (b)

Figure 6.19: 95% CL upper limits on $\sigma(tbH^+) \times BR(H^+ \to Wh(\to b\bar{b}))$, and the corresponding best-fit μ values as a function of m_{H^+} for an injected $\sigma = 0.04$ pb for $m_{H^+} = 800$ GeV.



(a) (b)

Figure 6.20: 95% CL upper limits on $\sigma(tbH^+) \times BR(H^+ \to Wh(\to b\bar{b}))$, and the corresponding best-fit μ values as a function of m_{H^+} for an injected $\sigma = 0.01$ pb for $m_{H^+} = 2$ TeV.

CHAPTER 7

Merged Topology

This chapter focuses on the search for merged $H^+ \to Wh(\to b\bar{b})$ decays. Analogous to the search for resolved $H^+ \to Wh(\to b\bar{b})$ decays, this search targets two decay channels: $H^+ \to Wh \to \ell\nu b\bar{b}$ ($\ell\nu b\bar{b}$ channel) and $H^+ \to Wh \to q\bar{q}b\bar{b}$ ($q\bar{q}b\bar{b}$ channel) (see Figure 4.1). In the merged topology, the angular separation between the H^+ boson decay products is small due to the large H^+ boson p_{T} such that the decay products of the hadronically decaying W/h boson can be reconstructed inside a large-radius (large-R) jet. Two kinds of neural networks (NNs) are used - one to identify the boosted $h \to b\bar{b}$ decays and the other to separate signal from backgrounds. Selection requirements on the kinematic properties of the final state particles and on the output score of a classification NN are used to define signal and control regions. The kinematic observable used for setting 95% CL upper limits on $\sigma(pp \to tbH^+) \times BR(H^+ \to Wh(\to b\bar{b}))$ is the reconstructed mass of the charged Higgs boson (m_{Wh}).

This chapter is structured as follows: Section 7.1 describes the boosted $X \to b\bar{b}$ tagging algorithm used for the identification of $h \to b\bar{b}$ decays and the tagger's signal efficiency calibration using $Z \to b\bar{b}$ + jets events at high Z boson transverse momentum. Section 7.2 gives an overview of the analysis strategy used to perform a search for merged $H^+ \to Wh$ decays. This section includes a description of the event selection requirements, an outline of the NN training in the $q\bar{q}b\bar{b}$ and $\ell\nu b\bar{b}$ channels, the definition of signal and control regions, and a brief description of the statistical model.

7.1 Boosted $X \to b\bar{b}$ tagging and its signal efficiency calibration

The boosted $X \to b\bar{b}$ tagger is an NN-based Higgs tagger trained to separate boosted $h \to b\bar{b}$ decays (signal) from boosted top-quark jets and multijet processes (backgrounds). The NN is based on the following input variables:

- large-R jet p_{T} and η,
- flavour information of up to three ghost-associated VR track-jets (see Section 3.4.3, Chapter 3) provided by the DL1r tagger[1].

[1] The flavour description is retrieved from a high-level deep neural network based single-b-tagging algorithm, DL1r [205], designed to identify VR track-jets containing single b-hadrons.

Details of the tagger's training are described in [229]. The tagger provides three output scores for each large-R jet corresponding to the probabilities of it being a Higgs boson - (p_{Higgs}) - initiated jet, multijet (p_{multijet}), and a top-quark - (p_{top}) - initiated jet. These output scores are combined in a single discriminant which is defined as:

$$D_{\text{X}\to\text{b}\bar{\text{b}}} = \ln\left(\frac{p_{\text{Higgs}}}{f_{\text{top}} \cdot p_{\text{top}} + (1 - f_{\text{top}}) \cdot p_{\text{multijet}}}\right) \tag{7.1}$$

where f_{top} determines the fraction of the top-quark background of the total background. A large-R jet is tagged as a Higgs jet candidate based on selection requirements on $D_{\text{X}\to\text{b}\bar{\text{b}}}$.

The $X \to b\bar{b}$ tagger performance is studied for large-R jets which satisfy the following event selection requirements:

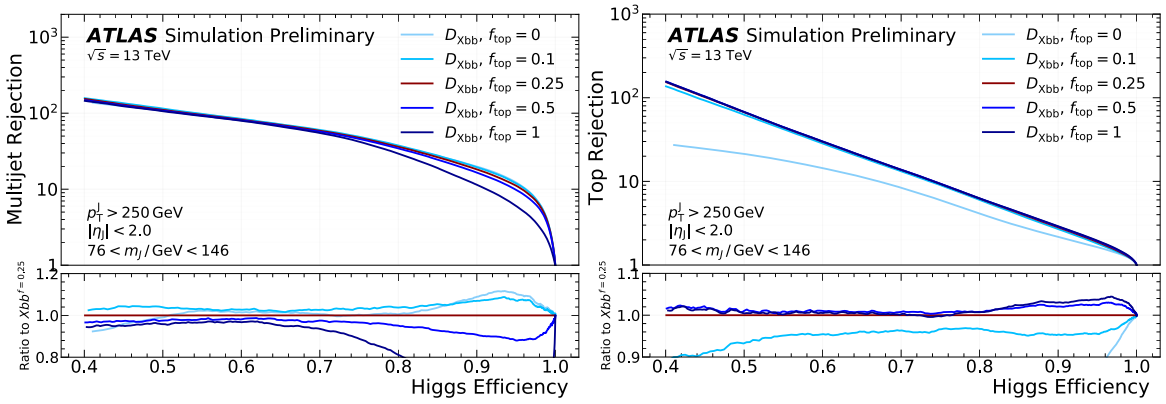
- $p_{\text{T}} > 250$ GeV,
- $|\eta| < 2$,
- masses within 76 GeV-146 GeV.

In order to study the $X \to b\bar{b}$ tagger performance in simulated events, generator-level information is used in addition to the above requirements to identify the correct Higgs and top-quark jet candidates. The Higgs (top-quark) jet candidates are required to have exactly one ghost-associated (see Section 3.4.3) generator-level Higgs boson (top-quark). The selected large-R jets in multijet events originate primarily from $g \to b\bar{b}$ splitting processes. The tagger performance is studied in terms of the signal tagging efficiency and background rejection. The tagging efficiency (ϵ) for Higgs jets is defined as the number of $X \to b\bar{b}$ tagged jets divided by the total number of large-R jets passing the aforementioned event selection requirements. Working points (WPs) of the $X \to b\bar{b}$ tagger are defined as the selection requirements on the $D_{\text{X}\to\text{b}\bar{\text{b}}}$ to achieve a desired Higgs jet tagging efficiency. Three WPs corresponding to 50%, 60%, and 70% Higgs jet tagging efficiencies with $f_{\text{top}} = 0.25$[2] are defined. The tagger threshold values for the three WPs are summarised in Table 7.1. This thesis uses the $X \to b\bar{b}$ tagger at 60% WP with $f_{\text{top}} = 0.25$. On the other hand, the background rejection factors are defined as the reciprocal value of the background mistagging efficiency ($1/\epsilon$) for boosted top-quark and multijet events. Figure 7.1 shows the background rejection as a function of the Higgs jet tagging efficiency for various f_{top} values.

Since the tagger is trained using simulated events, the tagging efficiency in simulation may differ from the tagging efficiency observed in data. To make use of this tagger in physics analyses, one needs to derive data-to-simulation efficiency correction factors (scale factors) to correct the tagger efficiency in simulation. The calibration of the $X \to b\bar{b}$ tagger follow a *process-based approach*[3] with the processes of interest being the Higgs signal, top-quark, and multijet processes. The signal efficiency calibration of the $X \to b\bar{b}$ tagger is performed using $Z \to b\bar{b} + \gamma$ and $Z \to b\bar{b}$ + jets events at high Z boson transverse momentum. The $Z \to b\bar{b}$ is a colour-singlet resonance with its mass close to the Higgs boson mass, such that the kinematic properties of $Z \to b\bar{b}$ and $h \to b\bar{b}$ events are expected to be similar [231]. Moreover, since the tagger is trained mass-independently [229], the calibration results can be readily applied to the $h \to b\bar{b}$ topology. The background mistagging

[2] $f_{\text{top}} = 0.25$ is the recommendation provided by the ATLAS Flavour Tagging Group [230] as it provides an optimal rejection for both multijet, and top-quark backgrounds.

[3] This is in contrast to a flavour-based approach usually followed for the calibration of small-R jets.



(a) (b)

Figure 7.1: (a) Multijet and (b) top-quark jet rejection as a function of the $h \to b\bar{b}$ tagging efficiency (denoted as "Higgs Efficiency"), for large-R jets with $p_{\mathrm{T}} > 250$ GeV. The performance of the $X \to b\bar{b}$ tagger is compared for $f_{\mathrm{top}} = 0, 0.1, 0.25, 0.5$ and 1. The efficiency and rejection are calculated with respect to jets that pass the p_{T}, η, and jet mass requirements given in [28].

Working point	50%	60%	70%
Threshold value	3.13	2.55	1.92

Table 7.1: The $X \to b\bar{b}$ tagger threshold values for all $X \to b\bar{b}$ WPs with $f_{\mathrm{top}} = 0.25$.

efficiency calibration for the hadronically decaying top-quarks is performed using a tag-and-probe method with semileptonic $t\bar{t}$ events. The leptonically decaying top-quark is used to select ('tag') the events. The calibration is applied on the hadronically decaying top-quark (the 'probe' object) [28]. However, the mistagging efficiency calibration for multijet processes is not provided by the ATLAS performance group [230] due to the observed post-fit mismodelling of the fit observable in data. The calibration scale factors are derived using dedicated techniques for this analysis as described in Section 7.2.4.

Keeping within the scope of the author's contribution in the presented topic, this section discusses a new method to calibrate the $X \to b\bar{b}$ tagger signal efficiency using $Z \to b\bar{b}$ + jets events. Section 7.1.1 describes the calibration methodology, followed by the details of the calibration analysis strategy in Section 7.1.2. Section 7.1.3 gives a brief overview of the systematic uncertainties and presents the results of the signal efficiency calibration of the $X \to b\bar{b}$ tagger.

7.1.1 Methodology

As mentioned earlier, data-to-simulation efficiency correction factors or scale factors (SFs) are needed to correct the tagger efficiency in simulation. The scale factors are defined as: $\mathrm{SF} = \epsilon^{\mathrm{data}}/\epsilon^{\mathrm{sim}}$ and are applied to all simulated events. ϵ^{data} is defined as the number of data events that pass the $X \to b\bar{b}$ tagger selection requirement ($N^{\mathrm{data}}_{\mathrm{passed}}$) divided by the total number of signal events in data ($N^{\mathrm{data}}_{\mathrm{total}}$). ϵ^{sim} is analogously defined as the number of simulated events that pass the $X \to b\bar{b}$ tagger

selection requirement ($N^{\text{sim}}_{\text{passed}}$) divided by the total number of signal events in simulation ($N^{\text{sim}}_{\text{total}}$). These efficiencies are thus expressed as:

$$\epsilon^{\text{data}} = N^{\text{data}}_{\text{passed}}/N^{\text{data}}_{\text{total}}, \qquad \epsilon^{\text{sim}} = N^{\text{sim}}_{\text{passed}}/N^{\text{sim}}_{\text{total}} \tag{7.2}$$

The equation for the SF can then be reformulated as:

$$\text{SF} = \frac{\epsilon^{\text{data}}}{\epsilon^{\text{sim}}} = \frac{\frac{N^{\text{data}}_{\text{passed}}}{N^{\text{data}}_{\text{total}}}}{\frac{N^{\text{sim}}_{\text{passed}}}{N^{\text{sim}}_{\text{total}}}} = \frac{\frac{N^{\text{data}}_{\text{passed}}}{N^{\text{sim}}_{\text{passed}}}}{\frac{N^{\text{data}}_{\text{total}}}{N^{\text{sim}}_{\text{total}}}} = \frac{\mu_{\text{post-tag}}}{\mu_{\text{pre-tag}}} \tag{7.3}$$

where $\mu_{\text{pre-tag}}$ and $\mu_{\text{post-tag}}$ denote the number of signal events in data divided by the number of signal events in simulation before and after $X \to b\bar{b}$ tagging, respectively. These are also referred to as the signal strengths before and after $X \to b\bar{b}$ tagging. The analysis methods used to measure both these signal strength parameters are summarised below.

The measurement of $\mu_{\text{post-tag}}$ is performed using events where at least one large-R jet is identified as a $Z \to b\bar{b}$ candidate. The $Z \to b\bar{b}$ candidate is determined by applying the $X \to b\bar{b}$ tagger at the 60% efficiency WP. A likelihood fit to the $Z \to b\bar{b}$ candidate mass is used to determine the signal strength. The dominant source of background for the $\mu_{\text{post-tag}}$ measurement is multijet events, with additional small contributions from $W(\to q\bar{q}')$ + jets and $t\bar{t}$ events.

$\mu_{\text{pre-tag}}$ is obtained using $Z \to l^+l^-$ ($l = e/\mu$) events because of low signal significance in $Z \to b\bar{b}$ events before $X \to b\bar{b}$ tagging due to high contamination from multijet background. $Z \to l^+l^-$ + jets events have a substantially smaller multijet background. The main backgrounds for this measurement are the leptonically decaying WZ and ZZ events [232]. The number of $Z \to b\bar{b}$ events in data before $X \to b\bar{b}$ tagging can be estimated at the leading order as:

$$N^{\text{data}}_{Z\to b\bar{b}} = N^{\text{sim}}_{Z\to b\bar{b}} \cdot \frac{N^{\text{data}}_{\ell\ell} - N^{\text{sim}}_{\text{bkg},\ell\ell}}{N^{\text{sim}}_{Z\to l^+l^-}}, \tag{7.4}$$

where $N^{\text{sim}}_{Z\to b\bar{b}}$ is the number of $Z \to b\bar{b}$ signal events in simulation before $X \to b\bar{b}$ tagging, $N^{\text{data}}_{\ell\ell}$ is the total number of selected events in data, $N^{\text{sim}}_{\text{bkg},\ell\ell}$ is the number of background events predicted by simulation, and $N^{\text{sim}}_{Z\to l^+l^-}$ is the number of $Z \to l^+l^-$ signal events predicted by simulation. $\mu_{\text{pre-tag}}$ can hence be determined as:

$$\mu_{\text{pre-tag}} = \frac{N^{\text{data}}_{\ell\ell} - N^{\text{sim}}_{\text{bkg},\ell\ell}}{N^{\text{sim}}_{Z\to l^+l^-}} \tag{7.5}$$

where $\mu_{\text{pre-tag}}$ depends only on the yields obtained in the $Z \to l^+l^-$ channel. There are no additional corrections applied in extrapolating the measurement from the $Z \to b\bar{b}$ to the $Z \to l^+l^-$ channel as a similar set-up is used to generate both simulated processes. Moreover, the kinematic distributions of $Z \to b\bar{b}$ and $Z \to l^+l^-$ decays are found to be similar as reported in [233]. The event selections for both the $Z \to b\bar{b}$ and $Z \to l^+l^-$ events used in the calibration analysis are described in the next section.

The SFs are measured as a function of the $Z \to b\bar{b}$ candidate p_{T}. Four bins are defined for the SF measurement: 200 GeV $< p_{\text{T}} <$ 450 GeV, 450 GeV $< p_{\text{T}} <$ 500 GeV, 500 GeV $< p_{\text{T}} <$ 600 GeV and

600 GeV $< p_{\mathrm{T}} <$ 1000 GeV. The SF in the first p_{T} bin is measured using $Z \to b\bar{b} + \gamma$ events and the SFs in the remaining p_{T} bins are measured using $Z \to b\bar{b}$ + jets events. p_{T} bins for the measurement using $Z \to b\bar{b}$ + jets events are defined based on having sufficient numbers of signal and background events in each bin. The measurement of the SFs in the high p_{T} regime using $Z \to b\bar{b} + \gamma$ events is not feasible due to the low number of signal events observed in data.

7.1.2 Calibration analysis strategy

Event selection for the $\mu_{\mathrm{post-tag}}$ measurement

Events with at least two large-R jets with $p_{\mathrm{T}} > 200$ GeV and $|\eta| < 2.0$, and no electrons or muons with $p_{\mathrm{T}} > 25$ GeV and $|\eta| < 2.5$ are selected. The leading-p_{T} large-R jet is required to fulfil the following requirements:

- $p_{\mathrm{T}} > 450$ GeV. This requirement ensures that the applied large-R jet triggers [234, 235] are 100% efficient,
- should contain at least two ghost-associated VR track-jets with $p_{\mathrm{T}} > 7$ GeV,
- mass > 50 GeV. This requirement excludes the region where the large-R jet mass calibration is not applicable.

Additional requirements are applied to the large-R jets to remove the the mismodelled regions and to reject backgrounds:

- $\frac{p_{\mathrm{T},1}-p_{\mathrm{T},2}}{p_{\mathrm{T},1}+p_{\mathrm{T},2}} < 0.15$, where $p_{\mathrm{T},1}$ ($p_{\mathrm{T},2}$) is the transverse momentum of the (sub-)leading large-R jet,
- $|\Delta y_{1,2}| < 1.2$, where $|\Delta y_{1,2}|$ is the rapidity difference of the leading and sub-leading large-R jets.

The leading-p_{T} large-R jet passing the $X \to b\bar{b}$ tagger requirement at the 60% efficiency WP is chosen as the $Z \to b\bar{b}$ candidate. Data-to-simulation comparisons for the $Z \to b\bar{b}$ candidate mass distribution for p_{T} bins corresponding to 450 GeV $< p_{\mathrm{T}} <$ 500 GeV, 500 GeV $< p_{\mathrm{T}} <$ 600 GeV, and 600 GeV $< p_{\mathrm{T}} <$ 1000 GeV are shown in Figure 7.2. One can observe a good agreement between data and simulated processes. Slight deviations are observed for large-R jet masses above 140 GeV for p_{T} bin corresponding to 450 GeV $< p_{\mathrm{T}} <$ 500 GeV. This disagreement could potentially arise from the mismodelled multijet background such that the flavour composition of the two b-tagged jets inside the large-R jet vary between simulation and data. The multijet background is hence not estimated using simulation, but using a data-driven procedure as described later.

Event selection for the $\mu_{\mathrm{pre-tag}}$ measurement

Events with at least two leptons of the same flavour and at least one large-R jet with $p_{\mathrm{T}} > 200$ GeV and $|\eta| < 2.0$ are selected. In case of two muons, a $p_{\mathrm{T}} > 27$ GeV requirement on both the muons is applied and the selected muons should have opposite charge. For electrons, a selection requirement of $p_{\mathrm{T}} >$ 25 GeV is applied with no opposite charge criteria due to a higher rate of charge misidentification. A lepton p_{T} balance requirement: $|p_{\mathrm{T}}^{\ell^+} - p_{\mathrm{T}}^{\ell^-}|/p_{\mathrm{T}}^{\ell^+\ell^-} < 0.8$ is also applied, where $p_{\mathrm{T}}^{\ell^+\ell^-}$ is the $Z \to l^+l^-$ candidate p_{T}. This requirement helps improve the data-to-simulation agreement at high $p_{\mathrm{T}}^{\ell^+\ell^-}$. In analogy to the $\mu_{\mathrm{post-tag}}$ measurement, the following requirements are applied:

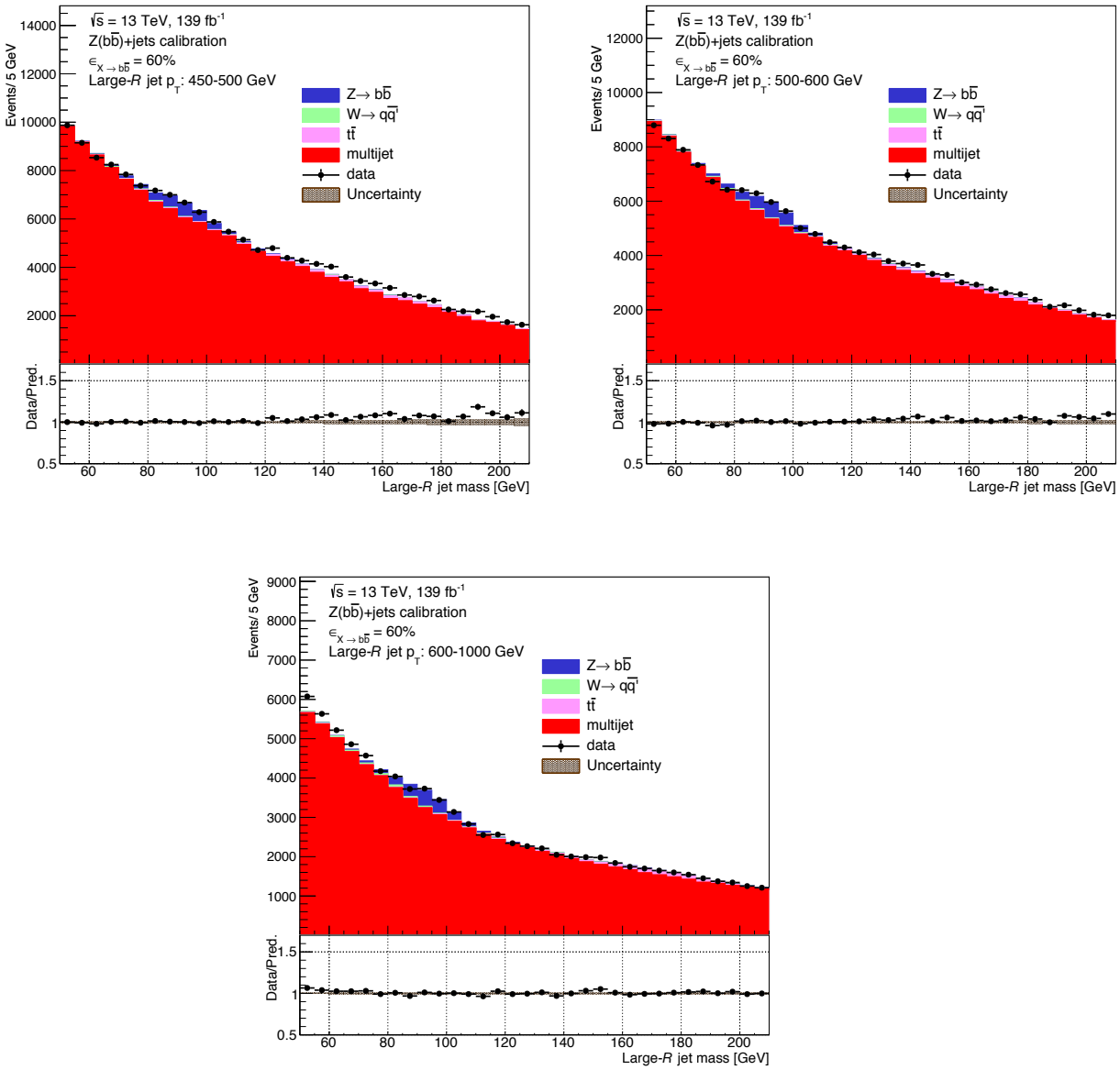


Figure 7.2: Data-to-simulation comparisons for the $Z \to b\bar{b}$ candidate mass distribution for p_{T} bins corresponding to $450 < p_{\mathrm{T}} < 500$ GeV, $500 < p_{\mathrm{T}} < 600$ GeV, and $600 < p_{\mathrm{T}} < 1000$ GeV. The uncertainty band shows only the statistical uncertainty of the simulated processes.

- $p_{\mathrm{T}}^{\ell^+\ell^-} > 450$ GeV and larger than the leading large-R jet p_{T}: $p_{\mathrm{T}}^{\ell^+\ell^-} > p_{\mathrm{T}}^{\text{lead. jet}}$,
- p_{T}-symmetry requirement: $\frac{p_{\mathrm{T}}^{\ell^+\ell^-} - p_{\mathrm{T}}^{\text{lead. jet}}}{p_{\mathrm{T}}^{\ell^+\ell^-} + p_{\mathrm{T}}^{\text{lead. jet}}} < 0.15$,
- rapidity difference: $|\Delta y_{\ell^+\ell^-,\text{lead. jet}}| < 1.2$.

The signal region is defined by a mass window of 66 GeV $< m_{\ell^+\ell^-} <$ 116 GeV.

Signal and background modelling

Techniques for signal and background modelling differ for the $\mu_{\text{post-tag}}$ and $\mu_{\text{pre-tag}}$ measurements. For the $\mu_{\text{post-tag}}$ measurement, the multijet background is determined using data-driven techniques. Other background processes and signal process are modelled using simulations. For the $\mu_{\text{pre-tag}}$ measurement, no dedicated modelling technique is used as only signal and background yields are needed to be extracted after applying the event selection requirements.

The $Z \to b\bar{b}$ candidate mass distribution is modelled using a double-sided Crystal Ball (DSCB) function [236]. The functional form is fitted to the simulated $Z \to b\bar{b}$ candidate events after applying the event selection requirements. The DSCB function is expressed as:

$$f\left(m|m_Z,\sigma_Z,\alpha_L,\alpha_H,n_L,n_H\right) = \begin{cases} \exp\left(-\frac{\alpha_L^2}{2}\right)\alpha_L\left[\frac{n_L}{\alpha_L}\left(\frac{n_L}{\alpha_L}-\alpha_L-\frac{m-m_Z}{\sigma_Z}\right)\right]^{-n_L}, & m < -\alpha_L \\ \exp\left[-\frac{1}{2}\left(\frac{m-m_Z}{\sigma_Z}\right)^2\right], & -\alpha_L \le m \le \alpha_H \\ \exp\left(-\frac{\alpha_H^2}{2}\right)\alpha_H\left[\frac{n_H}{\alpha_H}\left(\frac{n_H}{\alpha_H}-\alpha_H+\frac{m-m_Z}{\sigma_Z}\right)\right]^{-n_H}, & m > \alpha_H \end{cases} \tag{7.6}$$

where the first and the third case of the equation describes the tails and the second case describes the core of the distribution. m_Z and σ_Z are the mean and the standard deviation of the Gaussian core, respectively. α_L, n_L and α_H, n_H are the decay constants and normalisation of the low and the high mass tails, respectively. Figure 7.3 depicts the fit to the $Z \to b\bar{b}$ candidate mass distribution using a DSCB function for the p_{T} range of 450 GeV $< p_{\mathrm{T}} <$ 1000 GeV.

As mentioned earlier, the dominant source of background in the $Z \to b\bar{b}$ $\mu_{\text{post-tag}}$ measurement is multijet events with small contributions from $W(\to q\bar{q}')$ + jets and $t\bar{t}$ events. The $t\bar{t}$ events are split into two categories. The first category consists of $t\bar{t}$ events with a large-R jet containing the decay products of both top-quarks. The remaining events are included in the second category. The background shapes of both the $W(\to q\bar{q}')$ + jets and $t\bar{t}$ events are modelled using a DSCB function. The parameters of the DSCB function are derived by fitting the simulated events after applying the event selection requirements for different background processes.

The multijet background is modelled using an unbinned maximum-likelihood fit to the sidebands of the $Z \to b\bar{b}$ candidate mass distribution in data. The sidebands of the $Z \to b\bar{b}$ candidate mass distribution are defined as 50 GeV $< m_{b\bar{b}} <$ 70 GeV and 110 GeV $< m_{b\bar{b}} <$ 150 GeV[4]. The upper limit on the $Z \to b\bar{b}$ candidate mass range excludes the top-quark mass region. The fit is performed using classes of exponential and polynomial functions of 2$^{\text{nd}}$ to 5$^{\text{th}}$ order in p_{T} bins of the $Z \to b\bar{b}$

[4] The choice of the sidebands is motivated from the mass window requirement yielding least a signal significance, $S/\sqrt{B}$, from the different tested mass requirements.

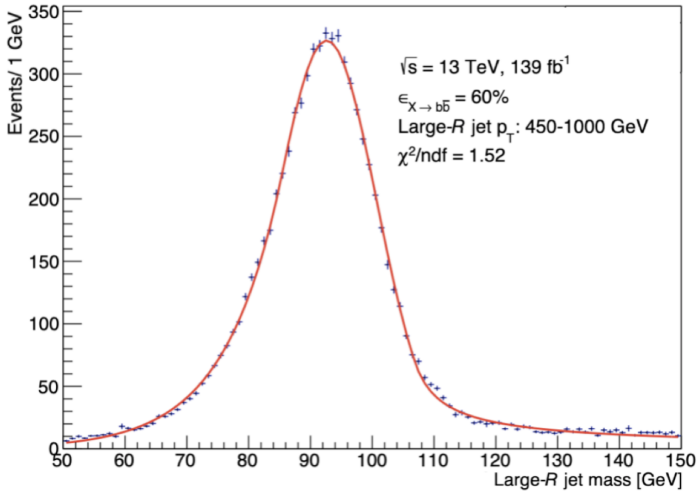


Figure 7.3: The fit to the $Z \rightarrow b\bar{b}$ candidate mass distribution using a DSCB function for a p_{T} range of 450 GeV < p_{T} < 1000 GeV.

candidate. To decide on the suitable number of free parameters (the order) out of the two tested classes of functions, an F-test is performed [237]. The test statistic $F_{p,q}$ is calculated as:

$$F_{p,q} = \frac{\chi_p^2 - \chi_q^2}{n_q - n_p} / \frac{\chi_q^2}{n - n_q}, \tag{7.7}$$

where χ_p^2 and χ_q^2 are computed in n bins of the two fits with n_p and n_q degrees of freedom. In the asymptotic limit, the test statistic $F_{p,q}$ follows a Fischer distribution, $\mathcal{F}(F|n_q - n_p, n - n_q)$. If the additional free parameter does not significantly improve the model, the hypothesis of the function with a smaller number of free parameters is rejected with $\mathcal{P}(F' \geq F) < 0.05$. The fits to the sidebands of the $Z \rightarrow b\bar{b}$ candidate mass distribution in data, using the functions resulting from an F-test, in different p_{T} bins are shown in Figures 7.4 and 7.5.

The final decision on the functional form selected out of the two suitable functions resulting from an F-test is taken based on a 'spurious signal test'. Any choice of the functional form can introduce a potential bias in the modelling of the backgrounds, and can lead to an artificial (spurious) signal. In order to quantify this spurious signal, the signal-plus-background model is used to fit the $Z \rightarrow b\bar{b}$ candidate mass distribution of the multijet background[5] in $Z \rightarrow b\bar{b}$ candidate p_{T} bins. The function with the smallest fitted signal strength (spurious signal strength) is chosen as the best function to model the multijet background. The best functions to describe the multijet background are summarised in Table 7.2.

[5] The multijet background is reweighted to match the data distribution using sideband regions.

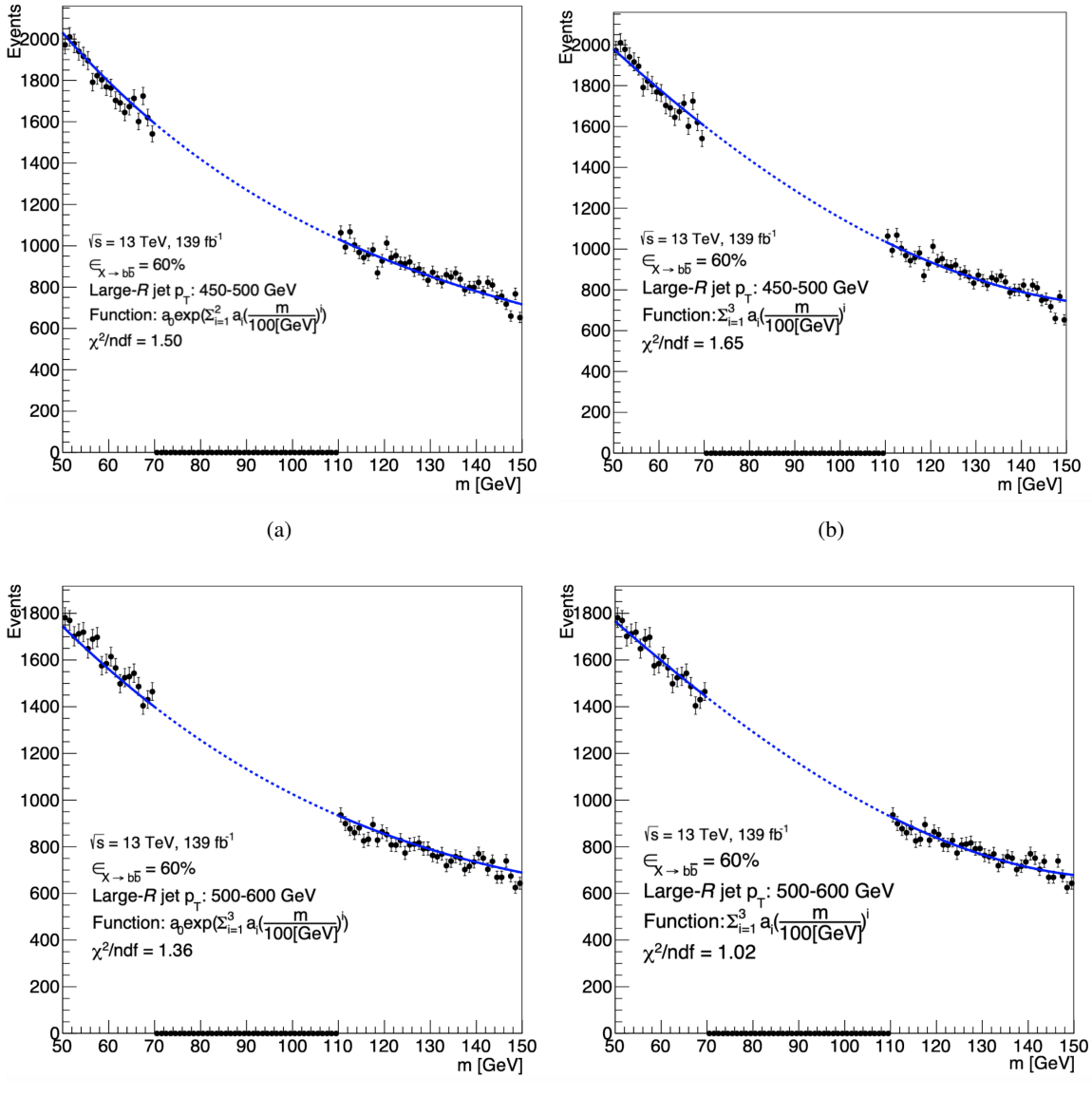
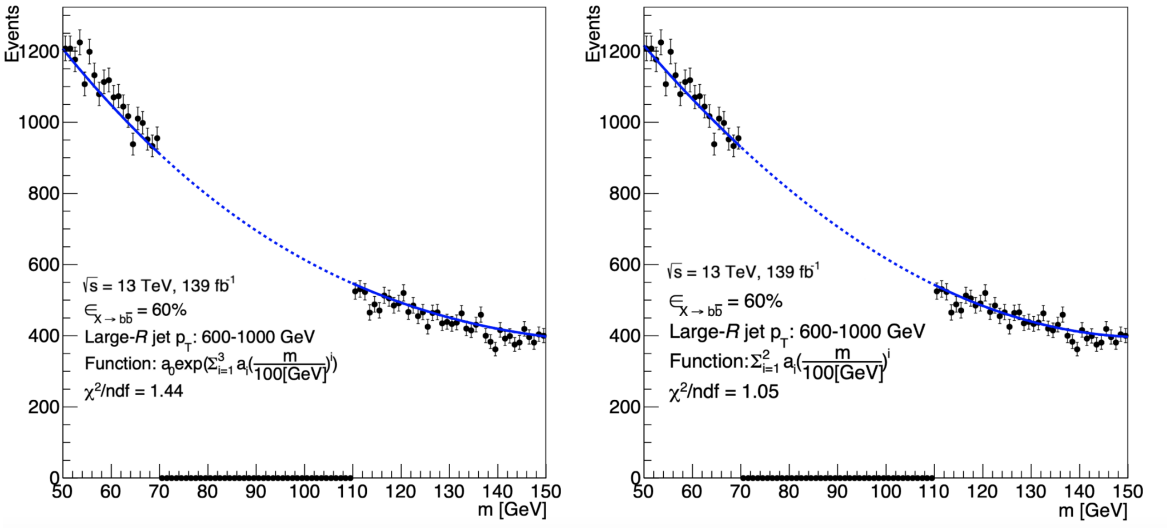


Figure 7.4: The fits to the sidebands of the $Z \to b\bar{b}$ candidate mass distribution in data using the functions resulting from an F-test for p_{T} bins corresponding to 450 GeV$< p_{\mathrm{T}} <$ 500 GeV and 500 $< p_{\mathrm{T}} <$ 600 GeV. The resulting exponential functions are shown on the left, and the polynomial functions are shown on the right. The function parameter m refers to the mass of the $Z \to b\bar{b}$ candidate.

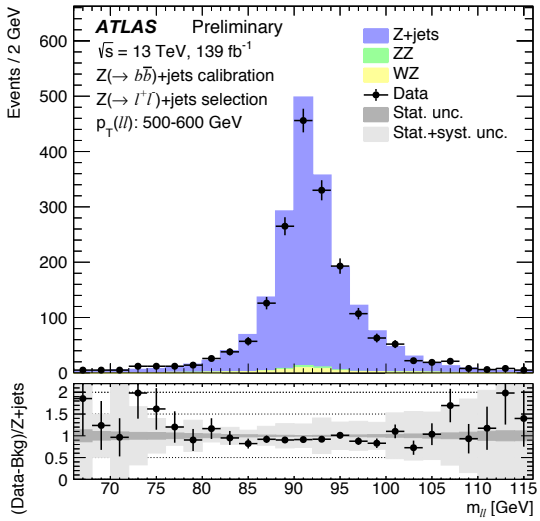


(a) (b)

Figure 7.5: The fits to the sidebands of the $Z \to b\bar{b}$ candidate mass distribution in data using the functions resulting from an F-test for p_{T} bins corresponding to 600 GeV$< p_{\mathrm{T}} <$ 1000 GeV. The resulting exponential functions are shown on the left, and the polynomial functions are shown on the right. The function parameter m refers to the mass of the $Z \to b\bar{b}$ candidate.

$Z \to b\bar{b}$ p_{T} bin	Best function
450 GeV $< p_{\mathrm{T}} <$ 500 GeV	$\sum_{i=0}^{3} a_i (\frac{m}{100[\text{ GeV}]})^i$
500 GeV $< p_{\mathrm{T}} <$ 600 GeV	$\sum_{i=0}^{3} a_i (\frac{m}{100[\text{ GeV}]})^i$
600 GeV $< p_{\mathrm{T}} <$ 1000 GeV	$a_0 \exp(\sum_{i=1}^{3} a_i (\frac{m}{100[\text{ GeV}]})^i)$

Table 7.2: Best functional forms used to describe the multijet background for three $Z \to b\bar{b}$ candidate p_{T} bins. a_i are the free parameters in the fit, and are determined in each p_{T} bin individually. m denotes the mass of the Z boson candidate.



(a)

Figure 7.6: The $Z \to l^+l^-$ candidate mass distribution in the Z boson candidate p_{T} bin 500 GeV < p_{T}^{Z} < 600 GeV [28].

The signal and background processes for the $\mu_{\mathrm{pre-tag}}$ measurement are modelled solely using simulated events. Figure 7.6 shows one of the most representative distributions for the $Z \to l^+l^-$ + jets process: $Z \to l^+l^-$ candidate mass distribution for 500 GeV< p_{T} < 600 GeV. The $Z \to l^+l^-$ candidate mass distribution for the remaining p_{T} bins are shown in Figure C.1 in Appendix C. Overall, a good agreement is observed between data and simulated processes.

7.1.3 Systematic uncertainties and results

The sources of systematic uncertainties considered for the calibration analysis are grouped into three main categories: experimental uncertainties on the physics objects, theoretical uncertainties on the modelling of the simulated processes. In addition, statistical uncertainties on the simulated processes are also considered.

The most relevant experimental uncertainties include the large-R jet energy scale, jet mass scale, and resolution uncertainties as described in [238, 239] for all simulated processes. Other uncertainties include the ones on the small-R jet energy scale and resolution. Wherever relevant, uncertainties on the trigger, reconstruction, identification, and isolation efficiencies, and on the momentum of electrons and muons are also taken into account (see Section 6.2.1).

An uncertainty on the integrated luminosity for the ATLAS full Run 2 dataset is considered as 0.83% [193], and is applied to the physics processes whose normalisation are taken from simulation. An uncertainty related to the pile-up simulation is also taken into account.

The uncertainties on the modelling of Z and W + jets events are evaluated by using alternative generator samples by varying the PDF, α_{s}, μ_F, and μ_R. A 10% normalisation uncertainty is assigned to the W + jets background, while the normalisation of the Z + jets signal is determined from the fit. An additional background modelling uncertainty (spurious signal uncertainty) associated with the choice

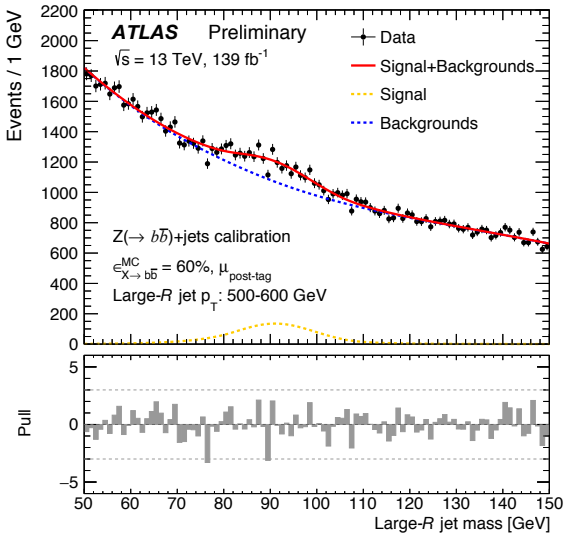


Figure 7.7: The $Z \to b\bar{b}$ candidate mass distribution after the fit and applying the $Z \to b\bar{b}$ + jets selection and the the $X \to b\bar{b}$ tagger requirement at the 60% WP for events with the large-R jet in 500 GeV< p_{T} < 600 GeVrange. The dashed yellow line represents the Z + jets signal process. The dashed blue line represents the sum of all background processes. The solid red line is the sum of all signal and background processes [28].

of the functional form for the multijet background is taken into account. This uncertainty is evaluated as the maximum between the smallest spurious signal strength and its associated statistical uncertainty. Due to the small contribution of the $t\bar{t}$ background in the analysis, only its dominant modelling uncertainties are considered which include the ones due to an alternative parton shower model and alternative matrix-element-to-parton-shower (ME-PS) matching. An overall 6% normalisation uncertainty on the $t\bar{t}$ production cross-section is furthermore applied. A systematic uncertainty on the choice of the large-R jet mass range that enters the likelihood fit for the $\mu_{\mathrm{post-tag}}$ measurement is also considered. To evaluate this uncertainty, $\mu_{\mathrm{post-tag}}$ is evaluated for three alternative mass ranges and the largest variation with respect to the mass range used in the default fit is taken as the systematic uncertainty.

$\mu_{\mathrm{post-tag}}$ in the $Z \to b\bar{b}$ + jets calibration is measured using the $Z \to b\bar{b}$ invariant mass distribution for the three large-R jet p_{T} bins as mentioned earlier. To extract the signal strength, an unbinned maximum-likelihood fit is performed. The fit is performed using the `RooFit` toolkit in the large-R jet mass range between 50 GeV and 150 GeV. The parameters describing the functional form for the multijet background are allowed to float independently in the likelihood function, and are determined directly from the fit to data. Figure 7.7 shows the post-fit $Z \to b\bar{b}$ candidate mass distribution with the $X \to b\bar{b}$ tagging requirement at the 60% WP for 500 GeV < p_{T} < 600 GeV. The post-fit $Z \to b\bar{b}$ candidate mass distributions for the remaining p_{T} bins are shown in Figure C.2 in the Appendix. The pulls defined as data minus the fitted model prediction divided by the data statistical uncertainty are mostly within 3 σ.

$\mu_{\mathrm{pre-tag}}$ (see Equation 7.5) is defined as the ratio of observed yields in data minus the expected background yields divided by the expected $Z \to l^+l^-$ signal yields. All yields are determined after the $Z \to l^+l^-$ selection requirements and in the Z boson candidate p_{T} bins.

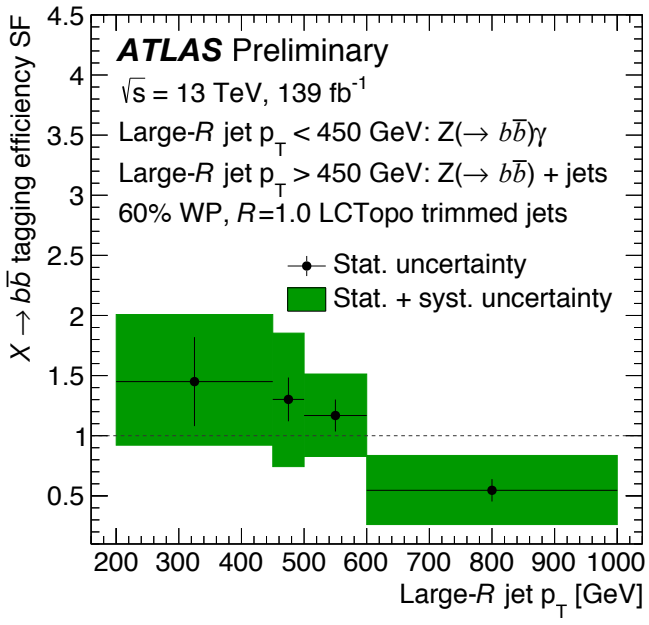


Figure 7.8: Signal efficiency scale factors for $X \to b\bar{b}$ tagger at the 60% efficiency WP [28].

The signal efficiency scale factors for the $X \to b\bar{b}$ tagger at the 60% efficiency WP are illustrated in Figure 7.8. As mentioned in Section 7.1.1, the SF for bin corresponding to 200 GeV$< p_{\mathrm{T}} <$ 450 GeVis measured using the $Z \to b\bar{b} + \gamma$ calibration method [28], and follows a similar strategy as the $Z \to b\bar{b}$ + jets calibration method. The SFs above 450 GeV are measured using the $Z \to b\bar{b}$ + jets calibration method. The SFs are found to be consistent with unity within the systematic uncertainties except for the bin corresponding to 600 GeV$< p_{\mathrm{T}} <$ 1000 GeV. The SF in this bin is found to be significantly below 1 which can potentially arise from insufficient background modelling due to a relatively smaller number of background events passing the selection requirements. However, studies reported in [231] suggest that a binned likelihood fit can help model the multijet background better than an unbinned fit, and is being used as a default method in the newer versions of this calibration analysis.

7.2 Analysis strategy used in the search for merged $H^+ \to Wh$ decays

7.2.1 Event selection

A common set of selection requirements on the reconstructed physics objects (see Section 3.4) are applied to both simulated events and data. The selection requirements are as following:

- Exactly one prompt e or μ,
- $E_{\mathrm{T}}^{\mathrm{miss}} > 30\,\mathrm{GeV}$,
- $\geq$ one large-R jet with exactly one $X \to b\bar{b}$ tagged jet at the 60% efficiency WP, with ΔR (large-R jet, e/μ) > 1.0,

- $\geq$ two small-R jets with ΔR (large-R jet, small-R jet) > 1.4.

The large-R jet passing the $X \rightarrow b\bar{b}$ tagging requirement at the 60% efficiency WP is taken as the h boson candidate. The $q\bar{q}b\bar{b}$ and $\ell\nu b\bar{b}$ channels are categorised based on the following requirements:

- $q\bar{q}b\bar{b}$: Presence of an additional large-R jet failing the $X \rightarrow b\bar{b}$ tagging requirement with its mass around the W boson mass: 50 GeV $< m_J <$ 110 GeV is taken as the W_{had} candidate, where m_J is the mass of the large-R jet,
- $\ell\nu b\bar{b}$: Absence of an additional large-R jet, W_{lep} candidate is reconstructed using a charged lepton and a neutrino candidate[6].

The H^+ boson candidate is then reconstructed from the h boson, and the $W_{\text{had}}/W_{\text{lep}}$ candidates in the $q\bar{q}b\bar{b}/\ell\nu b\bar{b}$ channels.

7.2.2 NN training in the $q\bar{q}b\bar{b}$ and $\ell\nu b\bar{b}$ channels

Two separate networks are trained for the two analysis channels with the aim of classifying events as either signal[7] or background. The training is performed using events passing the selection requirements described in Section 7.2.1. Simulated events from higher signal mass hypotheses, $m_{H^+} \geq 1.2$ TeV, are included in a single training. Only those signal events which are correctly labelled as "true $q\bar{q}b\bar{b}$" and "true $\ell\nu b\bar{b}$" are considered. An event is labelled as a "true $q\bar{q}b\bar{b}$" event if the generator-level W boson is ΔR-matched to the W_{had} candidate. Likewise, a "true $\ell\nu b\bar{b}$" event is the one where the generator-level W boson is ΔR-matched to the W_{lep} candidate. On the other hand, simulated events from all the background types (see Section 4.2.2) are included in the training. A two-fold cross-validation approach is used to help protect against any potential biases due to over-training, and to ensure that the network predictions are not applied to the events which were used for the training. Further details about the network architecture, and data processing can be found in [240].

NN training inputs in the $q\bar{q}b\bar{b}$ channel

In the $q\bar{q}b\bar{b}$ channel, seven variables built from reconstructed physics objects are used as inputs to the NN. These include the angular separations between the objects and the ratio of the transverse momentum of the objects. These variables are useful as the energy scale differences and separation between the physics objects can vary between signal and background. Analogous to the search for resolved $H^{\pm} \rightarrow Wh$ decays, the dependence on the H^+ boson mass is reduced by taking the ratio of the transverse momentum with respect to the reconstructed charged Higgs boson mass, m_{Wh}. The input variables are defined as following:

- The fraction of the transverse momentum of the reconstructed H^+ boson's decay products carried by the reconstructed W_{lep} candidate: $p_{\text{T}}^{W_{\text{lep}}} / \left(p_{\text{T}}^{W_{\text{had}}} + p_{\text{T}}^{h} \right)$,
- the angular separation between the lepton l and W_{had} candidates: $\Delta R\left(l, W_{\text{had}}\right)$,
- the angular separation between the lepton and h boson candidates: $\Delta R\left(l, h\right)$,

[6] The p_z component of the neutrino is reconstructed using the W boson mass constraint method (Section 4.5.1, Chapter 4).
[7] In the search for merged H^+ boson decays, simulated events from $m_{H^{\pm}} \geq 500$ GeV are used in the analysis.

- the pseudorapidity difference between the h boson and W_{had} candidates: $\Delta\eta\left(h, W_{\text{had}}\right)$,
- the azimuthal angle difference between the h boson and W_{had} candidates: $\Delta\phi\left(h, W_{\text{had}}\right)$,
- the ratio of the W_{had} transverse momentum to the mass of the reconstructed charged Higgs boson, m_{Wh}: $p_{\text{T}}^{\text{W}_{\text{had}}}/m_{\text{Wh}}$,
- the ratio of the h boson transverse momentum to the mass of the reconstructed charged Higgs boson, m_{Wh}: $p_{\text{T}}^{\text{h}}/m_{\text{Wh}}$.

NN training inputs in the $\ell\nu b\bar{b}$ channel

NN training in the $\ell\nu b\bar{b}$ channel is based on the similar input variables as used in the $q\bar{q}b\bar{b}$ channel. A leptonic top-quark candidate mass, reconstructed following the same procedure as described in Section 4.5.1, is used in addition to reduce the amount of $t\bar{t}$ + jets background. The input variables are defined as following:

- The fraction of the transverse momentum of the reconstructed H^+ boson decay products carried by the reconstructed W_{lep} candidate: $p_{\text{T}}^{W_{\text{lep}}}/\left(p_{\text{T}}^{W_{\text{had}}} + p_{\text{T}}^{h}\right)$[8],
- the reconstructed leptonic top-quark candidate mass: $m_{\text{top}}^{\text{lep}}$,
- the angular separation between the lepton and h boson candidates: $\Delta R\left(l, h\right)$,
- the pseudorapidity difference between the h boson and W_{lep} candidates: $\Delta\eta\left(h, W_{\text{lep}}\right)$,
- the azimuthal angle difference between the h boson and W_{lep} candidates: $\Delta\phi\left(h, W_{\text{lep}}\right)$,
- the ratio of the W_{lep} transverse momentum to the mass of the reconstructed charged Higgs boson, m_{Wh}: $p_{\text{T}}^{\text{W}_{\text{lep}}}/m_{\text{Wh}}$,
- the ratio of the h boson transverse momentum to the mass of the reconstructed charged Higgs boson, m_{Wh}: $p_{\text{T}}^{\text{h}}/m_{\text{Wh}}$.

Data-to-simulation comparisons for the classification NN output score (w_{NN}) overlaid with the corresponding signal distribution for m_{H^+} = 1 TeV for the two analysis channels is shown in Figure 7.9. This comparison is shown for events passing the selection requirements described in Section 7.2.1. A good agreement between data and background simulated process is observed within the systematic uncertainties. The signal distribution is seen to peak around higher w_{NN} values in contrast to the background distribution peaking at lower w_{NN} values reflecting a reasonable signal-to-background discrimination. Further details about the variable optimisation, variable ranking, and NN performance can be found in [240, 241].

[8] This variable is only built if there is a presence of an additional large-R jet under the $\ell\nu b\bar{b}$ event hypothesis, facilitating the reconstruction of W_{had} candidate.

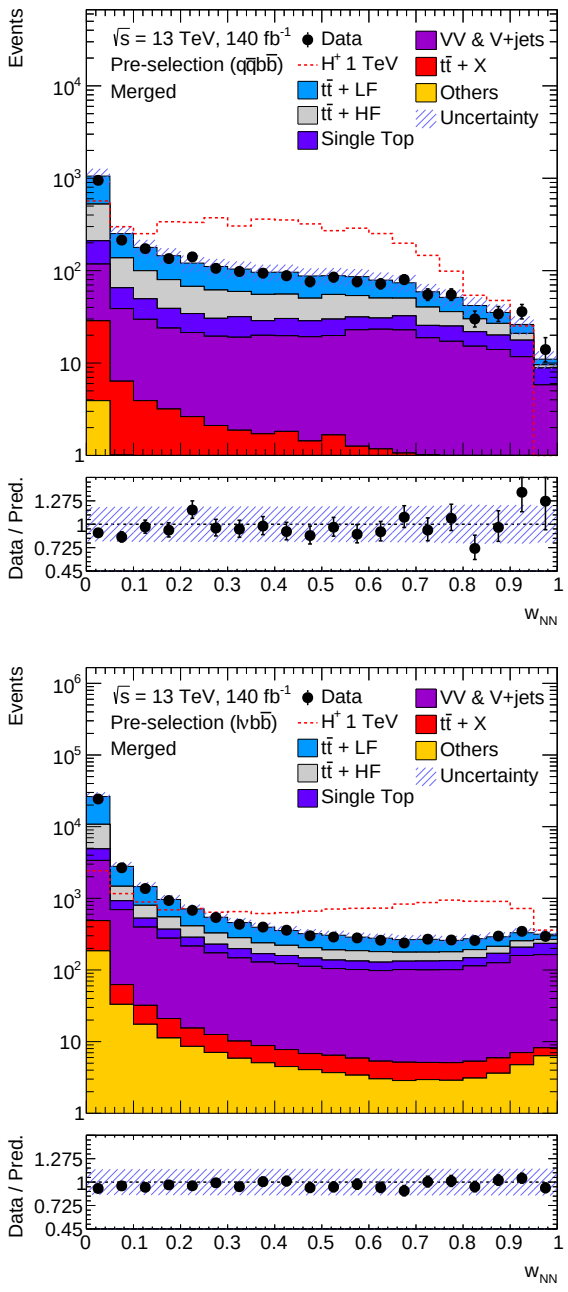


Figure 7.9: Data-to-simulation comparisons for w_{NN} for the two analysis channels. The corresponding signal distribution for $m_{H^+} = 1\,\mathrm{TeV}$ is also overlaid. This comparison is shown for events passing the selection requirements described in Section 7.2.1.

7.2.3 Signal and control regions

The definition of signal and control regions used in the search for merged $H^\pm \to Wh$ decays is based on the following criteria:

- The classification NN output score (w_{NN}) allocated to an event,
- The multiplicity of b-tagged small-R jets not originating from the H^+ boson. This criteria acts as a good discriminant for different SM background processes. Two event categories are defined in this regard, namely the zero b-tag category (0 b) and the at least one b-tag category (≥ 1 b),
- The reconstructed h boson candidate mass, $m_{b\bar{b}}$. The signal processes are expected to peak around the h boson mass contrary to the background processes. Events satisfying the h boson mass window requirement, 95 GeV$< m_{b\bar{b}} <$ 140 GeV, constitute the signal regions, whereas the ones falling outside this mass window requirement constitute the control regions.

A summary of the selection requirements used to define the signal and control regions[9] for both the channels is given in Table 7.3. Further details about the optimisation of signal and control regions can be found in [240, 241].

Table 7.3: A summary of the selection requirements used to define the signal and control regions in the search for merged $H^\pm \to Wh(\to b\bar{b})$ decays. The regions are defined for the zero and at least one b-tagged jets categories. All the requirements are applied to both these event categories, unless explicitly stated.

Region	**h mass window**	**w_{NN} ($\ell\nu b\bar{b}$)**	**w_{NN} ($q\bar{q}b\bar{b}$)**
High-NN score signal regions	$95 < m_h$ (GeV) < 140	$w_{\text{NN}} > 0.83$	$w_{\text{NN}} > 0.2$ (for 0 b category) $w_{\text{NN}} > 0.1$ (for ≥ 1 b category)
Medium-NN score signal regions	$95 < m_h$ (GeV) < 140	$0.4 < w_{\text{NN}} < 0.83$	-
Low-NN score signal regions	$95 < m_h$ (GeV) < 140	$w_{\text{NN}} < 0.4$	$w_{\text{NN}} < 0.2$ (for 0 b category) $w_{\text{NN}} < 0.1$ (for ≥ 1 b category)
Low-mass sideband control region	m_h (GeV) < 95	-	-
High-mass sideband control region	m_h (GeV) > 140	-	-

7.2.4 Statistical Model

The statistical model used to scrutinise the m_{Wh} distibution for the presence of an excess in data events over the SM background prediction in the search for merged H^+ boson decays is similar to the model used in the search for resolved H^+ boson decays (see Section 6.1) with the following main additions:

- Experimental systematic uncertainties related to large-R jets energy and mass scale, and resolution [238, 239] are incorporated in the statistical model for all simulated processes.

[9] The choice of the selection requirement on w_{NN}/m_h to define the signal regions follows a similar optimisation procedure as described in Section 5.2.

- Uncertainties associated with the $X \to b\bar{b}$ signal tagging efficiencies, and the corresponding mistagging efficiencies for boosted top-quarks and multijet events are also included for the relevant simulated processes. The $X \to b\bar{b}$ signal tagging scale factors and the corresponding uncertainties are presented in Section 7.1.3. The mistagging efficiency scale factors and the corresponding uncertainties for boosted top-quarks are calculated using a tag-and-probe method with semileptonic $t\bar{t}$ events as described in [28]. Uncertainties associated with the mistagging efficiency for multijet events are calculated exclusively for this analysis. These uncertainties are derived from events where there are no generator-level top-quark, $W/Z/h$ bosons in the vicinity of a large-R jet. The derivation methodology for these uncertainties can be found in [240].

- In addition to the overall normalisation factor of the $t\bar{t}$ + HF background, the overall normalisation factor of other dominant backgrounds, $t\bar{t}$ + LF and VV & V + jets (see Figure 7.9) are also allowed to float independently in a fit. Unlike the strategy followed in the search for resolved H^+ boson decays (Section 4.4), the $t\bar{t}$ + LF normalisation is not corrected by a dedicated $t\bar{t}$ reweighting procedure. The search for merged $H^+ \to Wh$ decays is more inclusive in regard of the small-R jet multiplicity, and is therefore less dependent on the modelling of the additional jets. Hence, the $t\bar{t}$ + LF normalisation is determined directly from the fit to data.

CHAPTER 8

Results

This chapter presents the results of the search for $H^+ \to Wh(\to b\bar{b})$ decays. The results obtained from fits under a background-only or a signal-plus-background hypothesis in the search for resolved $H^+ \to Wh(\to b\bar{b})$ decays are discussed in Section 8.1. The 95% CL upper limits on $\sigma(pp \to tbH^+) \times BR(H^+ \to Wh(\to b\bar{b}))$ as a function of signal mass hypothesis in the search for resolved H^+ boson decays and their combination with the limits obtained in the search for merged H^+ boson decays are shown in Section 8.2. Finally, Section 8.3 compares the $H^+ \to Wh(\to b\bar{b})$ search results with the $H^+ \to tb$ search which has the same H^+ boson production mode and final state.

8.1 Background-only and Signal-plus-background Fit Results

The distributions of the final discriminant, m_{Wh}, in various signal and control regions in the search for resolved H^+ boson decays are shown in Figures 8.1 to 8.4. These distributions are presented after the fit to data, under a background-only hypothesis, in all analysis regions. The observed data yields and background predictions after the fit to data are shown in Table 8.1. A good agreement between data and background prediction is observed. The best-fit value of the overall normalisation factor of the $t\bar{t}$ + HF background (k_{HF}) of 1.39 ± 0.18 is found to be consistent within the uncertainties with the values reported in various SM measurements [242, 243]. Similar results are obtained from fits, under a signal-plus-background hypothesis, for a range of signal mass values as shown in Table D.1 in Appendix D, confirming the stability of the statistical model. The best-fit k_{HF} value obtained from the fit under a background-only hypothesis is found to be consistent within 1σ with the best-fit k_{HF} values obtained from fits under the signal-plus-background hypothesis.

Table 8.1: The observed event yields and background predictions in various signal and control regions from a fit to data under a background-only hypothesis. The quoted uncertainties are the statistical and systematic uncertainties combined in quadrature. The uncertainties for the individual background processes are larger than the total background uncertainty due to correlations between the nuisance parameters in the fit.

	$t\bar{t}$ + LF	$t\bar{t}$ + HF	Single Top	VV & V + jets	$t\bar{t}$ + X	Others	Total	Data
$5j$ $3b$ CR ($q\bar{q}b\bar{b}$)	9000 ± 500	8400 ± 600	770 ± 190	500 ± 150	64 ± 11	5.0 ± 0.6	18720 ± 150	18737
$5j \geq 4b$ CR ($q\bar{q}b\bar{b}$)	470 ± 80	3640 ± 130	160 ± 50	69 ± 20	100 ± 15	5.3 ± 0.6	4450 ± 70	4449
$\geq 6j$ $3b$ CR ($q\bar{q}b\bar{b}$)	2200 ± 400	4100 ± 400	250 ± 80	180 ± 60	41 ± 8	6.0 ± 2.4	6790 ± 90	6788
$\geq 6j \geq 4b$ CR ($q\bar{q}b\bar{b}$)	330 ± 110	5140 ± 150	160 ± 60	87 ± 25	147 ± 18	22 ± 9	5890 ± 80	5889
$5j$ $3b$ Low-purity SR ($q\bar{q}b\bar{b}$)	57000 ± 4000	39000 ± 4000	3400 ± 900	1370 ± 120	530 ± 50	23.3 ± 1.3	100980 ± 330	100957
$5j \geq 4b$ Low-purity SR ($q\bar{q}b\bar{b}$)	330 ± 60	1650 ± 80	68 ± 24	27.9 ± 3.2	66 ± 10	2.46 ± 0.18	2140 ± 50	2145
$\geq 6j$ $3b$ Low-purity SR ($q\bar{q}b\bar{b}$)	41000 ± 6000	54000 ± 6000	2600 ± 700	1380 ± 140	1030 ± 100	78 ± 31	100500 ± 500	100485
$\geq 6j \geq 4b$ Low-purity SR ($q\bar{q}b\bar{b}$)	520 ± 140	6850 ± 190	190 ± 80	81 ± 8	290 ± 40	29 ± 13	7960 ± 90	7963
$5j$ $3b$ High-purity SR ($q\bar{q}b\bar{b}$)	5900 ± 700	5600 ± 800	470 ± 160	130 ± 10	117 ± 13	5.59 ± 0.28	12130 ± 140	12130
$5j \geq 4b$ High-purity SR ($q\bar{q}b\bar{b}$)	64 ± 16	356 ± 27	15 ± 5	4.2 ± 0.7	16.3 ± 2.9	0.66 ± 0.06	456 ± 21	456
$\geq 6j$ $3b$ High-purity SR ($q\bar{q}b\bar{b}$)	7900 ± 1400	14100 ± 1500	640 ± 230	236 ± 24	410 ± 40	26 ± 10	23260 ± 160	23258
$\geq 6j \geq 4b$ High-purity SR ($q\bar{q}b\bar{b}$)	190 ± 60	2850 ± 100	80 ± 40	28 ± 4	187 ± 26	16 ± 7	3340 ± 60	3343
$5j$ $3b$ CR ($l\nu b\bar{b}$)	820 ± 80	1310 ± 130	230 ± 90	220 ± 70	14 ± 4	1.22 ± 0.16	2600 ± 50	2599
$5j \geq 4b$ CR ($l\nu b\bar{b}$)	5.2 ± 2.2	34 ± 6	3.0 ± 1.9	5.0 ± 1.7	0.60 ± 0.15	0.05 ± 0.05	48 ± 6	49
$\geq 6j$ $3b$ CR ($l\nu b\bar{b}$)	710 ± 120	1870 ± 180	140 ± 70	240 ± 70	33 ± 7	5.0 ± 2.2	3000 ± 50	2991
$\geq 6j \geq 4b$ CR ($l\nu b\bar{b}$)	6.2 ± 2.7	121 ± 14	7 ± 5	8.9 ± 2.7	2.5 ± 0.4	0.68 ± 0.32	146 ± 12	147
5j 3b SR ($l\nu b\bar{b}$)	1380 ± 130	1060 ± 170	380 ± 80	165 ± 13	46 ± 5	2.86 ± 0.20	3030 ± 50	3026
5j $\geq 4b$ SR ($l\nu b\bar{b}$)	15.0 ± 2.8	130 ± 14	15 ± 6	8.1 ± 1.1	6.3 ± 0.9	0.44 ± 0.06	174 ± 13	175
$\geq 6j$ 3b SR ($l\nu b\bar{b}$)	1060 ± 170	2090 ± 220	200 ± 100	223 ± 21	98 ± 10	10 ± 4	3690 ± 60	3703
$\geq 6j \geq 4b$ SR ($l\nu b\bar{b}$)	17 ± 4	403 ± 32	24 ± 19	23.9 ± 2.6	21.9 ± 2.8	3.6 ± 1.7	494 ± 22	495

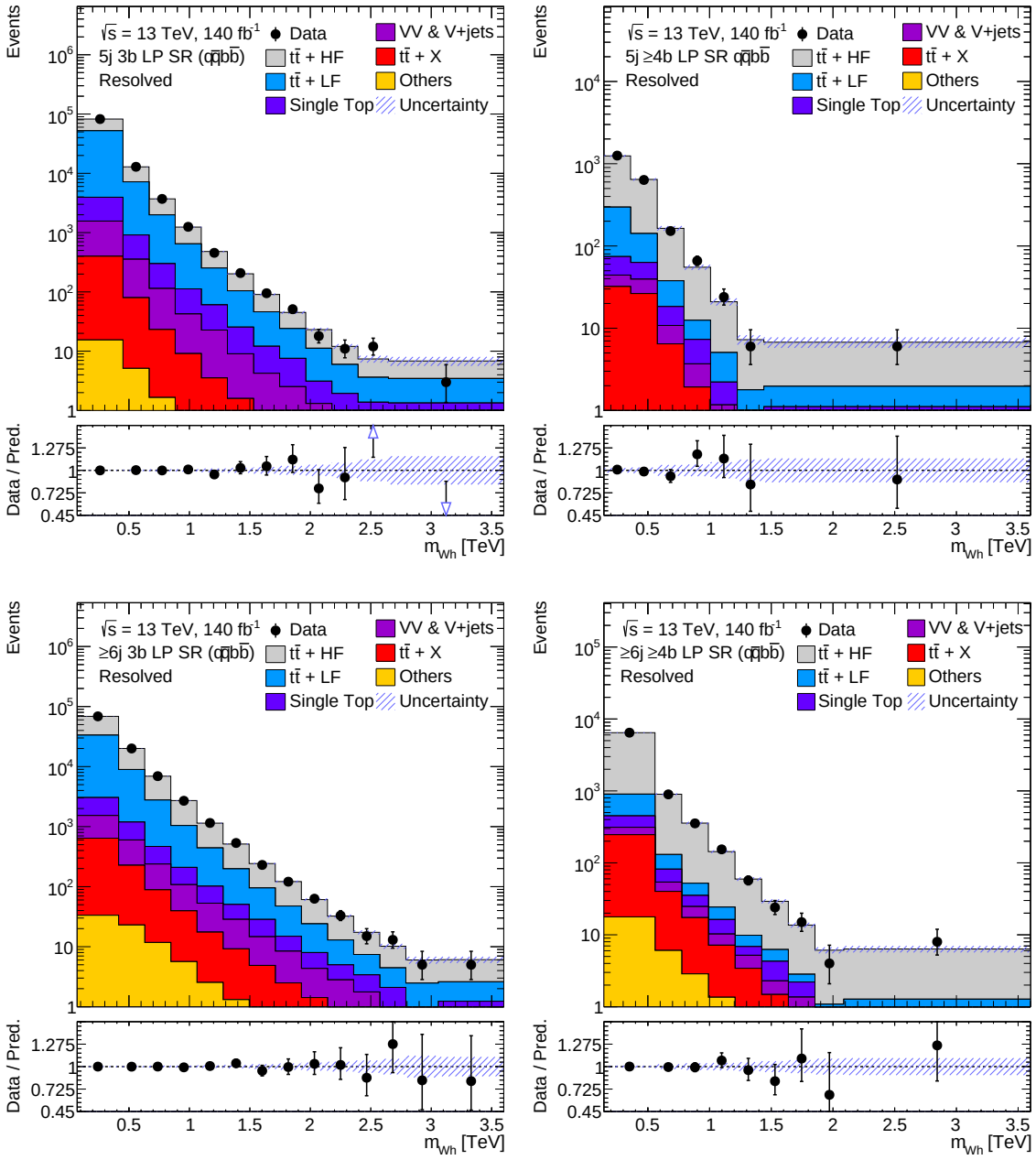


Figure 8.1: Data-to-simulation comparisons for the m_{Wh} observable in the low-purity (LP) signal regions of the resolved $q\bar{q}b\bar{b}$ channel from a fit to data, under a background-only hypothesis, in all analysis regions. The uncertainty band shows the post-fit statistical and systematic uncertainties for all the background processes added in quadrature.

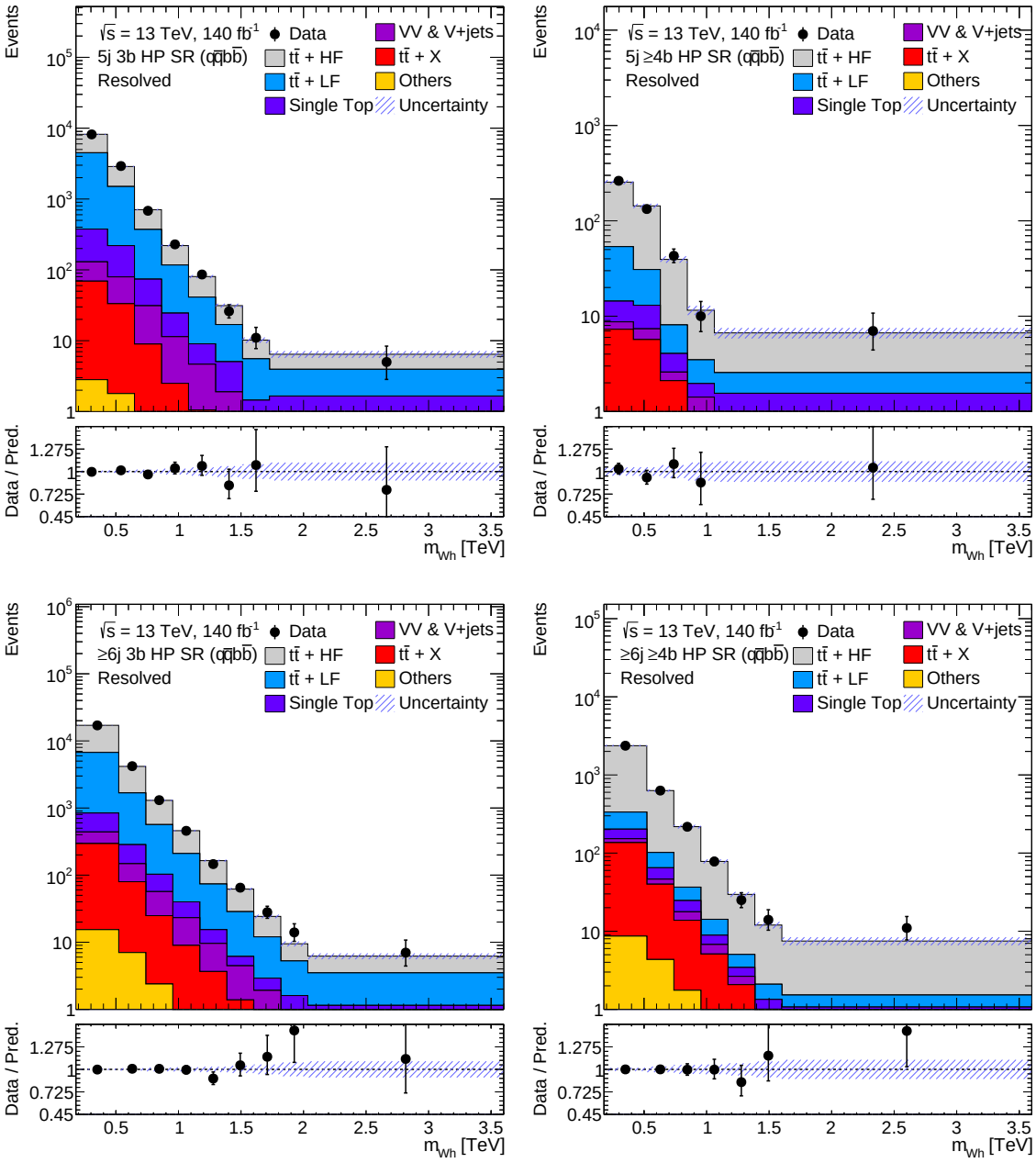


Figure 8.2: Data-to-simulation comparisons for the m_{Wh} observable in the high-purity (HP) signal regions of the resolved $q\bar{q}b\bar{b}$ channel from a fit to data, under a background-only hypothesis, in all analysis regions. The uncertainty band shows the post-fit statistical and systematic uncertainties for all the background processes added in quadrature.

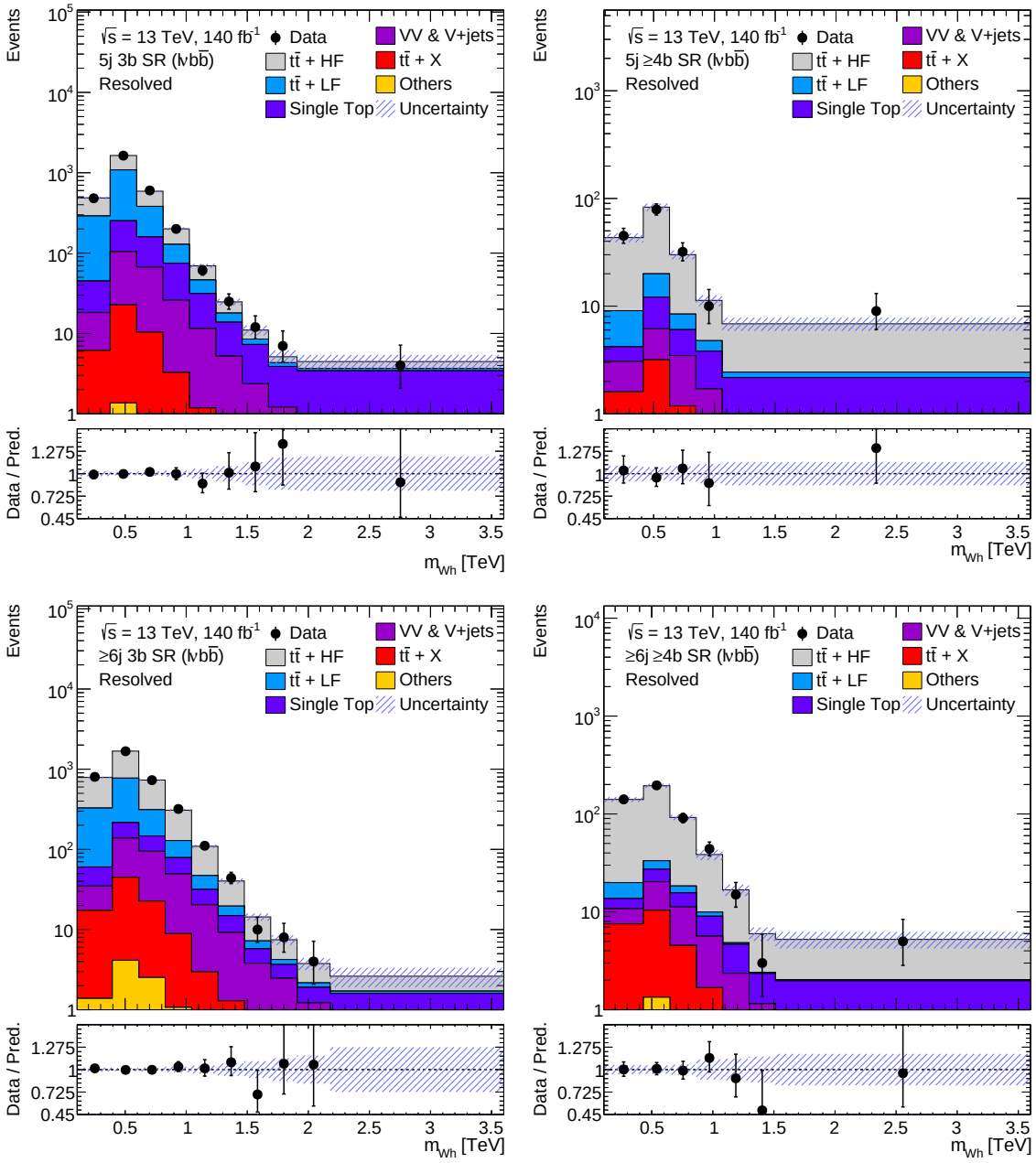
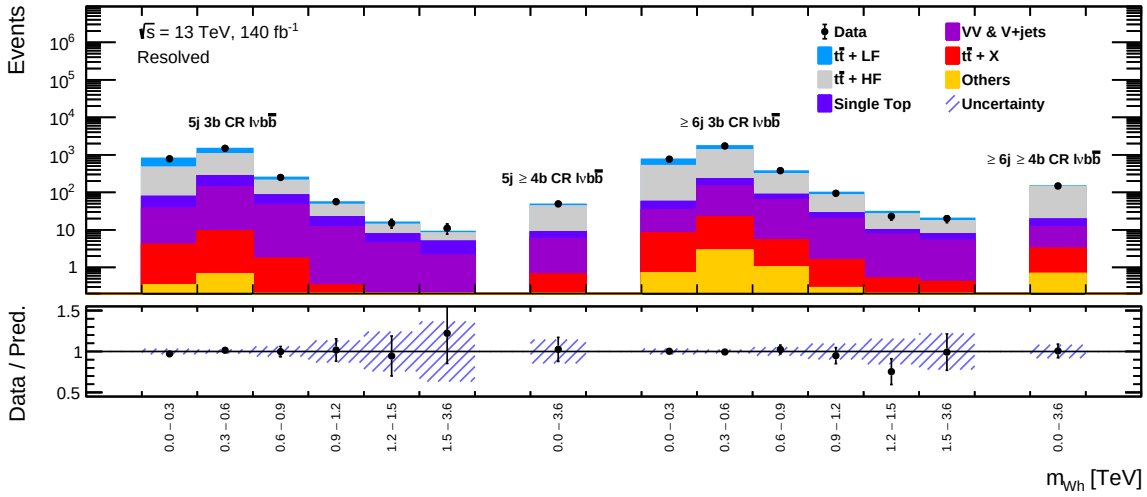
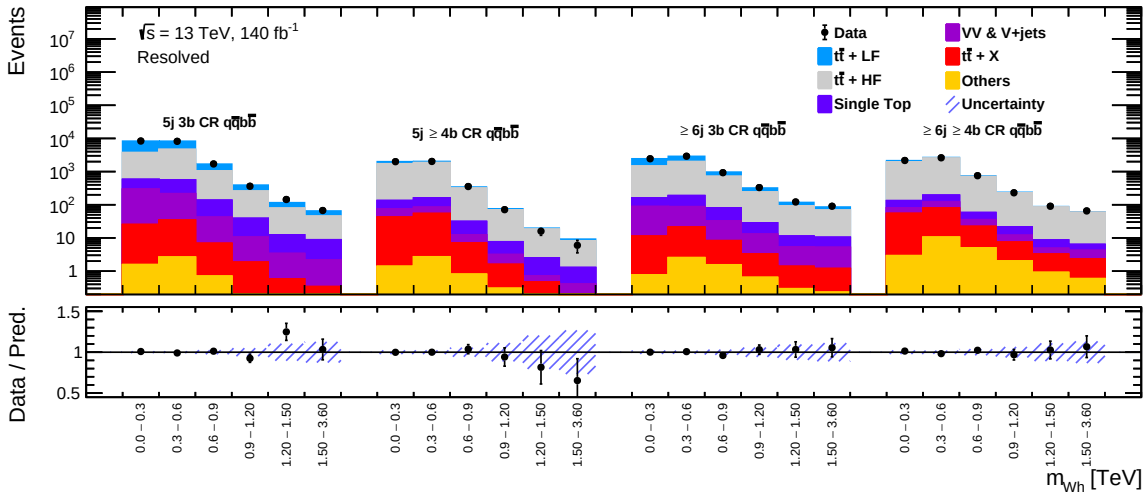


Figure 8.3: Data-to-simulation comparisons for the m_{Wh} observable in the signal regions of the resolved $\ell\nu b\bar{b}$ channel from a fit to data, under a background-only hypothesis, in all analysis regions. The uncertainty band shows the post-fit statistical and systematic uncertainties for all the background processes added in quadrature.



(a)



(b)

Figure 8.4: Data-to-simulation comparisons for the m_{Wh} observable in the control regions of the (a) resolved $\ell\nu b\bar{b}$ and (b) resolved $q\bar{q}b\bar{b}$ channels resulting from a fit to data, under a background-only hypothesis, in all analysis regions.

A breakdown of the relative contributions of the different categories of uncertainties[1] to the uncertainty in the best-fit signal-strength parameter, $\hat{\mu}$, for two signal mass hypotheses corresponding to $m_{H^+} = 400$ GeV and $m_{H^+} = 1.6$ TeV is given in Table 8.2. The contributions are obtained by fixing the relevant nuisance parameters, for a particular category of uncertainty, to their best-fit values obtained from a fit to data under a signal-plus-background hypothesis. The relative impact is determined as:

$$\sqrt{\frac{(\Delta\hat{\mu}_{\text{nom}})^2 - (\Delta\hat{\mu}_{\text{category}})^2}{(\Delta\hat{\mu}_{\text{nom}})^2}} \tag{8.1}$$

where $\Delta\hat{\mu}_{\text{nom}}$ is the uncertainty on $\hat{\mu}$ obtained from a fit to data where all the NPs are allowed to vary and fitted to their best values (nominal scenario), and $\Delta\hat{\mu}_{\text{category}}$ is the uncertainty obtained by fixing the relevant NPs to their best-fit values for a particular category of uncertainty. The contribution from data statistical uncertainties is determined from fits with all NPs fixed to their best-fit values. The dominant uncertainties for low H^+ boson masses (i.e. $m_{H^+} = 400$ GeV) are related to the modelling of $t\bar{t}$ + HF and $t\bar{t}$ + LF backgrounds, while at high H^+ boson masses (i.e. $m_{H^\pm} = 1.6$ TeV) the dominant uncertainties are related to the size of the data sample and non-$t\bar{t}$ modelling uncertainties. This observation confirms the expectation discussed earlier in Section 6.3.2.

[1] The statistical uncertainty of the simulated processes is placed under the category of modelling uncertainties.

Table 8.2: Breakdown of the relative contributions of the different categories of uncertainties to the uncertainty in $\hat{\mu}$ for the two signal mass hypotheses, $m_{H^+} = 400$ GeV and $m_{H^+} = 1.6$ TeV. The contributions are obtained by fixing the relevant NPs to their best-fit values in a fit to data, under a signal-plus-background hypothesis. The sum in quadrature of the individual components differs from the total uncertainty due to correlations between uncertainties in different categories.

$m_{H^+} = 400$ GeV		$m_{H^+} = 1.6$ TeV	
Category	Relative contribution	Category	Relative contribution
Modelling uncertainties			
$t\bar{t}$ + HF modelling	74%	Non-$t\bar{t}$ modelling	37%
$t\bar{t}$ + LF modelling	34%	Simulation statistical uncertainty	26%
Extrapolation	20%	$t\bar{t}$ + LF modelling	25%
Non-$t\bar{t}$ modelling	9%	Extrapolation	15%
Signal modelling	6%	Signal modelling	22%
$t\bar{t}$ + HF normalisation	5%	$t\bar{t}$ + HF modelling	10%
Non-closure	3%	$t\bar{t}$ + HF normalisation	2%
Simulation statistical uncertainty	3%	Non-closure	1%
Experimental uncertainties			
Small-R jets	16%	Small-R jets	27%
b-tagging	13%	b-tagging	10%
Pile-up	5%	Pile-up	7%
Luminosity	3%	Electrons	3%
Electrons	2%	Pile-up	3%
Muons	1%	Muons	2%
Missing transverse momentum	<0.1%	Missing transverse momentum	1%
Total uncertainty on simulated processes	93%	Total uncertainty on simulated processes	64%
Data statistical uncertainty	37%	Data statistical uncertainty	77%

8.2 95% CL Upper Limits

As no significant excess in data events over the SM prediction is found, the results of this search are expressed as 95% CL upper limits on the signal strength ($\sigma(tbH^+) \times BR(H^+ \to Wh(\to b\bar{b}))$). Figure 8.5 shows the 95% CL upper limits on $\sigma(pp \to tbH^+) \times BR(H^+ \to Wh(\to b\bar{b}))$ as a function of m_{H^+} for resolved $H^+ \to Wh(\to b\bar{b})$ decays. The observed limit ranges from 2.77 pb for m_{H^+} = 250 GeV to 3.80 fb for m_{H^+} = 3 TeV. The largest deviation from the SM prediction is found for m_{H^+} = 900 GeV, and corresponds to a significance of 0.9 σ with a p-value of 0.184[2]. Figure 8.6 shows the 95% CL upper limits on $\sigma(tbH^+) \times BR(H^+ \to Wh(\to b\bar{b}))$ as a function of m_{H^+} for the combination of resolved and merged $H^+ \to Wh(\to b\bar{b})$ decays[3]. A simple combination of the two searches is performed and the most stringent expected limit for each m_{H^+} hypothesis is selected. The search for resolved H^+ boson decays is more sensitive for $m_{H^+} \leq 900$ GeV, while the search for merged H^+ boson decays is more sensitive for $m_{H^+} > 900$ GeV. The observed limit ranges from 2.77 pb for m_{H^+} = 250 GeV to 1.18 fb for m_{H^+} = 3 TeV, with the largest deviation from the SM prediction observed at m_{H^+} = 900 GeV from the resolved H^+ boson decays.

An overlay of the 95% CL expected upper limits obtained in the searches for resolved and merged H^+ boson decays is presented in Figure D.5 in the Appendix. For the lowest probed H^+ boson mass of 500 GeV (common to both the searches), the expected limits obtained in the search for resolved H^+ boson decays are a factor of four times better than the ones obtained for merged H^+ boson decays. On the other hand, for the highest probed H^+ boson mass of 3 TeV, the expected limits obtained in the search for merged H^+ boson decays are a factor of four times better than the ones for resolved H^+ boson decays.

8.3 Comparison with the $H^+ \to tb$ search

The 95% CL expected upper limits obtained in the search for resolved $H^+ \to Wh(\to b\bar{b})$ decays are compared with the ones obtained from the search for resolved $H^+ \to tb$ decays. The production mode for both searches is the same: $pp \to \bar{t}bH^+$, and they share the same final state. The $H^+ \to tb$ search was also performed within the ATLAS collaboration [23] using the full Run 2 dataset, and currently shows the best limits from the LHC for this decay mode. This search categorises events based on jet and b-jet multiplicities, in analogy to the $H^+ \to Wh$ search, and uses a neural-network-based discriminant to separate signal from backgrounds and to determine the search sensitivity. Figure 8.7 shows a comparison[4] of the expected upper limits on $\sigma(pp \to tbH^+) \times BR(H^+ \to Y)$ ($Y = Wh, tb$) for the two H^+ boson decay modes. Overall, the limits from the $H^+ \to Wh$ search are up to nine times better than those from the $H^+ \to tb$ search. For low H^+ boson masses (m_{H^+} = 250 GeV and m_{H^+} = 300 GeV), the $H^+ \to tb$ search shows better limits. This is potentially due to a huge impact of the $t\bar{t}$ + jets background at low H^+ boson masses in the $H^+ \to Wh$ search, owing to the shape of the m_{Wh} observable in various analysis regions.

[2] The quoted p-value and significance are *local* in nature which corresponds to the chance of a statistical fluctuation being at least as large as the observed one at its specific location in contrast to a *global* p-value or significance which corresponds to a fluctuation anywhere in the analysis.

[3] The m_{Wh} distributions in the signal regions used in the search for merged H^+ boson decays from a background-only fit to data in all analysis regions are shown in Figures D.2 to D.4 in the Appendix.

[4] This comparison is performed with a caveat that the branching ratios for the two H^+ boson decays from theoretical predictions are not taken into account.

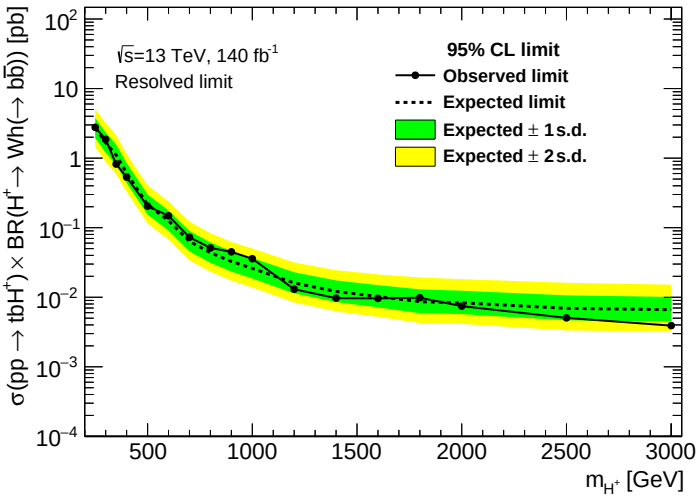


Figure 8.5: 95% CL upper limits on $\sigma(tbH^+) \times BR(H^+ \to Wh(\to b\bar{b}))$ as a function of m_{H^+} for resolved $H^+ \to Wh(\to b\bar{b})$ decays.

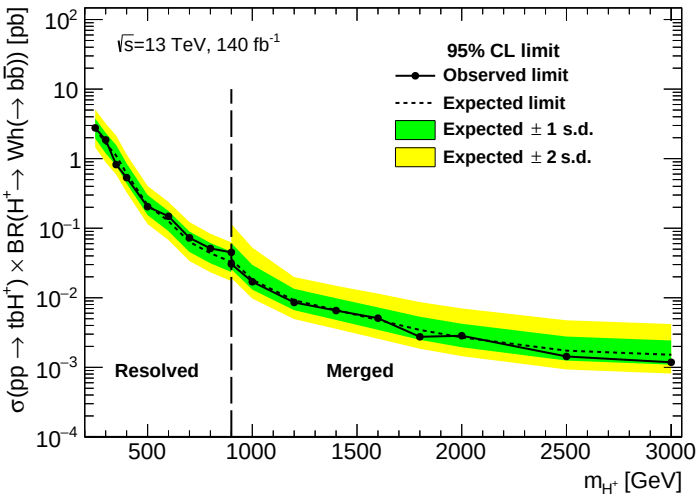


Figure 8.6: 95% CL upper limits on $\sigma(tbH^+) \times BR(H^+ \to Wh(\to b\bar{b}))$ as a function of m_{H^+} for the combination of resolved and merged $H^+ \to Wh(\to b\bar{b})$ decays. The dashed line separating the "resolved" and "merged" searches, at $m_{H^+} = 900$ GeV, is designated based on the choice of better limit from the two searches.

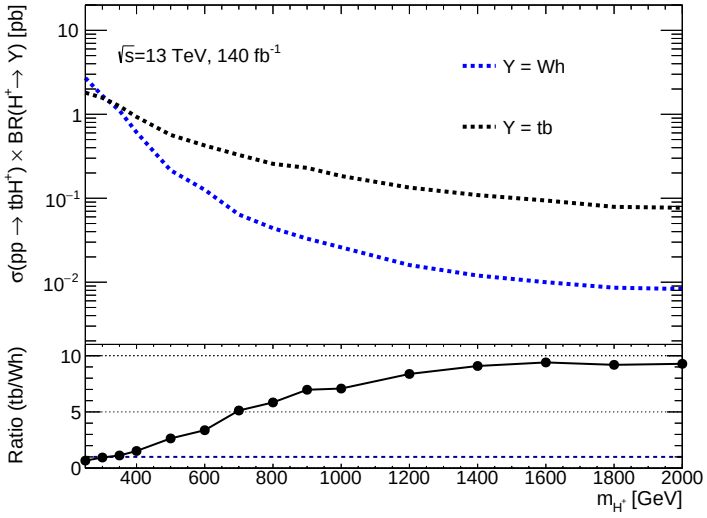


Figure 8.7: Comparsion of 95% CL expected upper limits on $\sigma(tbH^+) \times BR(H^+ \to Y)$ as a function of m_{H^+} for resolved $H^+ \to Wh$ and $H^+ \to tb$ decays, where Y refers to the Wh or tb final state.

CHAPTER 9

Summary and Outlook

A search for a heavy charged Higgs boson produced in association with a top- and a bottom-quark, and decaying into a W boson and a 125 GeV Higgs boson was performed for the first time at the LHC in the H^+ boson mass range from 250 GeV to 3 TeV. Two different scenarios were studied to ensure a high sensitivity for both low- and high-mass resonances. The author's main contribution focused on the search for resolved charged Higgs boson decays which was seen to be more sensitive for $m_{H^+} \leq 900$ GeV. This search probed final states with exactly one charged electron or muon, missing transverse momentum, and at least five small-R jets. Events were classified based on the kinematic requirements as well as the multiplicity of b-tagged jets per event. Boosted decision trees were used to reconstruct the four-momentum of the charged Higgs boson candidate, and hence its invariant mass. The search for merged charged Higgs boson decays was seen to be more sensitive for $m_{H^+} > 900$ GeV, and probed final states with exactly one charged electron or muon, missing transverse momentum, and at least one large-R jet. Recently developed $X \to b\bar{b}$ tagging techniques were used to reconstruct the boosted Higgs bosons, and neural networks were employed to separate between signal and background processes.

The reconstructed invariant mass of the charged Higgs boson was scrutinised for an excess in 140 fb^{-1} of 13 TeV proton-proton collision data collected from 2015 to 2018 with the ATLAS detector at the LHC. No significant excess of events above the SM prediction was observed, and 95% CL upper limits on $\sigma(pp \to tbH^+) \times BR(H^+ \to Wh(\to b\bar{b}))$ were set ranging from 2.77 pb for $m_{H^+} =$ 250 GeV to 1.18 fb for $m_{H^+} = 3$ TeV.

A potential area to explore for future iterations of this analysis can be towards training a multi-class classifier for distingishing between signals from the $\ell\nu b\bar{b}$ and $q\bar{q}b\bar{b}$ channels and backgrounds simultaneously in the search for resolved charged Higgs boson decays. This classifier can help reduce the search complexity. Other potential improvements are an optimisation of the search for resolved and merged charged Higgs boson decays in a statistically independent way. For the presented search, no overlap removal of events between the two search scenarios is taken into account. The fraction of events within the signal regions of the resolved charged Higgs boson decays passing the selection requirements for merged charged Higgs boson decays can range upto 50% in the studied charged Higgs boson mass range, as reported in [240]. The optimisation of signal regions after removing the overlapping events in both search scenarios would help reduce the background contamination and improve the signal efficiency, thus enhancing the search sensitivity. A dedicated comparison of the upper limits on $\sigma(pp \to tbH^+) \times BR(H^+ \to Wh(\to b\bar{b}))$ to the predictions of benchmark points

of the theoretical models presented in Chapter 2 can also be considered as a future extension of this project.

The quest for an extended Higgs sector opens up a wide scope for testing the Standard Model and looking for new physics. A continuous evolution of charged Higgs boson searches, within the ATLAS and CMS collaborations, involves the development of new and improved event reconstruction techniques to handle the challenges posed by different final states. Moreover, an increased amount of data from the future LHC runs might bring us closer to observing new particles from an extended Higgs sector. Such an observation could help answer some of the open questions of the Standard Model like the baryonic asymmetry of the universe, unravel the mysterious hierarchy problem, or provide a dark matter candidate.

APPENDIX A

Additional plots supporting the analysis strategy used in the search for resolved charged Higgs boson decays

A.1 Analytical fits to the correction factor used in $H_{\mathrm{T}}^{\mathrm{all}}$ reweighting

This section presents the fits performed for modelling the correction factor, C($H_{\mathrm{T}}^{\mathrm{all}}$), as described in Section 4.4.1, where the fits using an exponential-plus-sigmoid functional form were shown in Figure 4.6. The fits for the hyperbola-plus-sigmoid functional form fitted to the data points, describing C($H_{\mathrm{T}}^{\mathrm{all}}$), in bins of $H_{\mathrm{T}}^{\mathrm{all}}$ in different jet multiplicity regions are shown in Figure A.1, and the corresponding fits for the combination of second-order polynomial and a first-order exponentiated polynomial are shown in Figure A.2.

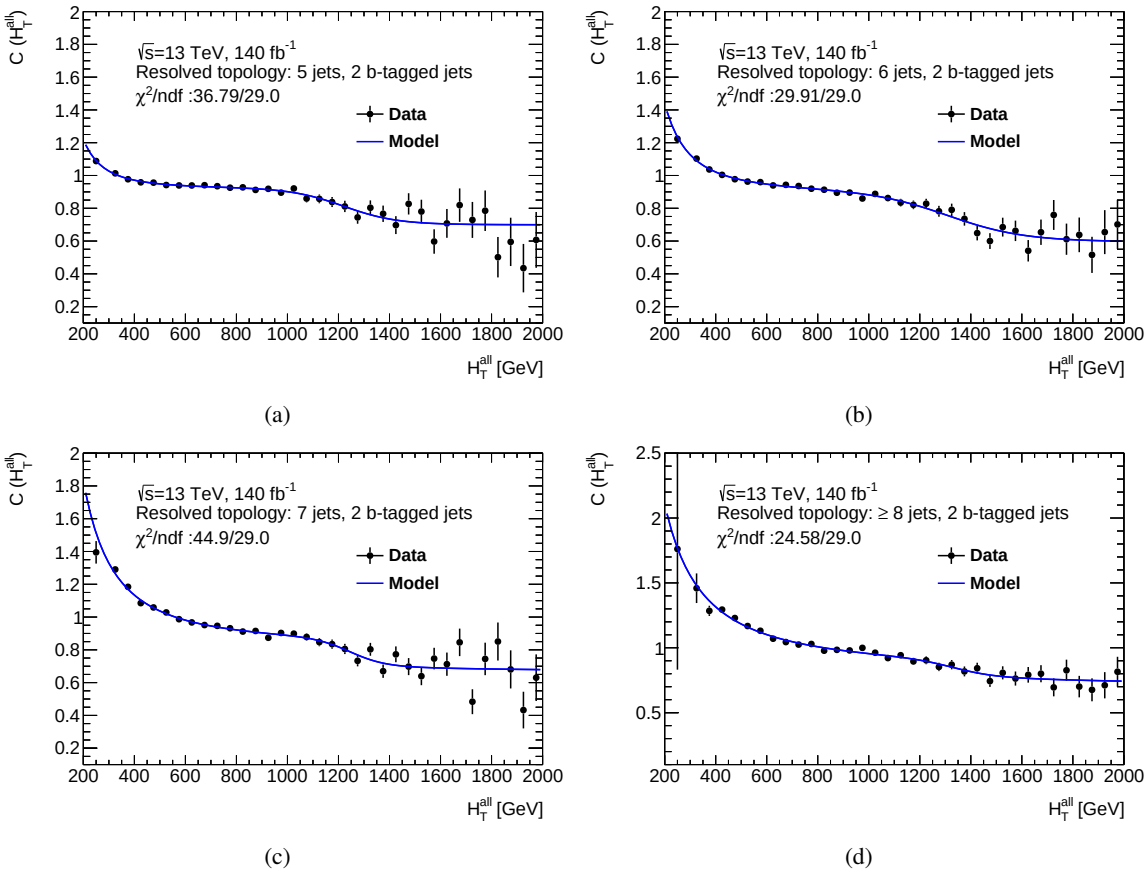


Figure A.1: (a)-(d) show the analytical fits to the correction factor, (C(H_T^{all})), in bins of H_T^{all} modelled by a hyperbola-plus-sigmoid functional form in different jet multiplicity regions (a: 5 jets, b: 6 jets, c: 7 jets, d: ≥ 8 jets) for events containing two *b*-tagged jets.

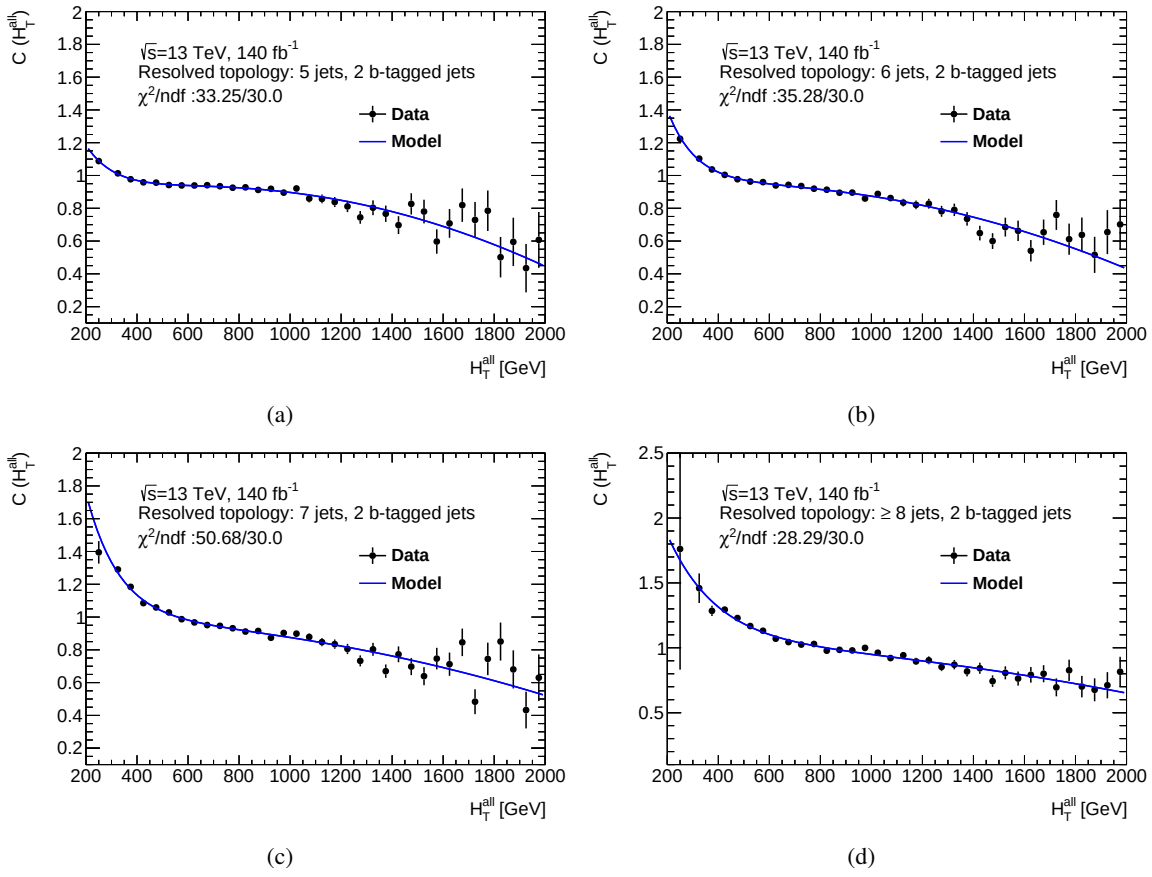


Figure A.2: (a)-(d) show the analytical fits to the correction factor, (C($H_{\mathrm{T}}^{\mathrm{all}}$)), in bins of $H_{\mathrm{T}}^{\mathrm{all}}$ modelled by a combination of a second-order polynomial and a first-order exponentiated poynominal in different jet multiplicity regions (a: 5 jets, b: 6 jets, c: 7 jets, d: $\geq$ 8 jets) for events containing two b-tagged jets.

A.2 Additional distributions on $t\bar{t}$ + jets reweighting validation

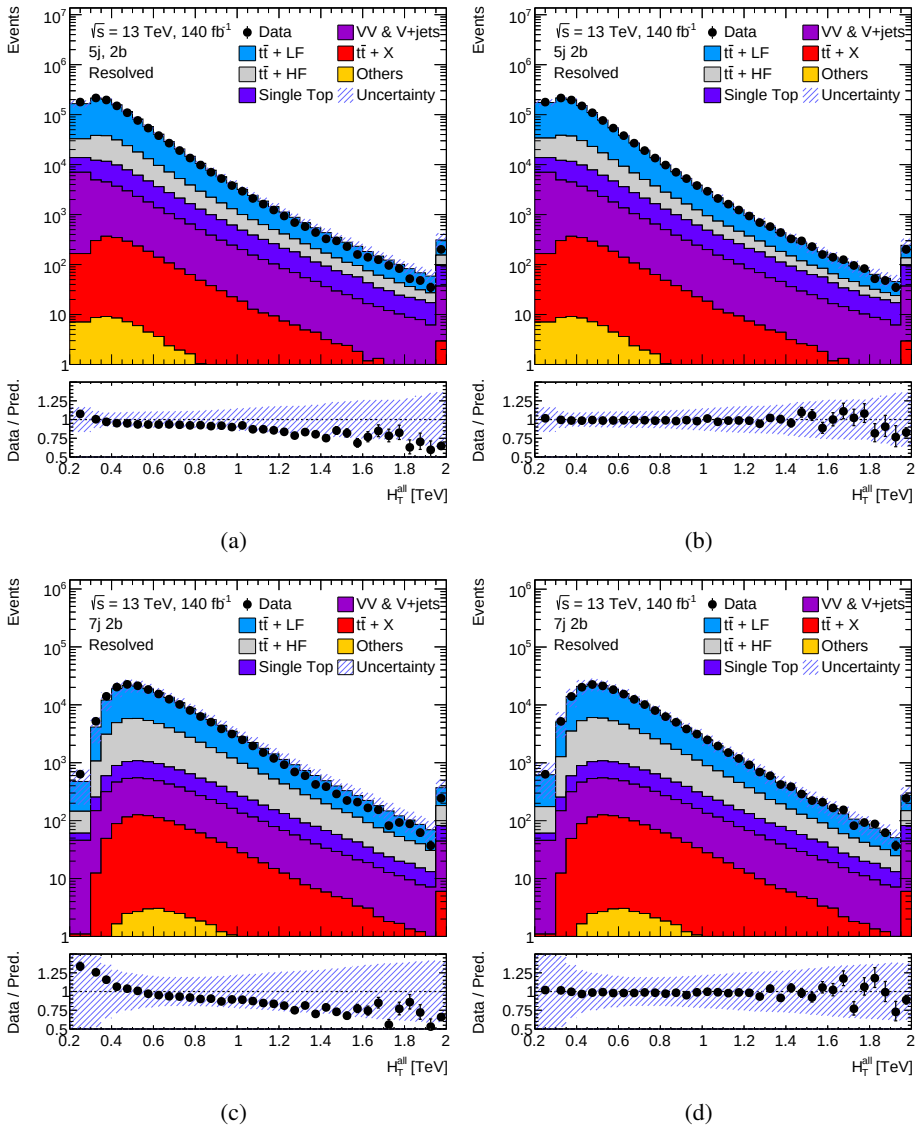
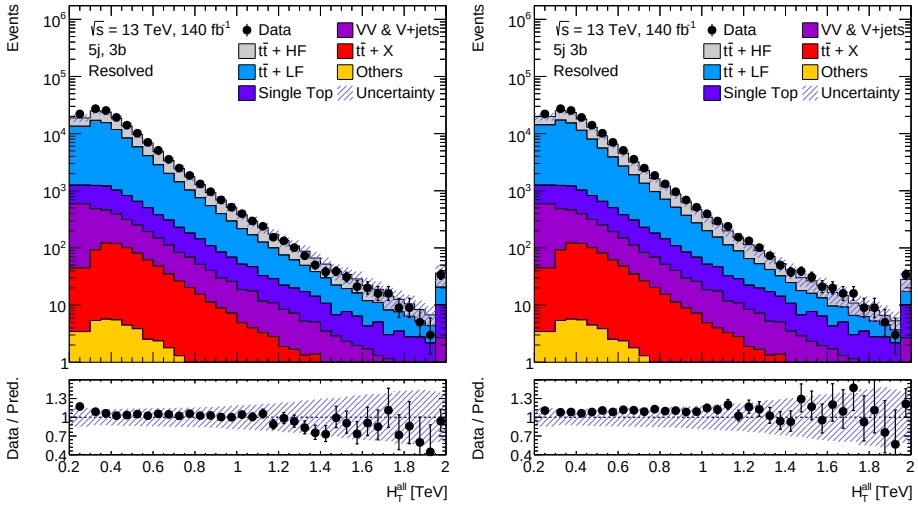
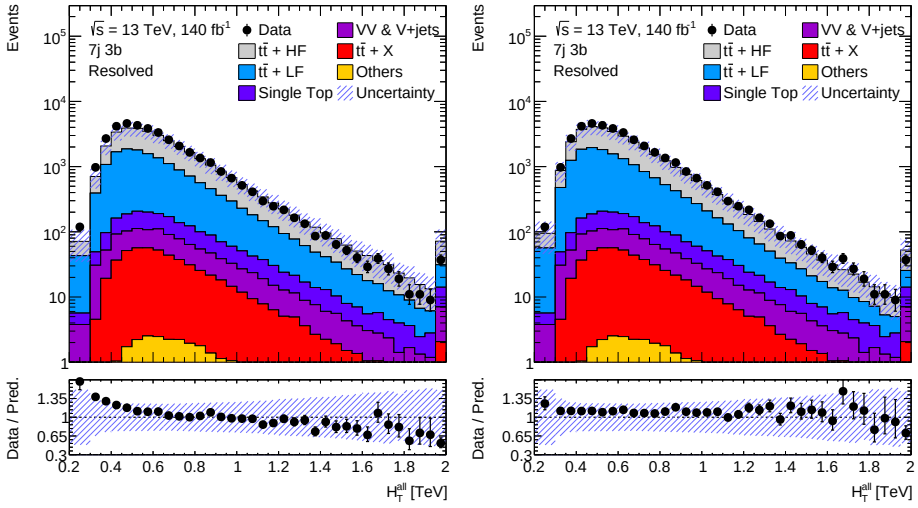


Figure A.3: $H_{\mathrm{T}}^{\mathrm{all}}$ distributions, for n_{jets} = 5 and 7, before (left) and after (right) the reweighting is applied in events containing two b-tagged jets.



(a) (b)



(c) (d)

Figure A.4: $H_{\mathrm{T}}^{\mathrm{all}}$ distributions, for n_{jets} = 5 and 7, before (left) and after (right) the reweighting is applied in events containing three b-tagged jets.

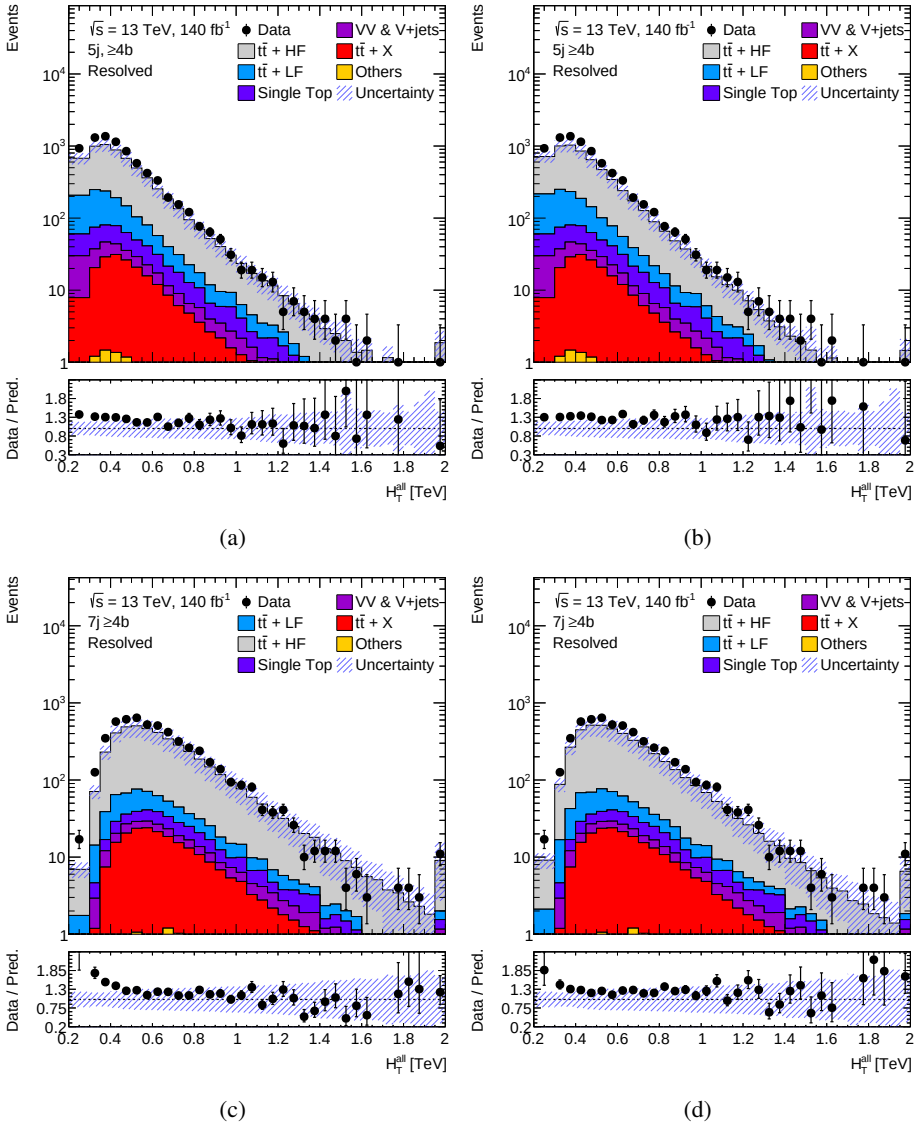


Figure A.5: $H_{\mathrm{T}}^{\mathrm{all}}$ distributions, for n_{jets} = 5 and 7, before (left) and after (right) the reweighting is applied in events containing at least four b-tagged jets.

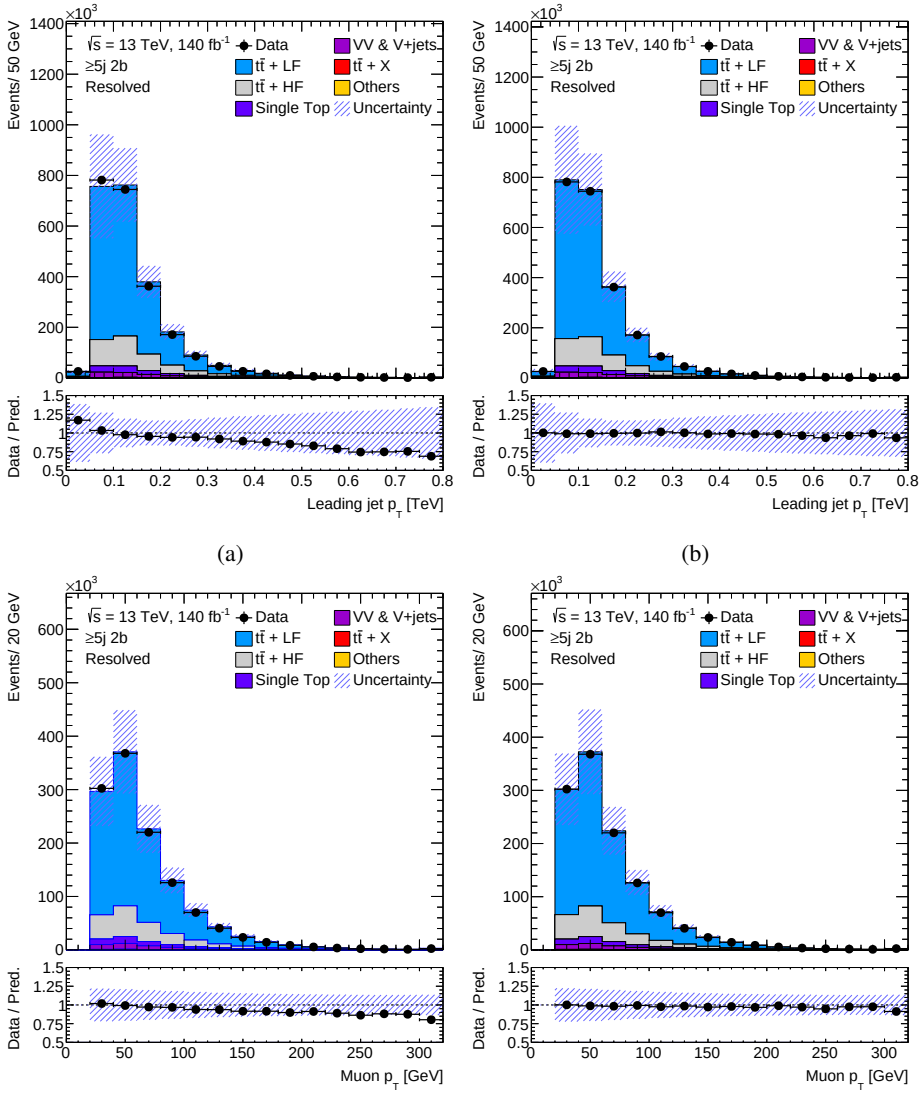


Figure A.6: The distributions for leading jet p_{T} and muon p_{T}, before (left) and after (right) the reweighting is applied in events containing at least five jets and two b-tagged jets.

A.3 Data-to-simulation comparison for $m_{\text{top}}^{\text{lep}}$ observable

Figure A.7 shows the data-to-simulation comparison for the $m_{\text{top}}^{\text{lep}}$ in pre-selection. The $t\bar{t}$ + LF and $t\bar{t}$ + HF backgrounds are corrected via the $H_{\text{T}}^{\text{all}}$ reweighting procedure (see Section 4.4.1). A good agreement between data and simulated backgrounds is observed within the uncertainties.

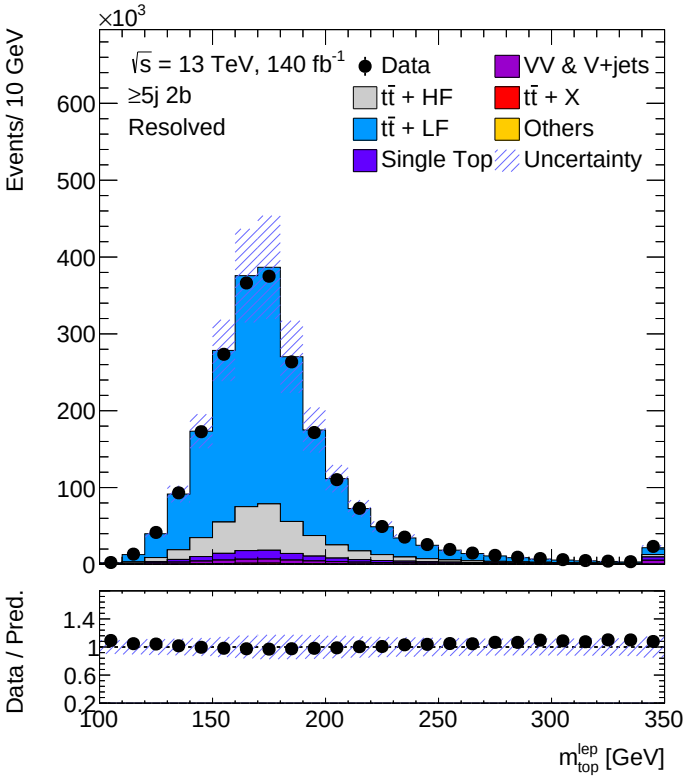


Figure A.7: Data-to-simulation comparison for the $m_{\text{top}}^{\text{lep}}$. The background uncertainty band shows the statistical and systematic uncertainties added in quadrature.

A.4 Study on jet bound used for constructing the input dataset for reconstruction BDT training

Figure A.8 shows the matching efficiency (see Equation 5.1) from the reconstruction BDT training, in the $q\bar{q}b\bar{b}$ channel, on input datsaset constructed using different number of jets arranged in decreasing order of their p_{T}. It can observed from this figure that using up to nine leading jets to construct the input dataset is suitable for the training.

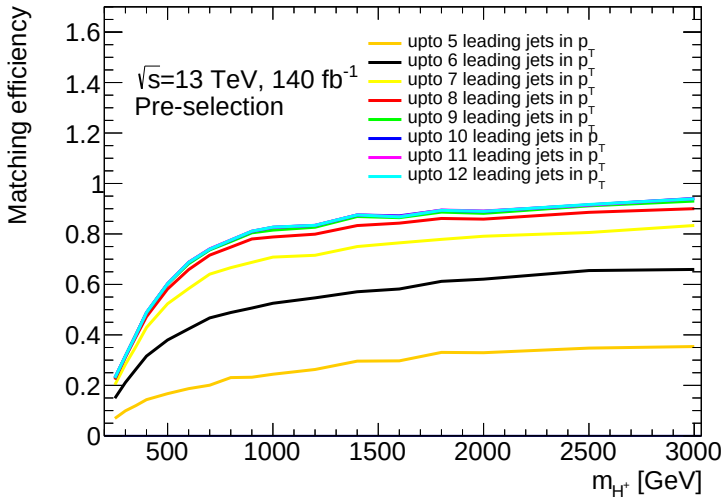
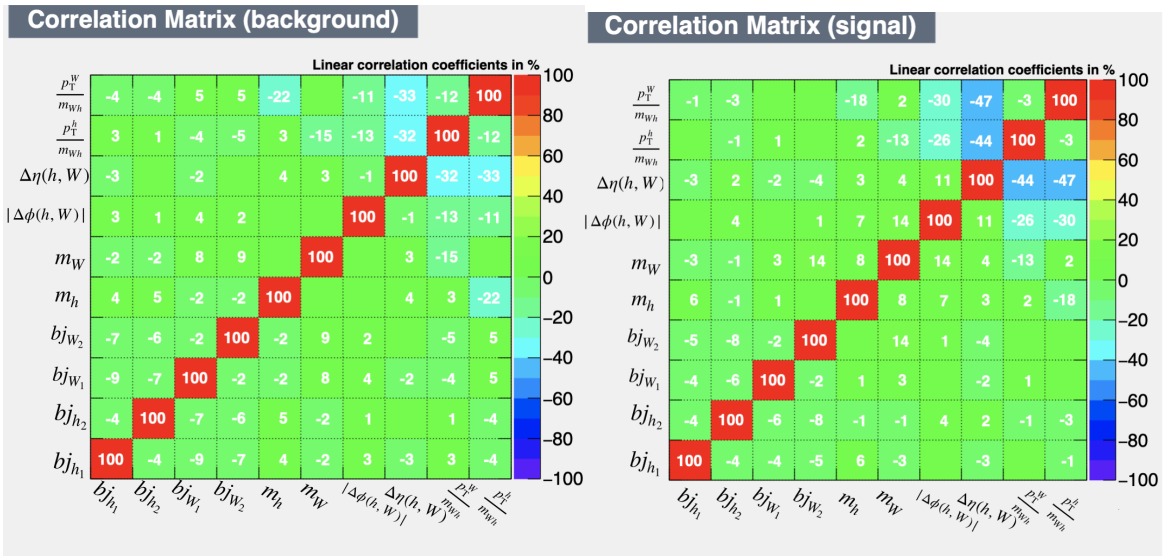


Figure A.8: The matching efficiency resulting from the $q\bar{q}b\bar{b}$ event reconstruction BDT training on the input datsaset constructed using different number of jets sorted in p_{T}.

A.5 Linear correlation coefficients between reconstruction BDT input variables in the $q\bar{q}b\bar{b}$ channel



(a) (b)

Figure A.9: Linear correlation coefficients between the reconstruction BDT input variables for the (a) signal and (b) combinatorial background in the $q\bar{q}b\bar{b}$ channel.

A.6 Data-to-simulation comparisons for reconstruction BDT input variables

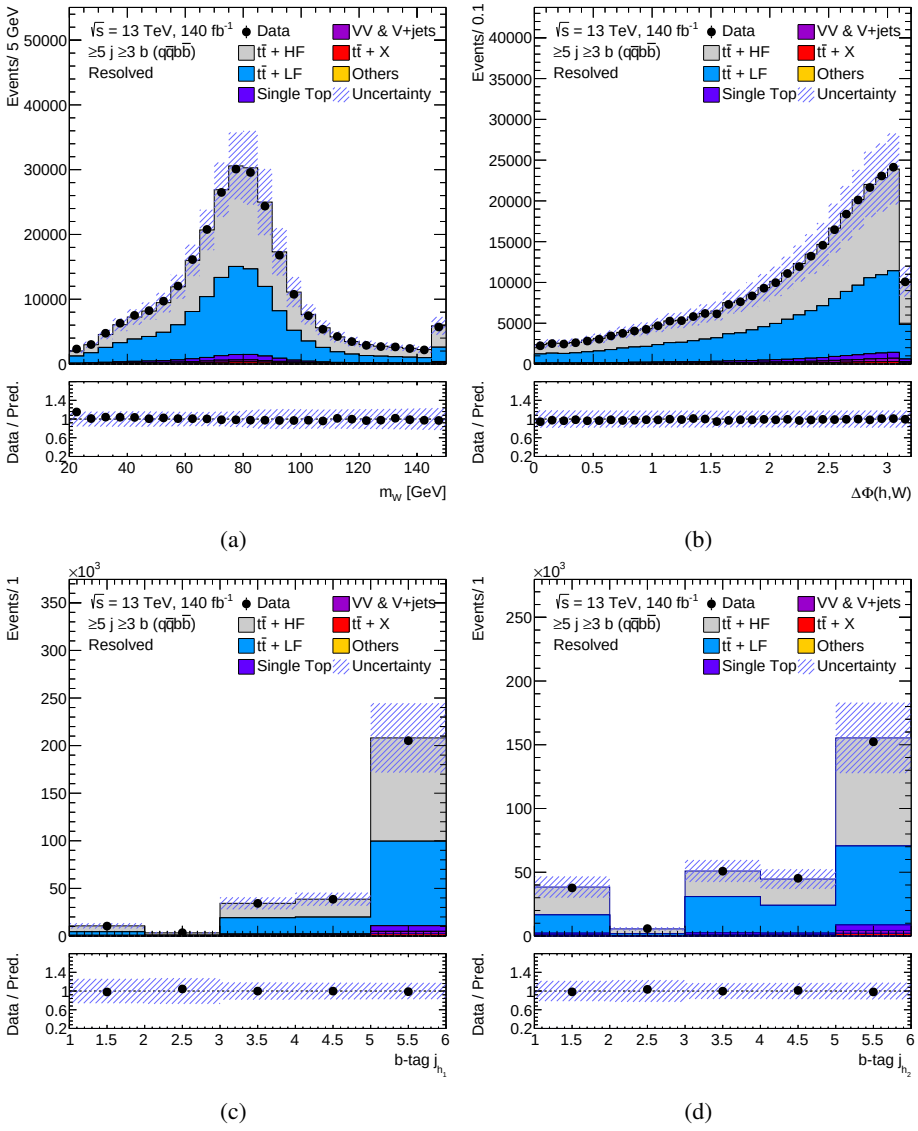
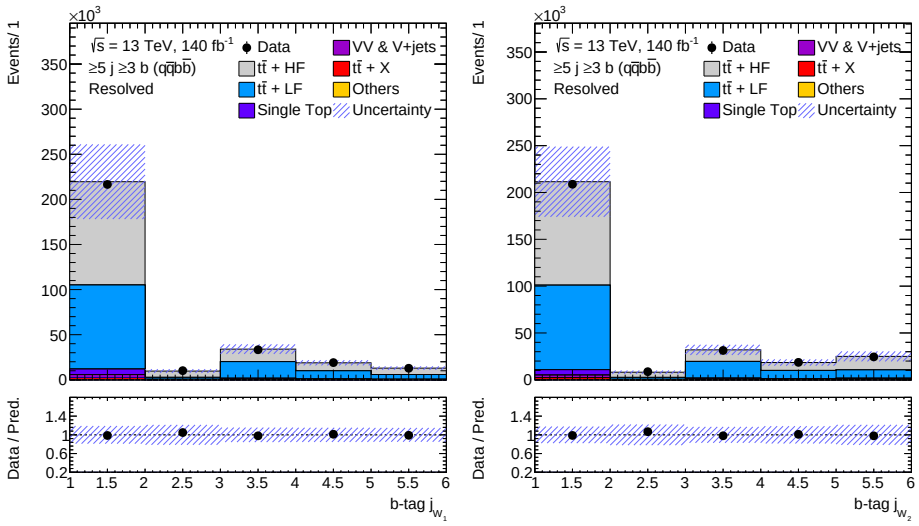


Figure A.10: Data-to-simulation comparisons for the BDT input variables in events containing at least three b-tagged jets in the $q\bar{q}b\bar{b}$ channel. (a): m_{W}, (b): $|\Delta\Phi(h,W)|$, (c): b-tag j_{h_1}, and (d): b-tag j_{h_2}.



(a) (b)

Figure A.11: Data-to-simulation comparison for the BDT input variables in events containing at least three b-tagged jets in the $q\bar{q}b\bar{b}$ channel. (a): b-tag j_{W_1}, (b): b-tag j_{W_2}.

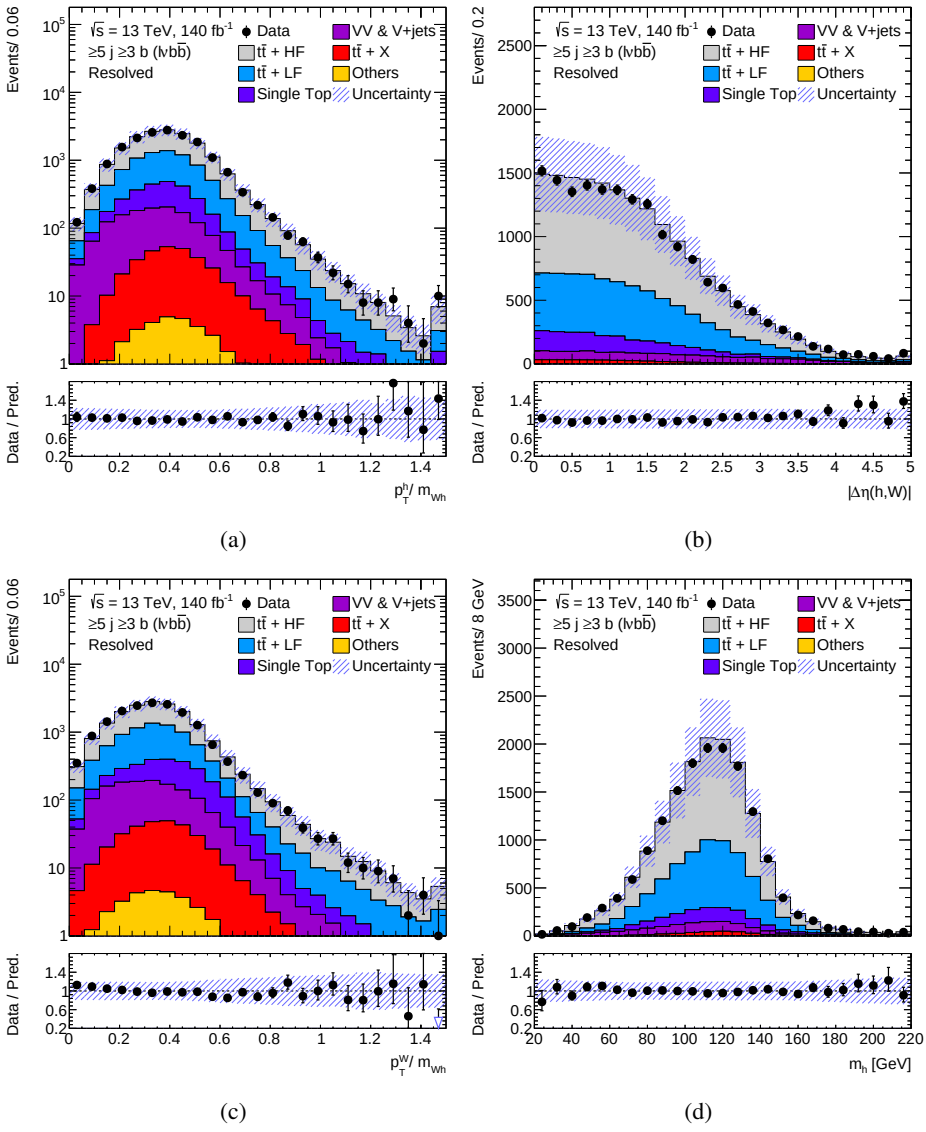
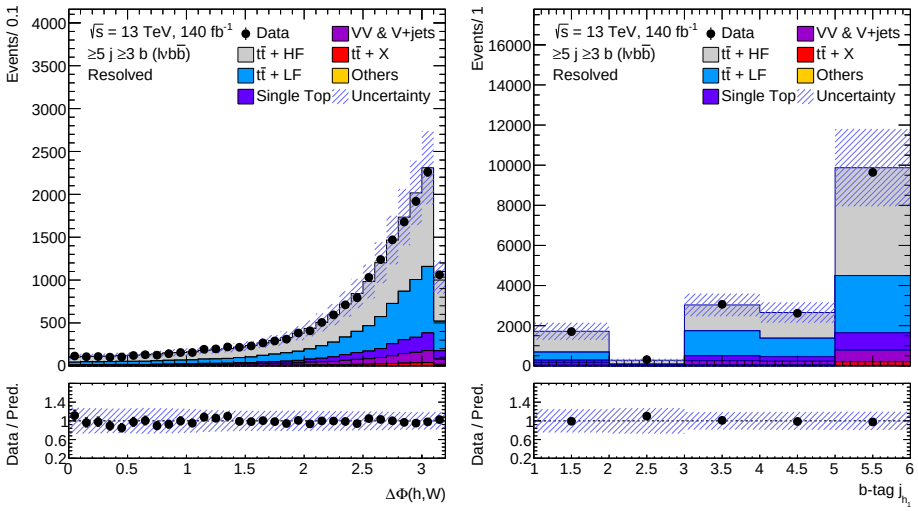
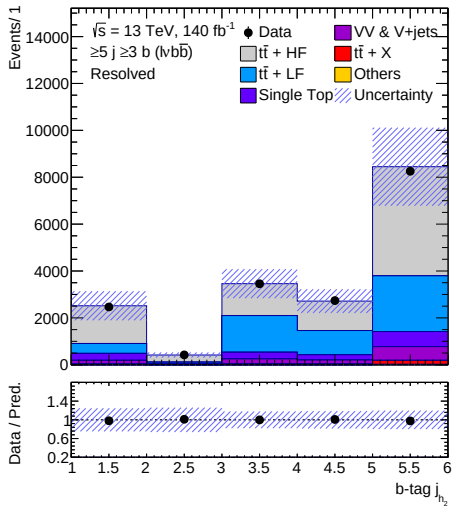


Figure A.12: Data-to-simulation comparisons for the BDT input variables in events containing at least three b-tagged jets in the $\ell\nu b\bar{b}$ channel. (a): $p_{\mathrm{T}}^{\mathrm{h}}/m_{\mathrm{Wh}}$, (b): $|\Delta\eta(h,W)|$, (c): $p_{\mathrm{T}}^{\mathrm{W}}/m_{\mathrm{Wh}}$, and (d): m_{h}.



(a) (b)



(c)

Figure A.13: Data-to-simulation comparisons for the BDT input variables in events containing at least three b-tagged jets in the $\ell\nu b\bar{b}$ channel. (a): $|\Delta\Phi(h,W)|$, (b): b-tag j_{h_1}, and (c): b-tag j_{h_2}.

APPENDIX B

Additional studies supporting the set-up of the statistical model used in the search for resolved charged Higgs boson decays

B.1 Relative acceptance uncertainties

b-tag multiplicity region	Uncertainty
$3b$ $\ell\nu b\bar{b}$ CR	0.178
$\geq 4b$ $\ell\nu b\bar{b}$ CR	0.223
$3b$ $\ell\nu b\bar{b}$ SR	0.166
$\geq 4b$ $\ell\nu b\bar{b}$ SR	0.294
3b $q\bar{q}b\bar{b}$ CR	0.195
$\geq 4b$ $q\bar{q}b\bar{b}$ CR	0.235
3b $q\bar{q}b\bar{b}$ SR	0.343
$\geq 4b$ $q\bar{q}b\bar{b}$ SR	0.227
3b $q\bar{q}b\bar{b}$ Low-purity SR	0.215
$\geq 4b$ $q\bar{q}b\bar{b}$ Low-purity SR	0.222

Table B.1: Pre-fit size of the cross-region extrapolation uncertainty - $5j \leftrightarrow \geq 6j$ - applicable for relevant analysis regions. The jet multiplicity region giving a lower yield of the $t\bar{t}$ + HF background is assigned the corresponding extrapolation uncertainty.

Region	Uncertainty
$5j\ 3b\ q\bar{q}b\bar{b}$	0.165
$\geq 6j\ 3b\ q\bar{q}b\bar{b}$	0.093
$5j\ \geq 4b\ q\bar{q}b\bar{b}$	0.106
$\geq 6j\ \geq 4b\ q\bar{q}b\bar{b}$	0.068

Table B.2: Pre-fit size of the cross-region extrapolation uncertainty - High-purity SRs ↔ Low-purity SRs - in the $q\bar{q}b\bar{b}$ channel applicable for relevant analysis regions. The region giving a lower yield of the $t\bar{t}$ + HF background is assigned the corresponding extrapolation uncertainty.

Region	Uncertainty
$5j\ 3b\ q\bar{q}b\bar{b}$	0.124
$\geq 6j\ 3b\ q\bar{q}b\bar{b}$	0.074
$5j\ \geq 4b\ q\bar{q}b\bar{b}$	0.056
$\geq 6j\ \geq 4b\ q\bar{q}b\bar{b}$	0.052

Table B.3: Pre-fit size of the cross-region extrapolation uncertainty - Low-purity SRs ↔ CRs - in the $q\bar{q}b\bar{b}$ channel applicable for relevant analysis regions. The region giving a lower yield of the $t\bar{t}$ + HF background is assigned the corresponding extrapolation uncertainty.

Region	Uncertainty
$5j\ 3b$ CR	0.053
$\geq 6j\ 3b$ CR	0.082
$5j\ \geq 4b$ CR	0.164
$\geq 6j\ \geq 4b$ CR	0.111
$5j\ 3b$ (High-purity) SR	0.15
$\geq 6j\ 3b$ (High-purity) SR	0.051
$5j\ \geq 4b$ (High-purity) SR	0.165
$\geq 6j\ \geq 4b$ (High-purity) SR	0.075

Table B.4: Pre-fit size of the cross-region extrapolation uncertainty - $\ell\nu b\bar{b} \leftrightarrow q\bar{q}b\bar{b}$ channel - applicable for relevant analysis regions. The region of an analysis channel giving a lower yield of the $t\bar{t}$ + HF background is assigned the corresponding extrapolation uncertainty.

Region	Uncertainty
$5j\ 3b\ \ell\nu b\bar{b}$	0.123
$\geq 6j\ 3b\ \ell\nu b\bar{b}$	0.114
$5j\ \geq 4b\ \ell\nu b\bar{b}$	0.264
$\geq 6j\ \geq 4b\ \ell\nu b\bar{b}$	0.115
$5j\ 3b\ q\bar{q}b\bar{b}$	0.333
$\geq 6j\ 3b\ q\bar{q}b\bar{b}$	0.158
$5j\ \geq 4b\ q\bar{q}b\bar{b}$	0.165
$\geq 6j\ \geq 4b\ q\bar{q}b\bar{b}$	0.125

Table B.5: Pre-fit size of the cross-region extrapolation uncertainty - (High-purity) SRs ↔ CRs - applicable for relevant analysis regions. The region giving a lower yield of the $t\bar{t}$ + HF background is assigned the corresponding extrapolation uncertainty.

Region	Uncertainty
$\ell\nu b\bar{b}$ 5j 3b CR	0.201
$\ell\nu b\bar{b}$ ≥6j 3b CR	0.272
$\ell\nu b\bar{b}$ 5j ≥4b CR	0.211
$\ell\nu b\bar{b}$ ≥6j ≥4b CR	0.310
$q\bar{q}b\bar{b}$ 5j 3b CR	0.262
$q\bar{q}b\bar{b}$ ≥6j 3b CR	0.329
$q\bar{q}b\bar{b}$ 5j ≥4b CR	0.148
$q\bar{q}b\bar{b}$ ≥6j ≥4b CR	0.129
$q\bar{q}b\bar{b}$ 5j 3b Low-pur. SR	0.288
$q\bar{q}b\bar{b}$ ≥6j 3b Low-pur. SR	0.266
$q\bar{q}b\bar{b}$ 5j ≥4b Low-pur. SR	0.199
$q\bar{q}b\bar{b}$ ≥6j ≥4b Low-pur. SR	0.126
$\ell\nu b\bar{b}$ 5j 3b SR	0.251
$\ell\nu b\bar{b}$ ≥6j 3b SR	0.176
$\ell\nu b\bar{b}$ 5j ≥4b SR	0.228
$\ell\nu b\bar{b}$ ≥6j ≥4b SR	0.109
$q\bar{q}b\bar{b}$ 5j 3b High-pur. SR	0.312
$q\bar{q}b\bar{b}$ ≥6j 3b High-pur. SR	0.233
$q\bar{q}b\bar{b}$ 5j ≥4b High-pur. SR	0.301
$q\bar{q}b\bar{b}$ ≥6j ≥4b High-pur. SR	0.137

Table B.6: Pre-fit size of the heavy-flavour composition ($t\bar{t}$ + ≥ 1b ↔ $t\bar{t}$ + ≥ 1c) uncertainty for the $t\bar{t}$ + HF background in various analysis regions. The component giving a lower yield is assigned the corresponding extrapolation uncertainty, and the other component is varied in such a way that the overall normalisation of the $t\bar{t}$ + HF background remains constant.

B.2 Decorrelation of constrained nuisance parameters

On performing a signal-plus-background fit using Asimov data in all analysis regions, 7 NPs were seen to be over-constrained (NP pull uncertainty < 0.5) indicating an over-estimation of the associated uncertainties. These uncertainties are the PS uncertainties for both the $t\bar{t}$ + LF and $t\bar{t}$ + HF backgrounds, the ME-PS matching uncertainties for the both $t\bar{t}$ + LF and $t\bar{t}$ + HF backgrounds, the 4FS vs. 5FS uncertainty for the $t\bar{t}$ + HF background, the FSR uncertainty for the $t\bar{t}$ + LF background, and the Wt DS uncertainty for the single-top background. As discussed in Section 6.2.2, the pre-fit impact of the modelling uncertainties of the $t\bar{t}$ + jets background on the m_{Wh} distribution in some analysis regions ranges up to 37% and up to 63% for Wt DS uncertainty. These large uncertainties in various analysis regions lead to over-constrained NPs from a signal-plus-background fit. A step-wise decorrelation of these uncertainties among analysis regions is performed in order to avoid an over-estimation of these uncertainties. A decorrelation of NPs is performed up to a stage where they are seen to be well constrained in a fit or their behaviour is understood.

The first step of the decorrelation procedure is to decorrelate the NPs into $q\bar{q}b\bar{b}$ and $\ell\nu b\bar{b}$ analysis channels (Step 1). A decorrelation of NPs into the two analysis channels is motivated from the varying modelling impact of the event generators in the two channels due to the selection requirement on the leptonic top-quark candidate mass. Figure B.1 shows the comparison of the pull distributions from a signal-plus-background fit using Asimov data for Step 1 (red) and Step 0 (black). Step 0 is the step prior to performing any decorrelation of NPs in various analysis regions.

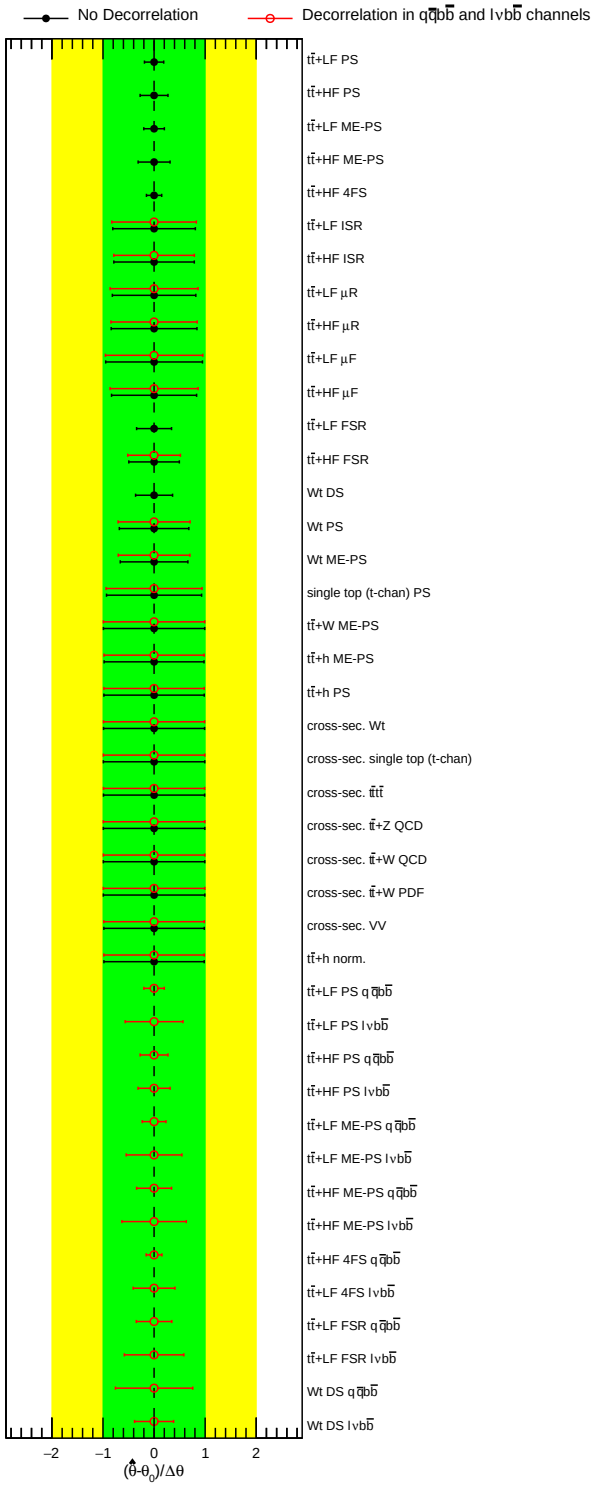


Figure B.1: Comparison of the pull distributions from a signal-plus-background fit using Asimov data for Step 1 (red) and Step 0 (black).

It can be observed from Figure B.1 that several NPs from Step 1 show over-constrained behaviour. These NPs are then decorrelated, based on the b-jet multiplicity i.e. in $3b$ and $\geq 4b$ regions (Step 2). A decorrelation of NPs in the two b-jet multiplicity regions is motivated from targeting different kinematic regions due to varying selection requirements on the reconstruction BDT score. Figure B.2 shows the comparison of the pull distributions from a signal-plus-background fit using Asimov data for Step 2 (red) and Step 1 (black).

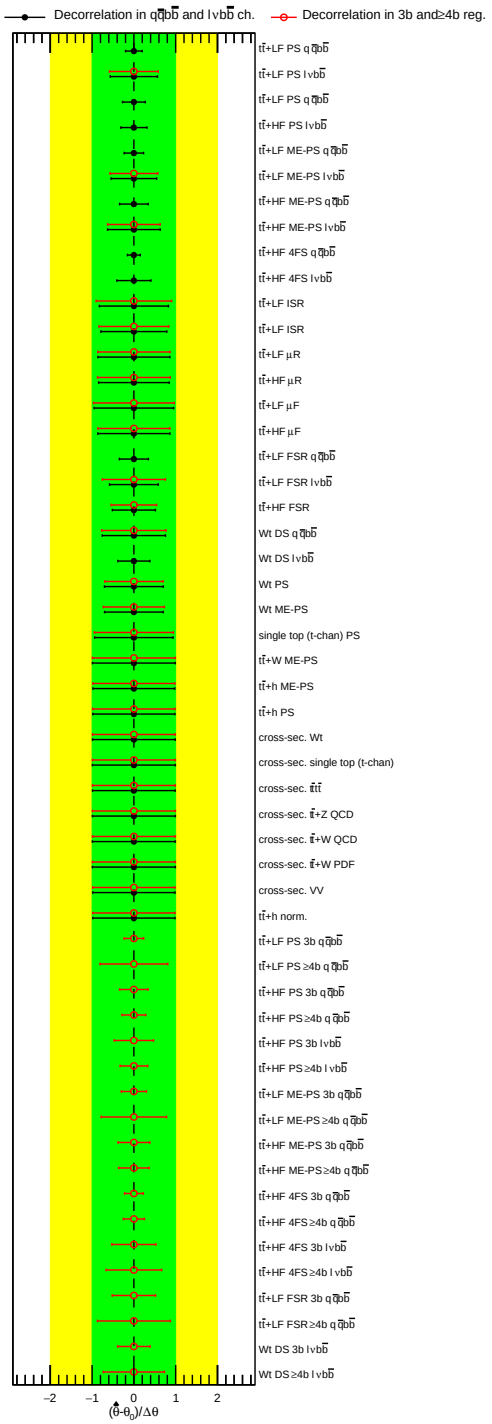


Figure B.2: Comparison of the pull distributions from a signal-plus-background fit using Asimov data for Step 2 (red) and Step 1 (black).

It can be observed from Figure B.2 that several NPs from Step 2 are further over-constrained. These NPs are then decorrelated based on the jet multiplicity i.e. in $5j$ and $\geq 6j$ regions (Step 3). A decorrelation of NPs in the two jet multiplicity regions is motivated from the varying modelling of the m_{Wh} distribution in data for the two regions. Figure B.3 shows the comparison of the pull distributions from a signal-plus-background fit using Asimov data for Step 3 (red) and Step 2 (black).

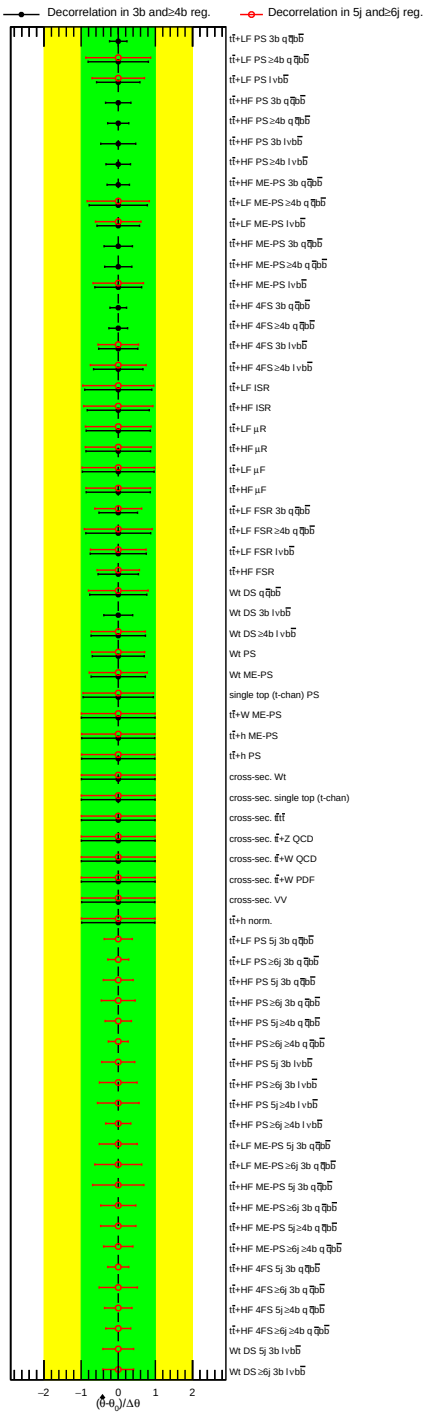


Figure B.3: Comparison of the pull distributions from a signal-plus-background fit using Asimov data for Step 3 (red) and Step 2 (black).

It can be observed from Figure B.3 that further decorrelation of some NPs is needed. This is performed in a final step (Step 4) by decorrelating NPs in individual analysis regions, or a particular analysis region split into a low-mass (L) region containing 70%-75% of the background events, and a high-mass (H) region containing the remaining background events. The latter split is performed only for cases where an NP associated with an individual analysis region shows an over-constrained behaviour. The low-mass/high-mass split is not performed for the Wt DS NP due to fewer background events in the m_{Wh} distribution. Figure B.4 shows the comparison of the pull distributions from a signal-plus-background fit using Asimov data for Step 4 (red) and Step 3 (black).

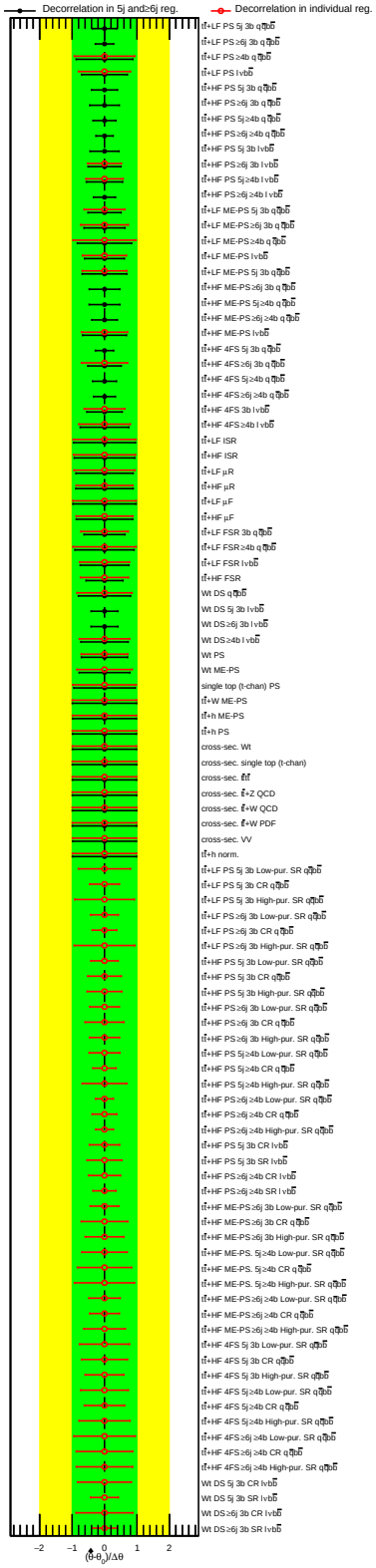


Figure B.4: Comparison of the pull distributions from a signal-plus-background fit using Asimov data for Step 4 (red) and Step 3 (black).

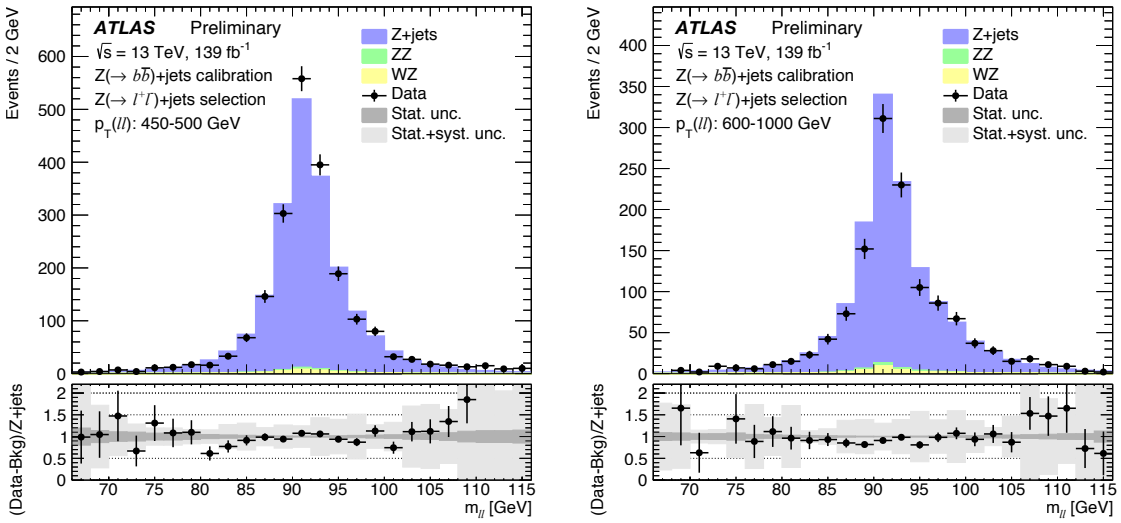
It can be observed from Figure B.4 that there are few remaining NPs which show an over-constrained behaviour. These over-constraints arise from the first two to three bins of the m_{Wh} distribution where the size of the uncertainty is up to 40% (see Figure 6.6). The step-wise decorrelation procedure leads to following independent components for the given uncertainties:

- $t\bar{t}$ + HF PS: 28
- $t\bar{t}$ + HF ME-PS: 13
- $t\bar{t}$ + LF PS: 11
- $t\bar{t}$ + LF ME-PS: 4
- $t\bar{t}$ + HF 4FS vs. 5FS: 12
- $t\bar{t}$ + LF FSR: 3
- Wt DS: 6

APPENDIX C

Additional plots supporting the studies related to the search for merged charged Higgs boson decays

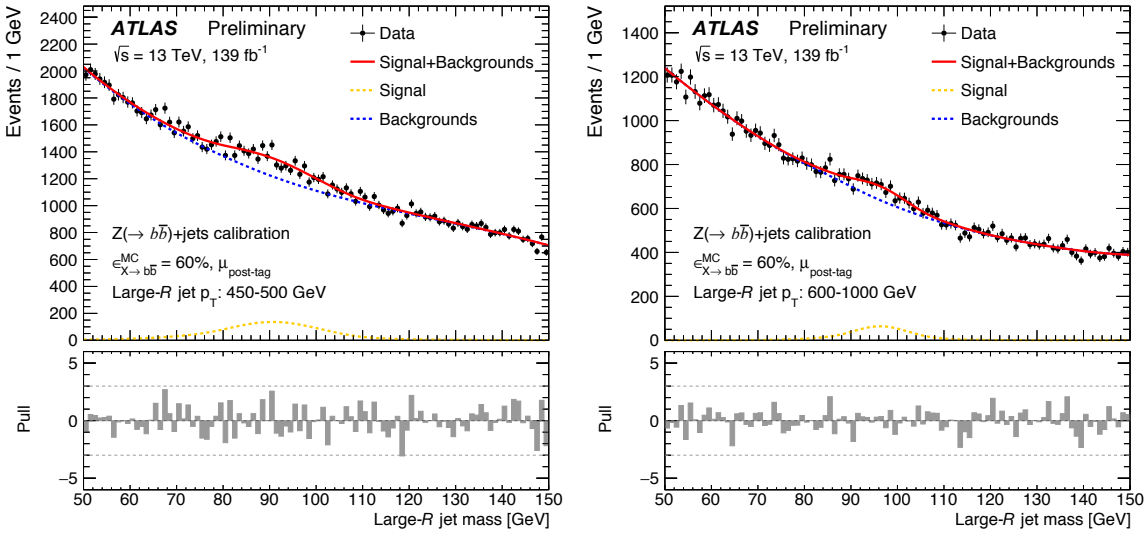
C.1 $\mu_{\text{pre-tag}}$: $Z \to l^+l^-$ candiate mass distributions



(a) (b)

Figure C.1: The $Z \to l^+l^-$ candidate mass distribution in two Z-boson candidate p_{T} bins: (a) 450 GeV $< p_{\mathrm{T}}^{Z} <$ 500 GeV and (b) 600 GeV $< p_{\mathrm{T}}^{Z} <$ 1000 GeV [28].

C.2 $\mu_{\text{post-tag}}$: Post-fit $Z \to b\bar{b}$ candidate mass distributions



(a) (b)

Figure C.2: The $Z \to b\bar{b}$ candidate mass distribution after the fit and applying the $Z \to b\bar{b}$ + jets selection and the $X \to b\bar{b}$ tagger requirement at the 60% working point for events with the large-R jet in (a) 450 GeV $< p_{\text{T}} <$ 500 GeV and (b) 600 GeV $< p_{\text{T}} <$ 1000 GeV range. The dashed yellow line represents the Z + jets signal process. The dashed blue line represents the sum of all background processes. The solid red line is the sum of all signal and background processes [28].

APPENDIX D

Additional studies supporting the analysis results

D.1 Resolved Topology

Table D.1 shows the best-fit k_{HF} values from a signal-plus-background fit to data in all analysis regions, for the full range of signal mass hypotheses.

Figure D.1 shows the local p-values, as a function of H^+ boson mass, obtained from a signal-plus-background fit to data in all analysis regions.

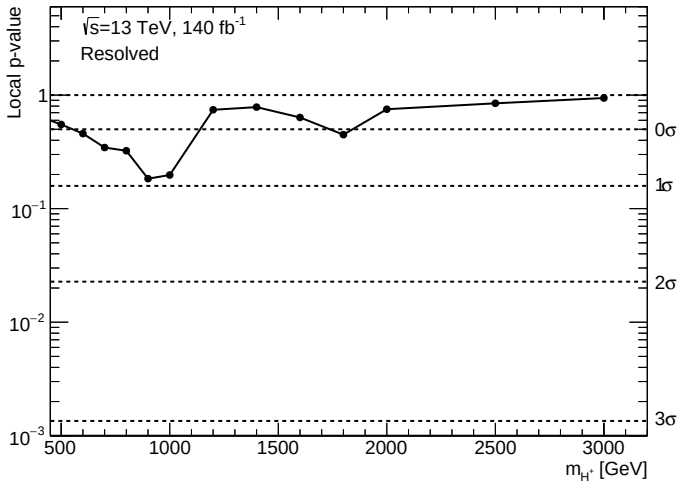


Figure D.1: Local p-value scan as a function of m_{H^+} in the search for resolved $H^+ \to Wh(b\bar{b})$ decays.

Table D.1: Best-fit k_{HF} values from a signal-plus-background fit to data in all regions.

m_{H^+} [GeV]	Best-fit k_{HF}
250	1.41 ± 0.19
300	1.38 ± 0.19
350	1.43 ± 0.19
400	1.41 ± 0.18
500	1.39 ± 0.17
600	1.40 ± 0.18
700	1.39 ± 0.18
800	1.38 ± 0.18
900	1.37 ± 0.18
1000	1.37 ± 0.18
1200	1.40 ± 0.17
1400	1.39 ± 0.17
1600	1.38 ± 0.18
1800	1.38 ± 0.18
2000	1.38 ± 0.18
2500	1.38 ± 0.18
3000	1.37 ± 0.18

D.2 Merged Topology

Figures D.2 to D.4 show the event distributions of the m_{Wh} in the signal regions used in the search for merged $H^+ \to Wh(b\bar{b})$ decays. These distributions are presented after the fit to data in all analysis regions under a background-only hypothesis. A good agreement between data and post-fit background prediction is observed.

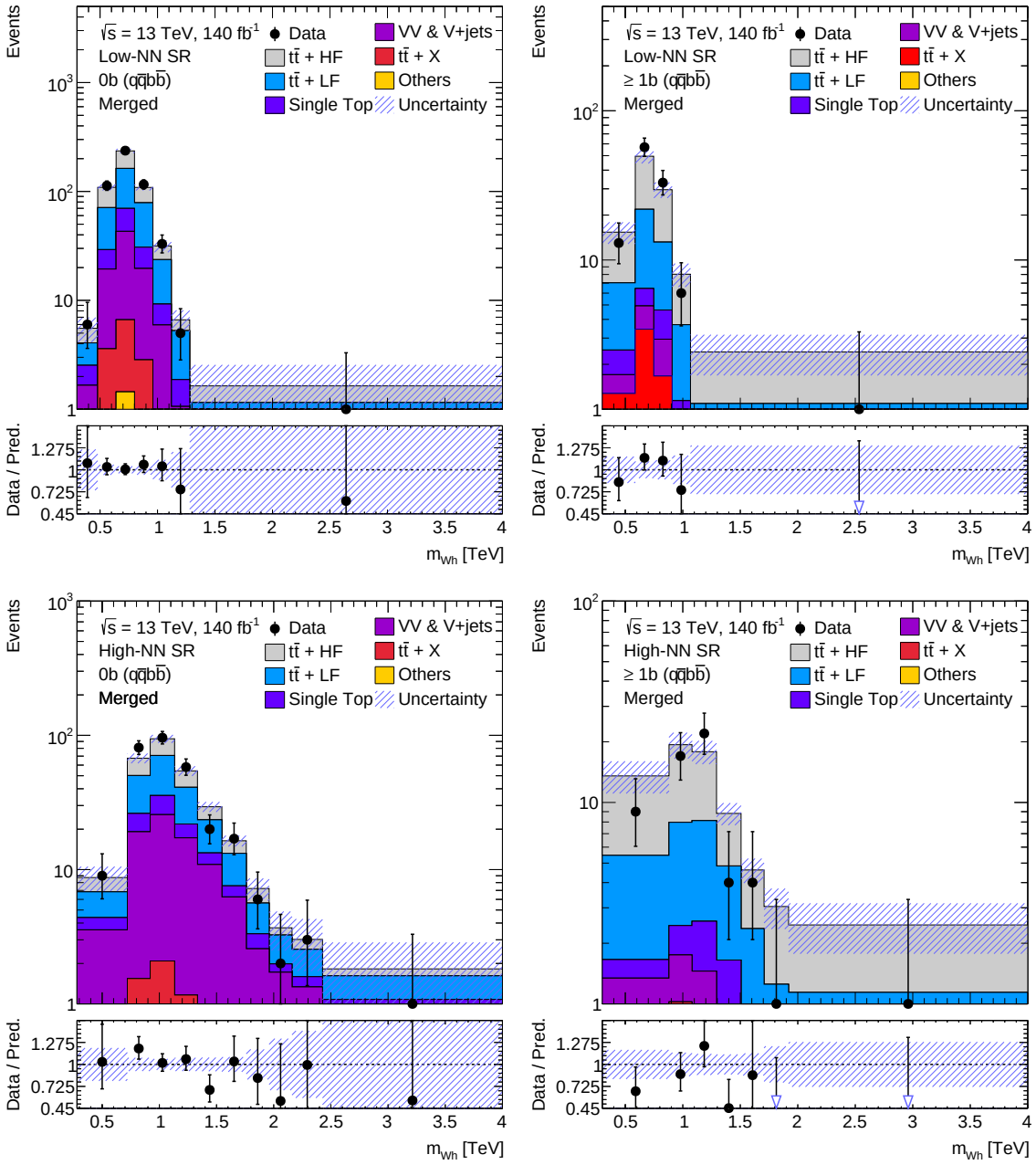


Figure D.2: Data-to-simulation comparisons for the m_{Wh} in the signal regions of the merged $q\bar{q}b\bar{b}$ channel resulting from a background-only fit to data in all analysis regions. The uncertainty band shows the post-fit statistical and systematic uncertainties for all the background processes added in quadrature.

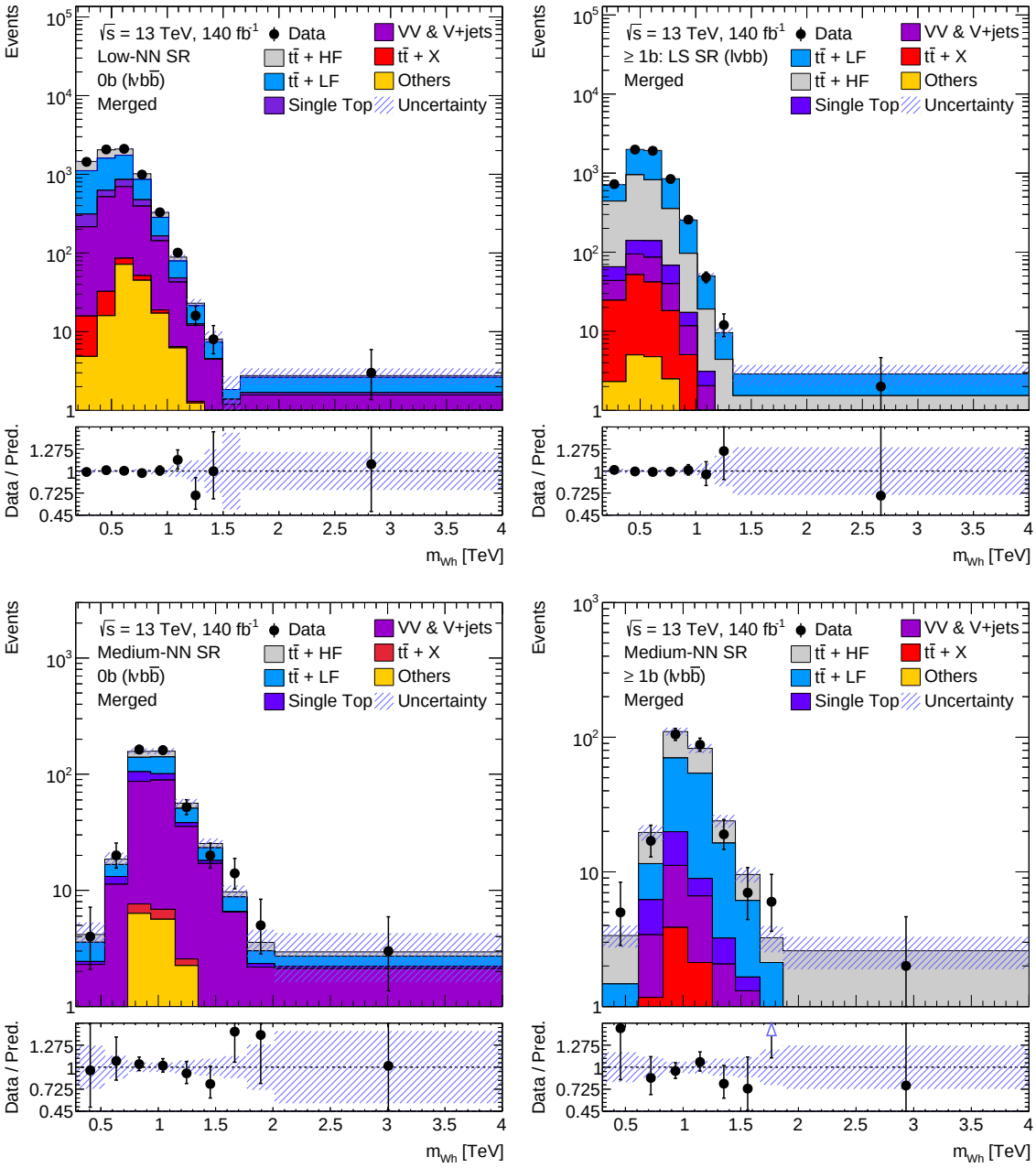


Figure D.3: Data-to-simulation comparisons for the m_{Wh} in the low-, and medium-NN score signal regions of the merged $\ell\nu b\bar{b}$ channel resulting from a background-only fit to data in all analysis regions. The uncertainty band shows the post-fit statistical and systematic uncertainties for all the background processes added in quadrature.

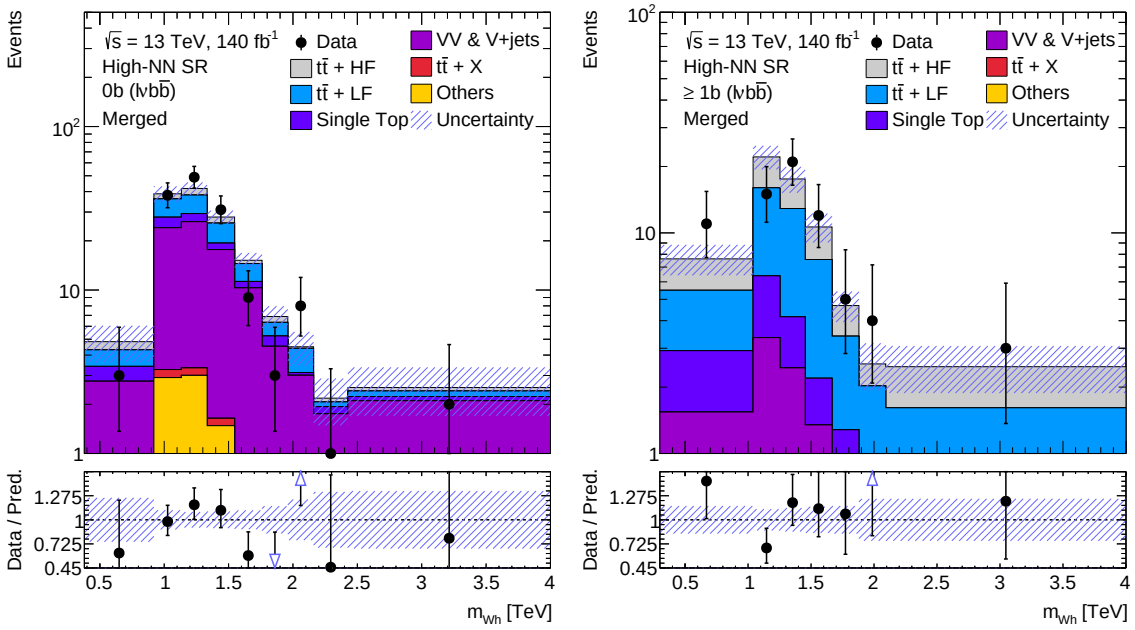


Figure D.4: Data-to-simulation comparisons for the m_{Wh} in the high-NN score signal regions of the merged $\ell\nu b\bar{b}$ channel resulting from a background-only fit to data in all analysis regions. The uncertainty band shows the post-fit statistical and systematic uncertainties for all the background processes added in quadrature.

D.3 95% CL upper limits

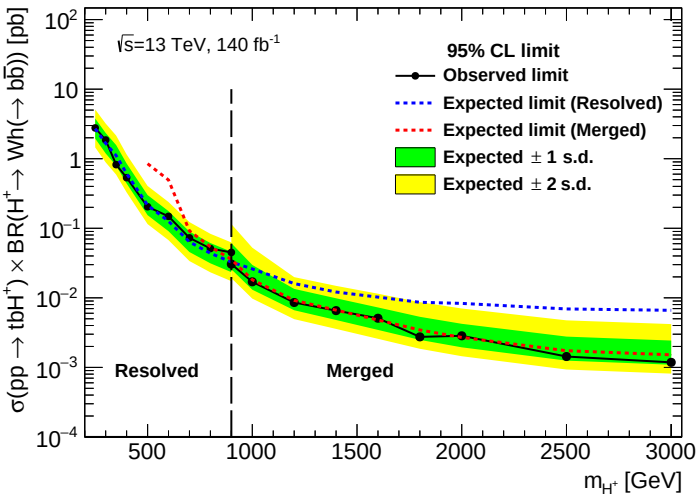


Figure D.5: 95% CL upper limits on $\sigma(pp \to tbH^+) \times BR(H^+ \to Wh(\to b\bar{b}))$ as a function of m_{H^+}, obtained in the search for resolved and merged H^+ boson decays. The results from the search for resolved H^+ boson decays are used up to $m_{H^+} \leq 900\,\text{GeV}$ and those from the search for merged H^+ boson decays are used for $m_{H^+} > 900\,\text{GeV}$.

Bibliography

[1] J.-P. Uzan, *The big-bang theory: construction, evolution and status*, Prog. Math. Phys. **76** (2021), arXiv: 1606.06112 [hep-ex] (cit. on p. 1).

[2] S. Bruning et al., *LHC Design Report Vol.1: The LHC Main Ring*, CERN-2004-003-V-1, 2004, URL: https://cds.cern.ch/record/782076 (cit. on p. 1).

[3] S. Weinberg, *The Making of the Standard Model*, Eur. Phys. J. C **34** (2004) 5, arXiv: hep-ph/0401010 (cit. on p. 1).

[4] A. Einstein, *"On the General Theory of Relativity"*, Math. Phys. (1915) 778 (cit. on p. 1).

[5] ATLAS Collaboration, *Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC*, Phys. Lett. B **716** (2012) 29 (cit. on pp. 1, 19).

[6] CMS Collaboration, *Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC*, Phys. Lett. B **716** (2012) 30 (cit. on pp. 1, 19).

[7] ATLAS Collaboration, *A detailed map of Higgs boson interactions by the ATLAS experiment ten years after the discovery*, Nature **607** (2022) 52, arXiv: 2207.00092 [hep-ex] (cit. on p. 1).

[8] CMS Collaboration, *A portrait of the Higgs boson by the CMS experiment ten years after the discovery*, Nature **607** (2022) 60, arXiv: 2207.00043 [hep-ex] (cit. on p. 1).

[9] G. C. Branco et al., *Theory and phenomenology of two-Higgs-doublet models*, Phys. Rep. **516** (2012) 1, arXiv: 1106.0034 [hep-ph] (cit. on pp. 1, 13).

[10] S. P. Martin, *A Supersymmetry Primer*, Adv. Ser. Direct. High Energy Phys. **21** (2010), arXiv: hep-ph/9709356 (cit. on p. 1).

[11] J. Camargo-Molina et al., *A new perspective on the electroweak phase transition in the Standard Model Effective Field Theory*, JHEP **2021** (2021) 127, arXiv: 2103.14022 [hep-ph] (cit. on p. 1).

[12] C-Y Chen et al., *Next-to-minimal two Higgs doublet model*, Phys. Rev. D **89** (2014) 075009 (cit. on pp. 2, 16).

[13] M. Muhlleitner et al., *The N2HDM under Theoretical and Experimental Scrutiny*, JHEP **1703** (2017) 94, arXiv: 1612.01309 [hep-ph] (cit. on p. 2).

[14] H. Georgi and M. Machacek, *Doubly charged Higgs bosons*, Nuc. Phys. B **262** (1985) 463 (cit. on pp. 2, 18).

[15] T. B. et al., *A 96 GeV Higgs boson in the N2HDM*, Eur. Phys. J. C **80** (2020) 2, arXiv: 1903.11661 [hep-ph] (cit. on pp. 2, 17).

[16] N. Ghosh et al., *Charged Higgs boson searches in the Georgi-Machacek model at the LHC*, Phys. Rev. D **101** (2020) 015029, arXiv: 1908.00396 [hep-ph] (cit. on pp. 2, 18).

[17] V. Keus et al., *CP Violating Two-Higgs-Doublet Model: Constraints and LHC Predictions*, JHEP **04** (2016) 048, arXiv: 1510.04028 [hep-ph] (cit. on p. 2).

[18] ATLAS Collaboration,
Search for charged Higgs bosons decaying via $H^\pm \to \tau^\pm \nu_\tau$ in the τ+jets and τ+lepton final states with 36 fb^{-1} of pp collision data recorded at $\sqrt{s}$=13 TeV with the ATLAS experiment, JHEP **09** (2018) 46, arXiv: 1807.07915 [hep-ex] (cit. on p. 2).

[19] CMS Collaboration, *Search for charged Higgs bosons decaying to charm and bottom quarks in proton-proton collisions at $\sqrt{s} = 8$ TeV*, JHEP **11** (2018) 115, arXiv: 1808.06575 [hep-ex] (cit. on p. 2).

[20] ATLAS Collaboration,
Search for a light charged Higgs boson in $t \to H^+ b$ decays with $H^+ \to cb$, in the lepton+jets final state in proton-proton collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector, JHEP **4** (2023) 51, arXiv: 2302.11739 [hep-ex] (cit. on pp. 2, 58).

[21] ATLAS Collaboration, *Search for a light charged Higgs boson in $t \to bH^\pm$ decays, with $H^\pm \to cs$, in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector*, 2024, arXiv: 2407.10096 [hep-ex] (cit. on p. 2).

[22] CMS Collaboration,
Search for a light charged Higgs boson decaying to a $c\bar{s}$ in pp collisions at $\sqrt{s} = 8$ TeV, JHEP **12** (2015) 178, arXiv: 1510.04252 [hep-ex] (cit. on p. 2).

[23] ATLAS Collaboration, *Search for charged Higgs bosons decaying into a top quark and a bottom quark at $\sqrt{s} = 13$ TeV with the ATLAS detector*, JHEP **06** (2021) 145, arXiv: 2102.10076 [hep-ex] (cit. on pp. 2, 58, 145).

[24] ATLAS Collaboration,
Search for a Charged Higgs Boson Produced in the Vector-boson Fusion Mode with Decay $H^\pm \to W^\pm Z$ using pp Collisions at $\sqrt{s} = 8$ TeV with the ATLAS Experiment, Phys. Rev. Lett. **114** (2015) 231801, arXiv: 1503.04233 [hep-ex] (cit. on p. 2).

[25] CMS Collaboration, *Search for a charged Higgs boson decaying into a heavy neutral Higgs boson and a W boson in proton-proton collisions at $\sqrt{s} = 13$ TeV*, JHEP **09** (2023) 032, arXiv: 2207.01046 [hep-ex] (cit. on p. 2).

[26] Higgs, Multi-Boson, and SUSY searches Working Group,
$H^+ \to WH$, $H \to bb$ (resolved and boosted), URL: https://atlas-glance.cern.ch/atlas/analysis/papers/details.php?ref_code=HMBS-2024-45 (cit. on p. 2).

[27] ATLAS Collaboration,
Search for a heavy charged Higgs boson decaying into a W boson and a Higgs boson in final states with leptons and b-jets in $\sqrt{s}$ = 13 TeV pp collisions with the ATLAS detector, 2024,
arXiv: 2411.03969 [hep-ex] (cit. on pp. 2, 44).

[28] ATLAS Collaboration, *Efficiency corrections for a tagger for boosted $H \to b\bar{b}$ decays in pp collisions at $\sqrt{s}$ = 13 TeV with the ATLAS detector*, ATL-PHYS-PUB-2021-035, 2021,
URL: https://cds.cern.ch/record/2777811
(cit. on pp. 2, 121, 129–131, 136, 179, 180).

[29] Wikimedia, *Standard Model*, (2016), URL: https://en.wikipedia.org/wiki/File:Standard_Model_of_Elementary_Particles.svg (cit. on p. 4).

[30] M. Tanabashi et al., *Review of Particle Physics*, Phys. Rev. D **1898** (2018) 274
(cit. on pp. 4, 10, 11).

[31] H. Georgi,
"Lie Algebras in Particle Physics: From Isospin to Unified Theories (Frontiers in Physics)",
Boulder, Colorado: Westview Press, 1999 (cit. on p. 5).

[32] M. Peskin and D. Schroeder, *"An Introduction to quantum field theory"*,
Addison-Wesley, 1995 (cit. on p. 5).

[33] F. Halzen and Alan D. Martin,
Quarks and Leptons: An Introductory Course in Modern Particle Physics, Wiley, 1984
(cit. on p. 5).

[34] M. Thomson, *"Modern particle physics"*, New York: Cambridge University Press, 2013
(cit. on pp. 5, 8).

[35] P. Collas and D. Klein, *"The Dirac Equation in Curved Spacetime"*, Springer, 2019
(cit. on p. 5).

[36] S. M. Bilenky and J. Hosek,
Glashow-Weinberg-Salam Theory of electroweak interactions and the neutral currents,
Phys. Rep. **90** (1982) 73 (cit. on p. 6).

[37] CERN Courier, *Weinberg, Salam and Glashow on physics*, 1980,
URL: https://cds.cern.ch/record/1730597/files/vol20-issue8-p350-e.pdf
(cit. on p. 6).

[38] R. A. Wilson, *A discrete model for Gell-Mann matrices*, 2024,
arXiv: 2401.13000 [hep-ph] (cit. on p. 7).

[39] R. L. Workman et al., *Review of Particle Physics*,
Progress of Theoretical and Experimental Physics **2022** (2022) 083CO1 (cit. on p. 7).

[40] A. Deur et al., *The QCD Running Coupling*, Prog. Part. Nuc. Phys. **90 1** (2016) 195,
arXiv: 1604.08082 [hep-ph] (cit. on p. 8).

[41] P. W. Higgs, *Broken Symmetries and the Masses of Gauge Bosons*,
Phys. Rev. Lett. **13** (1964) 508 (cit. on p. 8).

[42] F. Englert and R. Brout, *Broken Symmetry and the Mass of Gauge Vector Bosons*,
Phys. Rev. Lett. 13 **13** (1964) 321 (cit. on pp. 8, 9).

[43] J. Ellis et al., *A Historical Profile of the Higgs Boson*, 2015, arXiv: 1207.1303 [hep-ph] (cit. on p. 9).

[44] ATLAS Collaboration,
Observation of $H \to b\bar{b}$ decays and VH production with the ATLAS detector,
Phys. Lett. B **786** (2018) 59, arXiv: 1808.08238 [hep-ph] (cit. on p. 10).

[45] ATLAS Collaboration, *Cross-section measurements of the Higgs boson decaying into a pair of tau-leptons in proton-proton collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector*,
Phys. Rev. D **99** (2019) 58, arXiv: 1811.08856 [hep-ex] (cit. on p. 10).

[46] ATLAS Collaboration, *Direct constraint on the Higgs-charm coupling from a search for Higgs boson decays into charm quarks with the ATLAS detector*, Eur. Phys. J. C **82** (2022) 58, arXiv: 2201.11428 [hep-ex] (cit. on p. 10).

[47] ATLAS Collaboration, *Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC*, Phys. Lett. B **716** (2012) 24, arXiv: 1207.7214 [hep-ex] (cit. on p. 10).

[48] ATLAS Collaboration,
Observation and measurement of Higgs boson decays to WW^* with the ATLAS detector,
Phys. Rev. D **92** (2015) 69, arXiv: 1412.2641 [hep-ex] (cit. on p. 10).

[49] *LHC Higgs Cross Section Working Group*,
URL: https://twiki.cern.ch/twiki/bin/view/LHCPhysics/LHCHXSWG
(cit. on p. 11).

[50] C. M. Sommerfield, *Magnetic Dipole Moment of the Electron*, Phys. Rev. D **107** (1957) 2 (cit. on p. 10).

[51] A. D. Sakharov,
Violation of CP invariance, C asymmetry, and baryon asymmerty of the universe,
Sov. Phys. Usp. **34** (1991) 392 (cit. on p. 12).

[52] F. del Aguila et al., *Effective description of quark mixing*, Phys. Lett. B **492** (2000), arXiv: 0007160 [hep-ph] (cit. on p. 12).

[53] Super-Kamiokande Collaboration, *Atmospheric neutrino oscillation analysis with neutron tagging and an expanded fiducial volume in Super-Kamiokande I-V*,
Phys. Rev. D **109** (2023) 24, arXiv: 2311.05105 [hep-ex] (cit. on p. 12).

[54] C. B. Adams et al., *Axion Dark Matter*, 2022, arXiv: 2203.14923 [hep-ph] (cit. on p. 12).

[55] B. Carr et al., *Primordial black holes as dark matter*, Phys. Rev. D **94** (2016) 083504 (cit. on p. 12).

[56] S. Clesse and J. García-Bellido, *Seven hints for primordial black hole dark matter*, Elsevier **22** (2018) 146 (cit. on p. 12).

[57] P. D. Bolton et al., *Probes of Heavy Sterile Neutrinos*, 2022, arXiv: 2206.01140 [hep-ph] (cit. on p. 12).

[58] L. Roszkowski et al.,
WIMP dark matter candidates and searches - current status and future prospects,
Rep. Prog. Phys. **81** (2018) 84, arXiv: 1707.06277 [hep-ph] (cit. on p. 12).

[59] S. Weinberg, *Gauge hierarchies*, Phys. Lett. B **82** (1979) 387 (cit. on p. 13).

[60] C.H. Smith and G.G. Ross, *The Real Gauge Hierarchy Problem*, Phys. Lett. B **105** (1981) 38 (cit. on p. 13).

[61] V. A. Rubakov and M. E. Shaposhnikov, *Electroweak Baryon Number Non-Conservation in the Early Universe and in High Energy Collisions*, Phys. Usp. **39** (1996) 461, arXiv: 9603208 [hep-ph] (cit. on p. 13).

[62] M. Quiros, *Electroweak baryogenesis*, J. Phys. A **40** (2007) 6573 (cit. on p. 13).

[63] M. F. Sohnius, *Introducing supersymmetry*, Phys. Rep **128** (1985) 39 (cit. on p. 13).

[64] R. Ruiz, *Doubly Charged Higgs Boson Production at Hadron Colliders II: A Zee-Babu Case Study*, JHEP **10** (2022) 200, arXiv: 2206.14833 [hep-ph] (cit. on p. 13).

[65] A. Ringwald, *Axions and Axion-Like Particles*, 2014, arXiv: 1407.0546 [hep-ph] (cit. on p. 13).

[66] P. Fayet, *Supersymmetry and weak, electromagnetic and strong interactions*, Phys. Lett. B **64** (1976) 132 (cit. on p. 14).

[67] P. Fayet, *Spontaneously broken supersymmetric theories of weak, electromagnetic and strong interactions*, Phys. Lett. B **64** (1977) 584 (cit. on p. 14).

[68] C. Degrande et al., *Accurate predictions for charged Higgs production: closing the $m_{H^\pm} \sim m_t$ window*, Phys. Lett. B **772** (2017) 87, arXiv: 1607.05291 [hep-ph] (cit. on p. 14).

[69] D. Eriksson et al., *2HDMC - Two-Higgs-Doublet Model Calculator*, Comput. Phys. Commun **181** (2010) 189, arXiv: 0902.0851v2 [hep-ph] (cit. on p. 15).

[70] L. Biermann et al., *Double and Triple Higgs Production to probe the Electroweak Phase Transition*, 2024, arXiv: 2408.08043 [hep-ph] (cit. on p. 15).

[71] A. Tranberg and B. Wu, *Cold Electroweak Baryogenesis in the Two Higgs-Doublet Model*, 2012, arXiv: 1203.5012 [hep-ph] (cit. on p. 15).

[72] CMS Collaboration, *Search for a standard model-like Higgs boson in the mass range between 70 and 110 GeV in the diphoton final state in proton-proton collisions at $\sqrt{s} = 8$ TeV and 13 TeV.*, Phys. Lett. B **793** (2019) 320, arXiv: 2405.18149 [hep-ex] (cit. on p. 17).

[73] T. Taylor and D. Treille, *The Large Electron Positron Collider (LEP): Probing the Standard Model*, Adv. Ser. Direct. High Energy Phys. **27** (2017) (cit. on p. 17).

[74] ALEPH Collaboration, DELPHI Collaboration, L3 Collaboration, OPAL Collaboration, The LEP Working Group for Higgs Boson Searches, *Search for the Standard Model Higgs boson at LEP*, Phys. Lett. B **565** (2003) 61, eprint: 1808.08238 (cit. on p. 17).

[75] T. Biekötter et al., *A 96 GeV Higgs Boson in the N2HDM*, Eur. Phys. J. C **80** (2020) 44, arXiv: 1903.11661 [hep-ph] (cit. on p. 17).

[76] R. Boto et al., *Current bounds on the type-Z Z_3 three-Higgs-doublet model*, Phys. Rev. D **104** (2021) 095006, arXiv: 2106.11977 [hep-ph] (cit. on p. 17).

[77] N. G. Deshpande et al., *Left-right-symmetric electroweak models with triplet Higgs field*, Phys. Rev. D **44** (1991) 22 (cit. on p. 17).

[78] E. Ma and U. Sarkar, *Neutrino Masses and Leptogenesis with Heavy Higgs Triplets*, Phys. Rev. Lett. **80** (1998) 4 (cit. on p. 17).

[79] E. Ma and U. Sarkar, *Custodial symmetry, Georgi-Machacek model, and other scalar extensions*, Phys. Rev. D **105** (2022) 29, arXiv: 2111.14195 [hep-ph] (cit. on p. 18).

[80] ATLAS Collaboration, *The ATLAS Experiment at the CERN Large Hadron Collider*, JINST **3** (2008) S08003 (cit. on pp. 19, 23, 25).

[81] CMS Collaboration, *The CMS Experiment at the CERN LHC*, JINST **3** (2008) S08004 (cit. on p. 19).

[82] ALICE Collaboration, *The ALICE Experiment at the CERN LHC*, JINST **3** (2008) S08002 (cit. on p. 19).

[83] LHCb Collaboration, *The LHCb Detector at the LHC*, JINST **3** (2008) S08005 (cit. on p. 19).

[84] *Cern Control Centre*, URL: https://home.cern/tags/control-centre (cit. on pp. 20, 24).

[85] L. Arnaudon et al., *Linac4 technical design report*, CERN-AB-2006-084, 2006, URL: https://cds.cern.ch/record/1004186 (cit. on p. 20).

[86] K. Hanke et al., *Status of the upgrade of the CERN PS Booster*, CERN-ATS-2013-063, 2013, URL: https://cds.cern.ch/record/1566649 (cit. on p. 20).

[87] G. K. Green and E. D. Courant, *The Proton Synchrotron*, Springer Berlin Heidelberg, 1959 218 (cit. on p. 20).

[88] N. Doble et al., *The Super Proton Synchrotron (SPS): A Tale of Two Lives*, Adv. Ser. Direct. High Energy Phys. **27** (2017) 135 (cit. on p. 21).

[89] ATLAS Collaboration, *Luminosity determination in pp collisions at $\sqrt{s}$=13 TeV using the ATLAS detector at the LHC*, Eur. Phys. J. C **83** (2023) 982 (cit. on p. 21).

[90] G. Avoni et al., *The new LUCID-2 detector for luminosity measurement and monitoring in ATLAS*, JINST **13** (2018) (cit. on p. 21).

[91] A. Gorisek, *The ATLAS Beam Condition Monitor Commissioning*, CERN-2008-008, 2008, URL: https://cds.cern.ch/record/1158638/ (cit. on p. 21).

[92] ATLAS Collaboration, *Van der Meer scan luminosity measurement and beam-beam correction*, Eur. Phys. J. C **81** (2021) 26 (cit. on p. 21).

[93] ATLAS Collaboration, *Public ATLAS Luminosity Results for Run-2 of the LHC*, URL: https://twiki.cern.ch/twiki/bin/view/AtlasPublic/LuminosityPublicResultsRun2 (cit. on pp. 22, 23).

[94] J. Roloff, *Pileup Mitigation*, ATL-PHYS-SLIDE-2017-633, 2017, URL: https://cds.cern.ch/record/2276691/ (cit. on p. 22).

[95] ATLAS Collaboration, *ATLAS inner detector: Technical Design Report, 1*, CERN-LHCC-97-016, 1997, URL: https://cds.cern.ch/record/331063 (cit. on p. 23).

[96] ATLAS Liquid Argon group, *The ATLAS liquid argon electromagnetic calorimeter*, AIP Conf. Proc. **549** (2000) 874 (cit. on p. 23).

[97] ATLAS Collaboration, *Operation and performance of the ATLAS Tile Calorimeter in Run 1*, Eur. Phys. J. C **78** (2018) 987 (cit. on p. 23).

[98] ATLAS Collaboration, *ATLAS muon spectrometer: Technical Design Report*, (1997), URL: https://cds.cern.ch/record/331068 (cit. on p. 23).

[99] ATLAS Collaboration, *Expected Performance of the ATLAS Experiment - Detector, Trigger and Physics*, (2009), arXiv: 0901.0512 [hep-ex] (cit. on p. 25).

[100] H. Pernegger, *The Pixel Detector of the ATLAS Experiment for LHC Run-2*, ATL-INDET-PROC-2015-001, 2015, URL: https://cds.cern.ch/record/1985432 (cit. on p. 25).

[101] ATLAS Collaboration, *The ATLAS semiconductor tracker (SCT)*, Elsevier **541** (2005) 89 (cit. on p. 25).

[102] A. Vogel, *ATLAS Transition Radiation Tracker (TRT): Straw Tube Gaseous Detectors at High Rates*, Nucl. Instrum. Methods Phys. Res., A **732** (2013) (cit. on p. 25).

[103] ATLAS Experiment and CERN, *Computer generated image of the ATLAS inner detector*, URL: http://cds.cern.ch/record/1095926?ln=en (cit. on p. 26).

[104] J. Schieck, *Track-based alignment for the ATLAS Inner Detector Tracking System*, 2012, arXiv: 1110.4508 [hep-ex] (cit. on p. 26).

[105] Y. Takubo, *ATLAS IBL operational experience*, ATL-INDET-PROC-2016-012, 2016, URL: https://cds.cern.ch/record/2235541 (cit. on p. 25).

[106] ATLAS Experiment and CERN, *Computer generated image of the ATLAS calorimeter*, URL: http://cds.cern.ch/record/1095927 (cit. on p. 27).

[107] A. Correia, *The ATLAS Tile Calorimeter*, ATL-TILECAL-PROC-2015-002, 2015, URL: https://cds.cern.ch/record/2004868 (cit. on p. 27).

[108] ATLAS Collaboration, *Performance of the ATLAS Hadronic Endcap Calorimeter in beam tests: Selected results*, ATL-LARG-PUB-2022-001, 2022, URL: https://cds.cern.ch/record/2811731 (cit. on p. 27).

[109] ATLAS Liquid Argon Calorimeter Group, *Performance of the ATLAS Forward Calorimeters in First LHC Data*, J. Phys.: Conf. Ser. **293** (2011) 012041 (cit. on p. 28).

[110] H. J. Hilke and W. Riegler, *Gaseous Detectors*, Springer, 2020 91 (cit. on p. 28).

[111] M. Livan, *Monitored drift tubes in ATLAS*, Nucl. Instrum. Methods Phys. Res., A **384** (1996) (cit. on p. 28).

[112] ". A. et al.",
"Cathode strip chambers in ATLAS : Installation, commissioning and in situ performance",
2008 IEEE Nuclear Science Symposium Conference Record, 2008 2819 (cit. on p. 28).

[113] L. Massa, *The BIS78 Resistive Plate Chambers upgrade of the ATLAS Muon Spectrometer for the LHC Run-3*, JINST **15** (2020) C10026 (cit. on p. 28).

[114] E. P. Codina, *Small-Strip Thin Gap Chambers for the Muon Spectrometer Upgrade of the ATLAS Experiment*, Nucl. Instrum. Methods Phys. Res., A **824** (2016) (cit. on p. 28).

[115] ATLAS Experiment and CERN, *Computer generated image of the ATLAS Muons subsystem*, URL: http://cds.cern.ch/record/1095929?ln=en (cit. on p. 29).

[116] R. Barrue, *The ATLAS Trigger System*, ATL-DAQ-SLIDE-2023-541, 2023, URL: https://cds.cern.ch/record/2873720 (cit. on pp. 29, 30).

[117] ATLAS Collaboration, *Trigger Menu Information*, URL:
https://twiki.cern.ch/twiki/bin/viewauth/Atlas/TriggerMenuInformation
(cit. on p. 30).

[118] C. Ay et al., *Monte Carlo generators in ATLAS software*, J. Phys.: Conf. Ser. **219** (2010) 8 (cit. on p. 30).

[119] ATLAS Collaboration, *Optimizing the ATLAS Geant4 detector simulation software*, ATL-SOFT-PROC-2023-002, 2023, URL: https://cds.cern.ch/record/2851466 (cit. on p. 31).

[120] T. Gleisberg et al., *Event generation with SHERPA 1.1*, JHEP **02** (2009) 007, arXiv: 0811.4622 [hep-ph] (cit. on pp. 31, 33, 57).

[121] D. E. SOPER, "Basics of QCD Perturbation Theory", *Flavor Physics for the Millennium*, WORLD SCIENTIFIC, 2001 267,
URL: http://dx.doi.org/10.1142/9789812811509_0007 (cit. on p. 31).

[122] G. C. Fox and S. Wolfram, *A model for parton showers in QCD*, Nucl. Phys. B **168** (1980) 285 (cit. on p. 32).

[123] B. Andersson et al., *Parton Fragmentation and string dynamics*, Phys. Rep. **97** (1983) 145 (cit. on p. 32).

[124] J. Winter et al., *A modified cluster-hadronization model*, Eur. Phys. J. C **36** (2004) 381, arXiv: 0311085 [hep-ph] (cit. on p. 32).

[125] N. Armesto et al.,
Interference effects between the initial and final state radiation in a QCD medium,
J. Phys.: Conf. Ser. **446** (2013) 5 (cit. on p. 32).

[126] T. Sjoestrand and P. Z. Skands, *Multiple interactions and the structure of beam remnants*, JHEP **03** (2004) 053 (cit. on p. 32).

[127] T. Sjoerstrand et al., *A Brief Introduction to PYTHIA 8.1*,
Comput. Phys. Commun **178** (2008) 852, arXiv: 0710.3820 [hep-ph] (cit. on p. 33).

[128] J. Bellm et al., *Herwig 7.0/Herwig++ 3.0 release note*, Eur. Phys. J. C **76** (2016) 196, arXiv: 1512.01178 [hep-ph] (cit. on p. 33).

[129] 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 p. 33).

[130] 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. 33).

[131] ATLAS Experiment, *ATLAS Schematics*, URL: https://atlas.cern/Resources/Schematics (cit. on p. 34).

[132] ATLAS Collaboration, *Peformance of the ATLAS track reconstruction algorithms in dense environments in LHC Run 2*, Eur, Phy. J. C **77** (2017) 673, arXiv: 1704.07983 [hep-ex] (cit. on p. 34).

[133] X. Ai, *Tracking with A Common Tracking Software*, 2020, arXiv: 2007.01239v1 [physics.ins-det] (cit. on p. 34).

[134] S. Heer, *The secondary vertex finding algorithm with the ATLAS detector*, PoS **EPS-HEP2017** (2018) 762 (cit. on p. 34).

[135] F. Meloni, *Primary vertex reconstruction with the ATLAS detector*, JINST **11** (2016) (cit. on p. 34).

[136] ATLAS Collaboration, *Electron reconstruction and identification in the ATLAS experiment using the 2015 and 2016 LHC proton-proton collision data at $\sqrt{s}$=13 TeV*, Eur. Phys. J. C **79** (2019) 639, arXiv: 1902.04655 [physics.ins-det] (cit. on p. 35).

[137] ATLAS Collaboration, *ATLAS ElectronGamma Working Group*, URL: https://twiki.cern.ch/twiki/bin/view/AtlasProtected/ElectronGamma (cit. on p. 35).

[138] ATLAS Collaboration, *Performance of electron and photon triggers in ATLAS during LHC Run 2*, Eur. Phys. J. C **80** (2020) 47, arXiv: 1909.00761 [hep-ex] (cit. on p. 35).

[139] ATLAS Collaboration, *Muon reconstruction performance of the ATLAS detector in proton-proton collision data at $\sqrt{s}$=13 TeV*, Eur. Phys. J. C **79** (2016) 292, arXiv: 1603.05598 [hep-ex] (cit. on p. 36).

[140] ATLAS Collaboration, *ATLAS Muon Combined Performance Working Group*, URL: https://atlas-mcp.docs.cern.ch (cit. on p. 36).

[141] ATLAS Collaboration, *Performance of the ATLAS muon triggers in Run 2*, JINST **15** (2020) P09015, arXiv: 2004.13347 [hep-ex] (cit. on p. 36).

[142] R. Atkin, *Review of jet reconstruction algorithms*, J. Phys.: Conf. Ser. **645** (2015) 012008 (cit. on p. 36).

[143] ATLAS Collaboration, *Beam-induced and cosmic-ray backgrounds observed in the ATLAS detector during the LHC 2012 proton-proton running period*, JINST **11** (2016) 65, arXiv: 1603.09202 [hep-ex] (cit. on p. 37).

[144] ATLAS Collaboration, *Performance of pile-up mitigation techniques for jets in pp collisions at $\sqrt{s} = 8$ TeV using the ATLAS detector*, Eur. Phys. J. C **76** (2016) 581, arXiv: 1510.03823 [hep-ex] (cit. on p. 37).

[145] ATLAS Collaboration, *Selection of jets produced in proton-proton collisions with the ATLAS detector using 2011 data*, ATLAS-CONF-2012-020, 2012, URL: https://cds.cern.ch/record/1430034 (cit. on p. 37).

[146] ATLAS Collaboration, *Jet reconstruction and performance using particle flow with the ATLAS Detector*, Eur. Phys. J. C **77** (2017) 466, arXiv: 1703.10485 [hep-ex] (cit. on p. 37).

[147] ATLAS Collaboration, *Jet reconstruction and performance using particle flow with the ATLAS Detector*, Eur. Phys. J. C **77** (2017) 466 (cit. on p. 38).

[148] ATLAS Collaboration, *Jet energy scale measurements and their systematic uncertainties in proton-proton collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector*, Phys. Rev. D **96** (2017) 072002, arXiv: 1703.09665 [hep-ex] (cit. on p. 38).

[149] S. Marzani et al., *"Looking inside jets: an introduction to jet substructure and boosted-object phenomenology"*, 2019, arXiv: 1901.10342v2 [hep-ph] (cit. on p. 39).

[150] S. D. Ellis et al., *Recombination Algorithms and Jet Substructure: Pruning as a Tool for Heavy Particle Searches*, Phys. Rev. Lett (2009) 33, arXiv: 0912.0033v1 [hep-ph] (cit. on p. 39).

[151] D. Krohn et al., *Jet trimming*, JHEP **02** (2010) 084, arXiv: 0912.1342v2 [hep-ph] (cit. on p. 39).

[152] ATLAS Collaboration, *Performance of jet substructure techniques for large-R jets in proton-proton collisions at $\sqrt{s} = 7$ TeV using the ATLAS detector*, JHEP **09** (2013) 076, arXiv: 1306.4945 [hep-ph] (cit. on p. 40).

[153] ATLAS Collaboration, *In situ calibration of large-R jet energy and mass in 13 TeV proton-proton collisions with the ATLAS detector*, Eur. Phys. J. C **79** (2019) 135, arXiv: 1807.09477 [hep-ex] (cit. on p. 40).

[154] D. Krohn et al., *Jets with Variable R*, JHEP **0906** (2009) 059, arXiv: 0903.0392 [hep-ph] (cit. on p. 41).

[155] ATLAS Collaboration, *Variable Radius, Exclusive-k_T, and Center-of-Mass Subjet Reconstruction for Higgs($\rightarrow b\bar{b}$) Tagging in ATLAS*, ATL-PHYS-PUB-2017-010, 2017, URL: https://cds.cern.ch/record/2268678 (cit. on p. 41).

[156] M. Cacciari et al., *The Catchment Area of Jets*, JHEP **0804** (2008) 005, arXiv: 0802.1188 [hep-ph] (cit. on p. 41).

[157] ATLAS Collaboration, *Expected performance of the 2019 ATLAS b-taggers*, (2019), URL: http://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/PLOTS/FTAG-2019-005 (cit. on p. 41).

[158] ATLAS Collaboration,
ATLAS flavour-tagging algorithms for the LHC Run 2 pp collision dataset,
Eur. Phys. J. C **83** (2023) 681, arXiv: 2211.16345 [hep-ex] (cit. on pp. 41–43).

[159] ATLAS Collaboration, *Expected performance of the ATLAS b-tagging algorithms in Run-2*,
ATL-PHYS-PUB-2015-022, 2015, URL: https://cds.cern.ch/record/2037697
(cit. on p. 42).

[160] ATLAS Collaboration, *Performance of missing transverse momentum reconstruction with the ATLAS detector using proton–proton collisions at* $\sqrt{s} = 13$ *TeV*, Eur. Phys. J. C **78** (2018) 903,
arXiv: 1802.08168 [hep-ex] (cit. on pp. 44, 101).

[161] ATLAS Collaboration,
E_T^{miss} *performance in the ATLAS detector using 2015–2016 LHC pp collisions*,
ATLAS-CONF-2018-023, 2018, URL: https://cds.cern.ch/record/2625233
(cit. on pp. 44, 101).

[162] R. Esch, *Handbook Of Applied Mathematics*, Springer, Boston, MA, 1990 928 (cit. on p. 45).

[163] Y. Xu et al., *Artificial intelligence: A powerful paradigm for scientific research*,
The Innovation **02** (2021) 100179 (cit. on p. 45).

[164] A. Hoecker et al., *"TMVA - Toolkit for Multivariate Data Analysis"*, (2007),
arXiv: physics/0703039 [physics.data-an] (cit. on pp. 46, 48, 72).

[165] G. Tolkachev et al., *Multilayer Perceptron Neural Model for Particle Identification in MPD*,
Phys. Atom. Nuclei **86** (2023) 849 (cit. on p. 47).

[166] J. Lederer, *Activation Functions in Artificial Neural Networks: A Systematic Overview*, 2021,
arXiv: 2101.09957 [cs.LG] (cit. on p. 48).

[167] Y. Dodge, *"The Concise Encyclopedia of Statistics"*, Springer, 2008 425 (cit. on p. 49).

[168] E. L. Lehmann,
"The Fisher, Neyman-Pearson Theories of Testing Hypotheses: One Theory or Two?",
Taylor & Francis, 1993 (cit. on p. 49).

[169] G. Cowan, *"Statistical Data Analysis"*, Oxford University Press Inc., New York, 1998
(cit. on p. 49).

[170] G. Cowan et al., *Asymptotic formulae for likelihood-based tests of new physics*,
Eur. Phys. J. C **71** (2011) 32, arXiv: 1007.1727 [hep-ex] (cit. on p. 50).

[171] T. Junk, *Confidence level computation for combining searches with small statistics*,
Nucl. Instrum. Methods Phys. Res., A **434** (1999) 435 (cit. on p. 52).

[172] ATlAS Collaboration, *A study of heavy flavour quarks produced in association with top quark pairs at* $\sqrt{s} = 7$ *TeV using the ATLAS detector*, Phys. Rev. D **89** (2014) 072012,
arXiv: 1304.6386 [hep-ex] (cit. on p. 53).

[173] L. Pintucci, *Measurements of single top quark production processes with the ATLAS and CMS experiments*, (2024), arXiv: 2401.07794 [hep-ex] (cit. on p. 53).

[174] Z. Wu, *Measurements of W and Z boson production in ATLAS*, SciPost Phys. Proc. (2022) 031
(cit. on p. 53).

[175] M. Czakon et. al, *Dynamical scales for multi-TeV top-pair production at the LHC*, JHEP **71** (2017) 27, arXiv: 1606.03350 [hep-ph] (cit. on p. 53).

[176] ATLAS Collaboration, *Search for a heavy charged Higgs boson decaying into a W and a Higgs boson in final states with leptons and b-jets at $\sqrt{s}$=13 TeV*, ATL-COM-PHYS-2022-220, 2022, URL: https://cds.cern.ch/record/2805524/ (cit. on pp. 54, 83, 88, 114).

[177] S. Bansal, *MC16HPlustoWhValidation*, URL: https://twiki.cern.ch/twiki/bin/viewauth/AtlasProtected/MC16HPlustoWhValidation (cit. on p. 54).

[178] ATLAS Collaboration, *ATLAS Pythia 8 tunes to 7 TeV data*, ATL-PHYS-PUB-2014-021, 2014, URL: https://cds.cern.ch/record/1966419 (cit. on p. 54).

[179] A. Arhrib et al., *A Guidebook to Hunting Charged Higgs Bosons at the LHC*, Frontiers in Physics **8** (2020), URL: https://www.frontiersin.org/journals/physics/articles/10.3389/fphy.2020.00039 (cit. on p. 54).

[180] S. Frixione et al., *A Positive-Weight Next-to-Leading-Order Monte Carlo for Heavy Flavour Hadroproduction*, JHEP **0709** (2007) 126, arXiv: 0707.3088 [hep-ph] (cit. on p. 56).

[181] T. Sjoerstrand et al., *An Introduction to PYTHIA 8.2*, Comput. Phys. Commun **178** (2014) 45, arXiv: 1410.3012 [hep-ph] (cit. on pp. 56, 57).

[182] M. Cacciari et al., *Top-pair production at hadron colliders with next-to-next-to-leading logarithmic soft-gluon resummation*, 2011, arXiv: 1111.5869 [hep-ph] (cit. on p. 56).

[183] S. Frixione et al., *Single-top hadroproduction in association with a W boson*, JHEP **0807** (2008) 38, arXiv: 0805.3067 [hep-ph] (cit. on pp. 56, 105).

[184] S. Hoeche et al., *QCD matrix elements + parton showers: The NLO case*, JHEP **04** (2012) 14, arXiv: 1207.5030 [hep-ph] (cit. on p. 57).

[185] ATLAS Collaboration, *The Pythia 8 A3 tune description of ATLAS minimum bias and inelastic measurements incorporating the Donnachie–Landshoff diffractive model*, ATL-PHYS-PUB-2016-017, 2016, URL: https://cds.cern.ch/record/2206965 (cit. on p. 57).

[186] M. Czakon et al., *Top-pair production at the LHC through NNLO QCD and NLO EW*, JHEP **10** (2017) 22, arXiv: 1705.04105 [hep-ph] (cit. on p. 58).

[187] LHC Top Working Group, *ATLAS-CMS recommended predictions for top-quark-pair cross sections*, URL: https://twiki.cern.ch/twiki/bin/view/LHCPhysics/TtbarNNLO (cit. on p. 61).

[188] ATLAS Collaboration, *A search for $t\bar{t}$ resonances using lepton-plus-jets events in proton-proton collisions at $\sqrt{s} = 8$ TeV with the ATLAS detector*, JHEP **08** (2015) 148, arXiv: 1505.07018 [hep-ex] (cit. on p. 66).

[189] ATLAS Collaboration, *Search for pair production of Higgs bosons in the $b\bar{b}b\bar{b}$ final state using proton-proton collisions at $\sqrt{s} = 8$ TeV with the ATLAS detector*, JHEP **01** (2019) 030, arXiv: 1804.06174 [hep-ex] (cit. on p. 66).

[190] T. Gneiting and P. Vogel, *Receiver Operating Characteristic (ROC) Curves*, 2018, arXiv: 1809.04808 [stat.ME] (cit. on p. 73).

[191] A. D. Bukin, *Fitting function for asymmetric peaks*, 2008, arXiv: 0711.4449 [physics.data-an] (cit. on p. 81).

[192] W. Verkerke and D. Kirkby, *The RooFit toolkit for data modeling*, 2003, arXiv: physics/0306116 [physics.data-an] (cit. on p. 91).

[193] ATLAS Collaboration, *Luminosity determination in pp collisions at $\sqrt{s} = 13$ TeV using the ATLAS detector at the LHC*, ATLAS-CONF-2019-021, 2019, URL: https://cds.cern.ch/record/2677054 (cit. on pp. 100, 129).

[194] ATLAS Collaboration, *Measurement of the Inelastic Proton–Proton Cross Section at $\sqrt{s} = 13$ TeV with the ATLAS Detector at the LHC*, Phys. Rev. Lett. **117** (2016) 182002, arXiv: 1606.02625 [hep-ex] (cit. on p. 100).

[195] ATLAS Collaboration, *Electron reconstruction and identification in the ATLAS experiment using the 2015 and 2016 LHC proton–proton collision data at $\sqrt{s} = 13$ TeV*, Eur. Phys. J. C **79** (2019) 639, arXiv: 1902.04655 [hep-ex] (cit. on p. 100).

[196] ATLAS Collaboration, *Muon reconstruction performance of the ATLAS detector in proton–proton collision data at $\sqrt{s} = 13$ TeV*, Eur. Phys. J. C **76** (2016) 292, arXiv: 1603.05598 [hep-ex] (cit. on p. 100).

[197] ATLAS Collaboration, *Electron and photon performance measurements with the ATLAS detector using the 2015–2017 LHC proton–proton collision data*, JINST **14** (2019) P12006, arXiv: 1908.00005 [hep-ex] (cit. on p. 100).

[198] ATLAS Collaboration, *Performance of pile-up mitigation techniques for jets in pp collisions at $\sqrt{s} = 8$ TeV using the ATLAS detector*, Eur. Phys. J. C **76** (2016) 581, arXiv: 1510.03823 [hep-ex] (cit. on p. 100).

[199] ATLAS Collaboration, *Jet energy scale measurements and their systematic uncertainties in proton–proton collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector*, Phys. Rev. D **96** (2017) 072002, arXiv: 1703.09665 [hep-ex] (cit. on p. 100).

[200] ATLAS Collaboration, *Determination of jet calibration and energy resolution in proton–proton collisions at $\sqrt{s} = 8$ TeV using the ATLAS detector*, Eur. Phys. J. C **80** (2020) 1104, arXiv: 1910.04482 [hep-ex] (cit. on p. 101).

[201] ATLAS Collaboration, *ATLAS b-jet identification performance and efficiency measurement with $t\bar{t}$ events in pp collisions at $\sqrt{s} = 13$ TeV*, Eur. Phys. J. C **79** (2019) 970, arXiv: 1907.05120 [hep-ex] (cit. on p. 101).

[202] ATLAS Collaboration, *Measurement of b-tagging efficiency of c-jets in $t\bar{t}$ events using a likelihood approach with the ATLAS detector*, ATLAS-CONF-2018-001, 2018, URL: https://cds.cern.ch/record/2306649 (cit. on p. 101).

[203] ATLAS Collaboration, *Calibration of light-flavour b-jet mistagging rates using ATLAS proton–proton collision data at $\sqrt{s} = 13$ TeV*, ATLAS-CONF-2018-006, 2018, URL: https://cds.cern.ch/record/2314418 (cit. on p. 101).

[204] J.M. Berlingen, *High-p_{T} PCBT extrapolation uncertainties approval*, URL: https://indico.cern.ch/event/1134181/contributions/4770127/attachments/2403590/4111151/20220223_btag_highptPCBT.pdf (cit. on p. 101).

[205] ATLAS Collaboration, *Optimisation and performance studies of the ATLAS b-tagging algorithms for the 2017-18 LHC run*, ATL-PHYS-PUB-2017-013, 2017, URL: https://cds.cern.ch/record/2273281 (cit. on pp. 101, 119).

[206] R. D. Ball et al., *Parton distributions for the LHC Run II*, JHEP **04** (2015) 32, arXiv: 1410.8849 [hep-ph] (cit. on pp. 101, 105).

[207] S. Dulat et al., *New parton distribution functions from a global analysis of quantum chromodynamics*, Phys. Rev. D **93** (2016) 033006, arXiv: 1506.07443 [hep-ph] (cit. on p. 101).

[208] L. A. Harland-Lang et al., *Parton distributions in the LHC era: MMHT 2014 PDFs*, Eur. Phys. J. C **75** (2015) 204, arXiv: 1412.3989 [hep-ph] (cit. on p. 101).

[209] J. Butterworth et al., *PDF4LHC recommendations for LHC Run II*, J. Phys. G: Nucl. Part. Phys **43** (2016) 65, arXiv: 1510.03865 [hep-ph] (cit. on p. 102).

[210] ATLAS Collaboration, *ATLAS Pythia 8 tunes to 7 TeV data*, ATL-PHYS-PUB-2014-021, 2014, URL: https://cds.cern.ch/record/1966419 (cit. on p. 102).

[211] LHCTopWG, *NLO single-top channel cross sections*, 2017, URL: https://twiki.cern.ch/twiki/bin/view/LHCPhysics/SingleTopRefXsec (cit. on p. 105).

[212] A. D. Martin, W. J. Stirling, R. S. Thorne and G. Watt, *Parton distributions for the LHC*, Eur. Phys. J. C **63** (2009) 189, arXiv: 0901.0002 [hep-ph] (cit. on p. 105).

[213] A. D. Martin, W. J. Stirling, R. S. Thorne and G. Watt, *Uncertainties on α_S in global PDF analyses and implications for predicted hadronic cross sections*, Eur. Phys. J. C **64** (2009) 653, arXiv: 0905.3531 [hep-ph] (cit. on p. 105).

[214] M. Aliev et al., *HATHOR – HAdronic Top and Heavy quarks crOss section calculatoR*, Comput. Phys. Commun. **182** (2011) 1034, arXiv: 1007.1327 [hep-ph] (cit. on p. 105).

[215] P. Kant et al., *HatHor for single top-quark production: Updated predictions and uncertainty estimates for single top-quark production in hadronic collisions*, Comput. Phys. Commun. **191** (2015) 74, arXiv: 1406.4403 [hep-ph] (cit. on p. 105).

[216] 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 pp. 105, 106).

[217] S. Frixione, E. Laenen, P. Motylinski, C. White and B. R. Webber, *Single-top hadroproduction in association with a W boson*, JHEP **07** (2008) 029, arXiv: 0805.3067 [hep-ph] (cit. on p. 105).

[218] E. Bothmann et al., *Reweighting QCD matrix-element and parton-shower calculations*, Eur. Phys. J. C **76** (2016) 44, arXiv: 1606.08753 [hep-ph] (cit. on p. 105).

[219] ATLAS Collaboration, *Multi-boson simulation for 13 TeV ATLAS analyses*, ATL-PHYS-PUB-2016-002, 2016, URL: https://cds.cern.ch/record/2119986 (cit. on p. 106).

[220] 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 p. 106).

[221] R. Raitio and W. W. Wada, *Higgs Boson Production at Large Transverse Momentum in QCD*, Phys. Rev. D **19** (1979) 941 (cit. on p. 106).

[222] W. Beenakker et al., *NLO QCD corrections to $t\bar{t}H$ production in hadron collisions*, Nucl. Phys. B **653** (2003) 151, arXiv: hep-ph/0211352 [hep-ph] (cit. on p. 106).

[223] S. Dawson et al., *Associated Higgs production with top quarks at the large hadron collider: NLO QCD corrections*, Phys. Rev. D **68** (2003) 034022, arXiv: hep-ph/0305087 [hep-ph] (cit. on p. 106).

[224] Y. Zhang et al., *QCD NLO and EW NLO corrections to $t\bar{t}H$ production with top quark decays at hadron collider*, Phys. Lett. B **738** (2014) 1, arXiv: 1407.1110 [hep-ph] (cit. on p. 106).

[225] S. Frixione et al., *Electroweak and QCD corrections to top-pair hadroproduction in association with heavy bosons*, JHEP **06** (2015) 184, arXiv: 1504.03446 [hep-ph] (cit. on p. 106).

[226] LHCHXSWG1, *Recommended predictions for* $\mathrm{t\bar{t}V}$ *cross sections*, URL: https://twiki.cern.ch/twiki/bin/view/LHCPhysics/LHCHXSWGTTH#Plans_for_YR4 (cit. on p. 106).

[227] J. M. Campbell and R. K. Ellis, *$t\bar{t}W^{\pm}$ production and decay at NLO*, JHEP **07** (2012) 052, arXiv: 1204.5678 [hep-ph] (cit. on p. 106).

[228] $t\bar{t}$ resonance search team, *Full unblinding test results*, URL: https://indico.cern.ch/event/669913/contributions/2769795/attachments/1549339/2433688/ttres-fullunblind-smooth2-summary2.pdf (cit. on p. 109).

[229] ATLAS Collaboration, *Identification of Boosted Higgs Bosons Decaying Into $b\bar{b}$ With Neural Networks and Variable Radius Subjets in ATLAS*, ATL-PHYS-PUB-2020-019, 2020, URL: https://cds.cern.ch/record/2724739 (cit. on p. 120).

[230] ATLAS Collaboration, *Flavour Tagging Working Group*, URL: https://ftag.docs.cern.ch/ (cit. on pp. 120, 121).

[231] Y. He, *Identification of boosted Higgs bosons for new physics searches and improvement of the ATLAS tracker for the high luminosity phase of LHC*, CERN-THESIS-2022-207, 2022, URL: https://cds.cern.ch/record/2841330 (cit. on pp. 120, 131).

[232] ATLAS Collaboration, *Measurement of the associated production of a Higgs boson decaying to b quarks with a vector boson at high transverse momentum in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector*, ATLAS-CONF-2020-007, 2020, URL: https://cds.cern.ch/record/2715063 (cit. on p. 122).

[233] ATLAS Collaboration, *In-situ calibration of $X \to b\bar{b}$ tagger using $Z \to b\bar{b}$ + jets events*, ATL-COM-PHYS-2021-430, 2021, URL: https://cds.cern.ch/record/2774822 (cit. on p. 122).

[234] ATLAS Collaboration,
The performance of the jet trigger for the ATLAS detector during 2011 data taking,
Eur. Phys. J. C **76** (2016) 526, arXiv: 1606.07759 [hep-ex] (cit. on p. 123).

[235] ATLAS Collaboration, *Trigger Menu in 2018*, ATL-DAQ-PUB-2019-001, 2019,
URL: https://cds.cern.ch/record/2693402 (cit. on p. 123).

[236] ATLAS Collaboration,
Search for resonances in diphoton events at $\sqrt{s}$ = 13 TeV with the ATLAS detector,
JHEP **09** (2016) 001, arXiv: 1606.03833 [hep-ex] (cit. on p. 125).

[237] R. G. Lomax, *Statistical Concepts: A Second Course*, Lawrence Erlbaum Associates, 2007,
ISBN: 9780805858501, URL: https://books.google.de/books?id=p17rT373FNAC
(cit. on p. 126).

[238] ATLAS Collaboration, *In situ calibration of large-radius jet energy and mass in 13 TeV proton–proton collisions with the ATLAS detector*, Eur. Phys. J. C **79** (2019) 135,
arXiv: 1807.09477 [hep-ex] (cit. on pp. 129, 135).

[239] ATLAS Collaboration, *Measurement of the ATLAS Detector Jet Mass Response using Forward Folding with 80 fb^{-1} of $\sqrt{s}$ = 13 TeV pp data*, ATLAS-CONF-2020-022, 2020,
URL: https://cds.cern.ch/record/2724442 (cit. on pp. 129, 135).

[240] ATLAS Collaboration, *Search for a heavy charged Higgs boson decaying into a W boson and a SM-like Higgs boson at high transverse momenta in final states with one charged lepton and at least one large-R jet at $\sqrt{s}$=13 TeV with the ATLAS detector*, ATL-COM-PHYS-2024-166,
2024, URL: https://cds.cern.ch/record/2805524/
(cit. on pp. 132, 133, 135, 136, 149).

[241] S. Baines, *$H^+ \to Wh(\to b\bar{b})$ analysis status report*, URL:
https://indico.cern.ch/event/1275378/#2-h-whbb-analysis-status-repor
(cit. on pp. 133, 135).

[242] ATLAS Collaboration, *Measurement of Higgs boson decay into b-quarks in associated production with a top-quark in pp collisions at $\sqrt{s}$ = 13 TeV with the ATLAS Detector*,
JHEP **6** (2022) 62, arXiv: 2111.06712 [hep-ex] (cit. on p. 137).

[243] ATLAS Collaboration,
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,
Phys. Rev. D **97, 072016** (2018) 68, arXiv: 1712.08895 [hep-ex] (cit. on p. 137).

List of Figures

List of Tables

Acknowledgements

I am extremely greatful to Prof. Dr. Jochen Christian Dingfelder for giving me an opportunity to write my PhD thesis in his research group, and for his constant support and guidance throughout these five years. Further, I would like to extend my thank you to PD Dr. Tatjana Lenz for her regular supervision, and for the great help on reviewing this thesis. I would also like to thank Dr. Eckhard von Toerne for giving useful feedback on parts of this thesis. A big thanks to Prof. Dr. Florian Urs Bernlochner, Prof. Dr. Herbert Dreiner, and PD Dr. Christiane Dahl for agreeing to be members of my doctoral committee.

I would like to extend my thank you to all the current and former members of the Bonn ATLAS analysis working group for the fruitful discussions during the group meetings, lunches, and coffee breaks. I would like to acknowledge all the team members of the Release 21 $X \rightarrow b\bar{b}$ tagger calibration analysis using $Z \rightarrow b\bar{b}$ events, and $H^{\pm} \rightarrow W^{\pm}h(b\bar{b})$ analysis within the ATLAS collaboration. In particular, I would like to thank Reina Coromoto Camacho Toro, Yajun HE, Dominik Duda, Ulla Blumenschein, Imma Riu, Adrian Salvador Salas, Simon Grewe, Raphael Hulsken, and Sid Baines, for being immensely cooperating and motivating to work with. I would also like to thank the Bonn IT-support for immensely helping me with the technical aspects during my PhD.

I would like to extend my profound gratitude to my dear parents and my siblings for patiently listening to all the ups and downs during these years and encouraging me unfailingly. I owe you all! Lastly, I would like to thank all my friends in and outside Germany for keeping my mind off from work from time-to-time during these five years.