UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

Measurement of ggF and VBF Higgs boson production cross-sections in the $H \to WW^* \to e\nu\mu\nu$ decay channel from pp collisions collected with the ATLAS detector at the LHC

A DISSERTATION

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

Degree of

DOCTOR OF PHILOSOPHY

By

David Richard Shope
Norman, Oklahoma
2019

Measurement of ggF and VBF Higgs boson production cross-sections in the $H \to WW^* \to e\nu\mu\nu$ decay channel from pp collisions collected with the ATLAS detector at the LHC

A DISSERTATION APPROVED FOR THE
HOMER L. DODGE DEPARTMENT OF PHYSICS AND ASTRONOMY

BY THE COMMITTEE
CONSISTING OF

Dr. Michael Strauss, Chair

Dr. Phillip Gutierrez

Dr. Howie Baer

Dr. Kieran Mullen

Dr. Keith Strevett

Abstract

This thesis describes measurements of the inclusive Higgs boson production cross-sections via gluon-gluon fusion (ggF) and vector-boson fusion (VBF) through the $H \to WW^* \to e\nu\mu\nu$ decay mode, using proton-proton collision data collected at $\sqrt{s} = 13\,\text{TeV}$ by the ATLAS detector with the full dataset corresponding to an integrated luminosity of $36.1\,\text{fb}^{-1}$ recorded in the years 2015 and 2016. An overview of the theoretical concepts and experimental apparatuses as well as the data processing techniques used for these results are provided, followed by the general strategy and background estimation of the analysis. Particular detail is given for the estimation of backgrounds originating from misidentified leptons, which is one of the specific focuses of the author's work. Finally, the statistical treatment of the analysis is presented, where the product of the $H \to WW^*$ branching ratio times the ggF and VBF cross-sections are measured to be $11.4^{+1.2}_{-1.1}(\text{stat.})^{+1.8}_{-1.7}(\text{syst.})$ pb and $0.50^{+0.24}_{-0.22}(\text{stat.}) \pm 0.17(\text{syst.})$ pb, respectively, with no significant deviation from the Standard Model prediction being observed.

Acknowledgements

I would like to thank my advisor Dr. Mike Strauss not only for provided his guidance throughout my thesis, but for always giving me the freedom to pursue my interests - particularly when it necessitated fewer short-term results. I would also like to thank the OU High Energy physics group for giving me the invaluable opportunity of being stationed at CERN. I recognize the rarity of my length of stay and so worked to make the very most of every day. To each member of the ATLAS HWW analysis team past and present, thanks for fostering such a friendly and collaborative environment - it's truly been a pleasure working with you all. Special thanks go to my colleagues from Freiburg: Karsten, Carsten, and Ralf. Your expertise has been inspiring and the technical discussions we have had, particularly concerning the development of our analysis frameworks, are among the highlights of the latter half of my journey as a graduate student. Finally, I would most of all like to thank my parents for their wholehearted support as I pursued my dreams and for providing me with all of the opportunities needed from the very beginning to help me arrive at where I am today.

Contents

Chapter 1

Introduction

Throughout history, the question of what the universe is composed of at its most fundamental level is one which has both occupied the thoughts of many great minds and brought motivation to countless experiments, driven to put new theories to the test. Today, the Standard Model of Particle Physics represents the culmination of these efforts, with its remarkable predictory power and ability to withstand even the most well-formulated and focused scrutiny. The discovery of the Higgs boson on July 4, 2012 by the ATLAS and CMS collaborations at CERN marked the end of an era in which not all particles of the Standard Model had yet been found. However, there are still open issues which aren't addressed by the Standard Model. For example, while the electroweak and strong forces are well accomodated by the theory, the fourth and weakest force of gravity is still excluded.

In addition to searching explicitly for new particles, a second avenue for finding phenomena that are not explained by the Standard Model in its current form is the performance of precision measurements of the predictions made by the theory, since any significant deviations would come as well with an opportunity for revision. For example, the Higgs mechanism makes specific predictions about how strongly the Higgs boson couples to other elementary particles, which can be related to their mass. This thesis presents a detailed analysis which is one of the latest efforts in probing these Higgs couplings, utilizing pp collision data collected from the Large Hadron Collider (LHC) at $\sqrt{s} = 13\,\text{TeV}$ in 2015 and 2016 with the ATLAS detector. The two most dominant Higgs production modes at the LHC (gluon-gluon fusion and vector boson fusion) are considered, while the Higgs decay mode to two W bosons with subsequently decay into two leptons is targeted for study.

The content of this thesis is structured in the following way: Chapter 2 provides an introduction to the theoretical framework. Chapter 3 describes both the experimental apparatuses and data processing techniques which are used to collect and prepare the dataset for analysis. Chapter 4 introduces the general analysis strategy. Chapter 5 discusses the majority of the backgrounds in the analysis and how they are estimated, while chapter 6 details the specific estimation of backgrounds originating from misidentified leptons. Chapter 7 lists the various systematic uncertainties that are considered in the analysis. Chapter 8 outlines the statistical treatment of the analysis. Finally, chapter 9 provides a summary of the analysis results while chapter 10 offers some concluding remarks and prospects for future studies.

Chapter 2

Theoretical Overview

This chapter provides the theoretical context in which the rest of the work is described. A more in-depth analysis for many of these topics can be found in e.g. [81, 103, 106], which served as a basis for the following discussions. The Standard Model and its formulation are introduced in section 2.1. Relevant features of proton-proton collisions are described in section 2.2, while section 2.3 details the various ways in which the Higgs boson is produced and decays at the LHC.

2.1 The Standard Model of Particle Physics

The theories that incorporate all of our current understanding of the subatomic particles and their interactions are collected into a single framework known as the *Standard Model of Elementary Particle Physics* (SM). The main features of the SM have largely remained intact since the mid-1970s when experimental evidence of the electroweak theory formed by Glashow, Weinberg, and Salam (GSW, [77, 80]) began to emerge, as well as evidence of the quark model [75]. All SM particles are separated into two main categories (also displayed in Table 2.1 and Table 2.2):

- **Fermions** - identified as having 1/2 integer intrinsic spin e.g. 1/2, 3/2, ... and are the building blocks for matter in the universe
- **Bosons** - identified as having integer intrinsic spin e.g. 0, 1, 2, ... and are responsible for mediating the fundamental forces of nature[1]

The fermions come in three generations which differ primarily in mass and can further be sub-divided based on their charge under each fundamental force (and therefore the force interactions in which they participate). Quarks are the names given to the matter particles which have associated color charge (often referred to as red, green, and blue) under the strong force. For each generation, there exists a doublet of an up-type and down-type quark, for a total of 6 particles. In ascending order of mass, the up-type quarks are called the up u, charm c, and the top t, while the down-type quarks are called the down d, strange s, and bottom b. The up-type quarks carry an electric charge of $+\frac{2}{3}$, while the down-type quarks carry an electric charge of $-\frac{1}{3}$. The remainder of the fermions which lack color charge are known as leptons and also appear in generational doublets, each containing one

[1] The notable exception being the Higgs particle which doesn't mediate a force per se, but rather gives rise to particle mass through interactions with its associated field.

with an electrical charge of -1 (the electron e, the muon μ, and the τ-lepton) together with an associated neutrino which has no electrical charge (ν_e, ν_μ, or ν_τ). In addition, for each fermion above there exists an anti-particle with opposite charge and chirality.

Generation	1^{st}	2^{nd}	3^{rd}	Q
Quarks	$\begin{pmatrix} u \\ d \end{pmatrix}$	$\begin{pmatrix} c \\ s \end{pmatrix}$	$\begin{pmatrix} t \\ b \end{pmatrix}$	$+\frac{2}{3}$ $-\frac{1}{3}$
Leptons	$\begin{pmatrix} \nu_e \\ e \end{pmatrix}$	$\begin{pmatrix} \nu_\mu \\ \mu \end{pmatrix}$	$\begin{pmatrix} \nu_\tau \\ \tau \end{pmatrix}$	0 -1

Table 2.1: A listing of the SM fermions, along with their charge under the electromagnetic force. The quarks are also charged under the strong force, while all of the (left-handed) doublets are charged under the weak force.

While it can be very useful to picture these elementary objects as point-like particles, the theoretical foundation of the SM known as *Quantum Field Theory* (QFT) instead builds upon *fields* as the fundamental constituents with particles being interpreted simply as their local excitations. In such a framework, the dynamics of a system of fields are described by a Lagrangian ($\mathscr{L}$). By postulating that the Lagrangian of the SM is invariant under local gauge transformations of the $SU(3) \otimes SU(2) \otimes U(1)$ group, it follows that additional massless vector fields are required. Also referred to as gauge fields, they can be identified as the underlying fields associated with the force mediators. The strong interactions are mediated by eight massless gluon fields, while the electroweak interactions are mediated by the massive $W^{\pm}$ and Z together with the massless photon. In order for the weak force carriers to acquire mass, the $SU(2) \otimes U(1)$ symmetry must be spontaneously broken through what is known as the Higgs mechanism. During this process, a scalar field is added and partly results in the introduction of the massive Higgs boson which interacts directly with other massive particles. The Higgs boson is the most recent particle to have been discovered (its observation being announced in 2012 [8, 9]), completing the picture for the current SM as it was predicted half a century ago.

Name	Symbol	Q	Mass	Description
8 Gluons	g	0	0	Strong mediator
W^+ boson	W^+	$+1$	~ 80.4 GeV	Weak mediators
W^- boson	W^-	-1	~ 80.4 GeV	
Z boson	Z	0	~ 91.2 GeV	
Photon	γ	0	0	Electromagnetic mediator
Higgs boson	H	0	~ 125.2 GeV	Couples with massive particles

Table 2.2: A listing of the SM bosons, along with some of their properties. Altogether there are 12 force mediators in addition to the most recently discovered Higgs boson.

2.1.1 Local Gauge invariance

In non-relativistic quantum mechanics, it follows from observable quantities involving combinations of $\overline{\psi}\psi$ that the global phase of a wavefunction is arbitrary. Similarly in

quantum field theory, fields which pick up a global phase leave any Lagrangian containing them invariant under such a transformation. For example, the Dirac Lagrangian[2]

$$\mathscr{L} = \overline{\psi}(i\gamma^\mu \partial_\mu - m)\psi \tag{2.1}$$

describing the field of a free fermion with mass m is invariant under

$$\psi \to e^{i\theta}\psi \tag{2.2}$$

where θ is any real number, due to the fact that this implies $\overline{\psi} \to e^{-i\theta}\overline{\psi}$ as well. In general, the phase may also be different for separate points in spacetime (i.e. θ can be a function of x^μ). It is also convenient to define

$$\lambda(x) = -\theta(x)\frac{1}{Q} \tag{2.3}$$

where Q is the charge operator such that

$$\psi \to e^{-i\lambda(x)Q}\psi. \tag{2.4}$$

However, this new requirement does not leave the Dirac Lagrangian invariant. Under such a local phase transformation, the Lagrangian becomes

$$\mathscr{L} \to \mathscr{L} + \overline{\psi}\gamma^\mu Q\psi(\partial_\mu \lambda). \tag{2.5}$$

Working under the assumption that the complete Lagrangian should remain invariant, a new term must be added to cancel the extra one above. This amounts to writing a new Lagrangian

$$\mathscr{L} = \overline{\psi}(i\gamma^\mu \partial_\mu - m)\psi - (\overline{\psi}\gamma^\mu Q\psi)A_\mu \tag{2.6}$$

which is now invariant under local phase transformations, provided that A_μ is a new field that itself changes simultaneously according to

$$A_\mu \to A_\mu + \partial_\mu \lambda. \tag{2.7}$$

Due to the fact that A_μ is a vector field, its dynamics can be described with the Proca Lagrangian (similar to the Dirac Lagrangian for spin 1/2 particles) which prescribes an additional term that allows it to propagate alone:

$$\mathscr{L}_{Proca} = -\frac{1}{4}F^{\mu\nu}F_{\mu\nu} + \frac{1}{2}m_A^2 A^\nu A_\nu \tag{2.8}$$

While $F^{\mu\nu} \equiv (\partial^\mu A^\nu - \partial^\nu A^\mu)$ is invariant under Equation 2.7, a term such as $m_A^2 A^\nu A_\nu$ is not. Therefore, the second term cannot appear in a locally gauge invariant Lagrangian (implying the new field is massless). Indeed, A_μ can be identified as the four-vector potential that defines the photon particle with the last term in Equation 2.6 and the first term in Equation 2.8 recovering the Maxwell Lagrangian

$$\mathscr{L}_{Maxwell} = -\frac{1}{4}F^{\mu\nu}F_{\mu\nu} - \frac{1}{c}J^\mu A_\mu \tag{2.9}$$

[2] γ^μ for $\mu = 0, 1, 2, 3$ are known as the Dirac matrices, with $\gamma^0 = \sigma^3 \otimes I$, and $\gamma^j = i\sigma^2 \otimes \sigma^j$, where σ^j (for $j = 1, 2, 3$) denote the Pauli spin matrices.

with

$$J^\mu = c(\overline{\psi}\gamma^\mu Q\psi). \tag{2.10}$$

The notation of Equation 2.6 can be simplified by folding the extra term into the derivative so that the "covariant derivative" is defined by

$$\mathcal{D} \equiv \partial_\mu + iQA_\mu. \tag{2.11}$$

The resulting expression is the well-known Lagrangian of quantum electrodynamics:

$$\mathscr{L}_{QED} = \overline{\psi}(i\gamma^\mu \mathcal{D}_\mu - m)\psi - \frac{1}{4}F^{\mu\nu}F_{\mu\nu} \tag{2.12}$$

What is remarkable about this result is that purely through enforcing local phase invariance as a property of the Lagrangian describing free Dirac fields, a new massless vector field must be introduced. With the recipe for Equation 2.7 being reminiscent of the gauge freedom present in classical electrodynamics, Equation 2.4 and Equation 2.7 are often referred to as gauge transformations, while A^μ is described as a gauge field.

In a more general formulation, Equation 2.2 can be viewed as

$$\psi \to U\psi \tag{2.13}$$

where U is the unitary 1×1 matrix $e^{-i\lambda(x)Q}$. The collection of all possible matrices of this type form the group known as $U(1)$, where $\lambda(x)$ and Q are identifed as the real group parameters and the group generator, respectively. For this reason, such a symmetry is called "$U(1)$ gauge invariance". This formalism is extendable, such that *all* of the other fundamental interactions in the SM can be generated in a similar fashion through the requirement that Lagrangians obey local gauge invariance of more complex groups. For each generator of these symmetries, a new massless vector gauge field must be introduced through the covariant derivative so as to keep the resulting Lagrangian invariant.

2.1.2 Quantum Chromodynamics

The theory of strong interactions is known as Quantum Chromodynamics (QCD) and can be derived starting from the Lagrangian of free quark color fields q

$$\mathscr{L} = \overline{q}_i(i\gamma^\mu \partial_\mu - m)q_i \tag{2.14}$$

where the indices i represent the three color charges (red, green, and blue) and imposing local gauge invariance under $SU(3)$ group transformations

$$q_i \to e^{i\sum_{a=1}^{8}\alpha_a(x)\frac{\lambda_a}{2}}q_i \tag{2.15}$$

in which α_a and λ_a are identified as the $SU(3)$ group parameters and generators, respectively. λ_a are commonly referred to as the Gell-Mann matrices and don't commute with one another - specifically, $\left[\frac{\lambda_a}{2}, \frac{\lambda_b}{2}\right] = i\sum_{c=1}^{8} f_{abc}\frac{\lambda_c}{2}$ with f_{abc} being the SU(3) structure constants. Local gauge invariance of the Lagrangian is recovered (analogous to subsection 2.1.1) by introducing vector fields which transform as

$$G_\mu^a \to G_\mu^a - \frac{1}{g_s}\partial_\mu\alpha_a - \sum_{b,c=1}^{8} f_{abc}\alpha_b G_\mu^c \tag{2.16}$$

and associated covariant derivative

$$\mathcal{D}_\mu = \partial_\mu + ig_s \sum_{a=1}^{8} \frac{\lambda_a}{2} G_\mu^a \tag{2.17}$$

where g_s is the strong coupling constant and G_μ^a are the eight gluon fields associated with specific combinations of color and anticolor. Combining these developments together yields the Lagrangian of quantum chromodynamics:

$$\mathscr{L}_{QCD} = \overline{q}_i(i\gamma^\mu \mathcal{D}_\mu - m)q_i - \frac{1}{4} G_{\mu\nu}^a G_a^{\mu\nu}. \tag{2.18}$$

where the gluon field strength tensor has been defined as

$$G_{\mu\nu}^a = \partial_\mu G_\nu^a - \partial_\nu G_\mu^a - g_s \sum_{b,c=1}^{8} f_{abc} G_\mu^b G_\nu^c. \tag{2.19}$$

The appearance of the last term in $G_{\mu\nu}^a$ can be traced back to the non-vanishing commutations between the Gell-Mann matrices and gives rise to self couplings of the gluon fields. In general, groups which have generators that don't commute are referred to as non-Abelian and consequentially result in such self interaction terms.

2.1.3 Electroweak Theory

Beginning with the unification of the electromagnetic force through Maxwell's equations, electroweak symmetry is the second and most recent step in the progression towards the unification of all four forces. The Electroweak Theory (EW) is the part of the SM that describes both the electromagnetic and weak interactions in the same framework of a gauge quantum field theory, where the unification is accomplished through the collective transformation invariance under the electroweak gauge symmetry group $SU(2)_L \times U(1)_Y$, with 4 total generators. $SU(2)_L$ is the weak isospin group which is non-abelian and has 3 generators $T_{1,2,3} = \sigma_{1,2,3}/2$, where $\sigma_{1,2,3}$ are the Pauli spin matrices. The subscript $_L$ refers to the fact that only the left-handed chiral fermion fields are transformed by this group, which therefore must exist as doublets (while the right handed chiral fields exist as singlets). $U(1)_Y$ is the weak hypercharge group which is abelian and has 1 generator $Y/2$.

Since the $W^\pm$ and Z bosons are massive, they cannot directly serve as the gauge bosons of the underlying theory (a result from subsection 2.1.1). Rather, the unified interaction spectrum of electroweak gauge bosons contains entirely new particles - the massless $W_\mu^{1,2,3}$ bosons (with no electric charge) for the $SU(2)_L$ group and the massless B_μ boson for the $U(1)_Y$ group. The physical spectrum of electroweak gauge bosons is then composed of the massive $W^\pm$ and Z, along with the massless γ. In the SM, this is a result of the electroweak symmetry being broken spontaneously via the Higgs Mechanism. Mathematically, this is expressed as the electroweak group being broken down to the electromagnetic subgroup $SU(2)_L \times U(1)_Y \rightarrow U(1)_{em}$ where the generator of $U(1)_{em}$ is $Q = T_3 + \frac{Y}{2}$. During the symmetry breaking, the W_μ^i and B_μ bosons mix with each other to form the physical $W^\pm$, Z, and γ.

Under the electroweak gauge symmetry, the fermion fields of the SM transform according to the following recipe (analogous to Equation 2.4 and Equation 2.15):

1) $SU(2)_L : \psi_L \rightarrow e^{-i\overrightarrow{\theta}(x)\cdot\frac{\vec{\sigma}}{2}}\psi_L$, doublets; $\psi_R \rightarrow \psi_R$, singlets

2) $U(1)_Y : \psi \to e^{-i\beta(x)\frac{Y}{2}}\psi$

where $\psi_L = \frac{(1-\gamma^5)}{2}\psi$ and $\psi_R = \frac{(1+\gamma^5)}{2}\psi$ are the two chiral projections of ψ with $\gamma^5 \equiv i\gamma^0\gamma^1\gamma^2\gamma^3$. Continuing with the analogy, the normal derivative in Equation 2.1 can be replaced with the appropriate covariant derivative

$$\partial_\mu\psi \to \mathcal{D}_\mu\psi = (\partial_\mu + ig\overrightarrow{T}\cdot\overrightarrow{W}_\mu + ig'\frac{Y}{2}B_\mu)\psi \tag{2.20}$$

where g and g' are the gauge couplings for the $SU(2)_L$ and $U(1)_Y$ gauges, respectively. Due to the fact that $U(1)_{em}$ is a subgroup of $SU(2)_L \times U(1)_Y$, they bear a simple relation to the electromagnetic coupling e:

$$g = \frac{e}{\sin\theta_W} \quad ; \quad g' = \frac{e}{\cos\theta_W} \tag{2.21}$$

Here, θ_W is the weak mixing angle which prescribes the relative composition of the physical neutral electroweak bosons in terms of W^3_μ and B_μ. Adding propagation terms for the gauge fields and summing over all the fermions in the standard model, the Lagrangian of the EW theory can be written as

$$\mathscr{L}_{EW} = \sum_\psi i\overline{\psi}\gamma^\mu\mathcal{D}_\mu\psi - \frac{1}{4}W^i_{\mu\nu}W^{\mu\nu}_i - \frac{1}{4}B_{\mu\nu}B^{\mu\nu} \tag{2.22}$$

with

$$W^i_{\mu\nu} = \partial_\mu W^i_\nu - \partial_\nu W^i_\mu + g\epsilon^{ijk}W^j_\mu W^k_\nu \tag{2.23}$$

and

$$B_{\mu\nu} = \partial_\mu B_\nu - \partial_\nu B_\mu. \tag{2.24}$$

This Lagrangian is invariant under $SU(2)_L \times U(1)_Y$ gauge transformations. However, the fermion mass terms must be excluded in addition to those for each vector field, for now $m\overline{\psi}\psi = m(\overline{\psi}_L\psi_R + \overline{\psi}_R\psi_L)$ is not invariant under the symmetry group $SU(2)_L$. The consequence is that the fermion masses must also be generated through the process of electroweak symmetry breaking.

2.1.4 Spontaneous Symmetry Breaking and the Higgs Mechanism

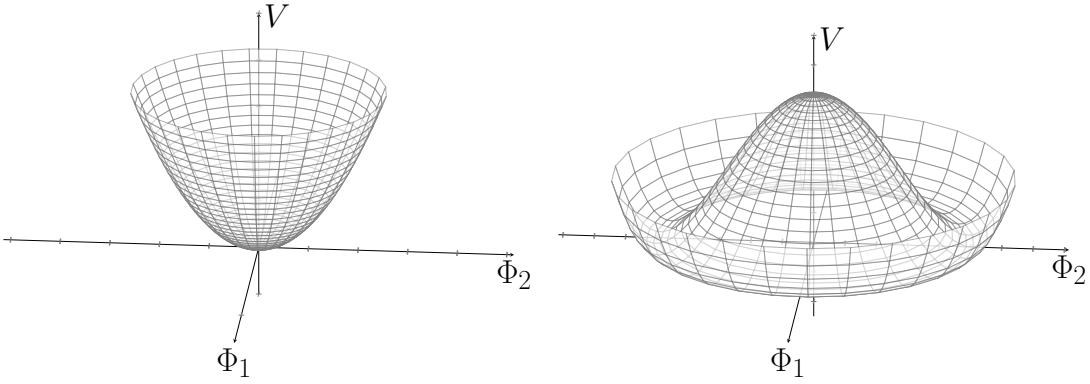
The process of electroweak symmetry breaking is carried out spontaneously within the SM. However, the phenomena of spontaneous symmetry breaking (SSB) is not unique to this system. In general, a physical system is said to exhibit a spontaneously broken symmetry if the underlying laws governing its dynamics respect such a symmetry but its ground state does not. One example of SSB outside the SM is the magnetization of ferromagnets. In these systems, which can be described as infinite sets of elementary spins and their interactions at a given temperature T, spacial rotations leave the dynamical equations invariant. A closer inspection of the ground state, on the other hand, reveals a more complex temperature dependent behavior. If T is above what's known as the Curie temperature T_C, the ground state will respect the symmetry of rotations in space. If $T < T_C$, then there will be a randomly preferred spin direction, effectively breaking rotational invariance. An average magnetization will be acquired, which is then said to be the order parameter of this SSB. The framework that successfully describes this phenomenon is known as the theory of Ginzburg-Landau.

A toy case

In QFT, a system experiences a spontaneously broken symmetry if the Lagrangian describing its dynamics is invariant under such a symmetry transformation, but the vacuum of the theory is not. To find the vacuum of the system, the state through which the expectation value of the Hamiltonian is a minimum must be computed. In order to demonstrate how massive bosons can appear in the gauge theory, consider the spontaneous breaking of the simplest gauge symmetry $U(1)$ applied to the Lagrangian

$$\mathscr{L} = (\mathcal{D}_\mu \Phi)^\dagger (\mathcal{D}^\mu \Phi) - \frac{1}{4} F^{\mu\nu} F_{\mu\nu} - V(\Phi) \tag{2.25}$$

where $\Phi = \frac{1}{\sqrt{2}}(\Phi_1 + i\Phi_2)$ is a complex scalar and $V(\Phi) = \mu^2 \Phi^\dagger \Phi + \lambda(\Phi^\dagger \Phi)^2$ with $\lambda > 0$ is a potential term of the Ginzburg-Landau form. A massless gauge field A_μ has also been implicitly introduced in Equation 2.25 so as to keep the Lagrangian invariant under the local transformation $\Phi \to e^{-i\alpha(x)}\Phi$ with $\mathcal{D}_\mu \equiv \partial_\mu + igA_\mu$ and $A_\mu \to A_\mu + \frac{1}{g}(\partial_\mu \alpha(x))$. In order to find the vacuum of the system, the potential must be minimized with respect to the components of Φ. It can be shown that the result depends on the sign of μ^2, in direct analogy to the temperature of a ferromagnet being above or below T_C. Both cases are illustrated in Figure 2.1.



(a) $\mu^2 > 0$, Symmetric phase

(b) $\mu^2 < 0$, Non-symmetric phase

Figure 2.1: In the symmetric phase (left), there is a unique vacuum at $< \Phi >= 0$ and it is $U(1)$ invariant. In the non-symmetric phase (right), there exist an infinite number of degenerate vacuum states that share the same $|< \Phi >|$ but which are all realized by the selection of a different complex phase. An arbitrary choice of the argument is what breaks the $U(1)$ symmetry.

For $\mu^2 < 0$, the degenerate vacua share in the value of $|< \Phi >| = \sqrt{\frac{-\mu^2}{2\lambda}} \equiv \frac{\nu}{\sqrt{2}}$ where ν represents a vacuum expectation value (vev), which can be interpreted as the order parameter of the SSB. The system as described in its non-symmetric phase matches the criteria for SSB since while the Lagrangian in Equation 2.25 is $U(1)$ gauge invariant, a particular vacuum is not.

There is a second motivation in finding the vacuum of a system, other than to look for SSB. For this it is convenient to work with the following uniquely real vacuum configuration:

$$arg< \Phi > = 0 \quad \Rightarrow \quad < \Phi_1 >= \sqrt{\frac{-\mu^2}{\lambda}} = \nu, \quad < \Phi_2 >= 0 \tag{2.26}$$

In order to work using a perturbative approach, it is important to start from the vacuum and interpret the fields as oscillations around it. Therefore, it is beneficial to define new field variables

$$\eta \equiv \Phi_1 - \nu, \quad \xi \equiv \Phi_2 \quad \text{s.t.} \quad <\eta>= 0; \quad <\xi>= 0 \tag{2.27}$$

and recast Equation 2.25 in terms of them. Unfortunately, when these steps are carried out, a non-physical interaction term $\sim g\nu A^\mu \partial_\mu \xi$ is inadvertently introduced in the Lagrangian. Read as an interaction, this would allow a ξ particle to transform into an A^μ particle. Such a term often suggests that the particles involved are not the fundamental particles of the theory. To find the physical states, this interaction term must be removed, which can be carried out by choosing the 'polar' coordinates

$$\Phi(x) = \frac{1}{\sqrt{2}}(\nu + \eta(x))e^{i\frac{\xi(x)}{\nu}} \tag{2.28}$$

so that the fields now describe small oscillations about the real vacuum. By fixing the gauge parameter to be $\alpha(x) = \frac{\xi(x)}{\nu}$, the ξ field can be transformed away altogether

$$\Phi(x) \to e^{-i\frac{\xi(x)}{\nu}}\Phi(x) = \frac{1}{\sqrt{2}}(\nu + \eta(x)) \tag{2.29}$$

$$A_\mu(x) \to A_\mu(x) + \frac{1}{g\nu}\partial_\mu \xi(x) \equiv B_\mu(x) \tag{2.30}$$

so that the Lagrangian of Equation 2.25 is rewritten in terms of the fields η and B_μ:

$$\begin{aligned}\mathscr{L} = &\tfrac{1}{2}(\partial_\mu \eta)^2 + \mu^2\eta^2 - \tfrac{1}{4}B_{\mu\nu}B^{\mu\nu} + \tfrac{1}{2}(g\nu)^2 B_\mu B^\mu \\ &+\tfrac{1}{2}g^2 B_\mu B^\mu \eta(2\nu + \eta) - \lambda\nu\eta^3 - \tfrac{1}{4}\lambda\eta^4\end{aligned} \tag{2.31}$$

In this form, it is clear that the Lagrangian does not respect $U(1)$ gauge symmetry in η. However, both Equation 2.25 and Equation 2.31 describe the same physical system - the difference being that in Equation 2.31, the physical content of the theory is directly manifest in a way that allows for perturbative analysis. The tradeoff is that by giving special treatment to a particular vacuum, the true $U(1)$ symmetry of the system is hidden.

In the process of moving from Equation 2.25 to Equation 2.31, a number of important results are demonstrated. First, through the identification with their counterparts in each corresponding free Lagrangian, terms that are second order in a field should be interpreted as giving mass to the particle that is generated by that field. Therefore, Equation 2.31 describes a massive scalar particle η with spin 0 and mass $m_\eta = \sqrt{2}|\mu|$, as well as a massive gauge boson particle B_μ with spin 1 and mass $m_{B_\mu} = g\nu$, in a way that preserves the gauge symmetry of the system. The intermediate scalar particle ξ, on the other hand, never acquires a mass term. In fact, the appearance of such a particle is ubiquitous among QFTs that exhibit SSB. It can be shown (Goldstone's Theorem [80]) that every spontaneously broken continuous symmetry of the Lagrangian directly entails one massless spin 0 boson, referred to as a Goldstone boson. Intuitively, this is related to the fact that there exists a flat direction in the potential.

Any fundamental theory constrained by current experimental evidence cannot contain, in its physical spectrum, massless scalar particles. When SSB takes place in a theory that is locally gauge invariant, the otherwise Goldstone bosons mix with an equal number of massless gauge bosons so that when the theory is built up from the vacuum, these gauge bosons become massive. This process is what is known as the *Higgs mechanism* [66, 83, 88, 89] and is extended to the case of the SM below.

Applied to the Standard Model

In theory, the Higgs mechanism can be employed to account for any number of massive gauge bosons as long as the scalar field Φ and its potential $V(\Phi)$ are sufficiently defined. In particular, the three massive electroweak gauge bosons W^+, W^-, and Z realized in nature can be accounted for if Φ and $V(\Phi)$ obey a few specific properties. First, there must be at least three degrees of freedom in Φ that become the longitudinal polarizations of the gauge bosons. Second, in order for the photon γ to remain massless, Φ must have electroweak quantum numbers such that it is gauge invariant only under the $U(1)_{em}$ subgroup. Finally, the component of Φ that acquires a vev can hold no electric charge if $U(1)_{em}$ invariance is to be maintained. While the combinations of Φ and $V(\Phi)$ that satisfy these constraints are still many, the SM utilizes the simplest among them.

In the SM, a single complex scalar $SU(2)$ doublet is introduced

$$\Phi = \begin{pmatrix} \Phi^+ \\ \Phi^0 \end{pmatrix} \tag{2.32}$$

with the electroweak quantum numbers $T = \frac{1}{2}$ and $Y = 1$, while $V(\Phi)$ is again a form of the Ginzburg-Landau potential:

$$V(\Phi) = \mu^2 \Phi^\dagger \Phi + \lambda (\Phi^\dagger \Phi)^2, \quad \lambda > 0 \tag{2.33}$$

Once these have been defined, the Higgs mechanism can proceed with infinite degenerate minima identified at

$$|< 0|\Phi|0 >| = \begin{pmatrix} 0 \\ \frac{\nu}{\sqrt{2}} \end{pmatrix}; \quad \nu \equiv \sqrt{\frac{-\mu^2}{\lambda}}. \tag{2.34}$$

With Φ being complex, it contains a total of 4 scalar fields. When the theory is built up from a particular vacuum state and the electroweak gauge bosons are rotated from the interaction eigenstates to the mass eigenstates, three of the degrees of freedom are given to the W^+, W^-, and the Z (making them massive). The fourth, having acquired a vev, becomes massive. This is what is commonly referred to as the SM Higgs boson.

By enforcing the Lagrangian to obey $SU(2)_L$ invariance, the fermions must also obtain a mass through the spontaneous breaking of electroweak symmetry. To accomplish this task, each fermion meant to acquire a mass is allowed to interact with Φ via Yukawa coupling from an additional term in the Lagrangian of the form

$$\mathscr{L}_{int} = \lambda_f \overline{f}_L \Phi f_R \tag{2.35}$$

where f stands for a particular quark or charged lepton and λ_f is its associated Yukawa coupling to Φ. During the symmetry breaking process, when new scalar fields are defined as oscillations about the vacuum, this term bifurcates; the result is a term that describes a Yukawa coupling of the fermion to the physical Higgs, while a second term describes an 'interaction' between the fermion and the vev of the Higgs field of the form

$$\mathscr{L}_{YW} = -\left(\lambda_f \frac{\nu}{\sqrt{2}}\right) \overline{f}_L f_R. \tag{2.36}$$

Being associated with the second term in the Dirac Lagrangian, $\mathscr{L}_{YW}$ gives mass to the associated fermion $m_f = \lambda_f \frac{\nu}{\sqrt{2}}$ which leads to the interpretation that a fermion's mass arises through how strongly it interacts with the Higgs vev. Incidentally, the coupling of a particle with the physical Higgs is also proportional to its mass, which provides a direct prediction that can be tested experimentally. However, with λ_f being an input parameter, the SM does not make any prediction as to the numerical value of m_f.

2.2 Hadronic Collision Phenomenology and Event Generation

While the formalism introduced in section 2.1 does well in describing particle interactions at a fundamental level, it doesn't directly contain the tools needed in order to make accurate predictions that can be tested experimentally. Particularly in the case of proton-proton collisions (which are used in this thesis), there exists a rich associated phenomenology. Figure 2.2 demonstrates the complex nature of the resulting debris for such an event. This section is dedicated to the considerations that must be made when performing theoretical calculations (specifically cross sections) and generating simulated events that are used to compare to experimental proton-proton collision data and is based on [84, 49, 40].

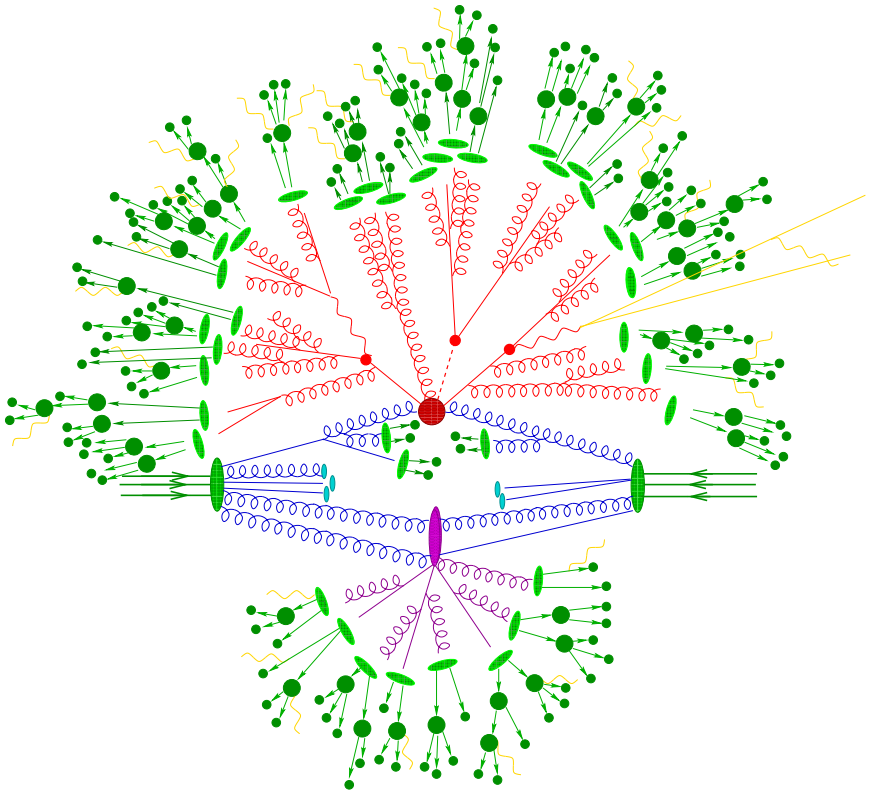


Figure 2.2: Color coded sketch of a proton-proton collision as simulated by a Monte Carlo event generator. The red indicates both the primary hard-scatter process as calculated via matrix element (exact fixed-order in perturbative QCD) and subsequent parton shower (approximate all-order in perturbative QCD). The purple indicates a secondary hard-scatter event representative of multiple parton interactions (MPI), one component of the underlying event. The light green blobs indicate the parton hadronization, while the dark green blobs indicate subsequent decay of the hadrons. Yellow lines also illustrate the radiation of soft photons. [90]

2.2.1 The Factorization Theorem

Protons are composite objects, being made up of quarks and gluons (collectively referred to as *partons*). Due to the governing low energy dynamics, their internal structure cannot be described in a perturbative fashion. However, the high energy interactions that are often

of interest are characterised by the presence of a hard scale Q in which perturbative QCD is still valid. For what follows the assumption is also made that when two protons collide, only a single pair of partons participate in the process of interest while the rest simply serve as *spectators*. Therefore, the cross section for a given process in a proton-proton collision can then be expressed as

$$\sigma_{pp\to X}(s) = \sum_{i,j} \int_0^1 \underbrace{dx_1 dx_2}_{\substack{\text{longitudinal} \\ \text{momentum} \\ \text{fractions}}} \underbrace{f_i(x_1, \mu_F^2) f_j(x_2, \mu_F^2)}_{\substack{\text{parton} \\ \text{distribution} \\ \text{functions}}} \underbrace{\hat{\sigma}_{ij\to X}(x_1 x_2 s, \alpha_S(\mu_R^2), \mu_F^2, \mu_R^2)}_{\substack{\text{partonic} \\ \text{cross section}}} \tag{2.37}$$

where i and j sum over the contributing parton types and the strong coupling constant is defined as $\alpha_S = g_s^2/4\pi$. This result is known as *the factorization theorem* and simplifies the task of computing $\sigma_{pp\to X}(s)$ down to having independent knowledge of the parton distribution functions and partonic cross sections, both of which are described below.

Parton Distribution Functions

Parton distribution functions (PDFs) provide the probability of finding a parton p inside the proton, carrying a fraction x of the total longitudinal momentum when probed with energy Q. In this way, they serve as a parametrization of our ignorance of what happens inside the proton below a given scale μ_F. Also called the factorization scale, the choice of μ_F effects both the PDF and partonic cross section result - although including higher orders reduces this effect, while carrying out all perturbative orders would remove it entirely.

Since PDFs aim to describe non-perturbative effects, they cannot be calculated analytically but rather must be derived from data. This is accomplished by performing a global fit to multiple experimental measurements, at some reference value of momentum transfer Q_0^2. The PDF for a different value Q^2 can then be obtained by using the DGLAP evolution equations [38], typically evaluated at next-to-next-to-leading-order. The resulting collection of PDFs is often referred to as a *PDF set*. An example of proton PDFs can be seen in Figure 2.3, for two separate momentum transfers Q^2. A convenient consequence of the factorization theorem is that it guarantees universality in the sense that PDFs extracted from one process are also valid for another.

Partonic Cross Sections

In the high energy (short distance) regime, the partonic cross section $\hat{\sigma}_{ij\to X}$ can be calculated as a perturbative series in terms of the strong coupling constant:

$$\hat{\sigma} = \alpha_S^k \left(\hat{\sigma}^{(0)} + \frac{\alpha_S}{\pi} \hat{\sigma}^{(1)} + \left(\frac{\alpha_S}{\pi} \right)^2 \hat{\sigma}^{(2)} + ... \right) \tag{2.38}$$

Every term in the expansion can be viewed as a collection of diagrams (often referred to as *Feynman diagrams*) which all contribute to the overall process and contain the same number of internal vertices. Through the Feynman rules (which are derivable from the underlying Lagrangian of the theory), it is possible to translate each diagram into a matrix element $\mathcal{M}$ that is then used to determine the corresponding cross section. This last step is typically performed by computer simulations using Monte Carlo techniques and are often referred to as event generators. A survey of different event generators that are in use is provided in [55].

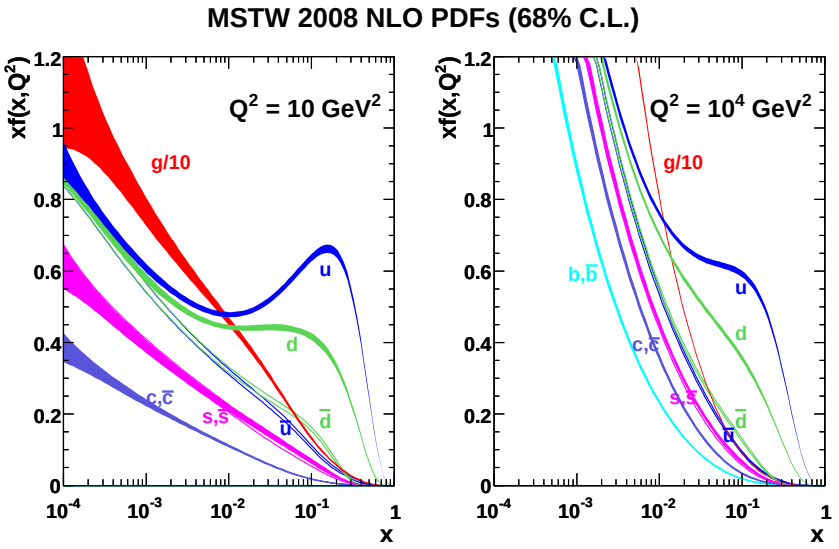


Figure 2.3: Proton PDFs for two separate momentum transfers of $Q^2 = 10$ GeV^2 (left) and $Q^2 = 10^4$ GeV^2 (right) as published by the PDF fitting group MSTW. Two $u-$ and one $d-$quark, often called the *valence* quarks, can be seen with larger probabilities for higher values of x. [97]

In practice, theory calculations only contain up to a fixed-order in the expansion since each following term contains both new loop contributions and real QCD emission which become progressively more time-consuming to compute. The first non-zero term in the series is considered leading-order (LO) and provides a first estimate of the cross section. The subsequent terms are denoted as next-to-leading-order (NLO), next-to-next-to-leading-order (NNLO), etc. and apply additional corrections.

Another consideration to make is that there also exist divergences which originate from both virtual loop diagrams and soft or collinear emissions. These are treated through a process known as *renormalization* in which the singularities are absorbed by reparameterizing the theory. Consequently, an energy-scale dependence μ_R is introduced into the partonic cross section result. Similarly to μ_F, it is reduced by incorporating more terms in the expansion. For this reason, both μ_F and μ_R are often used when estimating uncertainties due to a fixed-order calculation.

2.2.2 Parton Showers and Hadronization

Fixed-order perturbative calculations are reliable for hard and well-separated partons, but break down when considering many soft/collinear partons. However, the quarks and gluons produced in a high energy collision will continue radiating other QCD particles until their energy again reaches a non-perturbative scale. Therefore instead of explicitly calculating this process in exact terms, it is necessary to model these emissions (known as *parton showers*) in a procedural way through $1 \rightarrow 2$ branching and in doing so approximate *all* terms in the expansion. Event generators will typically also handle parton showers, although there are a few different ways in which they are implemented - for example the emissions can be either virtuality ordered, angular ordered, or handled through color dipoles. When combining the full chain of interactions, special attention must be taken so as not to double count particles. This is taken care of through dedicated merging and matching schemes (again implemented

differently across generators), often blurring the line between particles created as a result of matrix element calculation and parton shower.

At a scale near $\Lambda_{QCD} \sim 200$ MeV, perturbation theory breaks down and it is at this point that the low energy QCD partons are confined to color neutral (physically observable) composites through a process known as *hadronization*. There are two distinct hadronization models that are in use by event generators:

- **Lund string model** - based on the assumption that QCD is "Coulomb-like" at small distances ($1/r$ color field potential), while linear at longer distance through gluon self-attraction. An intense color field induces $q\bar{q}$ pair creation, leading to hadronization.
- **Cluster model** - based on the idea of pre-confinement in which past scales of $\sim$ few x Λ_{QCD} there is only a *local* redistribution of color, flavor and momentum flows leading to the construction of low-mass, color singlet clusters.

Illustrations for these models are shown in Figure 2.4. The choice of hadronization model can be important for some detector effects including jet response, heavy-flavor tagging, and lepton isolation - all of which are described in more detail in section 3.5.

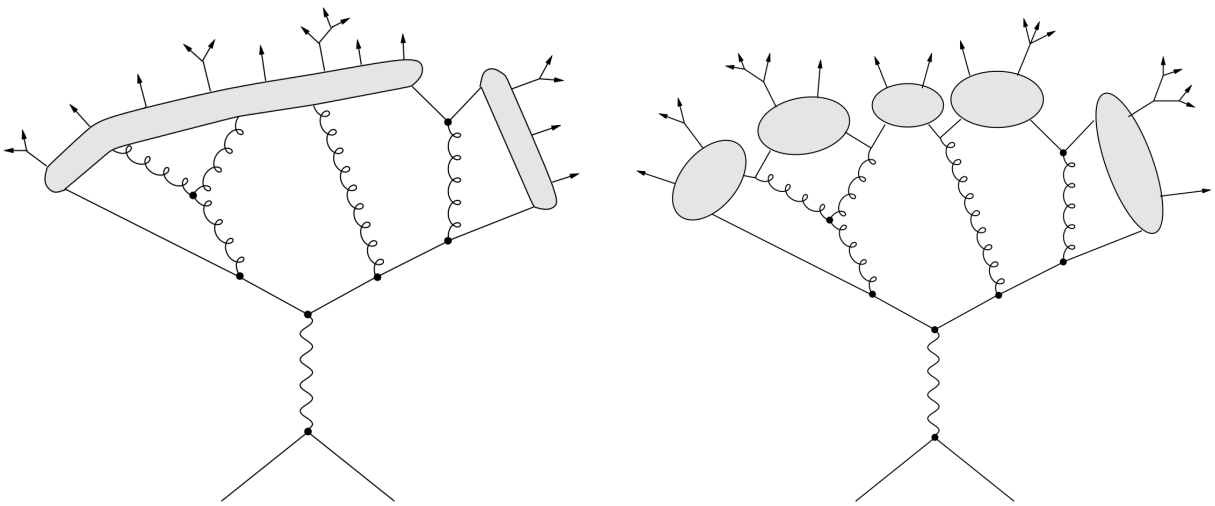


Figure 2.4: Illustrations of the two most popular hadronization models - the Lund string model (left) and the Cluster model (right). [91]

2.2.3 The Underlying Event and MC Tunes

Depending on the density profile of the proton, additional interactions inside the same collision known as *multiple parton interactions* (MPIs) can occur. These are typically assumed to be QCD $2 \to 2$ processes and can be important for modeling of the color flow and reconnection. Together with the fragmentation of beam remnants, these secondary effects are collectively referred to as *the underlying event*.

The underlying event, hadronization, and pileup simulation (described further in subsection 3.1.4) are all examples of semi-empirical models and as such require data to constrain a number of free parameters through dedicated studies whos results are known as Monte Carlo *tunes*. Reliable tunes can be essential for both precision measurements and discoveries.

2.3 Higgs Physics at the LHC

Due to its heavy mass and small production cross section, the Higgs boson had eluded observation until a powerful enough accelerator could be constructed. The Large Hadron Collider (LHC), with its ability to collide protons with a design center-of-mass energy of up to $\sqrt{s} = 14$ TeV, finally provided the necessary discovery potential. Given so much event activity, achieving sufficient sensitivity still however remained a challenge. In proton-proton collisions, the Higgs can be produced through a variety of different processes described in subsection 2.3.1. Once created, it very quickly decays before any hope of direct detection through one of the modes introduced in subsection 2.3.2.

2.3.1 Higgs Production

The leading Feynman diagrams illustrating the different means of Standard Model Higgs boson production at the LHC can be seen in Figure 2.5. The relative sizes of their associated cross sections for different center-of-mass energies are shown in Figure 2.6. The leading production mode is gluon fusion (ggF), occuring about an order of magnitude more often than other processes. It is accomplished via a virtual fermion loop for which the largest contributor is the top quark. The second largest production mode is vector boson fusion, which almost always contains the distinctive topology of two energetic forward jets that can be targeted for event selection. Both the ggF and VBF production cross sections are considered in the results described in this thesis. Higgs strahlung, or the production of Higgs bosons in association with a W or Z vector boson (VH), is less likely than VBF - but the presence of an additional vector boson in the final state provides an additional handle for filtering events. Finally, associated production with top (ttH) can occur - although in addition to its relatively low cross section, it can be challenging to study due the final state signature being less accessible.

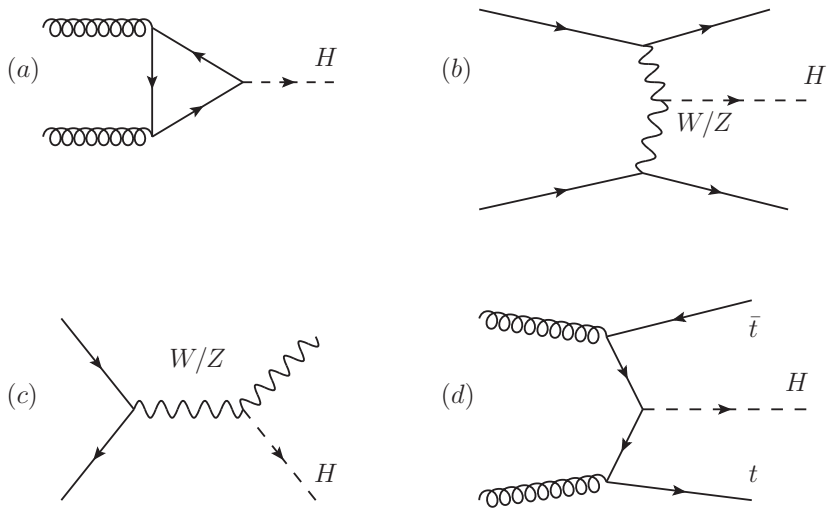


Figure 2.5: Leading Feynman diagrams for different production modes of the Higgs at the LHC in order of largest to smallest cross section. Gluon fusion (ggF) is shown in (a), vector boson fusion (VBF) is shown in (b), Higgs strahlung production (WH / ZH) is shown in (c), while associated production with top (ttH) is shown in (d). [76]

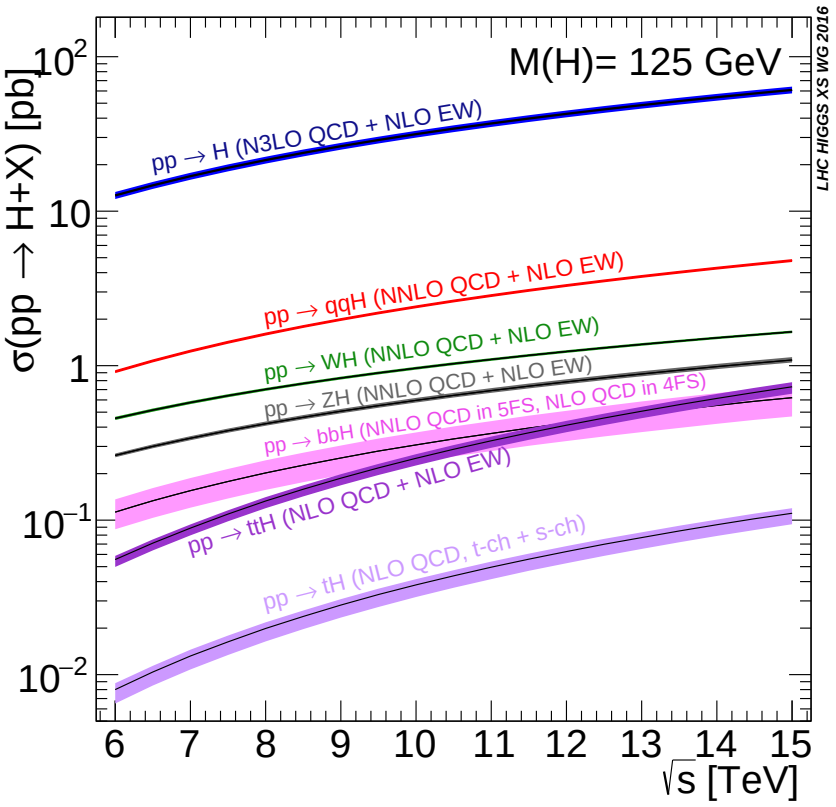


Figure 2.6: Standard Model Higgs production cross sections as determined by theory for different production modes at the LHC. The blue line denotes ggF, while the red line denotes VBF. [36]

2.3.2 Higgs Decay

The branching fractions of the Standard Model Higgs boson with a mass of 125 GeV are provided in Figure 2.7. The majority of Higgs bosons produced at the LHC will decay to a $b\bar{b}$ pair (occuring about 58% of the time). However, it is relatively difficult to study the Higgs using this channel due to the fact that the decay products are purely hadronic (providing little distinguishing features that separate themselves from the rest of the activity that can be found in a hadronic collision). Still, measurements are being performed by utilizing the signatures of the sub-dominant production modes - for example triggering on leptons originating from associated vector bosons [23, 15]. The second most common decay mode is $H \to WW$ (occuring about 21% of the time), which is the focus of the results presented in this thesis. In the case where the W bosons decay leptonically, this provides a means with which to select interesting events. Unfortunately a number of other SM processes can also give rise to a two lepton final state, leading to a host of backgrounds to consider. A similar situation exists for decay to τ-leptons [20, 14]. While there is a comparable branching fraction to gluons, this signature presents no features with which to distinguish it from backgrounds in a hadronic collision. Decay to $c\bar{c}$ can be treated similarly as for $b\bar{b}$, except that c-jets cannot be tagged as efficiently as b-jets. Decay to ZZ and $\gamma\gamma$ both profit from a clean background spectrum and therefore were the first channels in which the Higgs was observed. Finally, the two rarest decay modes $Z\gamma$ and $\mu\mu$ stand the most to gain from the increase in statistics that come with more integrated luminosity, although efforts in these channels are already underway [22, 10, 19, 11].

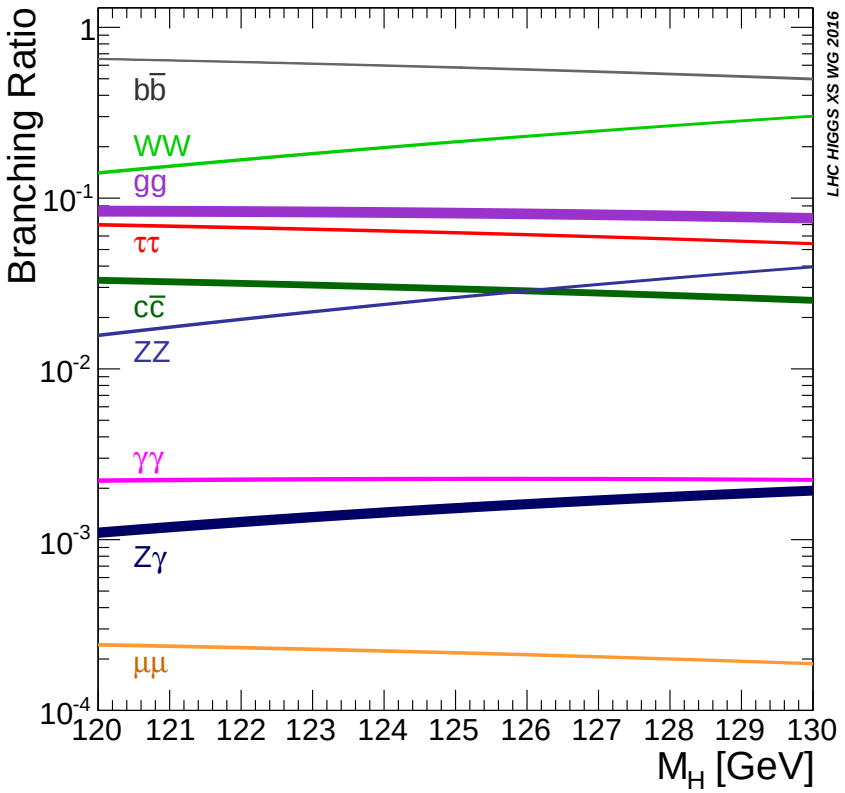


Figure 2.7: Standard Model branching fractions for different Higgs decay modes. [36]

Chapter 3

LHC and the ATLAS Experiment

The dataset used for the results presented in this thesis was collected with the ATLAS detector, one of the main experiments at the Large Hadron Collider (LHC) which is located at the European Organization for Nuclear Research (CERN) near Geneva, Switzerland. This chapter first introduces the experimental apparatuses in section 3.1 and section 3.2, deferring more detailed specifications to the relevant technical design reports [52, 53, 54, 3, 4]. Some aspects of the data flow from detector to tape storage are then provided in section 3.3, while the simulation of Monte Carlo generated events propagating through the detector and the subsequent detector response is discussed in section 3.4. Finally, section 3.5 describes the offline reconstruction of the physics objects used for the analysis presented in this thesis.

3.1 The Large Hadron Collider

The LHC is currently the world's largest and most powerful particle accelerator, capable of producing proton-proton collisions at an unprecedented center-of-mass energy of $\sqrt{s} = 13$ TeV[1]. It is housed inside a circular tunnel that is 26.7 km in circumference and an average of 100 m under the surface of the CERN facility, located on the border between France and Switzerland. The LHC is a synchrotron, keeping the protons at all times in counter-rotating beams with fixed well-defined orbits as it first accelerates them to their peak energy and then brings them into head-on collisions at the center of four detectors located around the ring: ATLAS, CMS, ALICE, and LHCb. Figure 3.1 illustrates the scale of the machine, spanning the distance between the Jura mountains to the west and Lake Geneva to the east.

3.1.1 LHC Magnet Configuration

In order for the protons to bend as they circle around the ring, they are guided by strong magnetic fields which are generated by 1232 separate 15 m long superconducting dipole magnets made out of cables composed of Niobium-titanium (NbTi) filaments - each serving to deflect the beams by 0.29 degrees as they pass through. During nominal operations, these dipole magnets are cooled to 1.9 K (a process taking weeks to complete) by surrounding them in superfluid helium in order to benefit from their resulting superconductive properties.

[1]The design parameter for the center-of-mass energy of the collisions is $\sqrt{s} = 14$ TeV. However, after an incident involving a faulty electrical connection between two of the accelerator's dipole magnets occured in the year 2008, the decision was made to temporarily operate at lower energies: $\sqrt{s} = 7/8$ TeV during Run 1 and $\sqrt{s} = 13$ TeV during Run 2.

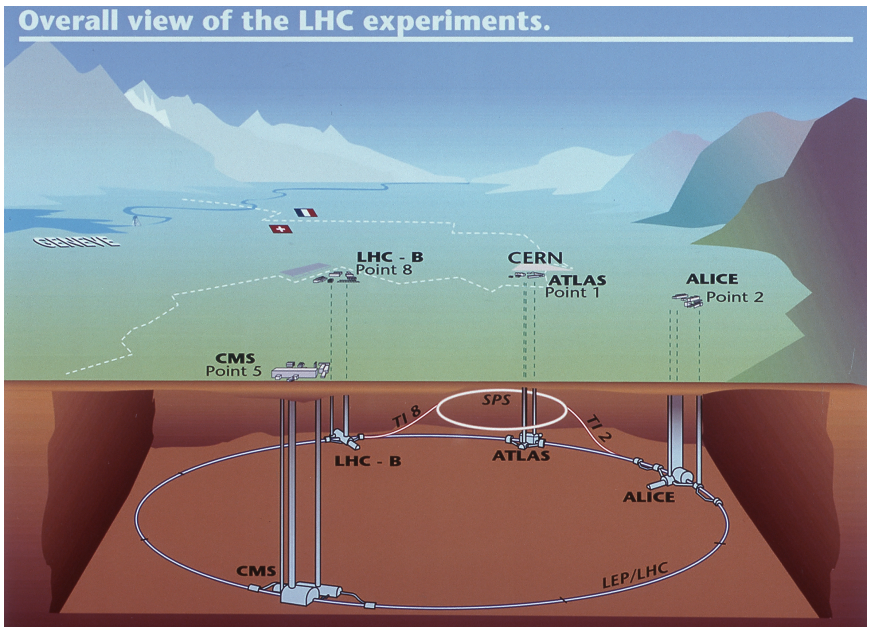


Figure 3.1: Display of the LHC machine and its four main experiments, along with their geographical locations. The SPS, which serves as the final injector in the pre-acceleration chain, is also in view to the south. [111]

Under peak performance, they are able to generate a current of over 11,000 amperes, corresponding to an induced magnetic field of 8.33 T.

In a homogeneous magnetic field, protons with varying initial conditions will fulfil oscillations around the design orbit in the transverse direction known as *betatron oscillations*. Quadrupole magnets, which offer a restorative force proportional to the distance from the design orbit, are therefore used in order to keep the protons constrained in the transverse directions. However, because such field configurations simultaneously cause focusing and defocusing in orthogonal directions, the LHC's magnets are arranged in a lattice of so-called 'FODO cells' - sequences which first focus in one direction and then defocus in the same direction, separated by non-focusing drift spaces. There are 8 primary points of interest spaced out around the ring, with arcs in between consisting of 23 FODO cells each. These sections can be seen in Figure 3.2.

Additional corrector magnets are also installed for making further precise adjustments. For example, the dipole magnets will generate dispersion due to the momentum spread of the protons with each one being bent to a slightly different degree. Known as 'chromaticity', this effect can be controlled through the use of sextupole magnets.

3.1.2 RF Cavities and Beam Parameters

The protons are accelerated up to collision energies by applying an alternating longitudinal electric field in dedicated radio frequency (RF) cavities, while the field strength of the dipole magnets is correspondingly increased so as to maintain a fixed orbital distance. On each turn, the phase of any single proton with respect to the RF waveform changes since in general its orbit will deviate slightly from the ideal frequency, giving rise to longitudinal oscillations known as *synchrotron oscillations*. With only one RF cavity section in the ring, synchrotron oscillations are much slower than betatron oscillations, often taking hundreds of turns to complete a full cycle. Partly due to the fluctuating nature of the electric field,

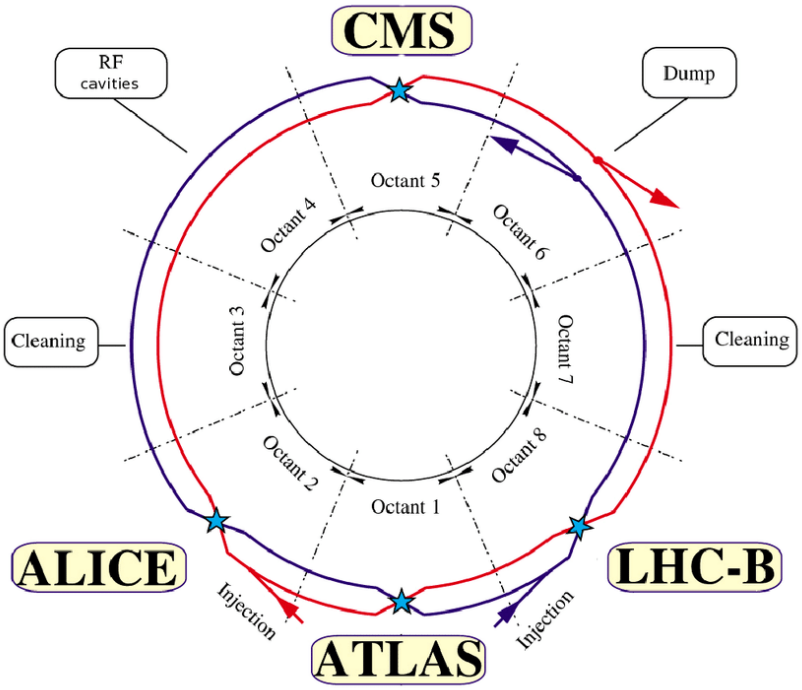


Figure 3.2: Sketch of the octagonal partitioning of the LHC. Aside from the sections that are dedicated to the four detector interaction points, there are two related to beam cleaning, one containing the RF cavity, and one for handling the beam dump. [67]

the counter-rotating beams are not composed of single streams of protons. Rather, they are made up of distinct *bunches*, each containing on the order of 100 billion protons. These bunches are constrained by the wavelength of the electrical pulse provided from the RF cavities to reside in stable regions of phase space called *buckets*.

Assuming that the distribution of protons in a single bunch follows a gaussian shape, the beam width can be parameterized by

$$\sigma_x = \sqrt{\epsilon\beta_x(s)}; \quad \sigma_y = \sqrt{\epsilon\beta_y(s)} \tag{3.1}$$

which are the one sigma intervals in the transverse plane where ϵ is the beam emittance and β is a periodic function that depends upon the location s along the accelerator. The emittance is an intrinsic beam property that is defined at its creation. It is proportional to the area of the phase space ellipse which contains inside of it all (or a defined percentage of) the protons in the bunch and can't be changed by focusing, although does decrease upon acceleration of the beams. The β function, on the other hand, is determined by the focusing properties of the quadrupole lattice and provides the envelope within which all protons oscillate. When the bunches are first injected into the LHC, the maximum beam width is $\sigma_{450\text{GeV}} = 1.1$ mm (still well within the distances of 19 mm and 23 mm to the beam pipe in the vertical and horizontal directions, respectively) with $\beta_{\text{max}} = 180$ m, while at collision energies it is closer to $\sigma_{7\text{TeV}} = 300$ μm due to the reduction in emittance.

3.1.3 LHC Operations

A series of progressively larger accelerators boost the protons before they are injected into the LHC. An overview of the accelerator complex which contains each of the stages is shown

in Figure 3.3. Taken from hydrogen atoms and having their electrons stripped, the protons begin their journey by being sent through the LINAC2 linear accelerator which accelerates them to 50 MeV over a length of 33 m. They are then injected into the first synchrotron known as the PS booster. Here the LINAC2 pulse is distributed over four stacked rings, each injecting over multiple turns to accumulate beam in the horizontal phase space - it is at this stage that the transverse brightness of the LHC beam is set. The PS booster also increases the proton energy up to 1.4 GeV. Afterwards, they are sent into the Proton Synchrotron (PS) where their energy is increased to 26 GeV. Here the longitudinal beam characteristics are defined, through a number of different bunch splitting techniques. Finally, the protons are sent into the Super Proton Synchrotron (SPS) where their energy is brought up to 450 GeV before being injected into the LHC.

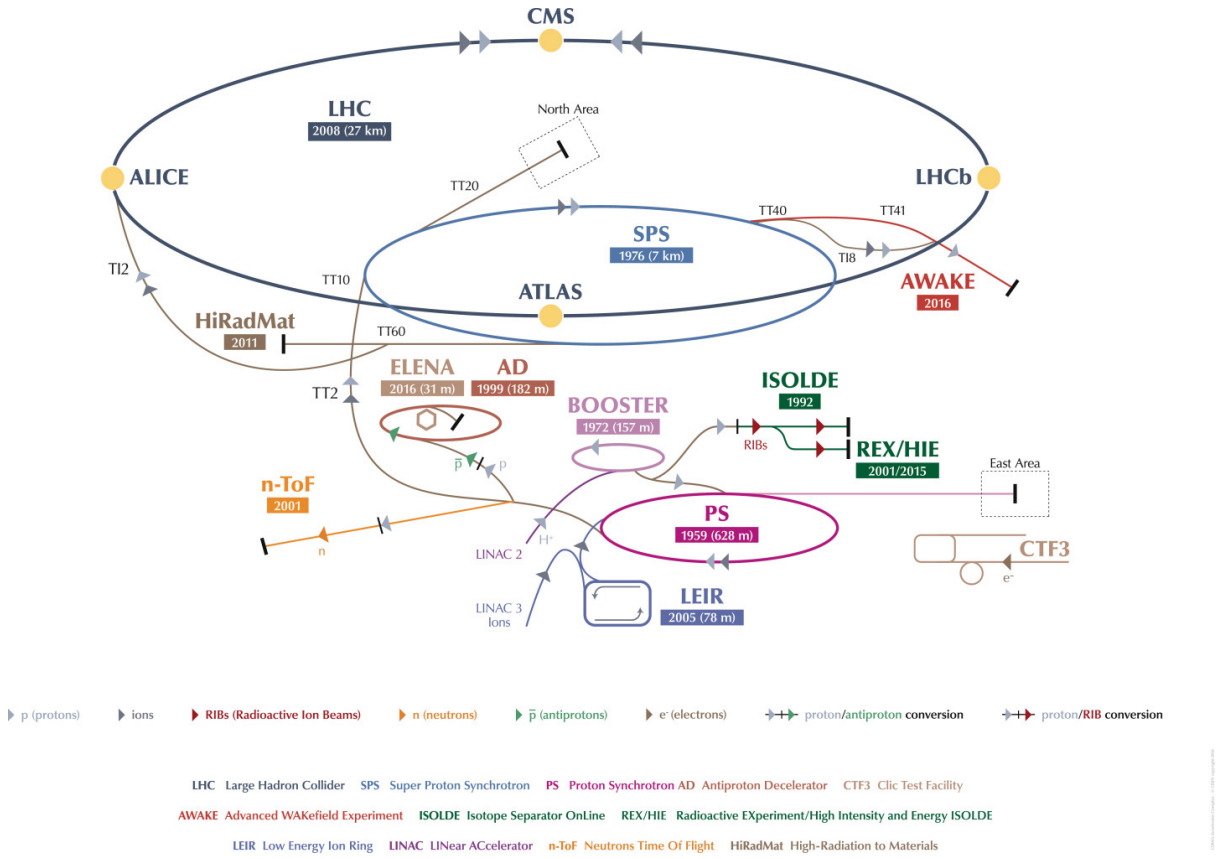


Figure 3.3: The CERN accelerator complex, illustrating the interconnections between accelerators along with many of the main experiments. [99]

Operation of the LHC takes place in fill cycles which ideally last for 10 or more hours. The protons are first injected into the ring in 2808 bunches per beam under normal conditions and the beam energies are then ramped up to 6.5 TeV in a process that collectively takes approximately 45 minutes. Once the protons reach their maximum energy, each beam will contain approximately 360 MJ of energy. Next, the proton beams are squeezed down to transverse widths of 16 μm at the center of the collision points. This is made possible by placing the detectors at the center of so-called 'mini-beta insertions' which are special symmetric drift spaces in the regular magnet lattice with an exceptionally small waist in the β function at the middle. On either end of the mini-beta insertion exist triplet quadrupole magnets which are responsible for the squeeze. It is here that due to a much higher required value in the β function, the beams reach their largest widths of 1.5 mm.

After a brief adjustment period, the counter-rotating beams are allowed to collide at the interaction points around the ring. For typical filling schemes the bunches are spaced by 25 ns, translating to 40 million bunch crossings per second. Under these stable beam conditions, collision data is taken until the beams deteriorate to the point that it becomes more efficient to dump them and refill.

3.1.4 Event Rate and Luminosity

Typically, physics experiments are interested in maximizing the rate of events that can be provided by a collider. In general, the event rate $\frac{dN}{dt}$ for a given process p can be written as

$$\frac{dN}{dt} = \mathcal{L} \times \sigma_p \tag{3.2}$$

or that it is proportional to the process cross section σ_p, where the constant of proportionality is the instantaneous luminosity $\mathcal{L}$, a measure of the number of collisions produced per second by the accelerator with units of $\text{cm}^{-2}\text{s}^{-1}$. The instantaneous luminosity that a collider is able to provide is one of the most important metrics used to characterize its performance and can be determine using

$$\mathcal{L} = \frac{f N_b N_1 N_2}{4\pi\beta^* \epsilon} \cdot F \tag{3.3}$$

where f is the revolution frequency, N_b is the number of bunches, N_1 and N_2 are the number of protons per bunch, β^* is the beta function at the collision point, ϵ is the beam emittance and F is a luminosity reduction factor which accounts for the geometric crossing angle of the beams [87]. By optimizing for more protons to collide in a smaller transverse cross section, the instantaneous luminosity can be increased. For instance, the LHC was able to exceed its design luminosity of $10^{34}\text{cm}^{-2}\text{s}^{-1}$ by as much as 40% in 2016 in part by operating with a smaller β^*.

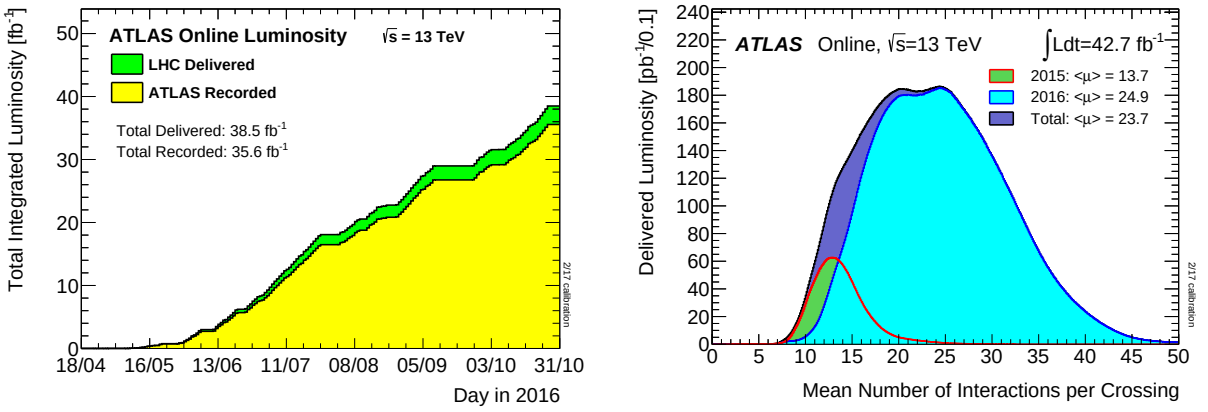
While the instantaneous luminosity is important, the ultimate figure of merit is the so-called integrated luminosity:

$$L = \int_0^T \mathcal{L}(t)dt \tag{3.4}$$

which is a measure of the total amount of data collected and is often expressed in units of 'barns' (where 1 barn = 10^{-24}cm^{-2}). Figure 3.4a shows the total luminosity collected in 2016 at $\sqrt{s} = 13$ TeV which provides the majority of the data analyzed for the results presented in this thesis.

The protons must be packed closely together at the interaction point in order to maximize the event rate. One side effect of this is that multiple pp collisions will typically occur even in a single bunch crossing, generating additional uncorrelated energy flow through the detector that is generally referred to as 'in-time' *pileup*[2]. The average number of interactions per bunch crossing (also denoted as μ) is typically used as a metric for pileup and scales linearly with the instantaneous luminosity. Figure 3.4b shows μ values during 2015 and 2016 data taking with an average of about 25, while in some cases reaching as high as 50. The modeling of pileup when simulating an event is very important in order to match the data and is described further in subsection 3.4.2.

[2]Since the spacing between bunch crossings is relatively small compared to the sensitive time window for some of the detector elements (in particular for calorimeter cells), pileup from previous and following bunch crossings can also contribute in which case it is referred to as 'out-of-time' pileup.



(a) (b)

Figure 3.4: (a) The total integrated luminosity delivered by the LHC during 2016 data taking in green and the amount recorded by the ATLAS detector in yellow. (b) The average number of interactions per bunch crossing during 2015 and 2016 data taking. [1]

3.2 The ATLAS Detector

The ATLAS (**A** **T**oroidal **L**HC **A**pparatu**s**) detector is a general purpose particle detector that is located 100 m underground at one of the four interaction points of the LHC. A cut-away view can be seen in Figure 3.5, also providing a sense of scale. It is currently the largest particle detector ever built with a length of 44 m, a diameter of 25 m, and a weight of 7000 tons. ATLAS is operated by a multi-national collaboration of around 3000 scientists and engineers, representing about 180 institutes and 38 countries throughout the world.

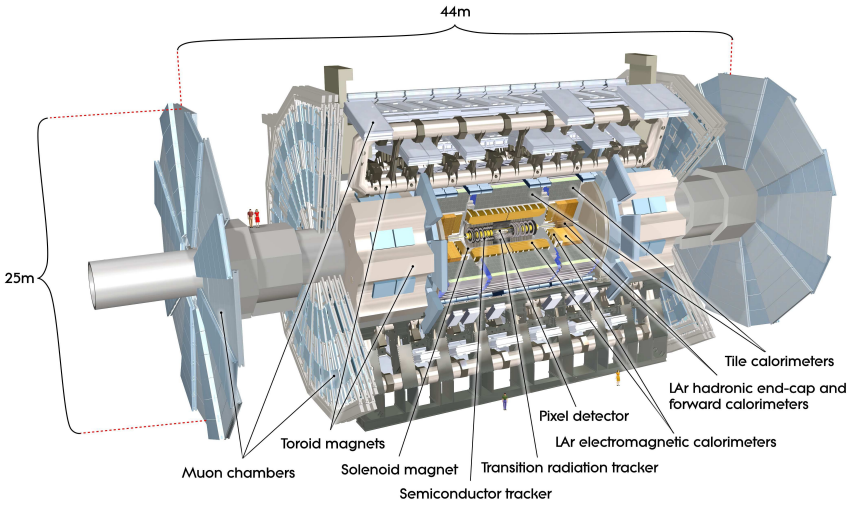


Figure 3.5: A cut-away view of the ATLAS detector. The largest components are shown, along with the overall dimensions and humans for comparison. [5]

The design considerations of ATLAS reflect the goals of its physics program which include measurements of the Higgs boson, further precision testing of Standard Model predictions, and a variety of searches for physics beyond the Standard Model. It covers a solid angle of nearly 4π (with the majority of lost particles being in the direction of the

beampipe) and is layered with different subsystems working together to identify all particles produced in a collision, while maintaining precise energy and momentum measurements even in the high density environments that are characteristic of a typical bunch crossing provided by the LHC. This section gives an overview of each of the subsystems starting from the innermost layer and working outwards.

In the ATLAS experiment, a right-handed Cartesian coordinate system is used where the origin is placed at the nominal interaction point. The axes are oriented such that the center of the LHC ring is in the x direction, the y direction points vertically upwards, and the z direction is along the beamline. The (x,y) plane defines the transverse direction, also frequently represented by the polar coordinates (r,ϕ). The azimuthal and polar angles ϕ and θ are measured from the x-axis counter-clockwise in the x-y-plane and from the z-axis, respectively. It is customary, however, to use the *pseudorapidity*

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

rather than θ directly, with the separation between two physics objects in the detector often being expressed with

$$\Delta R = \sqrt{\Delta\eta^2 + \Delta\phi^2}. \tag{3.6}$$

3.2.1 The Inner Detector

The inner detector (ID) is the closest of the systems to the interaction point and is designed to measure the trajectories of charged particles (or *tracks*) with a high spatial resolution within $|\eta| < 2.5$ as they bend in a magnetic field. A schematic overview of the ID can be seen in Figure 3.6. It is composed of three main sub-systems: the pixel detector, the semiconductor tracker (SCT) and the transition radiation tracker (TRT). A solenoid magnet surrounds the ID, generating an axial magnetic field of 2 T. Each of the sub-detectors are further divided into a barrel and two endcaps covering the central and forward regions, respectively. Figure 3.7 provides a cross-sectional sketch of the ID as a track passes through the barrel region of each of the three sub-detectors.

Pixel detector

The pixel detector comprises the innermost layers of ATLAS. It contains over 80 million individual silicon pixels, each with a size of 50 $\times$ 400 μm^2 in $R-\phi$ and z. These pixels are spread out over three layers in the barrel region and three discs in each endcap, providing space point measurements with an intrinsic resolution of 10 $\times$ 115 μm over the full pseudorapidity range spanning $|\eta| < 2.5$. The insertable b-layer (IBL) which is visible in Figure 3.7 now sits closest to the beampipe at a distance of only 27.5 mm and adds an additional 12 million readout channels. Installed during the long shutdown between 2013-2015 as a fourth layer in the barrel region, it improves the quality of impact parameter reconstruction for tracks, thereby improving vertex and b-tagging performance [60].

Semiconductor Tracker

Surrounding the pixel detector is the semiconductor tracker (SCT), adding another four cylindrical layers in the barrel and another nine disk layers in each endcap with 6.3 million

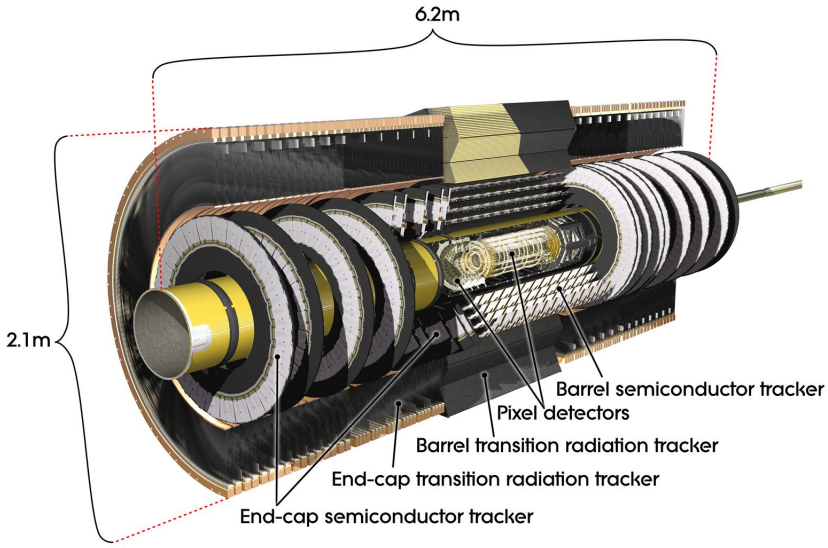


Figure 3.6: A schematic overview of the inner detector sub-systems in ATLAS. [5]

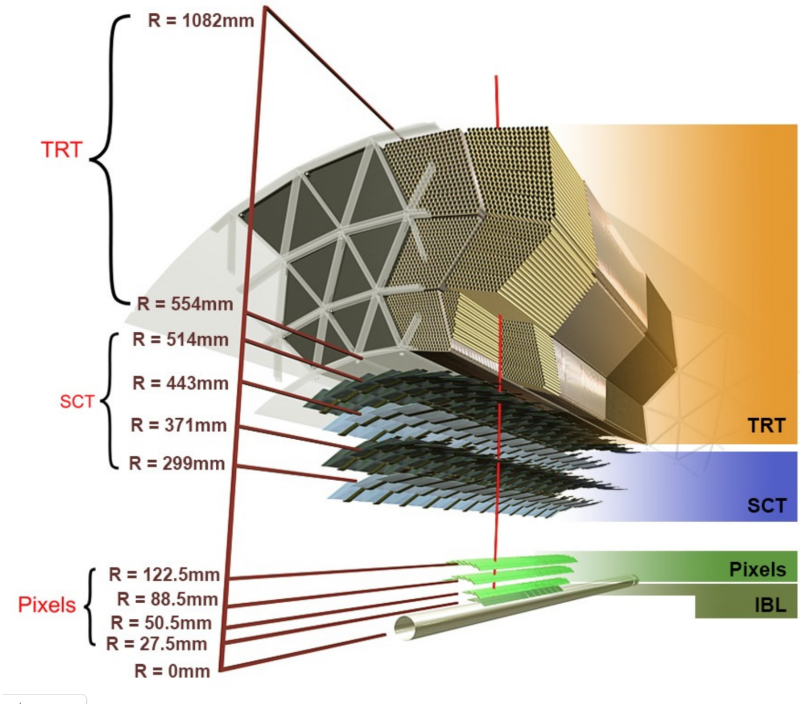


Figure 3.7: A cross-sectional sketch of the inner detector as a track passes through the barrel region of each of the three sub-detectors. [46]

readout channels and detector coverage also up to $|\eta| < 2.5$. Each SCT module is composed of a twin layer of silicon micro-strip sensors which are between 6-13 cm long and 80 μm wide, being rotated by 40 mrad with respect to one another in order to still provide a measurement in z (R) for the barrel (endcap) region. As a result, each SCT layer provides measurements with an intrinsic resolution of 17 μm $\times$ 580 μm.

Transition Radiation Tracker

The Transition Radiation Tracker (TRT) is the outer-most and largest of the ID sub-detectors, providing complementary track momentum measurements to those obtained from the silicon sensors of the pixel and SCT detectors. It is made from nearly 300,000 straw tubes that are 4 mm in diameter and filled with a Xenon-based gas mixture. The TRT provides only R - ϕ information and η coverage of $|\eta| < 2.0$. Measurements are made when charged particles ionize the gas mixture and the ions are collected by wires at the center of the straw tubes. The resulting drift time can be used to determine drift circles, with an intrinsic resolution of 130 μm per straw. The TRT also provides particle identification since the intensity of transition radiation that is emitted by a charged particle as it passes through different media can be used to determine its Lorentz factor and in so doing provide a measurement on its mass.

3.2.2 Calorimetry

The calorimeter system resides in the volume surrounding the inner detector and is designed to fully absorb electrons, photons and hadrons in order to provide a measurement of their energy. It covers the range $|\eta| < 4.9$ and contains two main components - an electromagnetic calorimeter and a hadronic calorimeter. The electromagnetic calorimeter forms an inner layer which is responsible for precisely measuring the energy of electrons and photons. The hadronic calorimeter sits around the electromagnetic calorimeter and is coarser, but provides enough stopping power also for hadronic jet activity so that the total energy-momentum balance of the collision can be measured. A schematic overview of the calorimeter system is shown in Figure 3.8. Both sub-system are sampling calorimeters, meaning that only a fraction of the energy of a particle is actually measured by an active layer. The total energy can then be inferred after careful calibration of the detector.

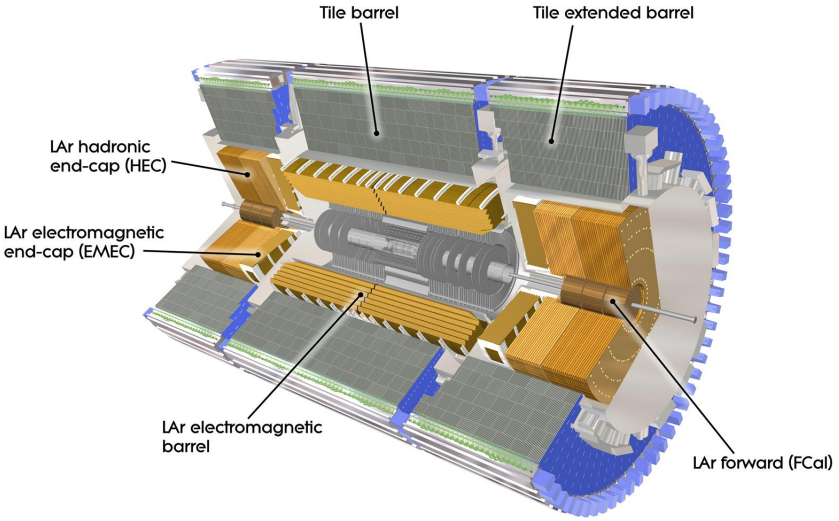
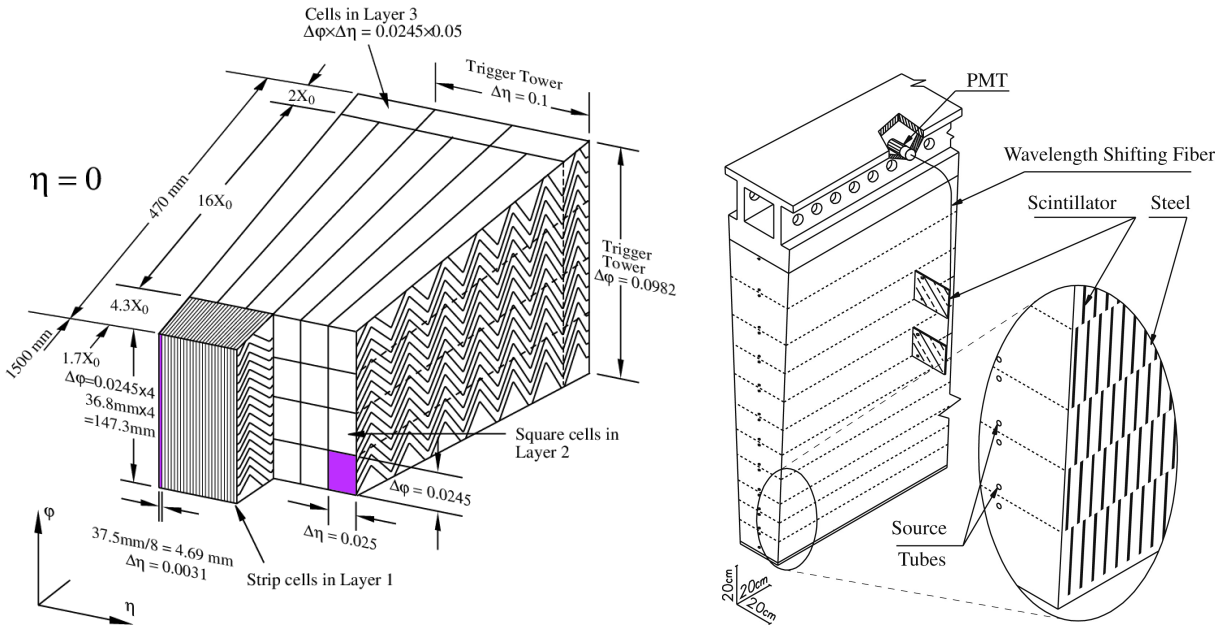


Figure 3.8: A schematic overview of the calorimeter system in ATLAS [5]



(a) (b)

Figure 3.9: (a) Cross-sectional view of a single EM calorimeter module in the barrel region. The three distinct layers can be seen, each containing a different degree of granularity. (b) Cross-sectional view of a single hadronic calorimeter module in the barrel region. The fiber readout connecting to a photomultiplier tube (PMT) is also visible. [5]

Electromagnetic Calorimeter

The electromagnetic (EM) calorimeter is a sampling calorimeter which uses liquid-argon (LAr) as an active layer and lead absorber plates as a passive layer. It consists of a barrel region covering $|\eta| < 1.475$ and two endcaps spanning $1.375 < |\eta| < 3.2$. The endcaps are mechanically divided into an outer and inner wheel covering $1.375 < |\eta| < 2.5$ and $2.5 < |\eta| < 3.2$, respectively. The overlap region between the barrel and endcaps suffers from a small reduction in performance and so it is common to veto electrons and photons which appear in this 'crack'. Each layer is arranged in an accordion shape geometry which provides a symmetry in ϕ without any azimuthal cracks.

The total thickness of the EM calorimeter is larger than 22 X_0 in the barrel and 24 X_0 in the endcaps, where X_0 is the *radiation length* and is defined as the mean distance over wich a high-energy electron loses all but $1/e$ of its energy due to bremsstrahlung. Over the region targeting precision physics which also defines the acceptance of the ID ($|\eta| < 2.5$), the EM calorimeter is segmented into three sections in depth. The endcap inner wheel, which falls outside this range, is segmented in only two section and has coarser lateral granularity. Figure 3.9a shows a cross-sectional view of one of the barrel modules of the EM calorimeter. The first layer, also known as the presampler, provides the highest granularity of $\delta\eta \times \delta\phi = 0.003 \times 0.1$ and is used to correct for energy loss before the EM calorimeter. The second and largest layer has a granularity of $\delta\eta \times \delta\phi = 0.025 \times 0.025$ and a thickness of 16 X_0, accounting for the majority of the energy measurement. The third layer is coarser with a granularity of $\delta\eta \times \delta\phi = 0.05 \times 0.025$ and a thickness of only 2 X_0.

Hadronic Calorimeter

The hadronic calorimeter is composed of three distinct sections: a tile calorimeter in the barrel region extending to $|\eta| < 1.7$, hadronic endcap calorimeters between $1.5 < |\eta| < 3.2$ and forward calorimeters covering the range $3.1 < |\eta| < 4.9$. The total calorimeter thickness for hadronic interactions is close to 9.7 λ in the barrel and 10 λ, where λ is the *nuclear interaction length* which is the distance over which the number of relativistic hadrons reduces by a factor of $1/e$. The granularity of the hadronic calorimeters is coarser than the EM calorimeter, with even a barrel segmentation of only $\delta\eta \times \delta\phi = 0.1 \times 0.1$.

The tile barrel covers the region $|\eta| < 1.0$, with two extended barrels ranging between $0.8 < |\eta| < 1.7$. It is a sampling calorimeter, using steel as an absorber and scintillating tiles as an active material. Figure 3.9b shows a cros-sectional view of one of the barrel modules of the hadronic calorimeter. The photons generated from the scintillating tiles are sent through wavelength-shifting fibers to photomultiplier tubes (PMTs) for readout. The hadronic endcap calorimeters (HECs) each consist of two independent liquid-argon-copper wheels for a total of four layers per endcap. The forward calorimeters (FCALs) each utilize one liquid-argon-copper and two liquid-argon-tungsten modules which serve to record measurements of EM and hadronic showers, respectively.

3.2.3 Muon Spectrometer

The muon spectrometer (MS) provides both the outermost layers and the largest volume to the ATLAS detector. It is built not only to give muon momentum measurements as they bend through the magnetic field generated by dedicated air-core toroid magnets, but also to allow for fast trigger decisions on events with high energy muons. The system reaches up to $|\eta| < 2.7$, with magnetic bending provided by a large barrel toroid over the range $|\eta| < 1.4$ and two smaller end-cap magnets for $1.6 < |\eta| < 2.7$. For $1.4 < |\eta| < 1.6$, also known as the transition region, deflection is provided by a combination of barrel and end-cap fields.

A schematic overview of the MS is shown in Figure 3.10. It contains four types of muon chamber systems, each based on the technology of gas ionization which then drifts charges to electrodes to be measured. For precise momentum measurement, Monitored Drift Tubes (MDTs) are placed over the full range of $|\eta| < 2.7$ with an average resolution of 80 μm. In the forward region between $2.0 < |\eta| < 2.7$, the MDTs are reduced in number and instead complemented by Cathode-Strip Chambers (CSCs). For fast triggering, Resistive Plate Chambers (RPCs) and Thin Gap Chambers (TGCs) are installed up to $|\eta| < 2.4$ in the barrel and endcaps, respectively.

3.3 Trigger and Data Acquisition

The rate at which the LHC provides pp collisions to the ATLAS detector is orders of magnitude larger than the rate at which they can practically be recorded. On the one hand, this is due to a latency with respect to the limited detector readout speed. On the other, there isn't near enough disk storage space available. Moreover, it is simply not necessary to record every event that is produced. As can be seen in Figure 3.11, most rare processes of interest occur at rates that are at least a factor of $\sigma/\sigma_{\text{TOT}} \sim 10^{-10}$ less common that the total inelastic pp cross section. Due to these reasons, ATLAS employs a trigger system which is responsible for making very fast yet informed decisions as to which events will be

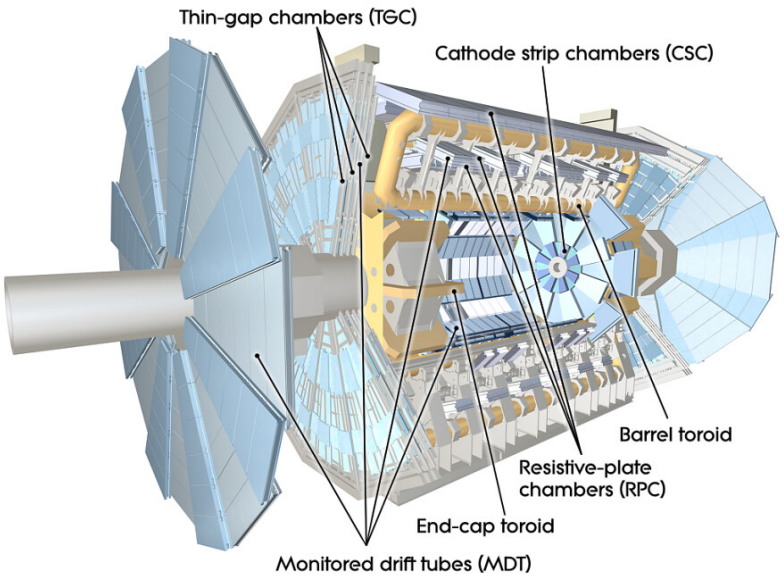


Figure 3.10: A schematic overview of the muon spectrometer in ATLAS [5]

recorded and a data acquisition (DAQ) system which handles the data flow logistics from the detector to disk.

The performance of the trigger and data acquisition systems is influenced by the number of readout channels from the detector and by the density of each collision environment - both of which saw an increase for Run 2 data taking[3]. This prompted an overhaul in the design of these systems and with it a simplification of the layering. Figure 3.12 provides an overview of the architecture for the trigger and data acquisition systems during Run 2, where the trigger system is now composed of only two tiers (down from three in Run 1):

- The **Level 1** (L1) trigger uses fast, custom-made hardware in order to arrive at a decision to pass on or reject the event within 2.5 μs. Regions of interest (RoIs) are determined from the EM and hadronic calorimeters with a granularity of $\delta\eta \times \delta\phi = 0.1 \times 0.1$ as well as the trigger chambers in the muon spectrometer. In addition to object multiplicity, it is capable of performing selections using topological requirements such as isolation, invariant mass, ΔR between two objects, and missing transverse momentum, reducing the rate of incoming data from 40 MHz down to 100 kHz.

- The **High-Level Trigger** (HLT) is a software-based layer which is housed in an on-site computing farm with ∼40k processing units. Seeded from the regions of interest provided by L1 and the with full availability of the detector readout, dedicated algorithms are run in sequences of feature extraction followed by hypotheses that are able to reconstruct the event inside the RoIs with near offline-like quality in order to make a final decision on whether or not to write out the event in a matter of no more than about 0.4 s.

The HLT algorithm sequences described above are often referred to as *trigger chains*, where each targets a specific signature for recording. The group of trigger chains which are

[3]The number of readout channels increased by 20% due to additional detectors such as the IBL and the Fast Tracker (FTK), while the collision density increased from the LHC exceeding its design luminosity of $10^{34}\text{cm}^{-2}\text{s}^{-1}$.

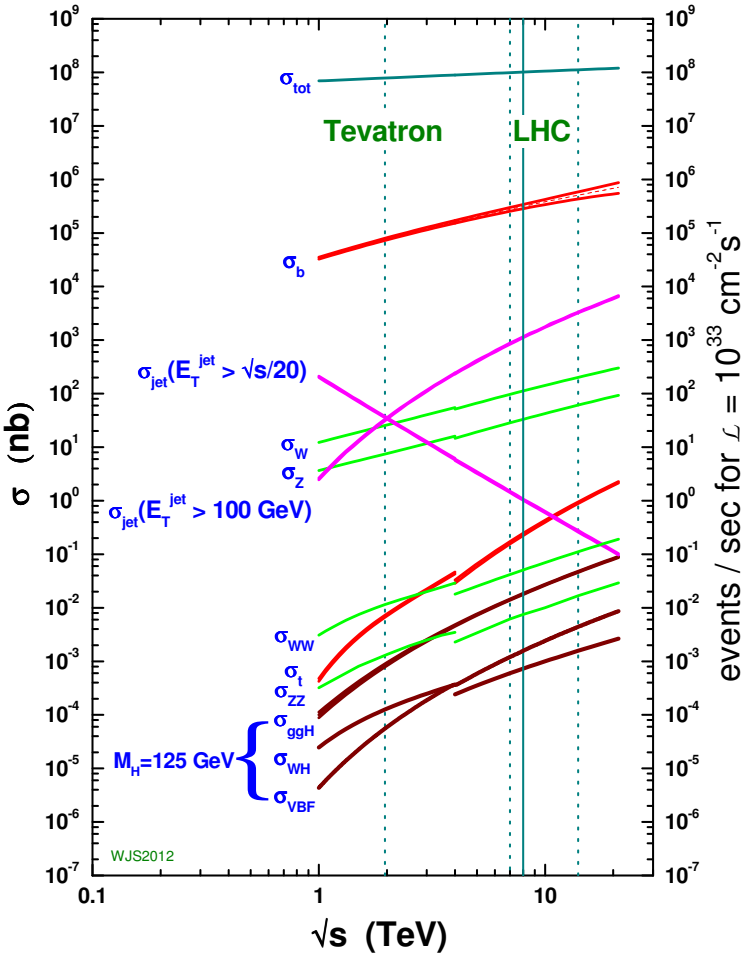


Figure 3.11: Cross sections for various processes in proton collisions as a function of the center of mass energy. The discontinuity between the Tevatron and LHC regimes is due to the switch from $p\bar{p}$ to pp collisions. [109]

active at any given time during operation of the detector is called the *trigger menu* and can therefore be viewed as an implementation of the physics program of ATLAS. During a typical data taking run, it is also common for some trigger chains to by adjusted by *prescales* such that only a subset of events that otherwise would have passed that chain are actually saved. When an event fires any trigger chain, it is written to disk through different inclusive *output streams* depending on which chain(s) passed. The output stream which is used for most physics analyses and for the one presented in this thesis is called the 'physics_Main' stream.

3.4 Detector Simulation and Digitization

The event generators described in section 2.2 provide their output in the form of final-state stable[4] particles from a single pp collision. In order to compare these Monte Carlo events to those taken from data, some additional processing which is described in this section must take place. After these steps are performed, however, the reconstruction of events proceeds

[4] A stable particle can be defined differently depending on the experiment. For ATLAS, stable particles are those which have lifetimes of $c\tau > 10$ mm.

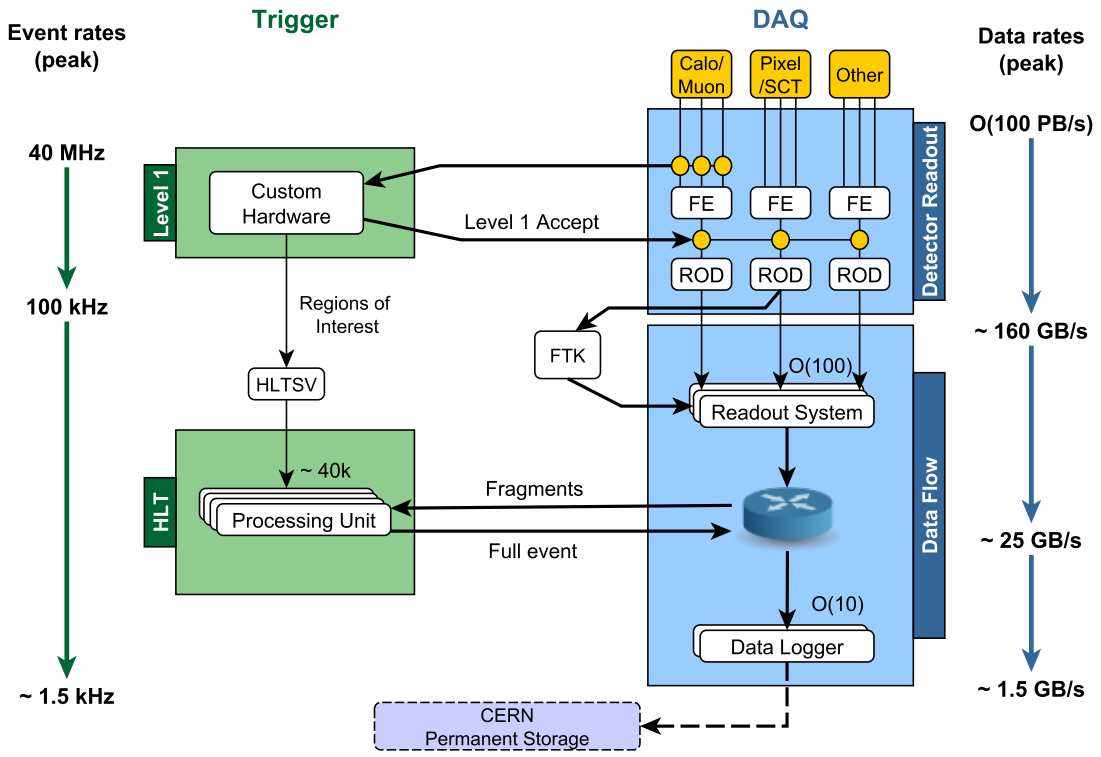


Figure 3.12: Architecture of the ATLAS trigger system in Run 2 [86]

in an identical way regardless of their source.

3.4.1 Simulation

The particles originating from the primary collision point in a real data event will interact with the detector as they fly outward, depositing their energy and creating new particles along the way. Therefore, these effects must also be simulated in Monte Carlo[5]. For this, one of two approaches is generally taken - either every detail is simulated using the toolbox Geant4 (GEometry ANd Tracking) [37], or multiple approximations are made in order to speed up the computation (e.g. longitudinal and transverse parameterization of showers in the calorimeter) in a scheme known as Atlfast-II (or AFII) simulation.

Detector simulation is the most computationally intensive link in the chain of Monte Carlo production. It proceeds in discrete steps (including the transportation of particles for which a map of the magnetic field is used in the case that they're charged) which rely heavily on numerical models, with the majority of time being spent moving electrons and photons around in the calorimeter since all particles are tracked until they either reach zero energy or exit the detector when running with Geant4. In order to reduce CPU time (particularly in the forward EM calorimeters), *frozen showers* are used in which low energetic particles get replaced by pre-simulated EM showers[6]. A comparison of CPU times for different simulation

[5]Other reasons for simulating the interaction of particles with detector elements include studying the performance of detectors before they are built and investigating potential radiation damage.

[6]The shower library is generated with Geant4 simulation and is used by default for full simulation

scenarios can be seen in Figure 3.13.

An additional complexity that must be considered when comparing to data is the accurate representation of the detector conditions at which time the data was taken. Some conditions, such as the masking or disabling of particular readout channels, can be accounted for after the simulation finishes. However, others like coarse misalignments of the detector and the size of the bunch crossing region must be incorporated during the simulation procedure.

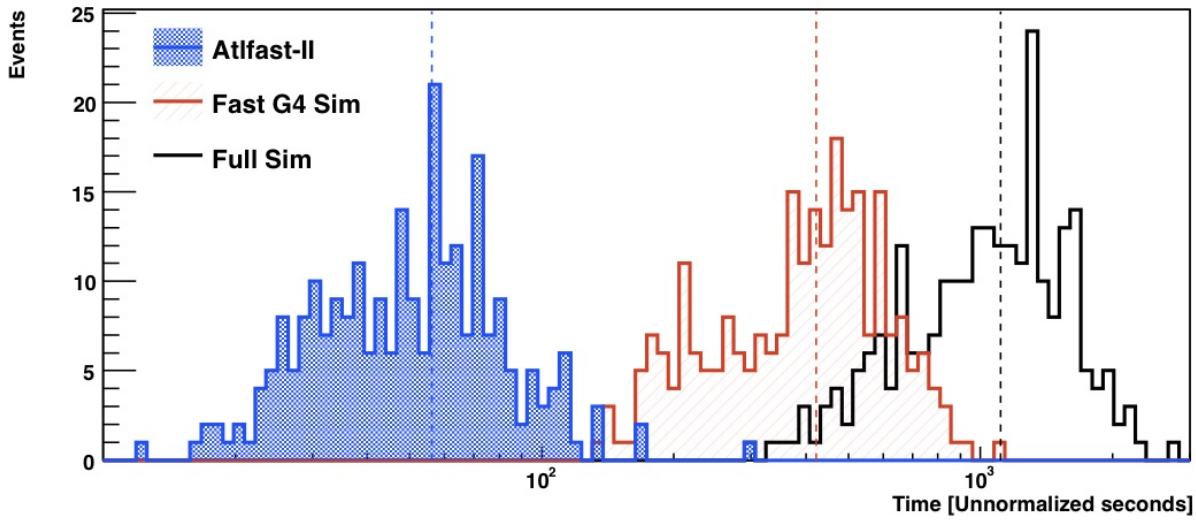


Figure 3.13: Distributions of CPU times for 250 $t\bar{t}$ events in Full Sim, Fast Geant4 Sim (using frozen showers) and AFII Sim. The vertical dotted lines mark the distribution averages. [29]

3.4.2 Pileup modeling

As mentioned briefly in subsection 3.1.4, data events will almost always contain more than one pp collision. The additional pileup that is created is known to have a significant impact on physics results and therefore must also be modeled when producing Monte Carlo. There are generally two approaches that are taken when modeling pileup - either the additional pp interactions are simulated in Monte Carlo, or real data events are overlaid on top of the simulated hard-scatter process.

Currently the most common approach is to simulate the additional pp interactions, using soft QCD processes that are usually generated with the PYTHIA 8 program [110] after careful tuning to data measurements. Each pp interaction is simulated separately and then they are combined[7] before being digitized. In order to determine how many should be included for a given event, a value of μ is first chosen based on the data conditions it will be compared to. A random number is then picked from a Poisson distribution with that μ as a mean to be added to the bunch crossing.

Typically, Monte Carlo production campaigns take place before the full corresponding dataset is available and so the μ that is used is only a best guess estimate from the information provided by the LHC operators. Also, during data taking the value of μ is not constant even over a single run. Partly due to these reasons, the Monte Carlo must be reweighted by scaling the selection efficiency of each event so that on average, the amount of

Monte Carlo production campaigns.

[7]This includes both in-time and out-of-time pp interactions within a [-800,800] ns window.

pileup reproduces that in data. In practice, this is achieved by matching the μ distribution in Monte Carlo with the μ distribution in data after scaling the data down by 9% since the A2 tune of PYTHIA 8 used herein results in the Monte Carlo over-estimating the hardness of each pp collision by this amount [7].

The second method for modeling pileup of overlaying real data events is less prevalent, although does appear in performance related work and specific analyses that are most sensitive to pileup effects. It is accomplished by first defining a data period to simulate and then selecting random events without a hard scatter from this period to combine with a simulated hard-scatter event using the same detector conditions. Some advantages to using this method include the non-reliance on generator tuning, the automatic matching of μ without the need for reweighting and the realistic detector noise and occupancy. However, these samples can only be made after the data is taken and most analyses are often designed to be as insensitive as possible to the modeling of effects for which it is advantageous to use data overlay. As pileup becomes more of an issue at higher luminosities, data overlay offers a potential alternative to purely simulated events.

3.4.3 Digitization

After energy deposits in the sensitive detector volumes have been simulated, dedicated digitization software then converts them into detector responses ("digits"), typically voltages or times on pre-amplifier outputs. Detector noise is modeled by first measuring the rates in data for a particular readout technology, storing the average amount of noise in a database. In order to determine the level of noise to add to a particular channel during digitization, the noise constant for that channel is multiplied by a Gaussian random number. Digit creation is then followed by a simulation of the RODs (Read Out Drivers) and triggers, to produce RDOs (Raw Data Outputs) that serve as input to the event reconstruction.

3.5 Event Reconstruction

Events passing at least one of the trigger chains mentioned in section 3.3 are later sent through a sequence of sophisticated algorithms in order to reconstruct in as much detail as possible the particles that participated in the collisions from the high density of hits and energy deposits measured in the detector. Each type of particle can be identified through a different signature that it leaves behind, as shown in Figure 3.14. An overview of the reconstruction of each physics object used for the analysis presented in this thesis is provided below, roughly following the order in which they are built.

3.5.1 Tracks and Primary Vertices

Tracks in the ID are reconstructed from hits in the pixel + SCT detectors and timing information in the TRT. The majority of tracks are found through an "inside-out" approach[8] which begins in the silicon layers and propagates out toward the TRT. As a first step, silicon hits that are in close proximity are combined into clusters. These clusters are then used to form three-dimensional measurements referred to as space-points. In the pixel detector,

[8] An "outside-in" approach is also employed in which track reconstruction starts in the TRT and propagates inward, but this contributes a significantly smaller amount of the total tracks. Some tracks are also "TRT standalone", having hits only in the TRT.

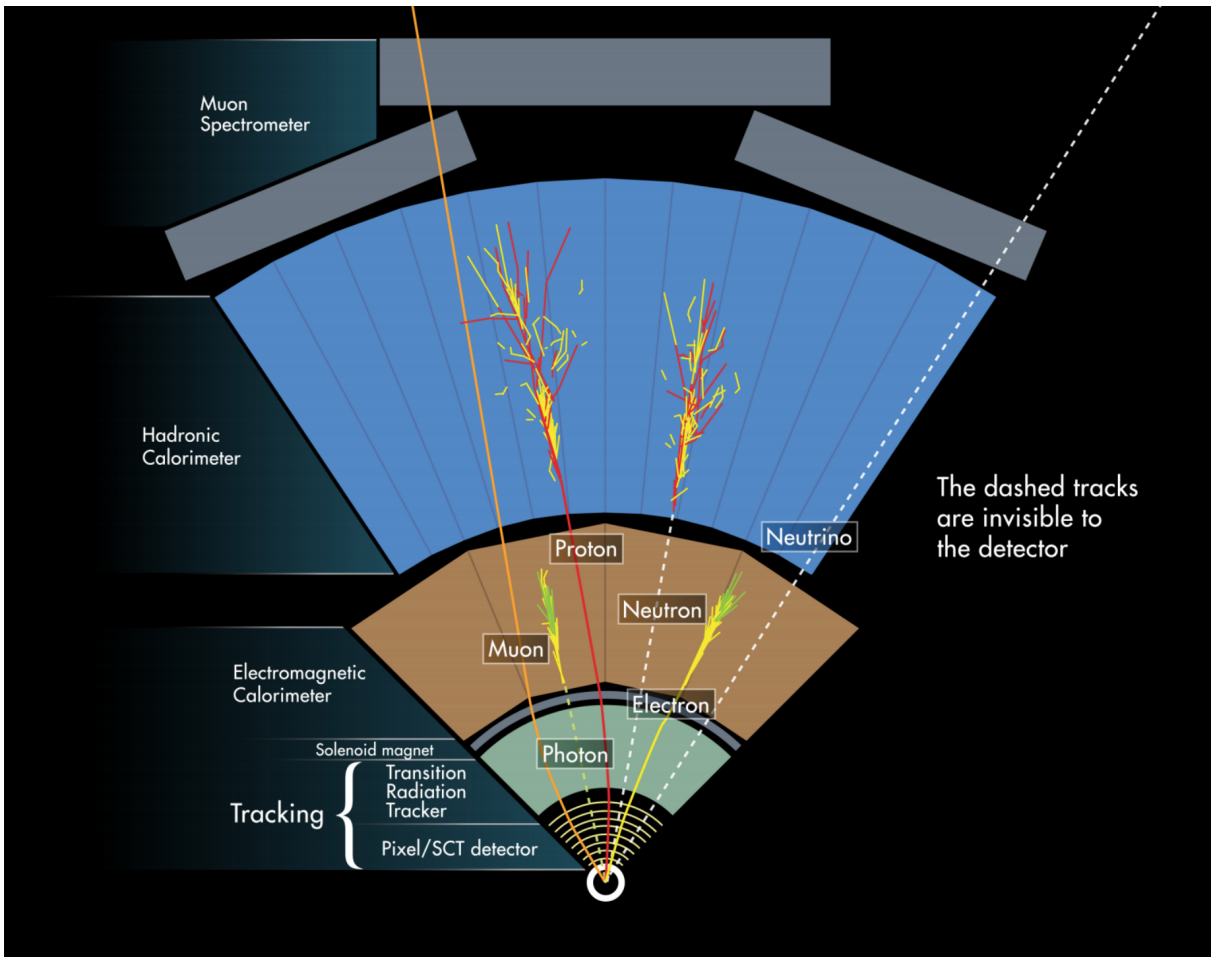


Figure 3.14: The different signatures for each type of particle shown in the transverse plane as they traverse the detector.

one space-point is identified for each cluster. However in the SCT, both stereo views of a strip layer must be combined to form a single space-point. Track seeds are then created from three space-points each in the silicon-detector layers. These seeds are then extended outward by applying a Kalman filter [71], following the most likely path using knowledge of the material and of the magnetic field configuration. The resulting track candidates with $p_{\mathrm{T}} > 400$ MeV are fit via the ATLAS Global χ^2 Track Fitter [64], at which time ambiguities from shared hits are also resolved by employing a track scoring mechanism which takes into account e.g. the number of hits and holes for each track, as well as their fit quality. Finally, the surviving track candidates are extended into the TRT through matches with drift-circles converted from the measured timing information.

For a charged particle traveling through a uniform magnetic field, five parameters are required in order to fully describe its trajectory. Although these parameters can be expressed a number of different ways, the most useful representation for physics analyses are typically the so-called "perigee parameters", or $\left(d_0, z_0, \phi, \theta, \frac{q}{p}\right)$ where q is the particle's charge, while d_0 and z_0 are the transverse and longitudinal impact parameters respectively, defining the point of closest approach that the particle takes with respect to the beamline. The perigee representation for track parameters are illustrated in Figure 3.15.

Primary vertices, defined each as the interaction point between two beam protons, are reconstructed using similar techniques as with tracks. First, a preselection containing tight

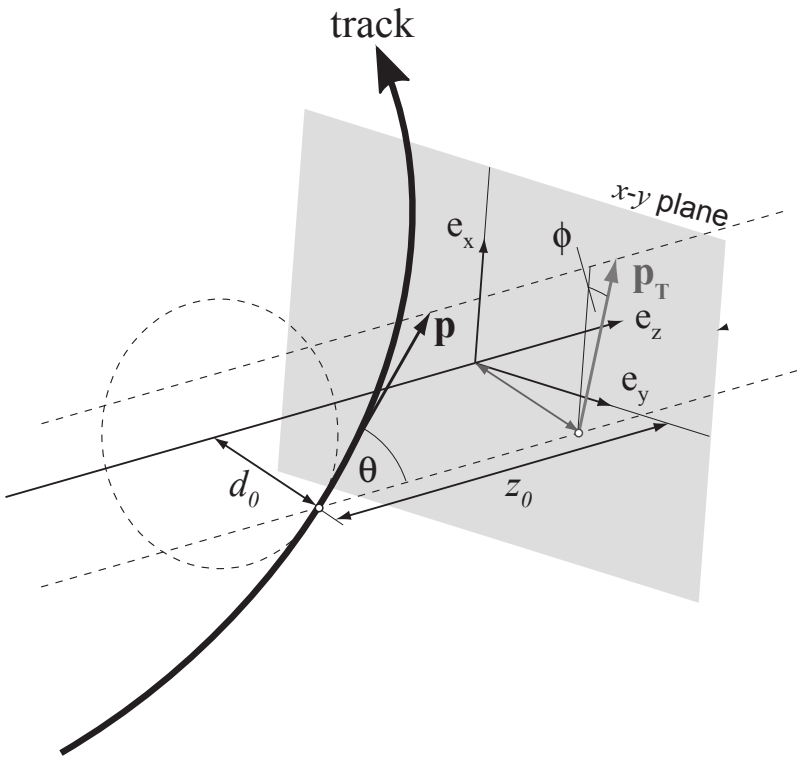


Figure 3.15: An illustration of the perigee parameters representing a track's trajectory, with the transverse (d_0) and longitudinal (z_0) impact parameters defining its point of closest approach with respect to the beamline. [63]

requirements is applied to the reconstructed tracks in order to reduce the amount which are fake or poorly measured and to ensure that tracks with very large impact parameters are not considered. Primary vertex reconstruction then proceeds through a "finding-before-fitting" approach in which they are found one at a time, beginning with a seed selected from a pool of tracks and fit with a Kalman filter using the impact parameters of associated tracks[9] and a constraint on the luminous region ("beam spot") location and size. Figure 3.16a illustrates the steps involved with this iterative vertex finding (IVF) method.

The Kalman filter is a least-squares estimator and therefore it is known to not be robust against outliers (in this case mis-associated tracks or mis-measured track errors). Due to this limitation, the tracks are weighted in the fit according to their χ^2 compatibility with the vertex. However, outliers can still have a disproportionate impact on the final vertex position during the first few iterations of the fit. To mitigate this effect, a scheme known as *adaptive vertex fitting* [72] is used in which track weights are also made to depend on a temperature through a thermodynamic annealing procedure:

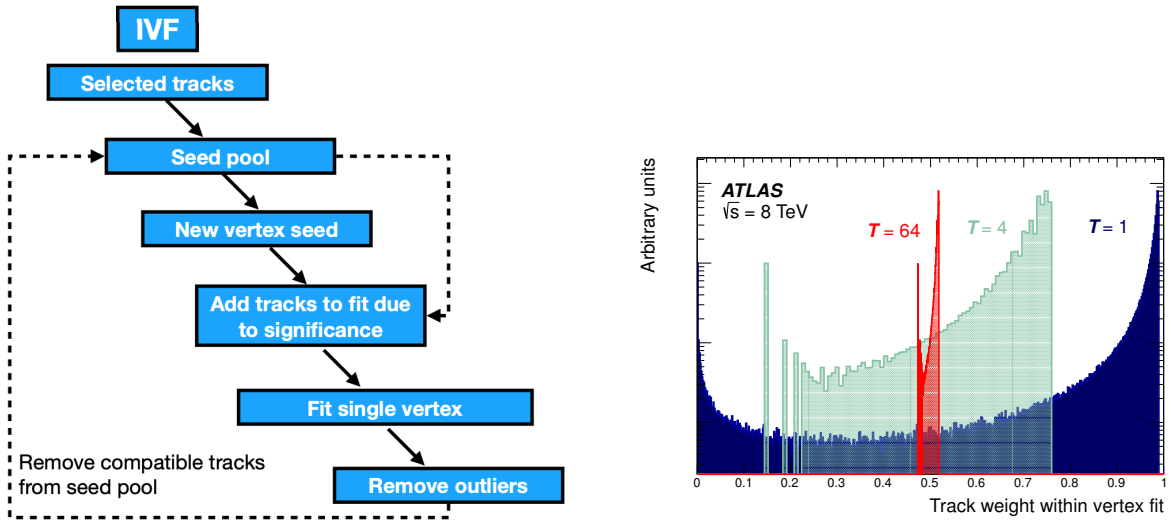
$$\omega(\chi^2, T) = \frac{1}{1 + e^{-\frac{1}{2}(\chi^2_{\text{cut}} - \chi^2)/T}} \tag{3.7}$$

where T is the temperature. Initially for high temperatures, the weight of each track will all

[9]Once a seed is found as the point of maximum track density along the beam axis, tracks that are within $12\sqrt{\sigma^2(d_0) + \sigma^2(z_0)}$ are assigned to the vertex fit.

be close to 1/2. During the iterative fit, the temperature is lowered such that the sensitivity of track weights to their compatibility with the vertex increases until finally for $T = 1$ compatible tracks within three standard deviations (if χ^2_{cut} is set to 9) will have weights close to 1 and incompatible tracks will have weights close to 0. Track weights corresponding to different temperatures are shown in Figure 3.16b. Tracks that end up with a weight larger than 0.01 or impact parameter significance less than 7 σ are removed from the pool of tracks that are available to find the next vertex seed.

Once all primary vertices have been found, the hard-scatter vertex is identified as the one which contains the largest sum of associated track p_{T} squared. Some analyses studying processes for which there are no charged particles directly appearing in the signal (such as $H \to \gamma\gamma$) must rely on different definitions, although the efficiency of reconstructing and identifying correctly the hard-scatter vertices used for the analysis presented in this thesis is nearly 100%. Also, the number of reconstructed primary vertices will in general be less than the number of pp interactions in any given bunch-crossing due to the merging of two sufficiently close-by interactions into one reconstructed vertex and is more pronounced for higher pileup. For this reason, vertexing algorithms that are more pileup robust have been developed and will be deployed during Run 3 of the LHC [57].



(a) (b)

Figure 3.16: (a) Schematic diagram of the steps involved in iterative vertex finding [57] (b) Track weights for various temperatures T, corresponding to different steps in a vertex fit [82]

3.5.2 Calorimeter Clusters

The fully reconstructed final state of a pp collision in ATLAS includes individually identified particles comprising electrons, photons, muons, and τ-leptons, in addition to hadronic jets and missing transverse momentum ($E_{\text{T}}^{\text{miss}}$). For each of these objects (aside from muons), calorimeter signals play a key role. Due to both electronic readout and pileup condition effects (the latter being more dominant for high bunch intensity), the LAr calorimeter is

subject to a significant amount of background noise[10]. These baseline fluctuations can be expressed by the standard deviation of their distribution and are central to measuring the energy deposited by incoming particles.

In ATLAS offline reconstruction, the collection of calorimeter cell signals into larger clusters proceeds via a three dimensional topological clustering algorithm [35]. The basic observable controlling this cluster formation is the cell signal significance $\varsigma^{\mathrm{EM}}_{\mathrm{cell}}$, which is defined as the ratio of the cell signal to its average (expected) noise $\sigma^{\mathrm{EM}}_{\mathrm{noise,cell}}$,

$$\varsigma^{\mathrm{EM}}_{\mathrm{cell}} = \frac{E^{\mathrm{EM}}_{\mathrm{cell}}}{\sigma^{\mathrm{EM}}_{\mathrm{noise,cell}}}, \tag{3.8}$$

where both $E^{\mathrm{EM}}_{\mathrm{cell}}$ and $\sigma^{\mathrm{EM}}_{\mathrm{noise,cell}}$ are measured on the electromagnetic (EM) energy scale[11].

Topological clusters (or *topo-clusters* for short) are formed starting from a calorimeter cell with a highly significant seed signal with $|\varsigma^{\mathrm{EM}}_{\mathrm{cell}}| > S$ (where S is the primary seed threshold). The cells neighboring a seed and satisfying $|\varsigma^{\mathrm{EM}}_{\mathrm{cell}}| > N$ (where N is the threshold for growth control) are then iteratively added to the cluster, while any neighboring cells with $|\varsigma^{\mathrm{EM}}_{\mathrm{cell}}| > P$ (where P is the principal cell filter) are then added as a final step. In this scheme, overlapping clusters will be merged, while an attempt to split the cluster is made if it is found to have several local maxima. Figure 3.17 shows an example of the final stage in topo-cluster formation for the first module in the FCAL calorimeter.

The configuration of $S = 4$, $N = 2$, $P = 0$ is optimized for ATLAS hadronic final state reconstruction, by removing cells with insignificant signals which are not in close proximity to cells with significant signals. Such cells with insignificant signals are considered noise and discarded from further jet, particle and $E^{\mathrm{miss}}_{\mathrm{T}}$ reconstruction. Once the topo-clusters have been found, their reconstructed observables (or "cluster moments") such as location, direction and internal signal distribution are determined - the last of which contains valuable information related to its origin, therefore dictating how it is to be calibrated.

3.5.3 Electrons

Electron reconstruction begins by associating EM clusters with ID tracks that are extrapolated to the calorimeter[12]. Tracks that are loosely matched to an EM cluster are refitted using the Gaussian Sum Filter (GSF) algorithm [6], which takes into account the non-linear effects of bremsstrahlung radiation and greatly improves their p_{T} and impact parameter measurements. The efficiency to reconstruct electrons which can be seen in Figure 3.19a is 99% in the central region, decreasing to $\sim 97\%$ in the endcap region for lower p_{T} [28]. The path of an electron traveling through the detector is illustrated in Figure 3.18.

The reconstruction procedure described above does well in efficiently identifying real electrons. However, other objects (such as hadronic jets) can also easily fake the signature of

[10]The tile calorimeter, on the other hand, shows much less sensitivity to pileup since most of the energy flow of soft particles is absorbed already in the LAr calorimeter.

[11]The EM energy scale reconstructs the energy deposited by electrons and photons correctly but does not include any corrections for the loss of signal for hadrons due to the non-compensating nature of the ATLAS calorimeters. More details on the different energy scales can be found in subsection 3.5.6 where jet calibration is discussed.

[12]The reconstruction of electrons is very closely connected to the reconstruction of photons, since both (sometimes referred to collectively as "EGamma" objects) leave energy deposits in the EM calorimeter. Photons can be identified as an EM cluster which is either not associated with an ID track or associated with an e^+e^- conversion vertex. Some EGamma objects are also left as ambiguous during reconstruction (if they could be either an electron or a photon).

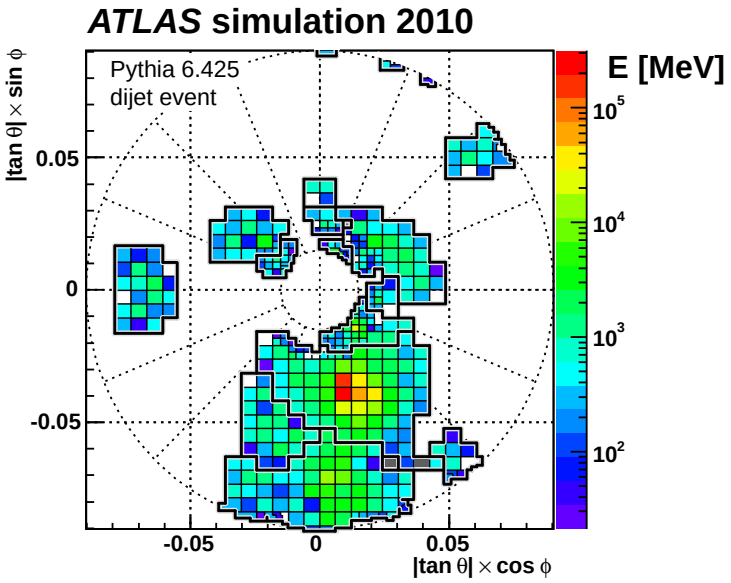


Figure 3.17: Final stage in topo-cluster formation once all cells have been added in the first module of the FCAL for a simulated dijet event with at least one jet entering the calorimeter. [35]

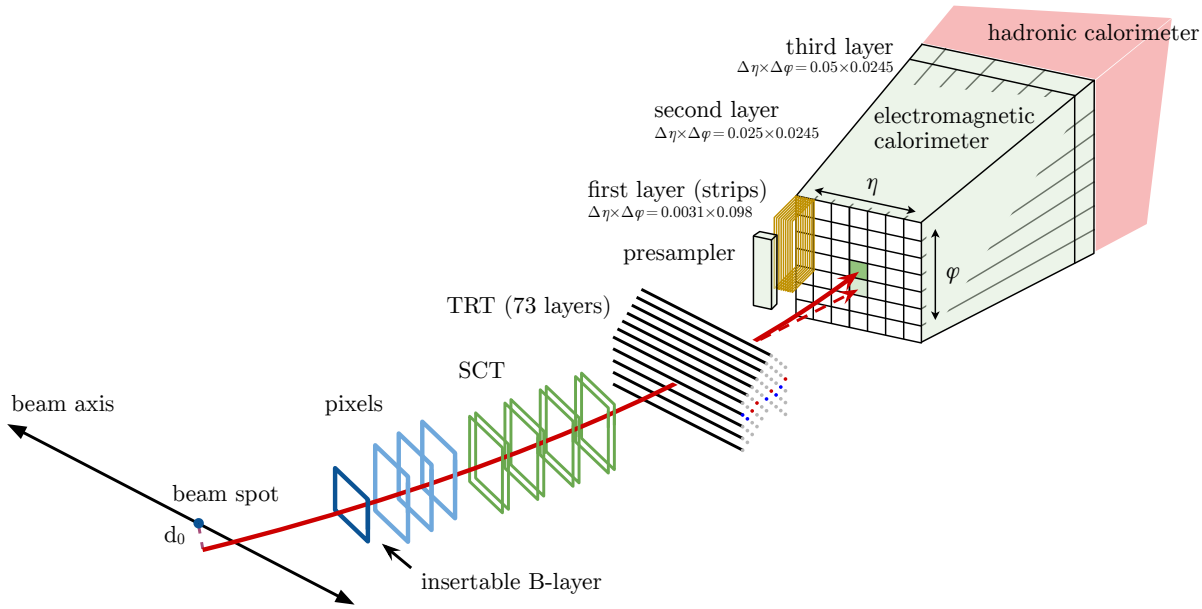
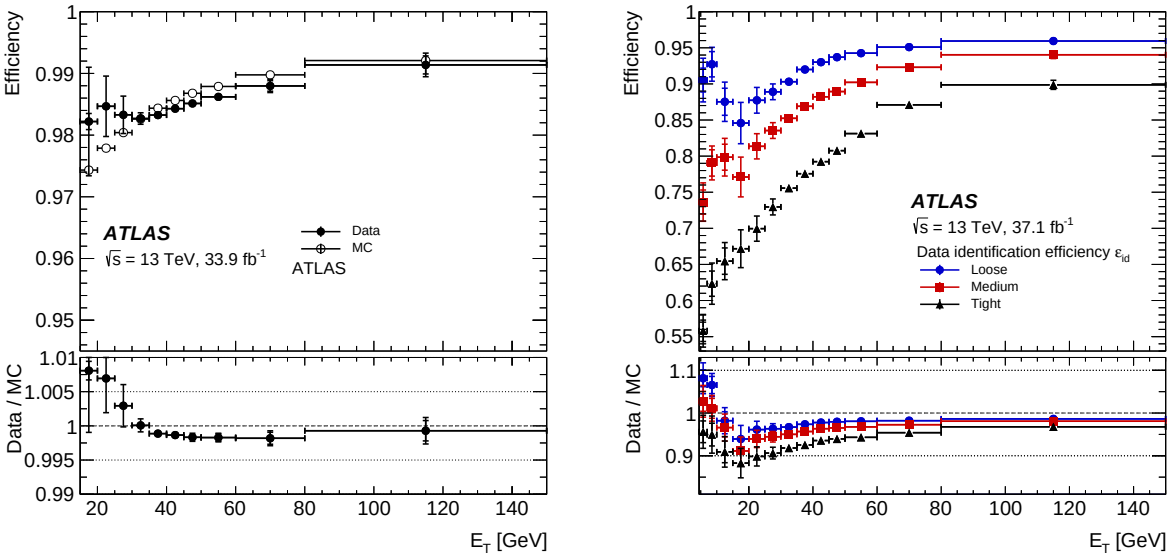


Figure 3.18: A schematic illustration of the path of an electron through the ATLAS detector. The dashed red trajectory indicates the path of a photon produced by the interaction of the electron with the material in the TRT. [28]

an electron and therefore mistakenly be reconstructed as one. In order to reduce the chance of this happening, additional identification working points have been developed. These working points all employ a likelihood-based method [28] which takes into account a number of discriminating variables such as shower shapes, track properties and track-cluster-matching in order to determine the likelihood that a given reconstructed electron is real. The three main identification working points are called LooseLH, MediumLH, and TightLH which offer identification efficiencies of 96%, 94% and 88% at high E_{T} respectively, and are also

shown in Figure 3.19b.



(a) (b)

Figure 3.19: Electron reconstruction efficiency (a) and electron identification efficiencies (b) derived from $Z \to e^+e^-$ events as functions of E_{T}. [28]

After reconstruction, the energy of an electron must be calibrated and in doing so correct for things such as instrumental effects, energy loss in front of the EM calorimeter, energy deposits not accounted for in clusters, etc. Once the electron energy has been calibrated, any residual difference in energy scale between data and Monte Carlo simulation can be defined as α_i where i corresponds to different regions in η, while any residual difference in energy resolution is assumed to contribute an extra constant term c_i:

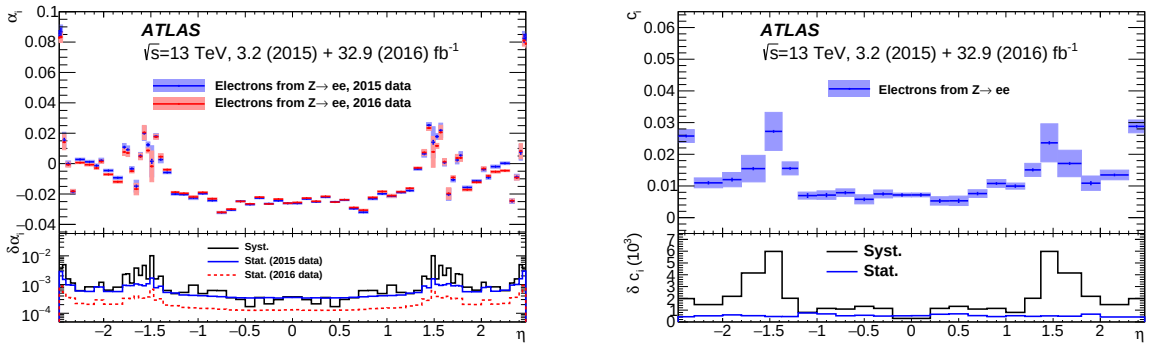
$$E^{\mathrm{data}} = E^{\mathrm{MC}}(1+\alpha_i) \tag{3.9}$$

$$\left(\frac{\sigma_E}{E}\right)^{\mathrm{data}} = \left(\frac{\sigma_E}{E}\right)^{\mathrm{MC}} \oplus c_i \tag{3.10}$$

The α_i and c_i parameters can be determined through a χ^2 minimization by comparing dielectron invariant mass distributions for $Z \to ee$ decays in both data and Monte Carlo, therefore providing a calibration uncertainty. Examples of typical values are shown in Figure 3.20.

3.5.4 Muons

The reconstruction of muons is first performed in the ID and MS independently. In the ID, muon tracks are identified just as for any other charged particle as described in subsection 3.5.1. In the MS, muon tracks are built starting with segments found by hit patterns inside each of the muon chambers. A global χ^2 fit is then applied and the track is accepted as a candidate if the χ^2 of the fit passes a certain selection criteria. When combining muon reconstruction from the ID and MS, four muon types are defined depending on which subdetectors are utilized. Combined (CB) muons are those for which an ID track was successfully matched with a muon track candidate in the MS and represent the bulk of all reconstructed muons. Extrapolated (ME) muons or 'stand-alone' muons are those



(a) (b)

Figure 3.20: Electron calibration uncertainties derived from data-to-MC comparisions in $Z \to ee$ events for (a) the energy scale corrections (α_i) and (b) the energy resolution corrections (c_i) as a function of η. [27]

which are based only on a track in the MS and are mainly used to extend the acceptance for muon reconstruction into the region $2.5 < |\eta| < 2.7$ which is not covered by the ID. Segment-tagged (ST) muons are those which are matched to an ID track, but only contain a segment in the MS which can happen for low p_{T} muons or in regions with poor detector coverage. Finally, calorimeter-tagged (CT) muons are those for which an ID track can be matched to an energy deposit in the calorimeter that is compatible with a minimum-ionizing particle.

Similar to electrons, a separate identification step for muons is also performed - although in this case it is based simply on applying a set of quality requirements which aim at suppressing backgrounds, most notably from pion and kaon decays. Four muon identification selections are provided, each targeted to address specific needs of different physics analyses:

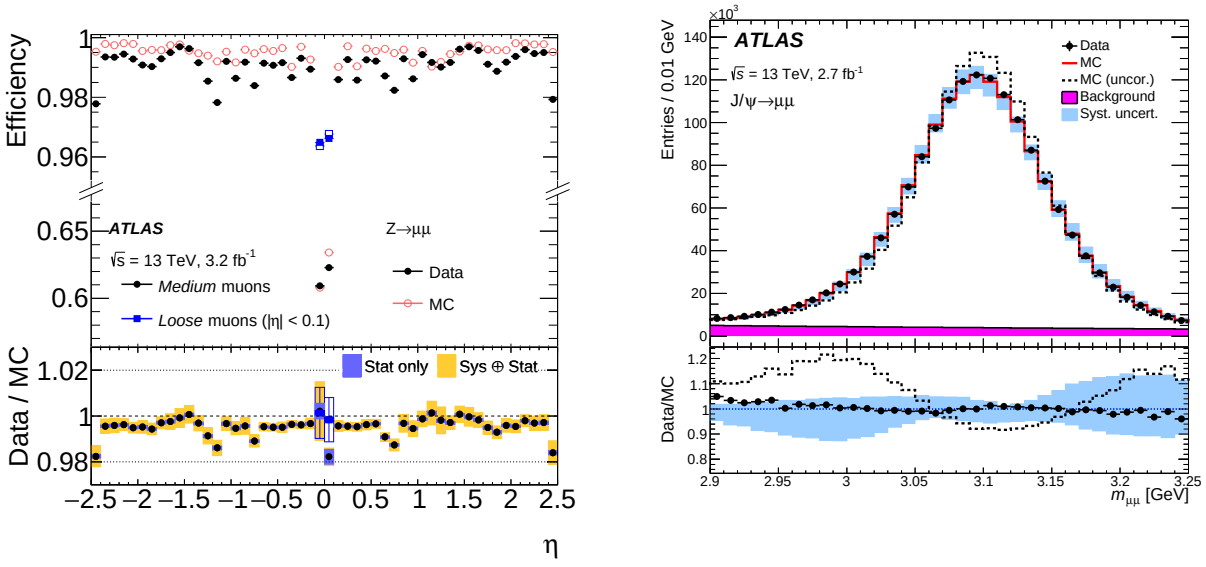
- **Loose muons** - designed to maximize reconstruction efficiency while also offering good-quality tracks.
- **Medium muons** - provide the default selection for muons in ATLAS and designed to minimize systematic uncertainties associated with muon reconstruction and calibration. Only CB and ME muons are used.
- **Tight muons** - selected with relatively lower efficiency so as to maximize the sample purity. Only CB muons are used.
- **High-p_{T} muons** - aim to maximize momentum resolution for tracks with $p_{\mathrm{T}} > 100$ GeV.

The efficiency for muon reconstruction is measured in $Z \to \mu\mu$ and $J/\psi \to \mu\mu$ decays for each identification working point and for medium quality muons is close to 99% for the majority of the range within the acceptance of the ID as can be seen in Figure 3.21a.

While the simulation of the ATLAS detector is generally quite accurate, there remain some small modeling imperfections[13] which translate into an observable mismatch between

[13]such as energy loss in the calorimeter and other materials, radial distortions of the detector, and inhomogeneities of the magnetic field

the muon momentum as measured in data and Monte Carlo. Referred to as "muon momentum calibration", both the scale and resolution of simulated muon p_{T} must be corrected to reproduce the same quantities as in data. Such corrections are typically below $\sim 0.1\%$. Figure 3.21b shows a dimuon invariant mass distribution of $J/\psi \to \mu\mu$ candidate events reconstructed with CB muons before and after calibration.



(a) (b)

Figure 3.21: (a) Reconstruction efficiency for medium quality muons as a function of η as measured in $Z \to \mu\mu$ events. (b) Dimuon invariant mass distribution of $J/\psi \to \mu\mu$ candidate events reconstructed with CB muons before (dashed) and after (solid red) calibration. [34]

3.5.5 Lepton Isolation

For many signal processes containing a lepton in the final state (e.g. leptonic decay of a vector boson), one of their distinguishing features is that they are typically well isolated from other activity. However, the identification criteria described so far do not explicitly make any isolation requirements. Leptons originating from heavy flavor decay (and therefore with relatively large amounts of close-by activity), for instance, will also often get successfully reconstructed. An effective way to remove such unwanted background is to apply an upper-limit on the additional tracks in the ID or energy deposits in the calorimeter that are allowed in some cone of radius $\Delta R = \sqrt{\Delta\eta^2 + \Delta\phi^2}$ around the lepton in question. Dedicated isolation working points will usually be defined which use track and calorimeter isolation variables by either cutting on them directly or through some parameterized p_{T} gradient.

Calorimeter-based Isolation

The calorimeter isolation variable $E_{\mathrm{T,cone}}^{\mathrm{isol}}$ is calculated first by summing the energies of all positive-energy (EM scale) topo-clusters, whose barycenters fall within the cone of radius ΔR. In the case of electrons, the core energy deposited by the candidate is then subtracted by removing cells included in a $\delta\eta \times \delta\phi = 0.125 \times 0.175$ rectangle around the candidate's

direction. This schema is depicted in Figure 3.22. Additional core leakage is also corrected for afterwards. In the case of muons, the estimated energy loss from their traversing the calorimeter is subtracted. Finally, the contributions from pileup and the underlying event to the isolation cone is estimated and also subtracted.

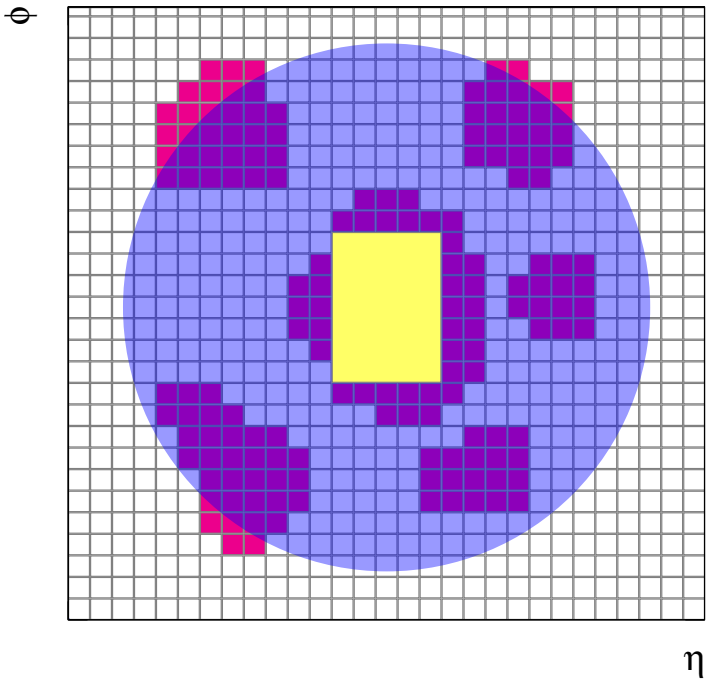


Figure 3.22: Visual depiction of the schema for calculating the calorimeter isolation variable in the case of electrons. The candidate electron is located at the center of the circle which represents the isolation cone. All topo-clusters whose barycenters fall within the isolation cone are also depicted. The 5×7 cell (covering an area of $\delta\eta \times \delta\phi = 0.125 \times 0.175$) represents the subtracted core. [28]

Track-based Isolation

The track isolation variable $p_{\mathrm{T}}^{\mathrm{isol}}$ is calculated using tracks with $p_{\mathrm{T}} > 1$ GeV and that satisfy basic track-quality requirements. In order to reduce the impact of pileup, a requirement on the track compatibility with the primary vertex can also be used. The $p_{\mathrm{T}}^{\mathrm{isol}}$ variable is then constructed by summing the transverse momenta of the tracks which lie in a cone of radius ΔR around the lepton track, excluding the candidate's own contribution. In the case of electrons, other tracks that fall within a $\delta\eta \times \delta\phi = 0.05 \times 0.1$ window of the candidate's EM calorimeter cluster are considered part of the candidate due to bremsstrahlung radiation and are therefore also removed from the computation of the track isolation variable. Due to the smaller tracker granularity compared with the calorimeter, the cone size for the track isolation variable can also be much narrower. In some cases (e.g. for boosted topologies) it is beneficial to use a variable-cone-size track isolation, $p_{\mathrm{T,var}}^{\mathrm{isol}}$, which progressively decreases the cone as the p_{T} of the candidate becomes larger:

$$\Delta R = \min\left(\frac{10\ \mathrm{GeV}}{p_{\mathrm{T}}\ [\mathrm{GeV}]}, R_{\mathrm{max}}\right), \tag{3.11}$$

where R_{max} is the maximum cone size (typically 0.2 to 0.4) and the value of 10 GeV is designed to maximize background rejection.

3.5.6 Jets

Quarks and gluons that are produced as final state particles in a pp collision will undergo parton showering and subsequent hadronization as described in subsection 2.2.2. Their experimental signature will therefore consist of a spray of particles with a common orientation (referred to as a *jet*) and can be identified as a collection of energy deposits in the calorimeter and associated tracks in the ID. The jets used for the analysis presented in this thesis are reconstructed using the anti-k_t algorithm [56] from uncalibrated (EM scale) topo-clusters with a distance parameter of $\Delta R = 0.4$ and as such are sometimes referred to as 'AntiKt4EMTopoJets'. The differences between most sequential recombination algorithms (of which the anti-k_t algorithm is a special case) lie in the definition of the distance measures d_{ij} (the distance between particles i and j) and d_{iB} (the distance between particle i and the beam B):

$$d_{ij} = \min\left(k_{ti}^{2p}, k_{tj}^{2p}\right)\frac{R_{ij}^2}{R^2} \tag{3.12}$$

$$d_{iB} = k_{ti}^{2p} \tag{3.13}$$

where $R_{ij}^2 = (y_i - y_j)^2 + (\phi_i - \phi_j)^2$ and k_{ti}, y_i and ϕ_i are the transverse momentum, rapidity and azimuth of particle i. The parameter R is some fixed distance measure which dictates the closest proximity that any two final jets can be located with respect to one another, while the parameter p governs the relative power of the energy versus geometrical (R_{ij}) scales with $p = -1$ defining the anti-k_t algorithm. The algorithm proceeds by iteratively finding the object pair with the smallest distance and merging them if they are both clusters or identifying one as a jet and removing the cluster from the list if the other is a beam until no unmatched clusters are left. One advantage of the anti-k_t algorithm is that it is known to be resilient with respect to soft radiation.

The four-momenta of the resulting jets are determined by adding the four-momenta of their associated constituents, which are assumed to be massless. The jet energies are also calibrated and corrected through "local hadronic cell weighting" (LCW) calibration [35, 16] (and in so doing moving them from the electromagnetic energy scale to the hadronic energy scale), in which the primary goal is to correct for the non-compensating calorimeter response - that is, to correct for the fact that the calorimeter signal for hadrons is smaller than the one for electrons and photons depositing the same energy. The LCW calibration also corrects for other effects such as signal loss due to the way in which the cells are clustered and energy loss from inactive material. The resulting uncertainties in the jet energy scale (JES) and jet energy resolution (JER) are among the largest systematic uncertainties in the results presented in this thesis.

As the multiplicity of pp collisions per bunch crossing seen by ATLAS continues to rise, the suppression of pileup jets becomes increasingly important. To this end, a dedicated observable called the jet-vertex-tagger (JVT) is constructed as a multivariate combination of the fractional transverse momentum of tracks within a jet that are associated with the hard-scatter primary vertex (the jet vertex fraction or JVF) corrected for the number of primary vertices in the event and the scalar p_{T} sum of the tracks in a jet originating from the hard-scatter primary vertex divided by the fully calibrated jet p_{T} [13]. Defined in this

way, jets from the hard-scatter primary vertex will tend towards larger values of JVT, while pileup jets will tend towards lower values of JVT.

3.5.7 Heavy Flavor Tagging

During the reconstruction of a jet, no attempt is made to determine whether it was initiated by a particular quark flavor or gluon. However, it is often very useful to identify jet origins. Specifically in the case of jets originating from b-quarks, it becomes feasible to classify (or *tag*) them due to the long lifetime of the b-hadrons which they contain. The b-hadrons will often travel on the order of a few millimeters before they decay, leaving behind tracks with relatively large impact parameters and secondary vertices which can be reconstructed.

The process of b-tagging is often left to sophisticated multivariate techniques. For the analysis described in this thesis, the MV2c10 algorithm is used, which offers a high-level boosted decision tree (BDT) discriminant based on a number of lower level taggers that utilize relevant quantities such as the track impact parameters or secondary vertices mentioned above. An example of the BDT output from the MV2c10 algorithm is shown in Figure 3.23. Multiple working points are also provided for a single algorithm. For instance, the working point MV2c10 85% corresponds to a fixed cut on the output discriminant such that close to 85%[14] of b-jets are successfully tagged [33].

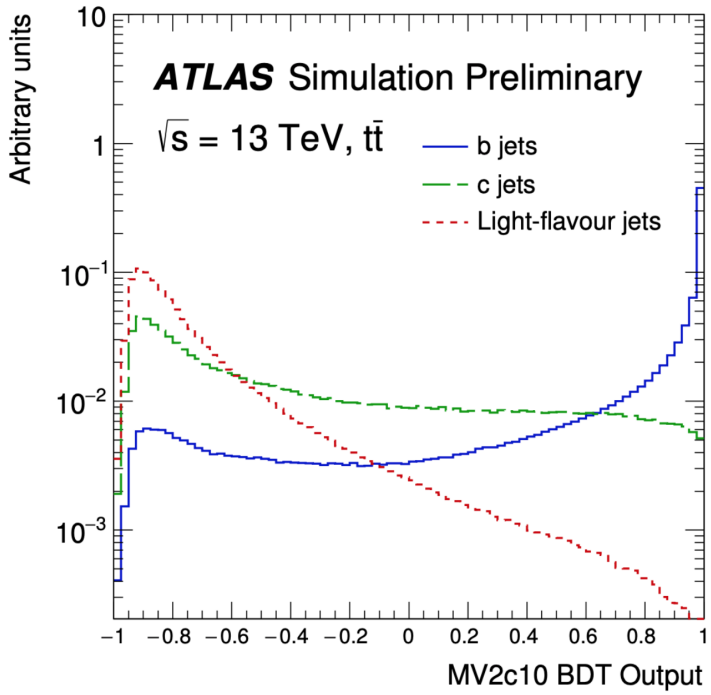


Figure 3.23: MV2c10 BDT output for different jet flavors evaluated with $t\bar{t}$ events. [21]

3.5.8 Missing Transverse Momentum

Particles which only interact via the weak force such as neutrinos or some potential beyond the standard model (BSM) signatures will travel through the detector and escape

[14] The b-tagging efficiency quoted in the working point serves as more of a guidance, since the performance is known to be strongly p_{T} dependant.

measurement altogether. In hadron colliders such as the LHC, the underlying event can carry away an arbitrary amount of momentum in the longitudinal direction. However, with the colliding protons having almost no momentum in the transverse plane, an imbalance of moment in this direction offers a reliable way to infer the presence of otherwise invisible particles.

The missing transverse momentum is an event-wide observable which can be defined as the negative vector sum of the transverse momenta of all visible objects which includes both a hard term and a soft term. The hard term contains all fully reconstructed (identified and calibrated) particles such as electrons, photons, muons and taus, in addition to all high p_{T} hadronic jets. The soft term, on the other hand, contains the rest of the measurements which aren't associated to any of the hard objects and can be determined from either the leftover clusters (CST) or remaining tracks with $p_{\mathrm{T}} > 500$ MeV that pass a set of quality requirements and originate from the primary vertex (TST) [17], the latter of which is used for the remainder of the thesis. The missing transverse momentum as defined above is also denoted as $E_{\mathrm{T}}^{\mathrm{miss}}$.

Real particles which escape detection are not the only sources of missing transverse momentum. Finite detector resolution, object miscalibration, and pileup effects all have the potential to enter the calculation as 'fake' sources. A dedicated object overlap removal procedure must also be performed when combining the hard objects, due to the fact that ATLAS reconstruction domains run independently[15]. One way to mitigate the contribution from pileup is to define the *track-based* missing transverse momentum (denoted as $p_{\mathrm{T}}^{\mathrm{miss}}$) with ID tracks replacing calorimeter jet measurements, since an association of the tracks to the primary vertex can then be imposed. However, this alternative definition of the missing transverse momentum has the disadvantage that it does not take into account neutral particles. Both $E_{\mathrm{T}}^{\mathrm{miss}}$ and $p_{\mathrm{T}}^{\mathrm{miss}}$ are used in the analysis described in this thesis.

[15]For instance, a given jet might also be reconstructed as an electron, a photon and/or a tau. Note however that overlap removal internal to the building of missing transverse momentum is distinct from the overlap removal at the final analysis level described later in subsection 4.3.3 since in the former case each track/cluster must be specifically accounted for.

Chapter 4

Analysis Strategy

This chapter details the analysis strategy for the cross section measurement of Higgs boson production and decay to a pair of W bosons via the two leading production modes, ggF and VBF, using ATLAS data collected in 2015 and 2016. An introduction is first given in section 4.1. The data and Monte Carlo samples used in the analysis are then described in section 4.2. The criteria by which physics objects are selected are provided in section 4.3. Many of the constructed observables used in the analysis are mentioned and motivated in section 4.4, while the event level selection is defined in section 4.5.

4.1 Introduction

While this analysis targets the $H \to WW^*$ decay mode, it does not consider all channels that are available from the subsequent decay of the W bosons. The total branching fraction of the W boson to leptons is $\mathcal{B}_{W\to e\nu/\mu\nu/\tau\nu} = 32.4\%$ [101], with the remainder being hadronic decays. Of the charged leptons, the electron is stable and the muon is at least stable enough to be measurable directly by the detector. However the τ will decay further, with the branching fraction to lighter leptons being $\mathcal{B}_{\tau\to e\nu/\mu\nu} = 35.2\%$ [101] where in this case the τ decay becomes nearly indistinguishable from a prompt electron or muon. Therefore in the context of this analysis, the lepton symbol ℓ refers only to a light lepton - either an electron e or a muon μ - with the contribution from $W \to \tau\nu_\tau \to \ell\nu_\tau\nu_\tau\nu_\ell$ being included implicitly. The effective decay branching fraction of a pair of W bosons to a pair of light leptons is thus 6.4%[1], which contains the final state studied in this analysis.

A clear detector signature is provided with two highly energetic and oppositely charged leptons, allowing signal-like events to be selected using dedicated lepton triggers and hadronic background events to be rejected at the price of a significant amount of statistical power. Furthermore, events with same-flavor opposite-charge leptons are not considered in this analysis due to the fact that these can easily occur through pair production from a Z/γ^* boson (the so-called "Drell-Yan" (DY) process)[2]. Rather, the focus is on events with different flavor leptons, which provide a higher sensitivity to the signal. The leptons in each event are distinguished not only by their flavor, but also by their p_{T} ordering such that the lepton with the highest p_{T} is referred to as "leading" and the lepton with the second highest p_{T} is referred to as "sub-leading". Expressed in this way, a distinction can be made between $e\mu$

[1] This number also includes the decay to same flavor leptons ($ee + \mu\mu$), two channels that are left out of this analysis as mentioned subsequently.

[2] The same-flavor opposite-charge final state was, however, included in the analysis of Run 1 data [18].

and μe events, in which the former contains a leading electron while the latter contains a leading muon.

One of the main signature differences between the ggF and VBF Higgs production modes is that ggF production only produces jets through parton radiation from the initial state partons whereas VBF production is characterized by the presence of two energetic jets with large separation in rapidity (As mentioned in subsection 2.3.1). Partly for this reason, the analysis is split into three separate categories based on the number of jets there are in the event with p_{T} above 30 GeV: events with zero jets and events with exactly one jet targeting the ggF production mode and events with at least two jets targeting the VBF production mode. A trait which is shared by all signal events, however, is the spin 0 nature of the SM Higgs boson, leading to spin correlations in the final state particles - in this case, the leptons being emitted in the same hemisphere, with the two neutrinos being emitted in the opposite one as illustrated in Figure 4.1. This feature can then be further exploited during event selection.

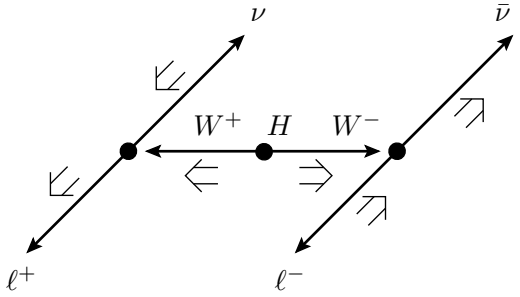


Figure 4.1: Spin correlations in the $H \to WW^* \to \ell\nu\ell\nu$ decay mode. The small arrows indicate the particles' directions of motion, while the larger double arrows indicate their spin projections. The spin 0 Higgs decays to W bosons with opposite spins and the spin 1 W bosons decay to leptons that have their spins aligned. The H and W decays are shown in the rest frame of the decaying particle.

4.2 Data and Monte Carlo Samples

4.2.1 Data Samples

The dataset used for the analysis presented in this thesis corresponds to an integrated luminosity of about 36.1 fb^{-1} taken from the physics_Main output stream in the years 2015 and 2016, with an uncertainty of 2.1% [30]. It is also required to pass a set of quality checks which are encoded by the ATLAS data preparation group into so-called "Good Run Lists" (GRLs) which flag each data taking run with whether or not it is good to use for physics.

A combination of single lepton triggers and one e-μ dilepton trigger are employed, as summarized in Table 4.1. The majority of events are captured by the single lepton triggers with p_{T} thresholds ranging between 24 GeV and 26 GeV for single-electron and between 20 GeV and 26 GeV for single-muon depending on the run period [26], while the dilepton trigger provides additional low p_{T} acceptance with a p_{T} threshold of 17 GeV for electrons and 14 GeV for muons. A second dilepton trigger with an even lower p_{T} threshold of 7 GeV for electrons was investigated, but found to give only a marginal gain in total trigger efficiency, as can be seen in Figure 4.2 which contains the trigger efficiencies under different configurations at the analysis level for each jet category.

Lepton	Level-1 trigger	High-level trigger
Year 2015		
e	20 GeV	24M OR 60M OR 120L GeV
μ	15 GeV	20i OR 50 GeV
$e\mu$	e: 15GeV, μ: 10 GeV	e17_lhloose_mu14
Year 2016		
e	20 GeV	24Ti OR 60M OR 140L GeV
μ	15 GeV	24i OR 50 GeV
$e\mu$	e: 15GeV, μ: 10 GeV	e17_lhloose_mu14
Year 2016: after D4		
e	20 GeV	26Ti OR 60M OR 140L GeV
μ	15 GeV	26i OR 50 GeV
$e\mu$	e: 15GeV, μ: 10 GeV	e17_lhloose_mu14

Table 4.1: Summary of trigger configurations used in the analysis. The minimum p_{T} requirements used for each trigger are shown, while the letters "T", "M" and "L" denote the Tight, Medium and Loose electron identification requirement, respectively. The letter "i" indicates an additional isolation requirement.

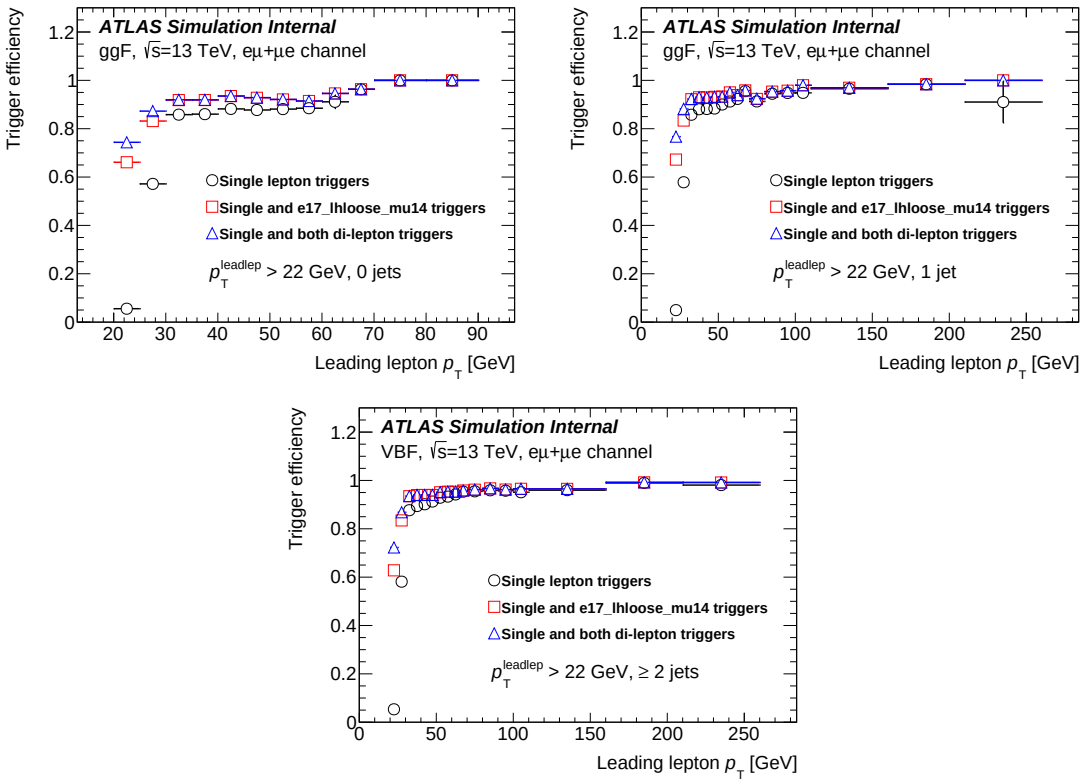


Figure 4.2: Trigger efficiencies as a function of p_{T} for different configurations in the 0 jet ggF (top left), 1 jet ggF (top right), and VBF (bottom) analyses after preselection and requiring a leading lepton with $p_{\mathrm{T}} > 22$ GeV.

4.2.2 Monte Carlo Samples

The Monte Carlo (MC) generators used to model the signal and background processes are listed in Table 4.2. In the majority of cases (with the exception of SHERPA), separate programs are used to generate the hard scattering process and to model the parton showering (PS), hadronization and the underlying event (UE). For instance, PYTHIA 8.210, PYTHIA 8.186 [110] or PYTHIA 6.428 [107] are used in the final three steps for the signal and also background from top processes. The CT10 and NNPDF3.0 parton distributions function (PDF) set [93] is used for the hard scattering process in POWHEG-BOX v2 [100] (most commonly interfaced with PYTHIA), with the AZNLO [31] tune being used for the diboson and signal processes and the A14 tune [12] being used for other processes. The hard scattering NLO predictions from SHERPA 2.2.1 [78] are calculated using the NNPDF 3.0 NNLO PDF set, together with a dedicated set of tuned parameters from the parton shower developed by the SHERPA authors [105]. Pileup is modeled using PYTHIA as detailed in subsection 3.4.2.

Process	Generator	$\sigma \cdot$ Br (pb)	Precision $\sigma_{\text{incl.}}$
ggF $H \to WW$	POWHEG NNLOPS	10.4	NNLO+NNLL
VBF $H \to WW$	POWHEG +PYTHIA 8	0.808	NNLO
WH $H \to WW$	POWHEG +PYTHIA 8 (MINLO)	0.293	NNLO
ZH $H \to WW$	POWHEG +PYTHIA 8 (MINLO)	0.189	NNLO
inclusive $Z/\gamma^\star \to \ell\ell$ ($40 \geq m_{\ell\ell} \geq 10$GeV)	SHERPA 2.2.1	6.80×10^3	NNLO
inclusive $Z/\gamma^\star \to \ell\ell$ ($m_{\ell\ell} \geq 40$GeV)	SHERPA 2.2.1	2.107×10^3	NNLO
$(W \to \ell\nu)\gamma$ ($p_T^\gamma > 7$GeV)	SHERPA 2.2.2	453	NLO
$(Z \to \ell\ell)\gamma$ ($p_T^\gamma > 7$GeV)	SHERPA 2.2.2	175	NLO
$t\bar{t}$ di-leptonic(e, μ, τ)	POWHEG +PYTHIA 8	87.6	NNLO+NNLL
Wt leptonic	POWHEG +PYTHIA 6	7.55	NLO
$q\bar{q}/g \to WW \to \ell\nu\ell\nu$	SHERPA 2.2.2	49.74	NLO
$Z^{(*)}Z^{(*)} \to 2\ell 2\nu$ ($m_{\ell\ell} \geq 4$GeV)	SHERPA 2.1	6.53	NLO
$gg \to 2\ell 2\nu$	SHERPA 2.1	0.87	NLO
$q\bar{q}/g \to \ell\nu\ell\ell$	SHERPA 2.1	11.9	NLO
$q\bar{q}/g, gg \to \ell\ell\ell\ell$	SHERPA 2.1	11.5	NLO
EW $WW + 2$ jets ($\ell\nu\ell\nu$)	SHERPA 2.1	0.012	LO
EW $WZ + 2$ jets ($\ell\nu\ell\ell$)	SHERPA 2.1	0.038	LO
EW $ZZ + 2$ jets ($\ell\ell\ell\ell$)	SHERPA 2.1	0.116	LO
EW $q\bar{q} \to (Z \to \tau\tau)q\bar{q}$	SHERPA	2.54	LO

Table 4.2: Summary of MC generators used to model the signal and background processes in the analysis, along with the corresponding cross sections to which they are normalized (the "Precision $\sigma_{\text{incl.}}$" column shows the accuracy of the cross sections). In the case of the signal, the Higgs mass is set to $m_H = 125$ GeV. In the case of a lepton decay filter being applied on W/Z bosons, the quoted cross section includes branching ratios and is inclusive in lepton flavor.

Signal

The processes considered as signal are the ggF and VBF Higgs production modes[3], with the $H \to WW^* \to e\nu\mu\nu$ decay mode featuring two charged opposite-sign opposite-flavor leptons in the final state. Other Higgs processes are either fixed to SM predictions and included as background (VH production and $H \to \tau\tau$ decay) or not considered due to their negligible

[3]The ggF Higgs production mode is considered as a background in the VBF signal region.

contributions ($t\bar{t}H$ and $b\bar{b}H$ associated production). All signal samples are generated with a Higgs boson mass of 125 GeV.

Higgs boson production via ggF is simulated with NNLO accuracy in QCD using the POWHEG-BOX v2 program [85] and normalized to a cross section calculated with next-to-next-to-next-to-leading-order accuracy in QCD [39]. Higgs boson production via VBF is also simulated using POWHEG-BOX v2, but with NLO accuracy in QCD. It is then normalized to a cross section calculated with NLO accuracy in QCD [62, 41] with an approximate NNLO correction applied [50].

Background

The main sources of Standard Model backgrounds which are present in the analysis include pairs of electroweak bosons which are generally divided into WW and Non-WW (WZ, $W\gamma^{(*)}$, ZZ), production of top-quarks ($t\bar{t}$ and Wt), W or Z bosons produced together with hadronic jets, and QCD multijet events.

The WW background is generated separately for the $qq \to WW$ and $gg \to WW$ production modes. The larger contribution of the two is the $qq \to WW$ process and is generated using SHERPA 2.2.2 [78, 79], with the matrix elements being calculated for up to one additional parton at NLO and up to three additional partons at LO accuracy. The loop-induced $gg \to WW$ process is simulated by SHERPA 2.1.1 with zero or one additional jets [61] and normalized to the NLO $gg \to WW$ cross section [59]. Interference of the two WW production modes are expected to have a negligible contribution in the analysis and are therefore not considered [58].

Top-quark pair production is generated using POWHEG with the POWHEG-BOX framework. A filter is applied, requiring that the W bosons decay leptonically. The samples are normalized to cross sections calculated at NNLO+NNLL accuracy [65]. Single top production is generated with POWHEG-BOX 2.0 and uses EVTGEN 1.2.0 [94] as an afterburner to more accurately model the properties of bottom and charm hadron decays.

Events with $W\gamma$ and $Z\gamma$ are modeled using SHERPA 2.2.2 at NLO accuracy, requiring the p_{T} of the γ to be larger than 7 GeV and to be a distance $\Delta R > 0.1$ from any lepton. The WZ and ZZ processes are generated with SHERPA 2.1, with WZ including also the contribution from $W\gamma^{(*)}$.

Z/γ^* DY production is generated using SHERPA 2.2.1 and is split according to $\max(H_T, p_T^V)$. Furthermore, the $Z \to \tau\tau$ process has an additional filter on the lepton (or hadron) p_{T} in order to better populate the analysis phase space. These samples are normalized to cross sections calculated at NNLO accuracy [98].

While the W+jets process is estimated through a purely data-driven method that is introduced in detail in section 6.1, Monte Carlo samples are still used in order to estimate the sample composition uncertainty of the method as described in section 6.7. To this end, V+jets processes are generated nominally from POWHEG MiNLO interfaced to PYTHIA 8 with the AZNLO tune and alternatively using the LO generator ALPGEN [96] v2.14 interfaced to PYTHIA 6.

4.3 Object Identification and Selection

4.3.1 Electrons and Muons

The electrons used in the analysis are required to have a transverse energy E_{T} greater than 15 GeV and pass "MediumLH" ("TightLH") selection as defined in subsection 3.5.3 for E_{T} greater (smaller) than 25 GeV. Their pseudorapidity must also be within the range of $|\eta| < 2.47$, excluding the transition region $1.37 < |\eta| < 1.52$ between the barrel and end-caps in the liquid argon calorimeter. A requirement (AUTHOR = 1) is added that selects electrons which were reconstructed unambiguously as electrons, which reduces the $W\gamma$ background by nearly half while maintaining a high signal efficiency.

The muons used in the analysis are obtained from the *combined muon* definition, as described in subsection 3.5.4. The muon candidates are then required to pass the "Tight" quality selection and also have $p_{\mathrm{T}} > 15\,\mathrm{GeV}$ and $|\eta| < 2.5$.

All leptons are required to originate from the hard-scatter primary vertex by requiring the absolute value of the longitudinal impact parameter to satisfy $|z_0 \sin\theta| < 0.5\,\mathrm{mm}$. In addition, the significance of the transverse impact parameter is required to be less than $|d_0|/\sigma_{d_0} < 5$ ($|d_0|/\sigma_{d_0} < 3$) for electrons (muons). The leptons are also required to be well isolated, where the working points have been optimized separately for each flavor by taking into account the signal and background efficiency, including specifically the W+jets process [47]. The full electron selection is summarized in Table 4.3, while the full muon selection is summarized in Table 4.4.

p_{T} range	Electron ID	Author	Isolation	Impact parameter
$< 25\,\mathrm{GeV}$	TightLH	1	FixedCutTrackCone40	$\lvert z_0 \sin\theta\rvert < 0.5\,\mathrm{mm}$, $\lvert d_0\rvert/\sigma_{d_0} < 5$
$> 25\,\mathrm{GeV}$	MediumLH	1	Gradient	

Table 4.3: Electron selection used in the analysis.

p_{T} range	Muon ID	Calo Isolation	Track Isolation	Impact parameter
$> 15\,\mathrm{GeV}$	"Tight"	$E_T^{\mathrm{cone20}}/p_{\mathrm{T}} < 0.09$	$p_{\mathrm{T}}^{\mathrm{varcone30}}/p_{\mathrm{T}} < 0.06$	$\lvert z_0 \sin\theta\rvert < 0.5\,\mathrm{mm}$, $\lvert d_0\rvert/\sigma_{d_0} < 3$

Table 4.4: Muon selection used in the analysis.

4.3.2 Jets and Missing Transverse Momentum

The jets used in the analysis are reconstructed using the anti-k_t algorithm with a distance parameters of $R = 0.4$ and calorimeter clusters for input as described in subsection 3.5.6. They are required to have $p_{\mathrm{T}} > 30\,\mathrm{GeV}$ and be within the range $|\eta| > 4.5$.

Jets with $p_{\mathrm{T}} < 60\,\mathrm{GeV}$ and $|\eta| < 2.4$ are also required to have a JVT value larger than 0.59 in order to suppress jets from pileup events. Furthermore, pileup jets are reduced in the forward region by applying a requirement on ForwardJVT which is a variable designed to identify pileup jets in the forward region ($|\eta| > 2.5$), outside tracking acceptance [92].

Jets containing b-hadrons are identified using the MV2c10 b-tagging algorithm as described in subsection 3.5.7 with the 85% efficiency working point and subsequently referred to as b-jets. A b-jet veto is applied in each signal region definition for jets with $p_{\mathrm{T}} > 20\,\mathrm{GeV}$, where jets in the range $20\,\mathrm{GeV} < p_{\mathrm{T}} < 30\,\mathrm{GeV}$ are referred to as "sub-threshold"

since they are below the nominal p_{T} requirement defining the jet categories of the analysis and are included in the b-veto in order to improve the suppression of the top backgrounds.

The missing transverse momentum is employed in the analysis using both of the two distinct definitions as introduced in subsection 3.5.8. On one hand, the track-based $p_{\mathrm{T}}^{\mathrm{miss}}$ enters as a cut in the ggF preselection to reduce backgrounds. On the other, the track-based soft-term $E_{\mathrm{T}}^{\mathrm{miss}}$ is used to build signal-sensitive variables such as m_{T}, $p_{\mathrm{T}}^{\mathrm{tot}}$, and $m_{\tau\tau}$ (each defined in section 4.4) since it offers superior resolution.

4.3.3 Overlap Removal

The overlap removal (OR) of physics objects used in the analysis addresses both the topic of duplication, i.e. the reconstruction of one true object as two separate objects, and the topic of isolation, i.e. the treatment of two separate but close-by objects. Due to the latter, there is a close connection between overlap removal and object identification. The following ordered steps are taken to remove object overlaps, with the rejected objects not contributing to subsequent steps:

- **electron-muon**: A duplication of a muon as an electron is possible if the muon radiates a hard photon, e.g. in the case of final state radiation (FSR) or bremsstrahlung, and the subsequent calorimeter energy deposit is matched with the muon track. When this happens, the two objects are typically less than 0.01 in ΔR and often share an ID track. If a combined muon shares an ID track with an electron, the electron is removed. Same flavor lepton overlaps are dealt with during reconstruction.

- **electron-jet**: Electrons can also be reconstructed as a jet because of their calorimeter energy deposits. In order to remove these duplicates, the jet is removed if $\Delta R(\mathrm{jet}, e) < 0.2$. For any surviving jets, the electron is removed if $\Delta R(\mathrm{jet}, e) < \min(0.4, 0.04 + 10\ \mathrm{GeV}/p_{\mathrm{T}}^{e})$ since it is likely to be the result of a real decay product (either light or heavy flavor) of the jet. Another reason to remove these electrons is that their reconstruction becomes bias due to the close-by jet.

- **muon-jet**: The duplication of muons as electrons also often comes with a duplication as a jet. Therefore if $\Delta R(\mathrm{jet}, \mu) < 0.2$, then the jet is removed if the jet has less than three associated tracks with $p_{\mathrm{T}} > 500$ MeV or both of the following conditions are met: the p_{T} ratio of the muon and jet is larger than 0.5 ($p_{\mathrm{T}}^{\mu}/p_{\mathrm{T}}^{\mathrm{jet}} > 0.5$) and the ratio of the muon p_{T} to the sum of p_{T} of tracks with $p_{\mathrm{T}} > 500$ MeV associated to the jet is larger than 0.7. For any surviving jets, the muon is removed if $\Delta R(\mathrm{jet}, \mu) < \min(0.4, 0.04 + 10\ \mathrm{GeV}/p_{\mathrm{T}}^{\mu})$ since it is likely to be the result of a real heavy flavor decay product of the jet.

4.4 Composite Observables

4.4.1 Background Rejection

The following observables are constructed in order to reduce the analysis backgrounds:

- $p_{\mathrm{T}}^{\ell\ell}$ - Transverse momentum of the dilepton system. Large values reflect that the leptons and neutrinos are emitted in opposite hemispheres, corresponding to the signal topology.

- $\Delta\phi(\ell\ell, E_{\mathrm{T}}^{\mathrm{miss}})$ - Azimuthal angle between the dilepton system and $E_{\mathrm{T}}^{\mathrm{miss}}$, providing an additional handle for enhancing the topology described above. Strongly peaked at back-to-back for signal and the majority of backgrounds.

- $m_{\tau\tau}$ - Invariant mass of the hypothetical $\tau\tau$ system under the assumption of the Collinear Approximation Method [104] in which the τ leptons are sufficiently boosted to force their decay products to be collinear. The energy fractions of the neutrinos can then be computed given that they are the only source of the observed $E_{\mathrm{T}}^{\mathrm{miss}}$. Cutting on such a variable is effective to suppress not only $Z \to \tau\tau$, but also $H \to \tau\tau$.

- $\max\left(m_{\mathrm{T}}^{\ell}\right)$ - The transverse mass of one of the two leptons,

$$m_{\mathrm{T}}^{\ell} = \sqrt{2\, p_{\mathrm{T}}^{\ell} \cdot E_{\mathrm{T}}^{\mathrm{miss}} \cdot \left(1 - \cos\Delta\phi\left(\ell, E_{\mathrm{T}}^{\mathrm{miss}}\right)\right)}, \tag{4.1}$$

  will usually have a large value if the process contains at least one real W boson. Selecting for these scenarios can be accomplished by applying a lower bound on the maximum of the two.

4.4.2 Topological Variables

The following observables are constructed in order to enhance the Higgs signal:

- $m_{\ell\ell}$ - Invariant mass of the leading and subleading leptons in the event. Requiring low $m_{\ell\ell}$ takes advantage of the initial spin zero state of the Higgs, with the leptonic decay products being more collimated in this case than the non-resonant WW background.

- $\Delta\phi_{\ell\ell}$ - Azimuthal angle between the leading and subleading leptons in the event. The Higgs signal tends toward smaller values for the same reason as above.

- m_{T} - Transverse mass, defined as

$$m_{\mathrm{T}} = \sqrt{\left(E_{\mathrm{T}}^{\ell\ell} + E_{\mathrm{T}}^{\mathrm{miss}}\right)^2 - \left|\mathbf{p}_{\mathrm{T}}^{\ell\ell} + \boldsymbol{E}_{\mathrm{T}}^{\mathrm{miss}}\right|^2} \tag{4.2}$$

  where

$$E_{\mathrm{T}}^{\ell\ell} = \sqrt{|\mathbf{p}_{\mathrm{T}}^{\ell\ell}|^2 + m_{\ell\ell}^2}. \tag{4.3}$$

  Rather than applying an explicit cut, it is used as a discriminating variable since the Higgs candidates are expected to peak in its distribution.

4.4.3 VBF Observables

Due to the distinguishable event topology of the VBF Higgs production mode, additional observables are constructed in order to discriminate the signal from background by serving as input to a multivariate analysis:

- m_{jj} - Invariant mass of the two leading jets in the event. Large values are characteristic of the VBF signal.

- Δy_{jj} - Gap in rapidity between the two leading jets in the event. A large separation is indicative of the forward jets produced by the VBF signal.

- $\sum_\ell C_\ell$ - Lepton η centrality, providing a way of quantifying the lepton positions with respect to the two leading jets in η:

$$\begin{aligned} \mathrm{OLV}_{l_0} &= 2 \cdot |\frac{\eta_{l_0} - \bar{\eta}}{\eta_{j_0} - \eta_{j_1}}| \\ \mathrm{OLV}_{l_1} &= 2 \cdot |\frac{\eta_{l_1} - \bar{\eta}}{\eta_{j_0} - \eta_{j_1}}| \\ \sum_\ell C_\ell &= \mathrm{OLV}_{l_0} + \mathrm{OLV}_{l_1} \end{aligned} \tag{4.4}$$

  where $\bar{\eta} = (\eta_{j_0} + \eta_{j_1})/2$ is the average η of the two leading jets. Positive values of OLV_l corresponds to the lepton being outside the leading jet rapidity gap, while negative values correspond to the lepton being within the leading jet rapidity gap.

- $\sum_{\ell,j} m_{\ell j}$ - Sum of the invariant masses of all four lepton-jet pairs. This variable can be helpful due the to relatively large opening angles between leptons and jets for the VBF signal.

- $p_{\mathrm{T}}^{\mathrm{tot}}$ - Total transverse momentum of objects in the event, defined as $\boldsymbol{p}_{\mathrm{T}}{}^{l1} + \boldsymbol{p}_{\mathrm{T}}{}^{\mathrm{l2}} + E_{\mathrm{T}}^{\mathrm{miss}} + \sum \boldsymbol{p}_{\mathrm{T}}{}^{\mathrm{jets}}$. This variable is larger for backgrounds with significant soft gluon radiation that carry away momentum in jets which don't pass the analysis level threshold.

4.5 Event Selection

The analysis is divided into three categories based on the number of reconstructed jets, with $N_{\mathrm{jet}} = 0$ and $N_{\mathrm{jet}} = 1$ targeting the measurement in the ggF production mode and $N_{\mathrm{jet}} \geq 2$ targeting the measurement in the VBF production mode. Aside from the difference in signal topology between ggF and VBF, these categories also take advantage of the differences in background composition between the jet bins, as can be seen in Figure 4.3. A summary of the event selections for each jet category is provided in Table 4.5. The rest of this section is devoted to describing them in more detail, beginning with a common preselection and followed by signal region specific requirements. For the VBF category, a boosted decision tree (BDT) is trained to separate signal and background which is also introduced below.

4.5.1 Preselection

The preselection common to all three jet categories (after the object-level selection detailed in section 4.3) is defined as:

- exactly two opposite sign and different flavor leptons ($e\mu + \mu e$)

- $p_{\mathrm{T}}^{\mathrm{lead}} > 22$ GeV, $p_{\mathrm{T}}^{\mathrm{sublead}} > 15$ GeV

- $m_{ll} > 10$ GeV in order to remove low mass meson resonances and DY events

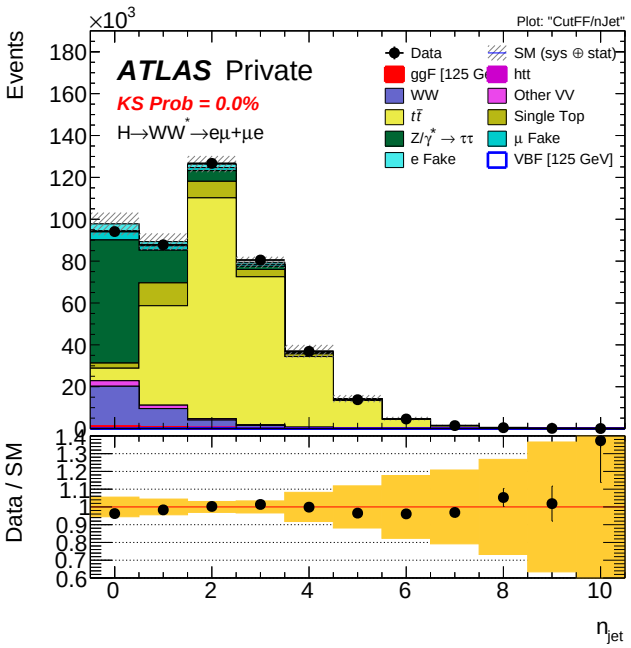


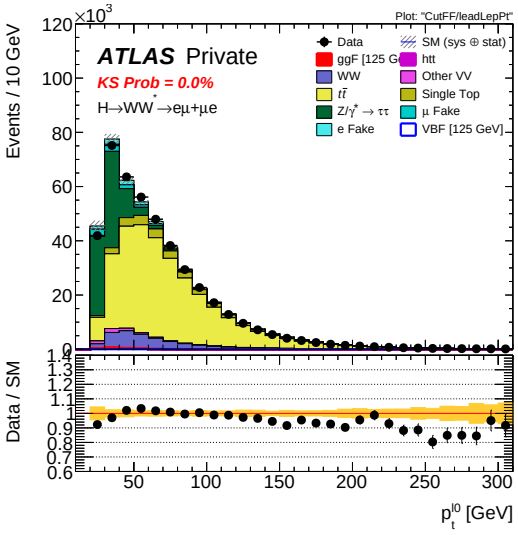
Figure 4.3: Distribution of jet multiplicity after the preselection. No normalization factors are applied to the background. The yellow band represents the MC statistical uncertainties and the main sources of detector systematics (JES/JER/$E_{\mathrm{T}}^{\mathrm{miss}}$/$b$-tagging/Leptons).

Category	$N_{\mathrm{jet},(p_{\mathrm{T}}>30\ \mathrm{GeV})}=0$ ggF	$N_{\mathrm{jet},(p_{\mathrm{T}}>30\ \mathrm{GeV})}=1$ ggF	$N_{\mathrm{jet},(p_{\mathrm{T}}>30\ \mathrm{GeV})}\geq 2$ VBF
Preselection	Two isolated, different-flavor leptons ($\ell=e,\mu$) with opposite charge $p_{\mathrm{T}}^{\mathrm{lead}}>22$ GeV , $p_{\mathrm{T}}^{\mathrm{sublead}}>15$ GeV $m_{\ell\ell}>10$ GeV $p_{\mathrm{T}}^{\mathrm{miss}}>20$ GeV		
Background rejection	$\Delta\phi(\ell\ell, E_{\mathrm{T}}^{\mathrm{miss}})>\pi/2$ $p_{\mathrm{T}}^{\ell\ell}>30$ GeV	$N_{b\text{-jet},(p_{\mathrm{T}}>20\ \mathrm{GeV})}=0$ $\max\left(m_{\mathrm{T}}^{\ell}\right)>50$ GeV $m_{\tau\tau}<m_Z-25$ GeV	BDT trained at this level
$H\to WW^*\to e\nu\mu\nu$ topology	$m_{\ell\ell}<55$ GeV $\Delta\phi_{\ell\ell}<1.8$		central jet veto outside lepton veto
Discriminant variable BDT input variables	m_{T}		BDT m_{jj}, Δy_{jj}, $m_{\ell\ell}$, $\Delta\phi_{\ell\ell}$, m_{T}, $\sum_{\ell}C_{\ell}$, $\sum_{\ell,j}m_{\ell j}$, $p_{\mathrm{T}}^{\mathrm{tot}}$

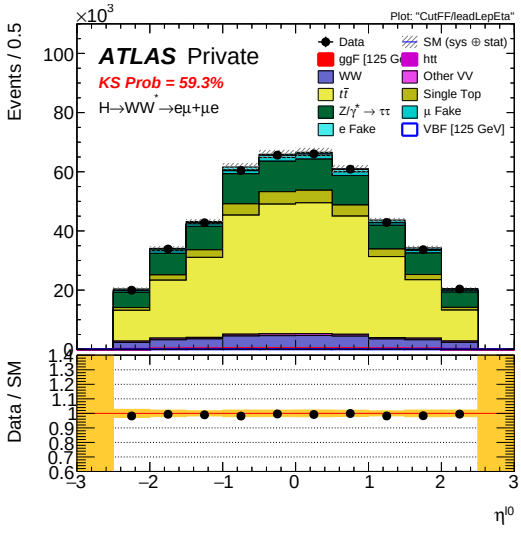
Table 4.5: Event selection criteria used to define the signal regions for both the ggF and VBF production modes.

The p_{T} and η distributions of the leading and subleading leptons in the event after these cuts have been applied are shown in Figure 4.4.

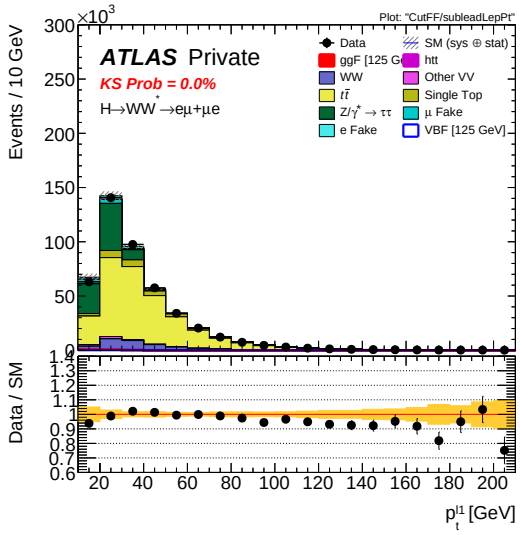
In addition to the above requirements, a cut of $P_{\mathrm{T}}^{\mathrm{miss}}>20\,\mathrm{GeV}$ is also imposed for the ggF categories ($N_{\mathrm{jet}}=0$ and $N_{\mathrm{jet}}=1$) in order to reject a significant fraction of $Z\to\tau\tau$ events. The reason for choosing $P_{\mathrm{T}}^{\mathrm{miss}}$ as opposed to $E_{\mathrm{T}}^{\mathrm{miss}}$ (both defined in subsection 3.5.8) is that the former gives better separation of the $Z\to\tau\tau$ background, as can be seen in Figure 4.5. Events are separated into the jet categories after the preselection, based on $N_{\mathrm{jet},(p_{\mathrm{T}}>30\ \mathrm{GeV})}$.



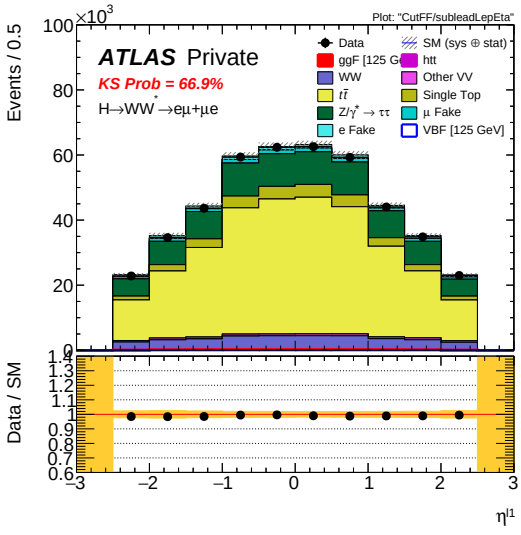
(a) Leading lepton p_{T}



(b) Leading lepton η

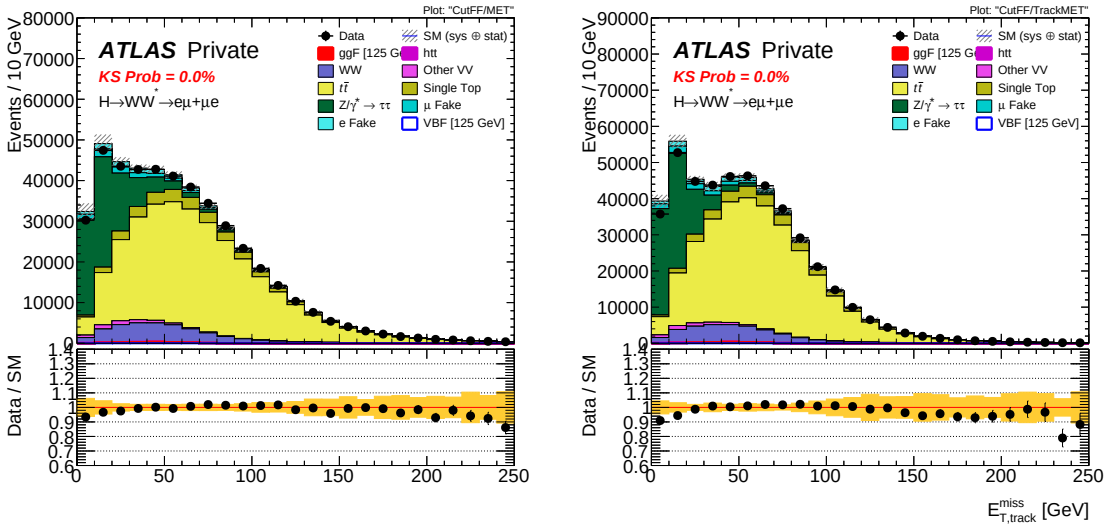


(c) Sub-leading lepton p_{T}



(d) Sub-leading lepton η

Figure 4.4: Distributions of the leading lepton (top) and sub-leading lepton (bottom) p_{T} (left) and η (right) after the common preselection cuts have been applied. The plots show the combination of $e\mu + \mu e$ channels, with reasonable agreement between data and MC observed. No normalization factors are applied to the background. The yellow band represents the MC statistical uncertainties and the main sources of detector systematics (JES/JER/$E_{\mathrm{T}}^{\mathrm{miss}}$/$b$-tagging/Leptons).



(a) $E_{\mathrm{T}}^{\mathrm{miss}}$ (b) $P_{\mathrm{T}}^{\mathrm{miss}}$

Figure 4.5: Distributions of $E_{\mathrm{T}}^{\mathrm{miss}}$ (left) and $P_{\mathrm{T}}^{\mathrm{miss}}$ (right) after the common preselection cuts have been applied. No normalization factors are applied to the background. The yellow band represents the MC statistical uncertainties and the main sources of detector systematics (JES/JER/$E_{\mathrm{T}}^{\mathrm{miss}}$/$b$-tagging/Leptons).

4.5.2 ggF Selection

For the ggF categories, a purely cut-based selection is employed. In the $N_{\text{jet}} = 0$ category, the background is dominated by non-resonant WW production whereas in the $N_{\text{jet}} = 1$ category, the background contains a significant contribution from $t\bar{t}$ and Wt processes. Thus, a separate selection exists for each case.

0 Jet Category

For the 0 jet category, the following cuts are made in order to reject background:

- $p_{\text{T}}^{\ell\ell} > 30\,\text{GeV}$, rejecting mainly $Z \to \tau\tau$ events
- $\Delta\phi(\ell\ell, E_{\text{T}}^{\text{miss}}) > \pi/2$, removing potentially pathological events in which the charged lepton system and the $E_{\text{T}}^{\text{miss}}$ are pointing in the same direction
- $N_{b\text{-jet},(p_{\text{T}}>20\text{ GeV})} = 0$, b-tagging veto on jets (including sub-threshold) rejecting mainly top events

Distributions of $p_{\text{T}}^{\ell\ell}$ and $\Delta\phi_{\ell\ell,E_{\text{T}}^{\text{miss}}}$ are shown in Figure 4.6, after selecting for the 0 jet category. Finally, the following cuts are made in order to exploit the $H \to WW^* \to e\nu\mu\nu$ signal topology:

- $m_{\ell\ell} < 55\,\text{GeV}$
- $\Delta\phi_{\ell\ell} < 1.8\,\text{GeV}$

Distributions of $m_{\ell\ell}$ and $\Delta\phi_{\ell\ell}$ are shown in Figure 4.7, after selecting for the 0 jet category. The full cutflow from preselection to signal region in the 0 jet category is provided in Table 4.6, which also shows the signal significance after each subsequent cut is imposed. A selection of kinematic variables are shown for the final 0 jet signal region (SR) in Figure 4.8.

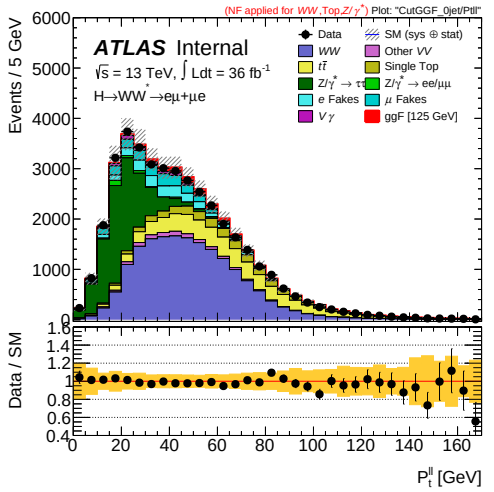
Selection	Higgs	WW	VZ/γ^*	$V\gamma$	Top	Z/DY	e-Fakes	μ-Fakes	Total Bkg	S/B	S/$\sqrt{S+B}$
$N_{\text{jet}} = 0$, $E_{\text{T}}^{\text{miss, track}} > 20$ GeV	819.4 ± 3.2	17669 ± 57	848 ± 14	516 ± 26	7839 ± 39	8362 ± 83	1914 ± 37	1876 ± 32	39024 ± 122	0.02	4.14
$\Delta\Phi_{\ell\ell,E_{\text{T}}^{\text{miss}}} > 1.57$	812.0 ± 3.2	17552 ± 57	798 ± 13	494 ± 25	7637 ± 38	7894 ± 81	1851 ± 36	1779 ± 31	38004 ± 120	0.02	4.16
$p_{\text{T}}^{\ell\ell} > 30$ GeV	692.6 ± 2.9	14236 ± 51	623 ± 12	265 ± 19	6925 ± 37	1170 ± 38	1297 ± 29	1097 ± 21	25613 ± 84	0.03	4.31
$M_{\ell\ell} < 55$ GeV	577.4 ± 2.7	3393 ± 23	194 ± 6	116 ± 12	1094 ± 14	165 ± 12	311 ± 13	322 ± 11	5596 ± 37	0.10	7.42
$\Delta\phi_{\ell\ell} < 1.8$	536.7 ± 2.6	3141 ± 22	182 ± 6	107 ± 11	1057 ± 14	25.4 ± 5.0	269 ± 13	255 ± 9.7	5036 ± 34	0.11	7.26
$N_{b\text{-jet}} = 0$	521.4 ± 2.6	3075 ± 22	175 ± 6	101 ± 11	545 ± 10	24.5 ± 4.7	260 ± 12	240 ± 9.3	4421 ± 32	0.12	7.49

Table 4.6: Cutflow for signal and background processes after each selection requirement applied to the ggF $N_{\text{jet}} = 0$ category. The numbers reflect both $e\mu + \mu e$ channels, where the uncertainty is statistical only.

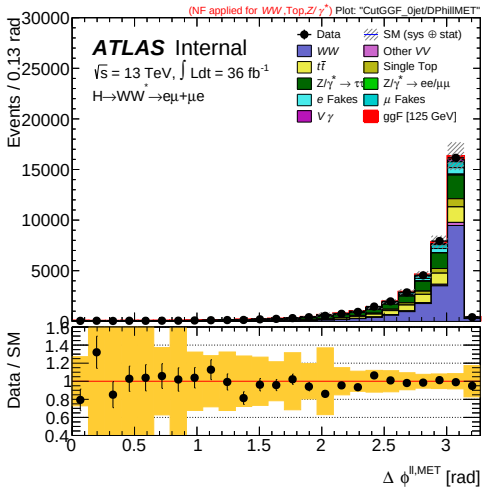
1 Jet Category

For the 1 jet category, the following cuts are made in order to reject background:

- $\max\left(m_{\text{T}}^{\ell}\right) > 50$ GeV, rejecting backgrounds without W bosons
- $m_{\tau\tau} < m_Z - 25$ GeV, $Z \to \tau\tau$ veto rejecting mainly $Z \to \tau\tau$ events
- $N_{b\text{-jet},(p_{\text{T}}>20\text{ GeV})} = 0$, b-tagging veto on jets (including sub-threshold) rejecting mainly top events

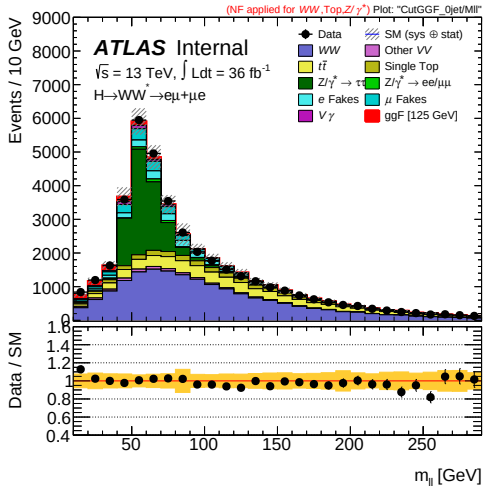


(a) $p_{\mathrm{T}}^{\ell\ell}$

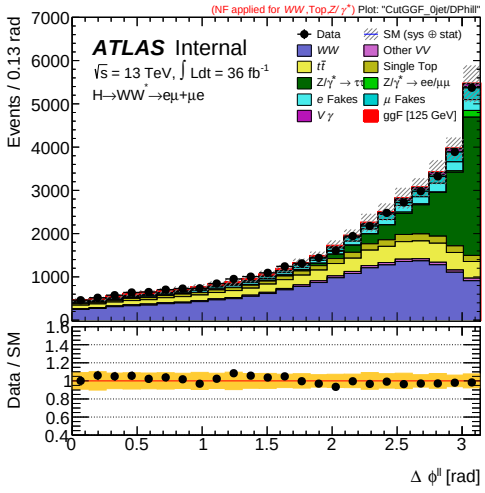


(b) $\Delta\phi_{\ell\ell,E_{\mathrm{T}}^{\mathrm{miss}}}$

Figure 4.6: Distributions of background rejection variables $p_{\mathrm{T}}^{\ell\ell}$ (left) and $\Delta\phi_{\ell\ell,E_{\mathrm{T}}^{\mathrm{miss}}}$ (right) after selecting for the 0 jet category, with $e\mu + \mu e$ channels combined. The normalization factors from Table 5.9 have been applied. The yellow band represents the MC statistical uncertainties and the main sources of detector systematics (JES/JER/$E_{\mathrm{T}}^{\mathrm{miss}}$/$b$-tagging/Leptons).



(a) $m_{\ell\ell}$



(b) $\Delta\phi_{\ell\ell}$

Figure 4.7: Distributions of signal topology enhancing variables $m_{\ell\ell}$ (left) and $\Delta\phi_{\ell\ell}$ (right) after selecting for the 0 jet category, with $e\mu + \mu e$ channels combined. The normalization factors from Table 5.9 have been applied. The yellow band represents the MC statistical uncertainties and the main sources of detector systematics (JES/JER/$E_{\mathrm{T}}^{\mathrm{miss}}$/$b$-tagging/Leptons).

Distributions of $\max\left(m_{\mathrm{T}}^{\ell}\right)$ and $m_{\tau\tau}$ are shown in Figure 4.9, after selecting for the 1 jet category. Finally, the same cuts are made to exploit the $H \to WW^* \to e\nu\mu\nu$ signal topology as in the 0 jet category:

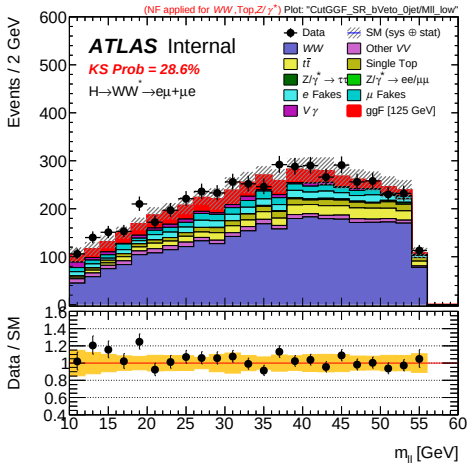
- $m_{\ell\ell} < 55\,\mathrm{GeV}$

- $\Delta\phi_{\ell\ell} < 1.8\,\text{GeV}$

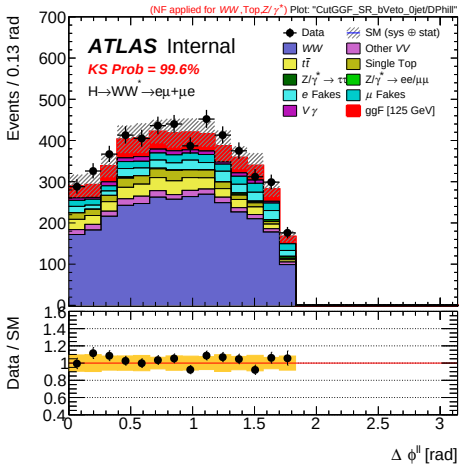
Distributions of $m_{\ell\ell}$ and $\Delta\phi_{\ell\ell}$ are shown in Figure 4.10, after selecting for the 1 jet category. The full cutflow from preselection to signal region in the 1 jet category is provided in Table 4.7, which also shows the signal significance after each subsequent cut is imposed. A selection of kinematic variables are shown for the final 1 jet signal region (SR) in Figure 4.11.

Selection	Higgs	WW	VZ/γ^*	$V\gamma$	Top	Z/DY	e-Fakes	μ-Fakes	Total Bkg	S/B	$S/\sqrt{S+B}$
$N_{\text{jet}} = 1$, $E_{\text{T}}^{\text{miss, track}} > 20$ GeV	569.6 ± 2.3	7934 ± 36	769 ± 14	446 ± 25	55803 ± 105	6138 ± 73	1630 ± 44	1559 ± 34	74279 ± 147	0.01	2.10
$N_{b\text{-jet}} = 0$	516.9 ± 2.2	7315 ± 34	686 ± 13	392 ± 23	8836 ± 40	5452 ± 68	1111 ± 31	1058 ± 26	24850 ± 99	0.02	3.27
$m_{\tau\tau} < m_Z - 25$GeV veto	357.9 ± 1.9	4812 ± 28	367 ± 9	130 ± 14	5810 ± 32	832 ± 25	513 ± 21	397 ± 15	12863 ± 58	0.03	3.13
$m_{\ell\ell} < 55$GeV	304.9 ± 1.7	1210 ± 14	131 ± 5	83 ± 10	1353 ± 15	305 ± 14	167 ± 10	171 ± 9.8	3420 ± 31	0.09	5.02
$\Delta\phi_{\ell\ell} < 1.8$	284.8 ± 1.7	1102 ± 13	117 ± 5	71 ± 9.5	1281 ± 15	66 ± 7	143 ± 10	132 ± 7.8	2911 ± 27	0.10	5.06

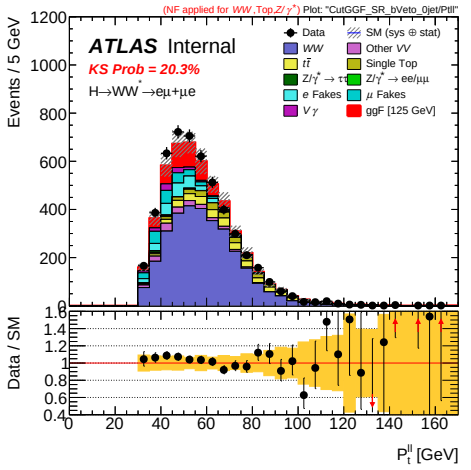
Table 4.7: Cutflow for signal and background processes after each selection requirement applied to the ggF $N_{\text{jet}} = 1$ category. The numbers reflect both $e\mu + \mu e$ channels, where the uncertainty is statistical only.



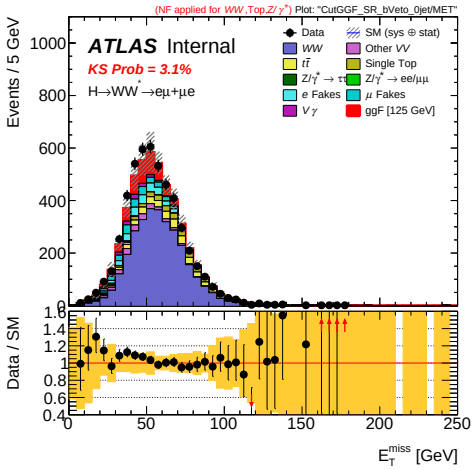
(a) $m_{\ell\ell}$



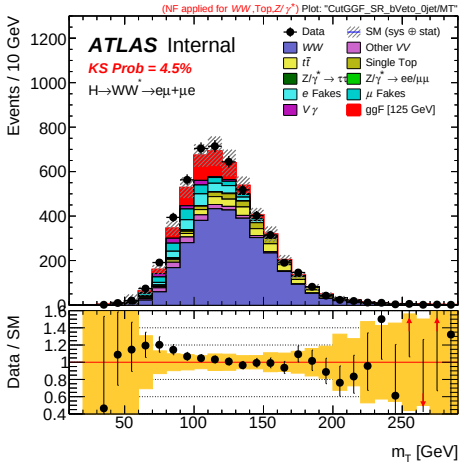
(b) $\Delta\phi_{\ell\ell}$



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

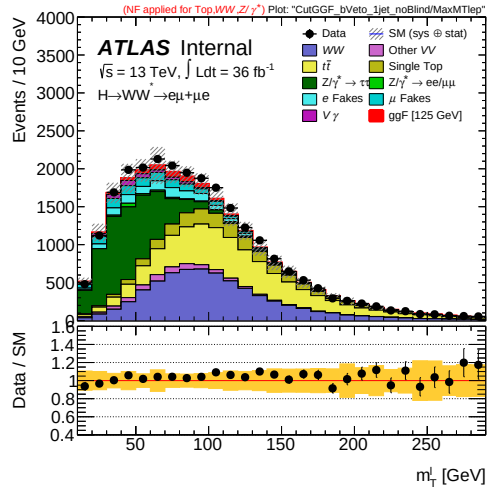


(d) $E_{\mathrm{T}}^{\mathrm{miss}}$

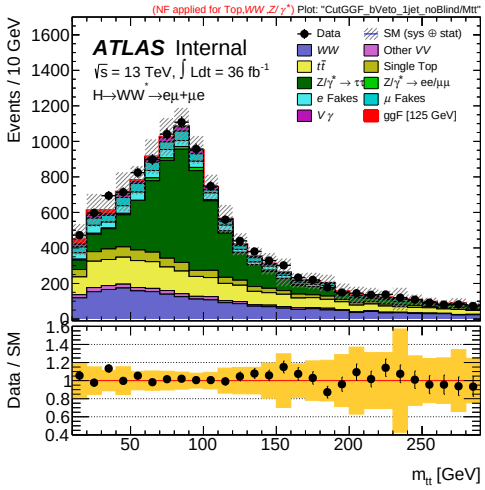


(e) m_{T}

Figure 4.8: Distributions of select kinematic variables in the 0 jet category signal region, with $e\mu + \mu e$ channels combined. The discriminating variable used in the final fit is m_{T} (bottom). The normalization factors from Table 5.9 have been applied. The yellow band represents the MC statistical uncertainties and the main sources of detector systematics (JES/JER/$E_{\mathrm{T}}^{\mathrm{miss}}$/$b$-tagging/Leptons).

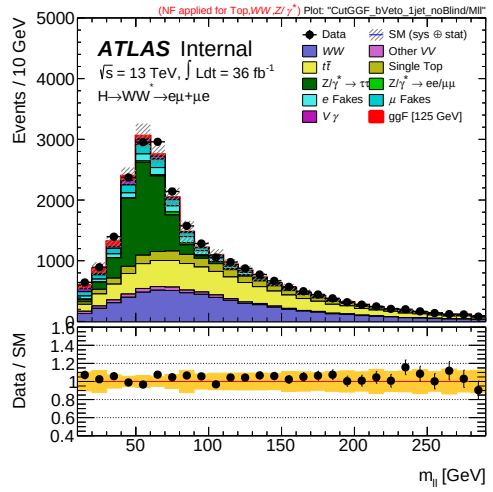


(a) $\max\left(m_{\mathrm{T}}^{\ell}\right)$

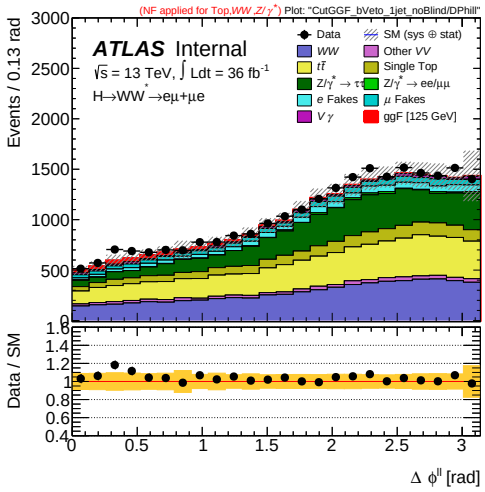


(b) $m_{\tau\tau}$

Figure 4.9: Distributions of background rejection variables $\max\left(m_{\mathrm{T}}^{\ell}\right)$ (left) and $m_{\tau\tau}$ (right) after selecting for the 1 jet category, with $e\mu + \mu e$ channels combined. The normalization factors from Table 5.9 have been applied. The yellow band represents the MC statistical uncertainties and the main sources of detector systematics (JES/JER/$E_{\mathrm{T}}^{\mathrm{miss}}$/$b$-tagging/Leptons).

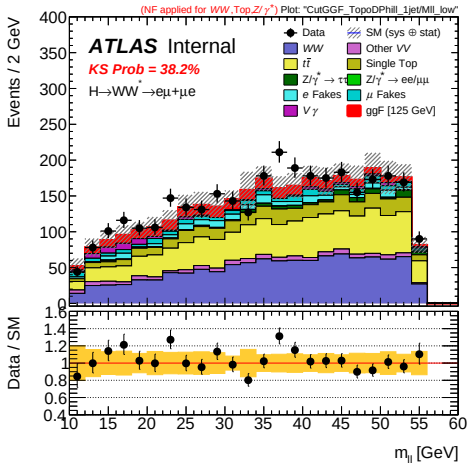


(a) $m_{\ell\ell}$

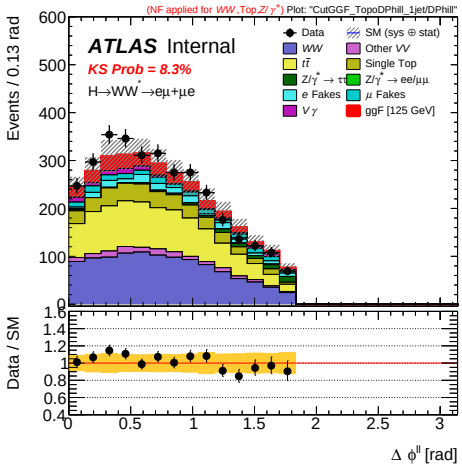


(b) $\Delta\phi_{\ell\ell}$

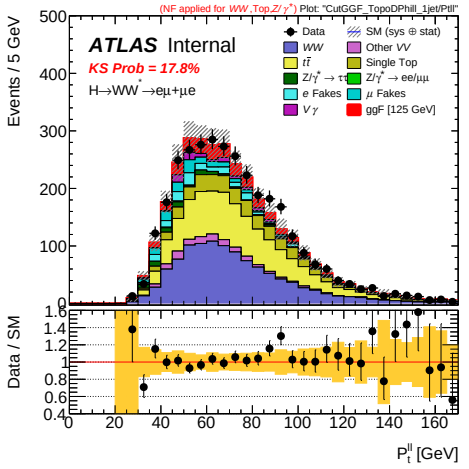
Figure 4.10: Distributions of signal topology enhancing variables $m_{\ell\ell}$ (left) and $\Delta\phi_{\ell\ell}$ (right) after selecting for the 1 jet category, with $e\mu + \mu e$ channels combined. The normalization factors from Table 5.9 have been applied. The yellow band represents the MC statistical uncertainties and the main sources of detector systematics (JES/JER/$E_{\mathrm{T}}^{\mathrm{miss}}$/$b$-tagging/Leptons).



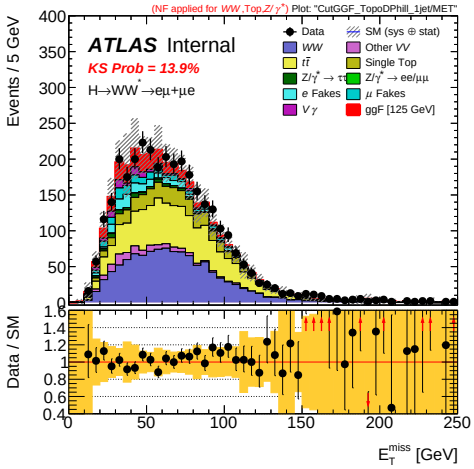
(a) $m_{\ell\ell}$



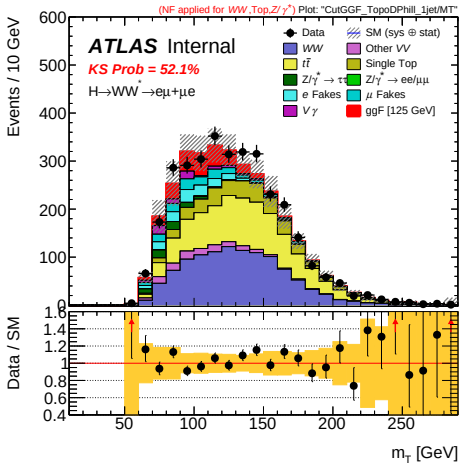
(b) $\Delta\phi_{\ell\ell}$



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



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



(e) m_{T}

Figure 4.11: Distributions of select kinematic variables in the 1 jet category signal region, with $e\mu + \mu e$ channels combined. The discriminating variable used in the final fit is m_{T} (bottom). The normalization factors from Table 5.9 have been applied. The yellow band represents the MC statistical uncertainties and the main sources of detector systematics (JES/JER/$E_{\mathrm{T}}^{\mathrm{miss}}$/$b$-tagging/Leptons).

4.5.3 VBF Selection and BDT

The VBF selection is driven by the distinctive event topology of the signal, with the presence of two energetic jets separated by a large gap in rapidity. Given the lack of statistics relative to the ggF categories, a boosted decision tree (BDT) [51, 2, 69] is employed in order to fully exploit the correlations between a number of discriminating variables.

After the common preselection defined in subsection 4.5.1 and the split into the $N_{\text{jet}} \geq 2$ category, a b-veto is then imposed:

- $N_{b\text{-jet},(p_{\text{T}}>20\ \text{GeV})} = 0$

It is at this point that the BDT is trained. The most relevant physics processes are included (Top, WW and $Z \to \tau\tau$) and the scikit-learn library is used as a back end [102]. The two leading jets in the event are considered the VBF tagged jets and used to construct the main discriminating variables introduced in subsection 4.4.3. In total, 8 variables are used as input: $\Delta\phi_{\ell\ell}$, $m_{\ell\ell}$, Δy_{jj}, m_{jj}, $p_{\text{T}}^{\text{tot}}$, m_{T}, $\sum_{\ell} C_{\ell}$, and $\sum_{\ell,j} m_{\ell j}$. Their distributions are shown in Figure 4.12 and Figure 4.13, in which the VBF signal has also been scaled up to demonstrate discriminatory power.

A grid scan is performed for determining the optimal BDT hyperparameters, with the result being displayed in Table 4.8. In order to prevent the BDT from being trained on the statistical fluctuations of the training sample (an effect known as "over-training"), the BDT is two-fold cross-validated with the BDT trained on even numbered events being applied to odd numbered events and vica versa. The training variables are also ranked in order of importance by counting in how many nodes each were used and weighting every instance with the resulting gain in separation and the number of events in the node. The ranking for the BDT trained on even numbered events is shown in Table 4.9, with m_{jj} and Δy_{jj} being the two most important variables.

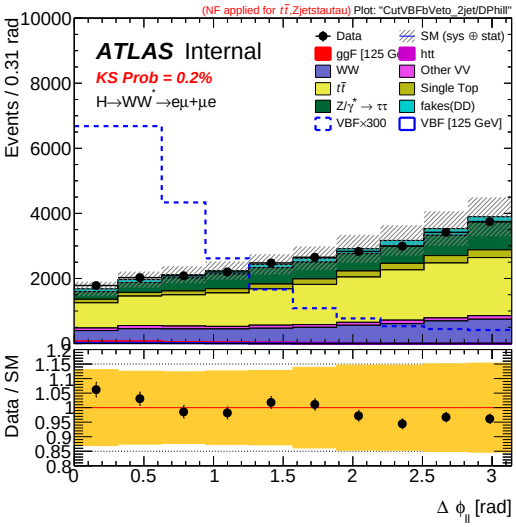
Parameter	Value	Range
Boosting algorithm	Gradient	–
Maximum tree depth	5	[3,16]
Number of trees	200	[200,100]
Minimum number of events required per mode	5%	[5%,20%]
Learning rate	0.1	–

Table 4.8: BDT hyperparameters used for the training.

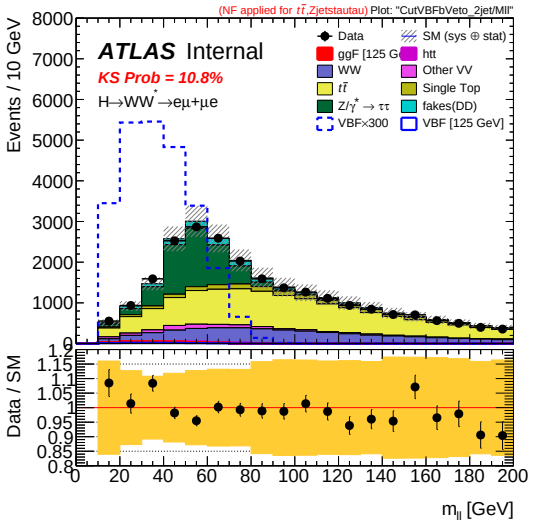
Once the BDT is trained, three additional cuts are placed in order to fully define the VBF signal region in which the BDT is then applied:

- CJV (Central Jet Veto): Events with any jets that have $p_{\text{T}} > 20\,\text{GeV}$ and lie between the two VBF tagged jets in η are rejected
- OLV (Outside Lepton Veto): The two charged leptons must have rapidities which lie between the two VBF tagged jets' rapidity gap
- $m_{\tau\tau} < m_Z - 25$ GeV: $Z \to \tau\tau$ veto

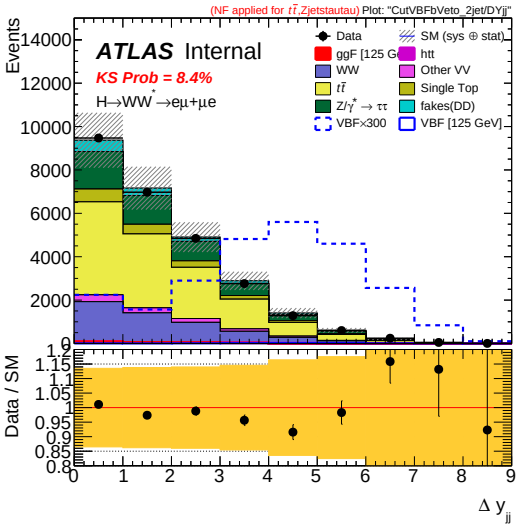
The BDT output in the VBF signal region is shown in Figure 4.14, with a clear separation between signal and background. The BDT output serves as the discriminating variable in the final fit for the VBF category.



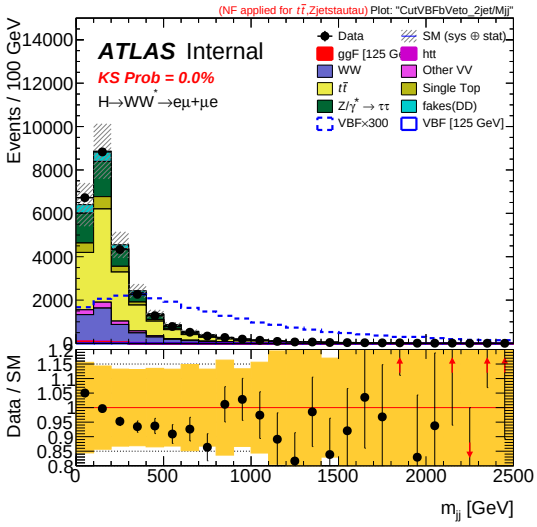
(a) $\Delta\phi_{\ell\ell}$



(b) $m_{\ell\ell}$

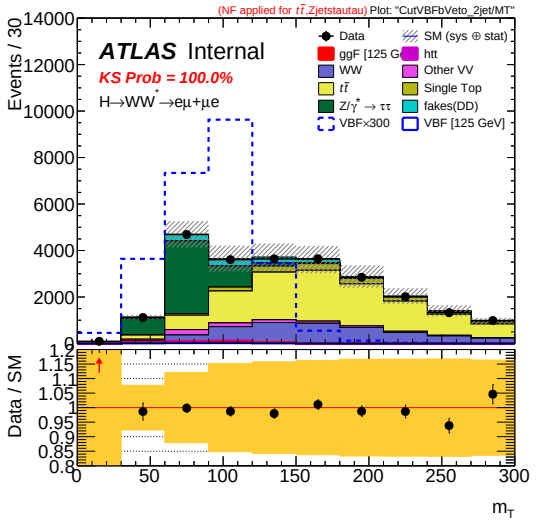


(c) Δy_{jj}

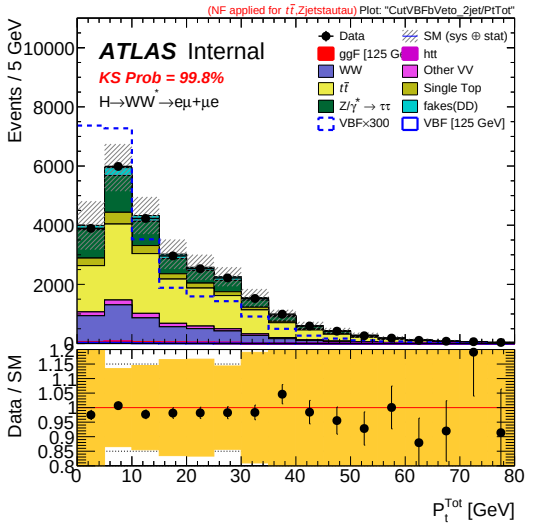


(d) m_{jj}

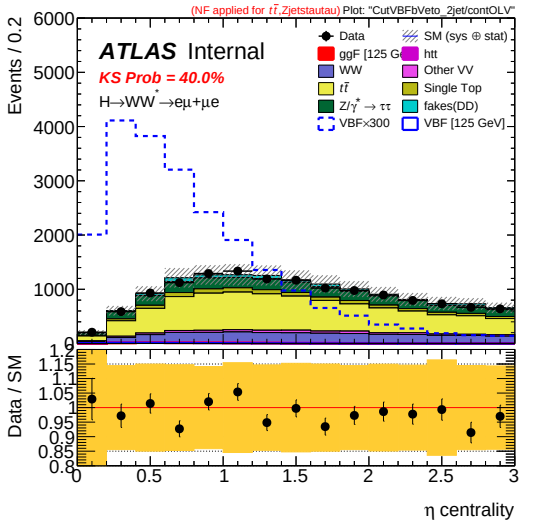
Figure 4.12: Distributions of $\Delta\phi_{\ell\ell}$, $m_{\ell\ell}$, Δy_{jj}, and m_{jj} after the b-jet veto. The VBF signal is scaled by a factor of 300 in order to demonstrate the discriminatory power of each variable. The normalization factors from Table 5.9 have been applied. The yellow band represents the MC statistical uncertainties and the main sources of detector systematics (JES/JER/$E_{\mathrm{T}}^{\mathrm{miss}}$/$b$-tagging/Leptons).



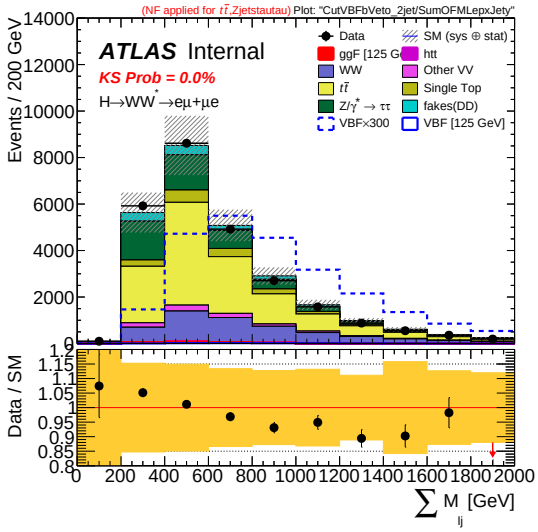
(a) m_{T}



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



(c) $\sum_{\ell} C_{\ell}$

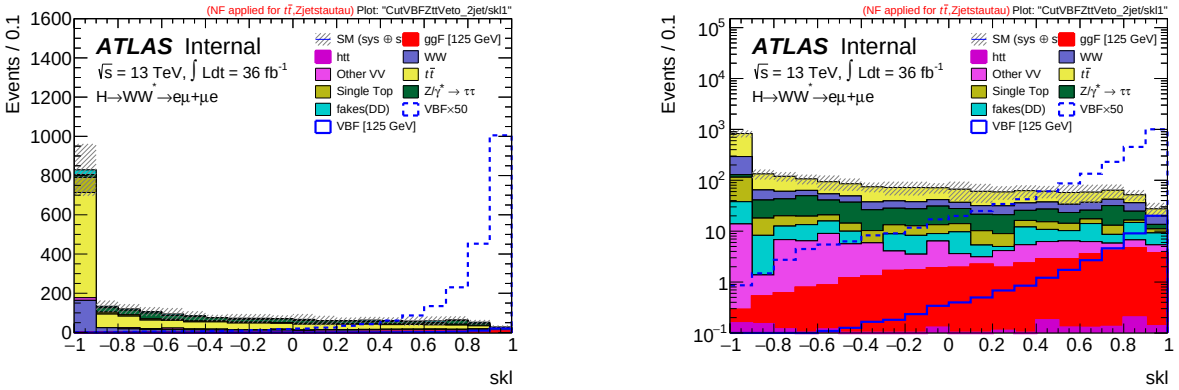


(d) $\sum_{\ell,j} m_{\ell j}$

Figure 4.13: Distributions of m_{T}, $p_{\mathrm{T}}^{\mathrm{tot}}$, $\sum_{\ell} C_{\ell}$, and $\sum_{\ell,j} m_{\ell j}$ after the b-jet veto. The VBF signal is scaled by a factor of 300 in order to demonstrate the discriminatory power of each variable. The normalization factors from Table 5.9 have been applied. The yellow band represents the MC statistical uncertainties and the main sources of detector systematics (JES/JER/$E_{\mathrm{T}}^{\mathrm{miss}}$/$b$-tagging/Leptons).

Ranking	Variable	Importance [%]
1	m_{jj}	19
2	Δy_{jj}	16
3	$m_{\ell\ell}$	14
4	m_{T}	14
5	$\sum_{\ell} C_{\ell}$	13
6	$\Delta\phi_{\ell\ell}$	10
7	$\sum_{\ell,j} m_{\ell j}$	8
8	$p_{\mathrm{T}}^{\mathrm{tot}}$	7

Table 4.9: Ranking of the BDT input variables for the BDT trained on even numbered events. The result for the second BDT is similar.



(a) BDT Output (lin)

(b) BDT Output (log)

Figure 4.14: The BDT distribution in the VBF signal region with linear (left) and logarithmic (right) scale. The VBF signal is scaled by a factor of 50 for visibility.

Chapter 5

Background Estimation

This chapter describes the background processes that are present in the analysis and how they are estimated. An overview is first given in section 5.1, followed by the specifics of each jet category in section 5.2, section 5.3, and section 5.4. Finally, a summary is provided by section 5.5.

5.1 Overview

A multitude of background processes contribute to the final signal region yields in all jet categories which include WW, top ($t\bar{t}$ and Wt), WZ, $W\gamma^{(*)}$, ZZ, W+jets, QCD, $Z \to \tau\tau$, and $Z \to ee/\mu\mu$ events. These backgrounds can be broadly categorized based on their final state properties which allow them to pass the signal region selection.

- All of $t\bar{t}$, Wt, and WW contain two W bosons in the final state, similar to the signal. The presence of b-tagged jets can be used to reject processes containing top quarks. The spin correlation kinematics shown in Figure 4.1 can be used to reject WW background.
- $Z/\gamma^* \to \tau\tau$, and the "Non-WW diboson" processes WZ, $W\gamma^{(*)}$, and ZZ, collectively referred to as VV, have a smaller cross section but also a softer subleading lepton which leads to kinematics similar to the signal.
- W+jets and multijet production via QCD processes have a high cross section and enter the signal region when a jet produces an object that is reconstructed as an isolated lepton. These events also have similar kinematics to the signal since leptons produced by jets tend to be soft.

Some processes (specifically VV) have both the benefits of a small cross section and sufficient modeling by the Monte Carlo generator and theory predictions. Therefore, no control region is used for these backgrounds. Other backgrounds that have more significant contributions to the analyses are estimated from data as much as possible. For WW, top, and $Z \to \tau\tau$ separate control regions and normalization factors are introduced for each jet category with the only exception being the case of WW for the VBF analysis due to limited purity. The W+jets process is treated in a unique way using a data-driven fake factor method. A summary of the background estimation strategy is provided in Table 5.1.

The normalization factors quoted throughout this thesis unless stated otherwise are derived using a simple matrix inversion method in which a system of linear equations is

solved, simultaneously taking into account the yields of each background to be normalized in each of the control regions after the rest of the Monte Carlo processes have been subtracted from data.

Channel	WW	Top	Z/γ^*	VV	W+jets
$N_{\text{jet}} = 0$	CR	CR	CR	MC+VR	Data
$N_{\text{jet}} = 1$	CR	CR	CR	MC+VR	Data
$N_{\text{jet}} \geq 2$	MC+VR	CR	CR	MC+VR	Data

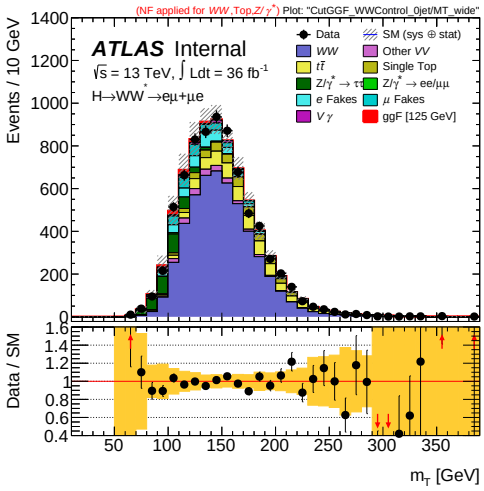
Table 5.1: Summary of the strategy for the treatment of the major backgrounds in each jet category. The estimations are split into three types: normalized from a dedicated control region (CR); data-driven approach (Data); and normalized with Monte Carlo, but agreement with data checked in a validation region (MC+VR).

5.2 ggF 0 Jet Background

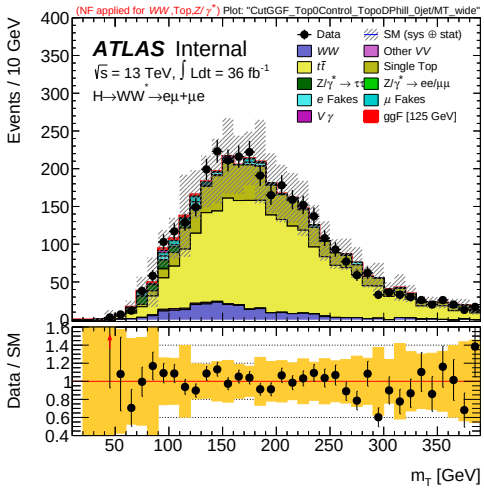
In the $N_{\text{jet}} = 0$ ggF analysis, all significant backgrounds are either estimated using a fully data-driven approach or normalized from data in an appropriate control region. The WW, top, and $Z \to \tau\tau$ backgrounds are normalized using control regions that are orthogonal to the signal region and are described in more detail throughout the remainder of this section. The non-WW diboson backgrounds ($W\gamma^{(*)}$, WZ, and ZZ) are estimated using purely Monte Carlo and are compared with data in a same-sign validation region which contains identical cuts as the signal region but requiring leptons with the same charge. The W+jets background is estimated using the data-driven fake factor method and is described in more detail in chapter 6. The event yields for different processes in each of the $N_{\text{jet}} = 0$ control regions as well as the same sign validation region are provided in Table 5.2. The m_{T} distributions for each region are also shown in Figure 5.1.

Selection	Higgs	WW	VZ/γ^*	$V\gamma$	Top	Z/DY	e-Fakes	μ-Fakes	Total Bkg	Data
WW CR	89.7 ± 1.1	5050 ± 30	204 ± 7.2	80.5 ± 11.1	1086 ± 14	338 ± 19	385 ± 16	308 ± 11	7452 ± 45	7461
Top CR	18.7 ± 0.6	249 ± 8.2	20 ± 2.5	10.1 ± 4	2978 ± 25	50 ± 6.4	56.4 ± 8	50.8 ± 5.7	3415 ± 29	3399
$Z \to \tau\tau$ CR	141 ± 1.2	926 ± 12	158 ± 6.1	726 ± 35	66.4 ± 3.6	40792 ± 136	867 ± 35	1977 ± 44	45511 ± 152	45463
SS VR	1.5 ± 0.2	13.5 ± 1.4	181 ± 5.7	127 ± 13	3.4 ± 0.7	1.9 ± 2.5	113 ± 7	69 ± 4.5	509 ± 17	581

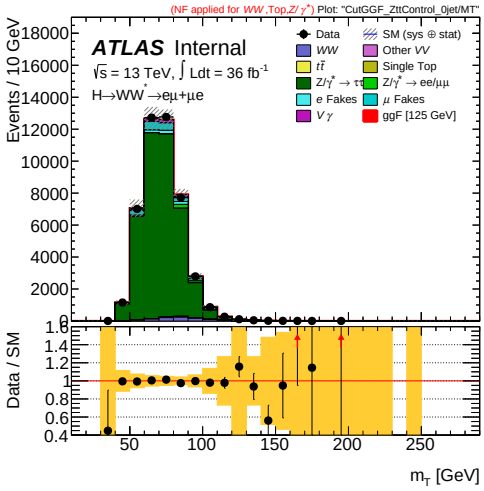
Table 5.2: Event yields for the control and validation regions in the $N_{\text{jet}} = 0$ category. The normalization factors from Table 5.9 have been applied. The uncertainties are statistical only.



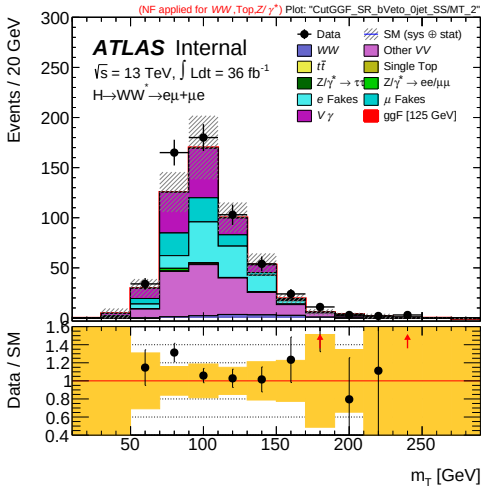
(a) $N_{\text{jet}} = 0$ WW CR



(b) $N_{\text{jet}} = 0$ Top CR



(c) $N_{\text{jet}} = 0$ $Z \to \tau\tau$ CR



(d) $N_{\text{jet}} = 0$ SS VR

Figure 5.1: Distributions of m_{T} for each control and validation region in the $N_{\text{jet}} = 0$ category. The normalization factors from Table 5.9 have been applied. The uncertainty bands represent the quadratic sum of the statistical and experimental systematic rate uncertainties.

5.2.1 WW Control Region

Continuum $WW \to l\nu l\nu$ production is the dominant background in the $N_{\text{jet}} = 0$ category. The following requirements define the $N_{\text{jet}} = 0$ WW control region,

- $55 < m_{\ell\ell} < 110$ GeV
- $\Delta\phi_{\ell\ell} < 2.6$
- No b-tagged jets (including sub-threshold), $N_{b\text{-jet},(p_{\text{T}}>20\text{ GeV})} = 0$

where orthogonality with the signal region is achieved through the higher dilepton invariant mass window.

The control region has a WW purity of $\sim 68\%$, with a resulting normalization factor using the simple matrix inversion method described in section 5.1 of 1.11 ± 0.03 (stat.) that is also quoted as a summary in Table 5.9. The significant deviation of the normalization factor from unity may be due to missing NNLO contributions to the cross section, since the WW sample used in the analysis has only NLO accuracy. Good agreement between data and Monte Carlo is achieved after the normalization factor is applied, as can be seen from Figure 5.1a and further from a number of other kinematic distributions in section A.1. Theory uncertainties are also evaluated on the extrapolation from the CR to SRs as well as on the modeling of the m_{T} shape used in the fit and are summarized in section 7.2.

5.2.2 Top Control Region

Due to the resulting production of a b-quark, processes including a top quark ($t\bar{t}$ and Wt) will have one or more jet in the final state and are therefore not expected to be a major background for the $N_{\text{jet}} = 0$ category. However, some top events still enter into the final signal region due to their relatively large cross sections and because the nominal jet p_{T} threshold is set to 30 GeV, allowing sub-threshold jets within the range $20 < p_{\text{T}} < 30$ GeV to still be present. The following requirements define the $N_{\text{jet}} = 0$ top control region,

- At least one b-tagged sub-threshold jet, $N_{b\text{-jet},(20\text{ GeV}<p_{\text{T}}<30\text{ GeV})} > 0$
- $\Delta\phi(\ell\ell, E_{\text{T}}^{\text{miss}}) > \pi/2$
- $p_{\text{T}}^{\ell\ell} > 30$ GeV
- $\Delta\phi_{\ell\ell} < 2.8$

where orthogonality with the signal region is achieved by requiring at least one b-tagged sub-threshold jet. The control region has a top purity of 87%. The relative fractions of $t\bar{t}$ and Wt events in the top and WW control regions as well as the signal region for the $N_{\text{jet}} = 0$ analysis are shown in Table 5.3.

The resulting normalization factor using the simple matrix inversion method described in section 5.1 is 1.02 ± 0.02 (stat.) and is also quoted as a summary in Table 5.9. Theory uncertainties are evaluated on the extrapolation from the control region to the signal region, as well as on the modeling of the m_{T} shape used in the fit as described in section 7.2. In addition to the m_{T} distribution shown in Figure 5.1b, more kinematic variable distributions for the $N_{\text{jet}} = 0$ Top CR are provided in section A.2.

Selection	$t\bar{t}$	Wt	$t\bar{t}/Wt$
Top CR	67.4 %	19.8 %	3.4
WW CR	9.05 %	5.38%	1.7
Signal Region	7.4 %	5.0 %	1.5

Table 5.3: The percentages of $t\bar{t}$ and Wt events as well as their ratio in the Top CR, WW CR and the signal region in the $N_{\text{jet}} = 0$ analysis for the different lepton flavor channels combined.

5.2.3 $Z \to \tau\tau$ Control Region

Background from $Z \to \tau\tau$ can occur if both taus decay leptonically. The following requirements define the $N_{\text{jet}} = 0$ $Z \to \tau\tau$ control region (excluding the preselection cut on $p_{\text{T}}^{\text{miss}}$),

- $\Delta\phi_{\ell\ell} > 2.8$
- No b-tagged jets (including sub-threshold), $N_{b\text{-jet},(p_{\text{T}}>20\text{ GeV})} = 0$
- $m_{\ell\ell} < 80$ GeV

where orthogonality with the signal region is achieved by reversing the cut on $\Delta\phi_{\ell\ell}$. The control region has a $Z \to \tau\tau$ purity of 90%, with a $Z \to ee/\mu\mu$ component of the total Z/DY yield of only 0.5% due to the different lepton flavor requirement at preselection.

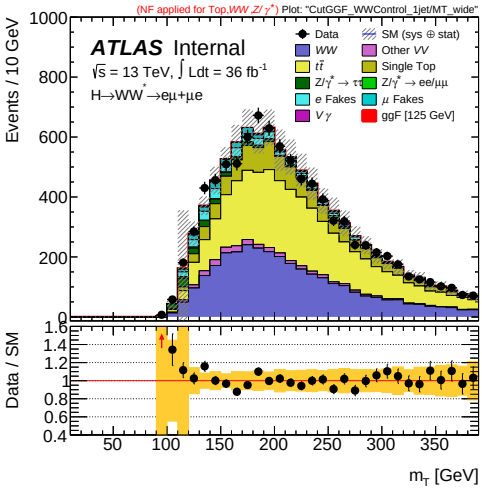
The resulting normalization factor using the simple matrix inversion method described in section 5.1 is 0.89 ± 0.01 (stat.) and is also quoted as a summary in Table 5.9. The normalization factor in this case is affected by residual misalignments in the inner detector which can distort the track parameter measurements for the leptons originating from the secondary decay vertex of the τ decay. Good agreement between data and Monte Carlo is achieved after the normalization factor is applied, as can be seen from Figure 5.1c and additional kinematic distributions in section A.3.

5.3 ggF 1 Jet Background

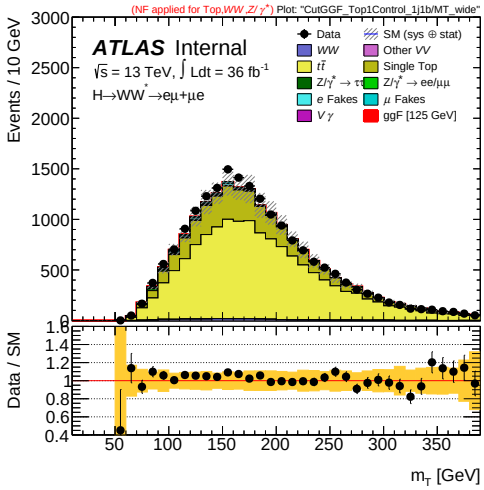
In the $N_{\text{jet}} = 1$ ggF analysis, all significant backgrounds are either estimated using a fully data-driven approach or normalized from data in an appropriate control region. The WW, top, and $Z \to \tau\tau$ backgrounds are normalized using control regions that are orthogonal to the signal region and are described in more detail throughout the remainder of this section. The non-WW diboson backgrounds ($W\gamma^{(*)}$, WZ, and ZZ) are estimated using purely Monte Carlo and are compared with data in a same-sign validation region which contains identical cuts as the signal region but requiring leptons with the same charge. The W+jets background is estimated using the data-driven fake factor method and is described in more detail in chapter 6. The event yields for different processes in each of the $N_{\text{jet}} = 1$ control regions as well as the same sign validation region are provided in Table 5.4. The m_{T} distributions for each region are also shown in Figure 5.2.

Selection	Higgs	WW	VZ/γ^*	$V\gamma$	Top	Z/DY	e-Fakes	μ-Fakes	Total Bkg	Data
WW CR	6.8 ± 0.2	3781 ± 25	257 ± 7.5	54 ± 8.7	4914 ± 30	199 ± 16	378 ± 18	197 ± 10.3	9780 ± 48	9784
Top CR	22 ± 0.5	238 ± 6.5	26 ± 3.2	7.0 ± 2.4	18678 ± 61	70 ± 6.9	217 ± 20	181 ± 13.3	19418 ± 66	19428
$Z \to \tau\tau$ CR	64 ± 0.7	313 ± 6.8	50 ± 3.6	87 ± 11	276 ± 7	2571 ± 41	99 ± 12	168 ± 12.9	3566 ± 47	3571
SS VR	1.1 ± 0.1	7.7 ± 1.0	114 ± 4.7	48 ± 7	9 ± 1.1	2.8 ± 1.8	67 ± 6	46.7 ± 3.8	296 ± 11	347

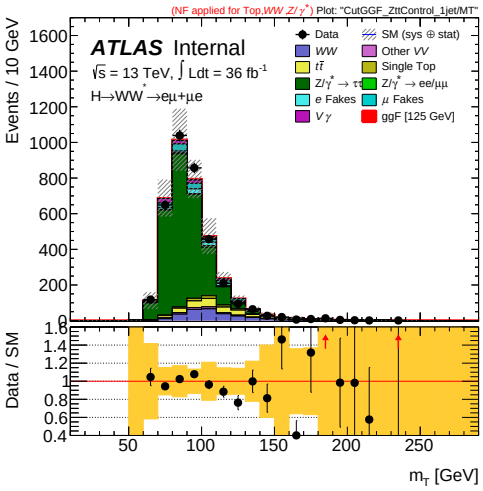
Table 5.4: Event yields for the control and validation regions in the $N_{\text{jet}} = 1$ category. The normalization factors from Table 5.9 have been applied. The uncertainties are statistical only.



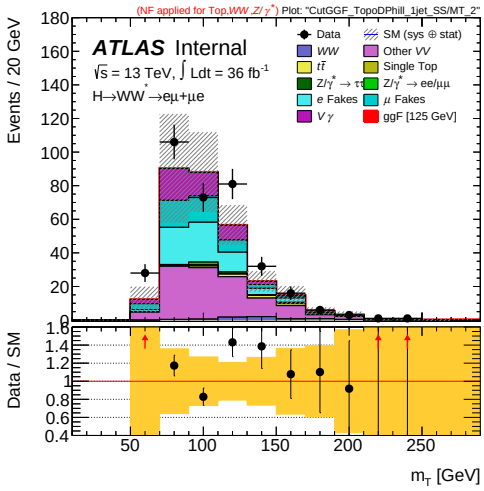
(a) $N_{\text{jet}} = 1$ WW CR



(b) $N_{\text{jet}} = 1$ Top CR



(c) $N_{\text{jet}} = 1$ $Z \to \tau\tau$ CR



(d) $N_{\text{jet}} = 1$ SS VR

Figure 5.2: Distributions of m_{T} for each control and validation region in the $N_{\text{jet}} = 1$ category. The normalization factors from Table 5.9 have been applied. The uncertainty bands represent the quadratic sum of the statistical and experimental systematic rate uncertainties.

5.3.1 WW Control Region

The WW background is normalized analogously to the $N_{\text{jet}} = 0$ analysis through a dedicated control region, in this case after additional cuts aiming to reject top and $Z \to \tau\tau$ backgrounds. The following requirements define the $N_{\text{jet}} = 1$ WW control region,

- $m_{\ell\ell} > 80$ GeV
- $|m_{\tau\tau} - m_Z| > 25$ GeV
- No b-tagged jets (including sub-threshold), $N_{b\text{-jet},(p_{\text{T}}>20\text{ GeV})} = 0$
- $\max\left(m_{\text{T}}^{\ell}\right) > 50$ GeV

where orthogonality with the signal region is achieved through the higher dilepton invariant mass requirement. The control region has a WW purity of only 39% - however, the ratio between WW and top events is very close to the ratio in the signal region. The resulting WW normalization factor using the simple matrix inversion method described in section 5.1 is 1.00 ± 0.04 (stat.), which is also quoted as a summary in Table 5.9. Theory uncertainties are evaluated on the extrapolation from the CR to the SRs as well as on the modeling of the m_{T} shape used in the fit and are summarized in section 7.2. In addition to the m_{T} distribution shown in Figure 5.2a, more kinematic variable distributions for the $N_{\text{jet}} = 1$ WW CR are provided in section A.1.

5.3.2 Top Control Region

In the $N_{\text{jet}} = 1$ category, the top background is also normalized through a dedicated control region. The following requirements define the $N_{\text{jet}} = 1$ top control region,

- Exactly one b-tagged nominal jet, $N_{b\text{-jet},(p_{\text{T}}>30\text{ GeV})} = 1$
- No b-tagged sub-threshold jets, $N_{b\text{-jet},(20\text{ GeV}<p_{\text{T}}<30\text{ GeV})} = 0$
- $\max\left(m_{\text{T}}^{\ell}\right) > 50$ GeV
- $Z \to \tau\tau$ veto, $m_{\tau\tau} < m_Z - 25$ GeV

where orthogonality with the signal region is achieved by requiring at least one b-tagged jet with $p_{\text{T}} > 30$ GeV. A b-veto is also applied to sub-threshold jets in order to ensure a similar $t\bar{t}/Wt$ ratio in the control and signal regions as can be seen in Table 5.5.

Selection	$t\bar{t}$	Wt	$t\bar{t}/Wt$
Top CR	74.4 %	21.9 %	3.4
WW CR	37.85 %	12.24 %	3.1
Signal Region	32 %	11.9 %	2.7

Table 5.5: The percentages of $t\bar{t}$ and Wt events as well as their ratio in the Top CR, WW CR and the signal region in the $N_{\text{jet}} = 1$ analysis for the different lepton flavor channels combined.

The control region has a top purity of about 96%, with a resulting normalization factor using the simple matrix inversion method described in section 5.1 of 1.04 ± 0.01 (stat.) that is also quoted as a summary in Table 5.9. Theory uncertainties are evaluated on the

extrapolation from the control region to the signal region, as well as on the modeling of the m_{T} shape used in the fit as described in section 7.2. In addition to the m_{T} distribution shown in Figure 5.2b, more kinematic variable distributions for the $N_{\mathrm{jet}} = 1$ Top CR are provided in section A.2.

5.3.3 $Z \to \tau\tau$ Control Region

The following requirements define the $N_{\mathrm{jet}} = 1$ $Z \to \tau\tau$ control region,

- Fail $Z \to \tau\tau$ veto, $m_{\tau\tau} > m_Z - 25$ GeV
- No b-tagged jets (including sub-threshold), $N_{b\text{-jet},(p_{\mathrm{T}}>20\ \mathrm{GeV})} = 0$
- $m_{\ell\ell} < 80$ GeV
- $\max\left(m_{\mathrm{T}}^{\ell}\right) > 50$ GeV

where orthogonality with the signal region is achieved by inverting the $Z \to \tau\tau$ veto. The control region has a $Z \to \tau\tau$ purity of 73%.

The resulting normalization factor using the simple matrix inversion method described in section 5.1 is 0.88 ± 0.02 (stat.) and is also quoted as a summary in Table 5.9. The same residual misalignment of the inner detector as was described in the $N_{\mathrm{jet}} = 0$ case is also reflected in this normalization factor. Good agreement between data and Monte Carlo is achieved after the normalization factor is applied, as can be seen from Figure 5.2c and additional kinematic distributions in section A.3.

5.4 VBF Background

In the $N_{\mathrm{jet}} \geq 2$ VBF analysis, backgrounds originate from very similar sources as in the ggF analysis and so the approach to their estimation is comparable, with only a few exceptions. The top and $Z \to \tau\tau$ backgrounds are still normalized from dedicated control regions - however, the WW background estimate relies only on Monte Carlo prediction due to the lack of a region sufficiently pure in WW production. Instead, a validation region is used to check the WW modeling. These regions are described in more detail throughout the remainder of this section. The W+jets background is estimated using the data-driven fake factor method and is described in more detail in chapter 6. The contribution of QCD events faking two leptons was found to be non-negligible in the VBF analysis and is therefore also accounted for, the description of which can be found in section 6.8. Finally, ggF+2 jets is also an important background that must be considered in the VBF analysis, with a similar amount of events as the VBF signal expected to be present in the signal region. The ggF+2 jets background is estimated with POWHEG + PYTHIA 8 NNLOPS.

5.4.1 WW Validation Region

The production of WW in association with 2 jets can be divided into two distinct types of processes - those containing only electroweak vertices (EW $WW + 2$ jets) and those containing a QCD vertex (QCD $WW + 2$ jets). Although the cross section of the QCD $WW + 2$ jets processes is more than an order of magnitude larger than the one for EW

$WW + 2$ jets, the two contributions are comparable in the VBF phase space of high m_{jj} and Δy_{jj}.

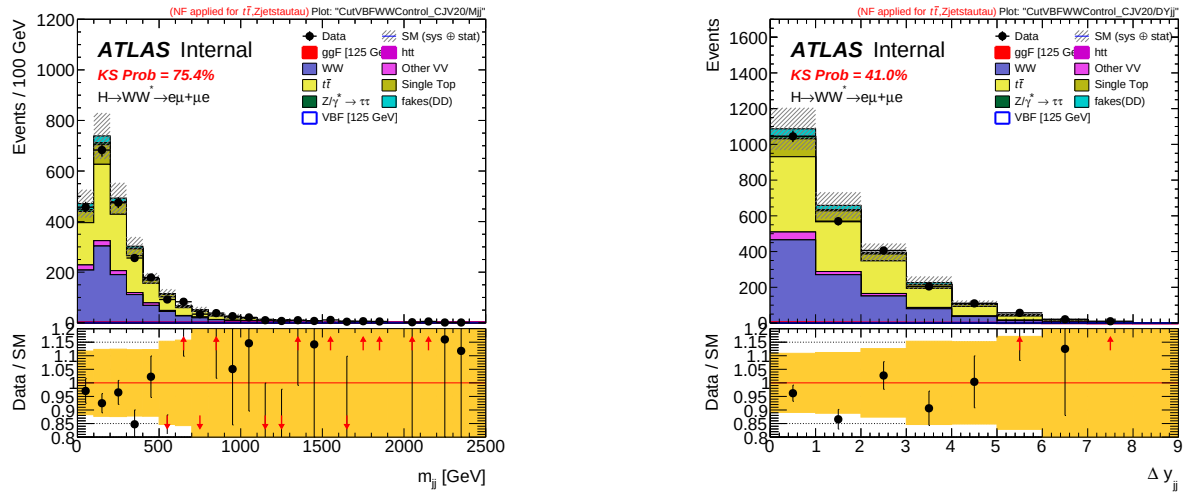
Due to a large $t\bar{t}$ contamination in regions which would otherwise be more enriched in WW, no WW control region is established in the VBF analysis. Instead, the Monte Carlo prediction of WW is applied directly, with a validation region being used to check the agreement with data. The following requirements define the $N_{\text{jet}} \geq 2$ WW validation region,

- $m_{\text{T}} > 130$ GeV
- $m_{\text{T2}} > 160$ GeV
- No b-tagged jets (including sub-threshold), $N_{b\text{-jet},(p_{\text{T}}>20\text{ GeV})} = 0$

where the m_{T2} variable is defined as

$$m_{T2} = \min_{p_T^1 + p_T^2 = p_T}(\max(m_{\text{T}}^2(p_T^1, p_T^a), m_{\text{T}}^2(p_T^2, p_T^b))) \tag{5.1}$$

and represents a lower bound on the parent particle's mass [95]. The WW purity of the validation region is 40%, with supplementary optimization studies detailed in [47]. The modeling of m_{jj} and Δy_{jj} in the validation region is shown in Figure 5.3.



(a) m_{jj} (b) Δy_{jj}

Figure 5.3: Distributions of m_{jj} and Δy_{jj} in the VBF WW VR. The normalization factors from Table 5.9 have been applied. The yellow band represents the MC statistical uncertainties and the main sources of detector systematics (JES/JER/$E_{\text{T}}^{\text{miss}}$/$b$-tagging/Leptons).

5.4.2 Top Control Region

In the VBF analysis, the top background is normalized through a dedicated control region. The following requirements define the $N_{\text{jet}} \geq 2$ top control region:

- Exactly one b-tagged jet (including sub-threshold), $N_{b\text{-jet},(p_{\text{T}}>20\text{ GeV})} = 1$
- Central jet veto (CJV)

- Outside lepton veto (OLV)
- $Z \to \tau\tau$ veto, $m_{\tau\tau} < m_Z - 25$ GeV

where orthogonality with the signal region is achieved by replacing the b-veto with a requirement of exactly one b-tagged jet[1]. The control region maintains a high top purity of $\sim 96\%$.

A cutflow from the VBF preselection to the top control region is provided in Table 5.6. The modeling in the control region of the two most important BDT variables m_{jj} and Δy_{jj}, as well as the BDT output, are shown in Figure 5.4. The resulting normalization factor using the simple matrix inversion method described in section 5.1 is found to be 1.02 ± 0.01 (stat.), with additional theory uncertainties on both rate and shape (described further in section 7.2) being applied to the extrapolation factor from the top control region to the signal region.

$\sqrt{s} = 13TeV$, $\mathcal{L} = 36fb^{-1}$,	VBF [125 GeV]	ggF [125 GeV]	htt	WW	Other VV	Top	Z/γ^*	fakes(DD)	Total Bkg	Data	Data/MC
Top CR: 2-jets	97.21 ± 0.39	269.55 ± 1.65	72.83 ± 0.72	6094.95 ± 16.93	1160.79 ± 27.02	247173.47 ± 224.15	7156.20 ± 64.44	5777.99 ± 94.57	274178.14 ± 254.49	264515	0.99 ± 0.00
Top CR: 1 b-jets	11.97 ± 0.15	40.30 ± 0.69	9.86 ± 0.29	897.23 ± 6.37	197.73 ± 9.84	87268.13 ± 128.80	1181.08 ± 26.94	2250.64 ± 55.60	92791.69 ± 143.54	92007	1.00 ± 0.00
Top CR: CJV	8.43 ± 0.13	26.50 ± 0.57	6.79 ± 0.24	564.32 ± 5.32	114.20 ± 7.13	58571.69 ± 105.47	780.95 ± 24.00	1467.64 ± 45.36	62136.86 ± 117.80	61802	1.00 ± 0.00
Top CR: OLV	5.81 ± 0.10	6.02 ± 0.27	2.29 ± 0.13	85.33 ± 2.00	20.30 ± 2.77	11258.63 ± 46.61	138.13 ± 15.73	271.73 ± 20.31	11870.76 ± 53.38	11722	0.99 ± 0.01
Top CR: Ztautau Veto	4.97 ± 0.10	5.30 ± 0.25	0.30 ± 0.06	51.46 ± 1.66	11.34 ± 1.78	7368.62 ± 37.60	51.61 ± 5.39	184.20 ± 16.41	7725.83 ± 41.50	7668	1.00 ± 0.01

Table 5.6: Cutflow from the VBF preselection to the top control region. Only the statistical errors are shown.

5.4.3 $Z \to \tau\tau$ Control Region

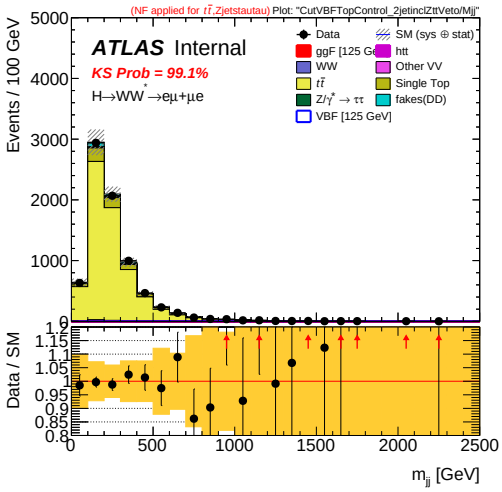
In the VBF analysis, the $Z \to \tau\tau$ background is normalized through a dedicated control region. The following requirements define the $N_{\text{jet}} \geq 2$ $Z \to \tau\tau$ control region:

- $|m_{\tau\tau} - m_Z| \leq 25$ GeV
- No b-tagged jets (including sub-threshold), $N_{b\text{-jet},(p_{\text{T}}>20\text{ GeV})} = 0$
- $m_{\ell\ell} < 80$ GeV
- Central jet veto
- Outside lepton veto

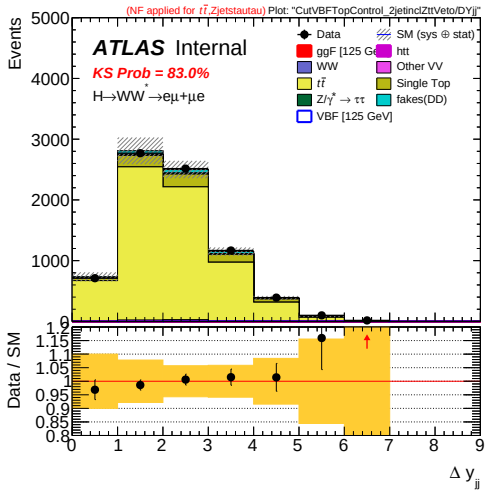
where orthogonality with the signal region is achieved by inverting the $Z \to \tau\tau$ veto. The control region has a $Z \to \tau\tau$ purity of $\sim 74\%$. The modeling of various kinematic variables has also been checked in a looser control region containing more statistics [47] and decent agreement between data and Monte Carlo is observed.

A cutflow from the VBF preselection to the $Z \to \tau\tau$ control region is provided in Table 5.7. The modeling in the control region of the two most important BDT variables m_{jj} and Δy_{jj}, as well as the BDT output, are shown in Figure 5.5. The resulting normalization factor using the simple matrix inversion method described in section 5.1 is found to be 0.97 ± 0.07 (stat.), with additional theory uncertainties on both rate and shape (described further in section 7.2) being applied to the extrapolation factor from the $Z \to \tau\tau$ control region to the signal region.

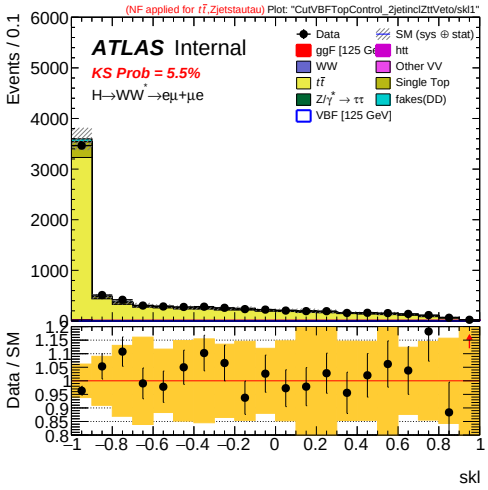
[1]The reason for requiring exactly one b-tagged jet as opposed to being inclusive in b-tagged jets is to keep the flavor composition of tagged jets closer to that in the signal region.



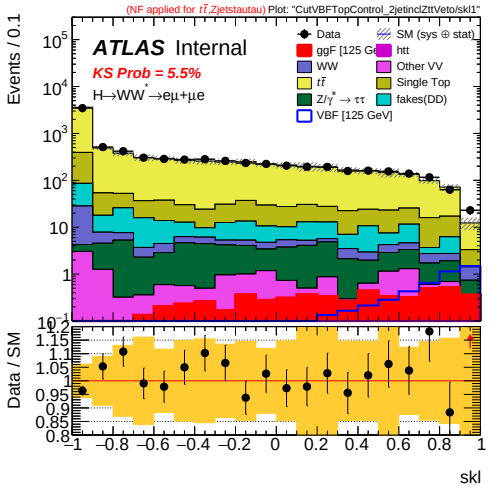
(a) m_{jj}



(b) Δy_{jj}



(c) BDT Output (lin)



(d) BDT Output (log)

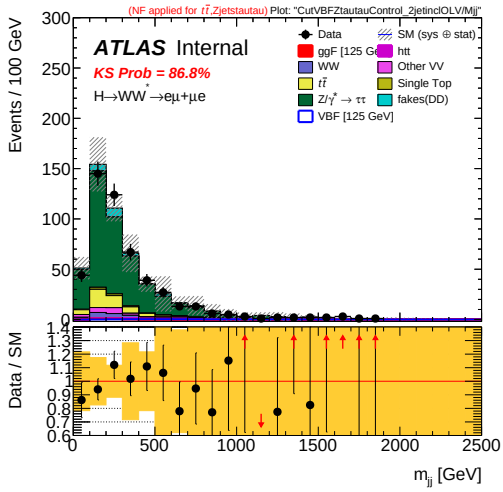
Figure 5.4: Distributions of m_{jj}, Δy_{jj}, and BDT output in the VBF Top CR. The normalization factors from Table 5.9 have been applied. The yellow band represents the MC statistical uncertainties and the main sources of detector systematics (JES/JER/$E_{\mathrm{T}}^{\mathrm{miss}}$/$b$-tagging/Leptons).

$\sqrt{s} = 13TeV$, $\mathcal{L} = 36fb^{-1}$,	VBF [125 GeV]	ggF [125 GeV]	htt	WW	Other VV	Top	Z/γ^*	fakes(DD)	Total Bkg	Data	Data/MC
Ztt CR:2-jets	97.21 ± 0.39	269.55 ± 1.65	72.83 ± 0.72	6094.95 ± 16.93	1160.79 ± 27.02	247173.47 ± 224.15	7156.20 ± 64.44	5777.99 ± 94.57	274178.14 ± 254.49	264515	0.99 ± 0.00
Ztt CR: bVeto	84.10 ± 0.36	222.86 ± 1.47	61.61 ± 0.65	5095.85 ± 15.55	925.16 ± 24.48	14027.31 ± 50.28	5749.73 ± 57.90	1289.38 ± 38.38	32791.72 ± 92.42	26229	0.96 ± 0.01
Ztt CR: $\lvert mtt - MZ\rvert < 25$	7.09 ± 0.10	16.89 ± 0.39	20.03 ± 0.35	439.70 ± 4.42	187.55 ± 11.14	1203.43 ± 14.31	2509.86 ± 34.34	185.32 ± 17.02	5027.87 ± 42.96	4223	0.93 ± 0.02
Ztt CR: $M_l l < 75/80$ GeV	6.92 ± 0.10	16.42 ± 0.39	17.00 ± 0.32	151.30 ± 2.56	140.82 ± 10.35	379.53 ± 7.95	2463.97 ± 33.86	123.22 ± 15.17	3456.22 ± 39.54	2970	0.90 ± 0.02
Ztt CR: CJV	5.24 ± 0.09	11.93 ± 0.33	12.80 ± 0.27	107.10 ± 2.17	98.84 ± 8.34	245.64 ± 6.18	1847.60 ± 30.88	90.97 ± 13.16	2531.91 ± 35.30	2194	0.91 ± 0.02
Ztt CR: OLV	4.01 ± 0.08	2.82 ± 0.16	5.32 ± 0.15	21.53 ± 1.02	19.45 ± 2.88	56.46 ± 3.00	381.81 ± 15.15	25.60 ± 7.32	535.98 ± 17.41	501	0.97 ± 0.05

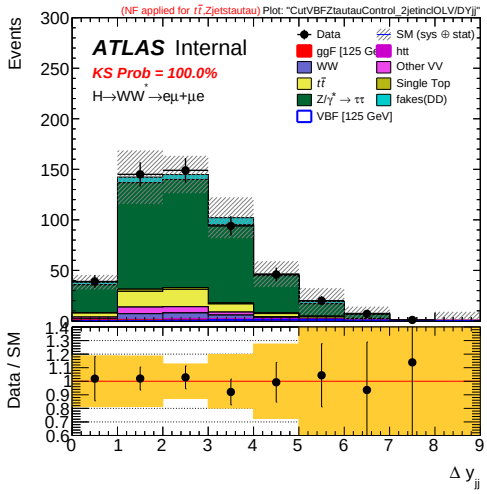
Table 5.7: Cutflow from the VBF preselection to the $Z \to \tau\tau$ control region. Only the statistical errors are shown.

5.5 Summary

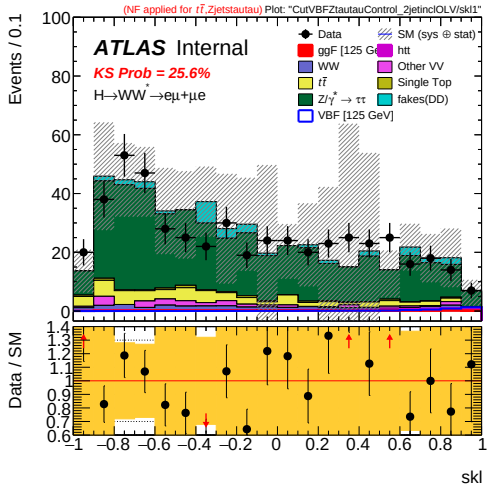
The complete definition of control regions described in this chapter and the resulting normalization factors using matrix inversion are summarized in Table 5.8 and Table 5.9, respectively. The normalization factors are obtained separately for each jet category, with



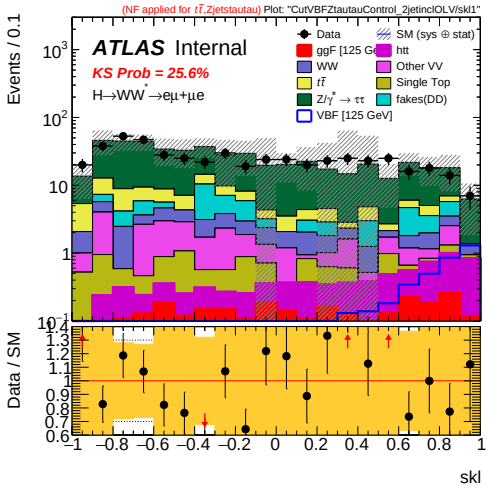
(a) m_{jj}



(b) Δy_{jj}



(c) BDT Output (lin)



(d) BDT Output (log)

Figure 5.5: Distributions of m_{jj}, Δy_{jj}, and BDT output in the VBF $Z \to \tau\tau$ CR. The normalization factors from Table 5.9 have been applied. The yellow band represents the MC statistical uncertainties and the main sources of detector systematics (JES/JER/$E_{\mathrm{T}}^{\mathrm{miss}}$/$b$-tagging/Leptons).

the statistical uncertainty being propagated through the usage of toy datasets[2]. While the matrix inversion method can provide a powerful way to quickly obtain background normalization factors, it is not used when determining the parameter of interest. Instead, the final normalization factors enter when performing a full simultaneous fit over all signal and control regions as unconstrained nuisance parameters during likelihood maximization.

[2]This entails randomly generating matrices that are within the uncertainties of their inputs. The mean and standard deviation of the resulting normalization factors are then quoted as the final values.

CR	$N_{\text{jet},(p_{\text{T}}>30\text{ GeV})} = 0$ ggF	$N_{\text{jet},(p_{\text{T}}>30\text{ GeV})} = 1$ ggF	$N_{\text{jet},(p_{\text{T}}>30\text{ GeV})} \geq 2$ VBF
WW	$55 < m_{\ell\ell} < 110$ GeV $\Delta\phi_{\ell\ell} < 2.6$	$m_{\ell\ell} > 80$ GeV $\lvert m_{\tau\tau} - m_Z \rvert > 25$ GeV	
	$N_{b\text{-jet},(p_{\text{T}}>20\text{ GeV})} = 0$		
		$\max\left(m_{\text{T}}^{\ell}\right) > 50$ GeV	
$t\bar{t}/Wt$	$N_{b\text{-jet},(20\text{ GeV}<p_{\text{T}}<30\text{ GeV})} > 0$	$N_{b\text{-jet},(p_{\text{T}}>30\text{ GeV})} = 1$ $N_{b\text{-jet},(20\text{ GeV}<p_{\text{T}}<30\text{ GeV})} = 0$	$N_{b\text{-jet},(p_{\text{T}}>20\text{ GeV})} = 1$
	$\Delta\phi(\ell\ell, E_{\text{T}}^{\text{miss}}) > \pi/2$	$\max\left(m_{\text{T}}^{\ell}\right) > 50$ GeV	central jet veto
	$p_{\text{T}}^{\ell\ell} > 30$ GeV	$m_{\tau\tau} < m_Z - 25$ GeV	
	$\Delta\phi_{\ell\ell} < 2.8$		outside lepton veto
Z/γ^*		$N_{b\text{-jet},(p_{\text{T}}>20\text{ GeV})} = 0$	
		$m_{\ell\ell} < 80$ GeV	
	no $p_{\text{T}}^{\text{miss}}$ requirement		central jet veto
		$\max\left(m_{\text{T}}^{\ell}\right) > 50$ GeV	outside lepton veto
	$\Delta\phi_{\ell\ell} > 2.8$	$m_{\tau\tau} > m_Z - 25$ GeV	$\lvert m_{\tau\tau} - m_Z \rvert \leq 25$ GeV

Table 5.8: Summary of the criteria used to define the control regions in each of the jet categories, starting from the preselection stage.

Control Regions	WW	Top	$Z \to \tau\tau$
$N_{\text{jet}} = 0$	1.11 ± 0.03	1.02 ± 0.02	0.89 ± 0.01
$N_{\text{jet}} = 1$	1.00 ± 0.04	1.04 ± 0.01	0.88 ± 0.02
$N_{\text{jet}} \geq 2$	–	1.02 ± 0.01	0.97 ± 0.07

Table 5.9: Summary of the background normalization factors obtained through matrix inversion from each control region, separately for each jet category. The quoted uncertainty is statistical only.

Chapter 6

Misidentified Leptons

This chapter is dedicated to the estimation of analysis backgrounds originating from objects being misidentified as prompt leptons. An introduction to this category of background along with the fake factor method which is used to estimate it is first provided in section 6.1 and section 6.2. A W+jets control region is defined in section 6.3, which together with fake factors derived from either a Z+jets or dijets sample as described in section 6.4 and section 6.5 respectively, provide the W+jets yield in the signal region. The flavor composition of the misidentified leptons and related corrections are discussed in section 6.6, while systematics associated with the fake factor estimate are described in section 6.7. Finally, considerations for the background from two misidentified leptons are detailed in section 6.8.

6.1 Introduction

One of the benefits associated with studying physics signatures containing final state leptons is the significant background rejection due to the excellent lepton identification of the ATLAS detector. Events containing only QCD interactions can be heavily suppressed by imposing sufficiently tight lepton identification criteria whereby the suppression of jets in ATLAS often approaches the level of 10^{-5}. However, the remaining jets that are misidentified as leptons populate the non-gaussian tails of the detector response, making them notoriously difficult to model using simulation. Furthermore, despite the low misidentification rate, sizable contributions of misidentified leptons can remain after selecting for signal due to the large production cross section of QCD jets at the LHC.

There exist a variety of sources for misidentified leptons. In the following, however, these objects are also referred to broadly as either "fake leptons" or simply just "fakes". For electrons, they can originate for instance from charged hadrons, conversion of photons[1], or semi-leptonic heavy-flavor decays. In the latter two cases, although an actual electron is present in the final state from a secondary process, it is still considered fake in the sense that it is not produced in isolation as the result of a hard electroweak scattering event (which, in contrast, is referred to as a "prompt" lepton). For muons, nearly all fakes originate from either semi-leptonic heavy-flavor decay or meson decay in flight - both of which contain a real non-prompt muon.

With a signal selection requiring two identified leptons used for the analysis presented in this thesis, the majority of misidentified leptons appear through W+jet processes in

[1]The photons originating in turn from e.g. bremsstrahlung as well as initial or final state radiation.

which one prompt lepton originates from the W boson decaying leptonically in association with one or more jets where one of them is misidentified as a second prompt lepton. A contribution is also present from pure QCD processes in which two jets are simultaneously misidentified as prompt leptons. However, the number of these events can be reduced with selections targeting final states containing high-energy neutrinos, for example by requiring a large missing transverse momentum. Rather than relying on the Monte Carlo modeling of fake leptons, the data-driven fake factor method is used in this analysis to estimate the W+jets background and is introduced below.

6.2 The Fake Factor Method

The total number of events in the signal region with two fully identified leptons can be expressed as

$$N^{\mathrm{SR}}_{\mathrm{id+id}} = N^{\mathrm{p}}_{\mathrm{id+id}} + N^{\mathrm{f}}_{\mathrm{id+id}} \tag{6.1}$$

where the superscripts 'p' and 'f' denote the true type of lepton - prompt or fake, respectively. In order to estimate the contribution from $N^{\mathrm{f}}_{\mathrm{id+id}}$, a W+jets control region is established with an enhanced rate of fake leptons. This is accomplished by requiring only one lepton to be fully identified, while the other is "anti-identified" or "Anti-ID", failing the full identification but satisfying a looser set of criteria. Table 6.1 provides a summary of the full ID and additional Anti-ID definitions in the analysis. The isolation requirements are removed and fully identified leptons are rejected in the Anti-ID definition. For electrons, the less restrictive identification working point 'LHLoose' is used instead of 'LHMedium' or 'LHTight'. For muons, the quality is reduced from 'Tight' to 'Medium' and the d_0 significance cut is loosened.

Analogous to the signal region, the total number of events in the W+jets control region can be written as

$$N^{W\text{+jets CR}}_{\mathrm{id+}\not{\mathrm{id}}} = N^{\mathrm{f}}_{\mathrm{id+}\not{\mathrm{id}}} + N^{\mathrm{p}}_{\mathrm{id+}\not{\mathrm{id}}} \tag{6.2}$$

where this time a strikethrough in 'id' represents an Anti-ID lepton. The W+jets background in the signal region is then estimated by scaling the number of events in the W+jets control region by the *fake factor*, which is defined as the ratio of fully ID leptons (N_{id}) and Anti-ID leptons ($N_{\not{\mathrm{id}}}$):

$$f^{id}_{\not{id}} \equiv \frac{N_{\mathrm{id}}}{N_{\not{\mathrm{id}}}}. \tag{6.3}$$

The fake factor is measured in a separate fake-enriched data sample (using the same definitions of ID and Anti-ID) and applied to the W+jets control region. It can be viewed simply as an extrapolation factor, providing an ID+ID fake yield given the number of ID+Anti-ID events with one fake lepton:

$$\begin{aligned} N^{\mathrm{f}}_{\mathrm{id+id}} &= N^{\mathrm{f}}_{\mathrm{id+}\not{\mathrm{id}}} \times f^{id}_{\not{id}} \\ &= (N^{\mathrm{data}}_{\mathrm{id+}\not{\mathrm{id}}} - N^{\mathrm{EW\ MC}}_{\mathrm{id+}\not{\mathrm{id}}}) \times f^{id}_{\not{id}}. \end{aligned} \tag{6.4}$$

ID electron	Anti-ID electron	ID muon	Anti-ID muon
$p_{\mathrm{T}} > 15$ GeV $\lvert\eta\rvert < 2.47$,excluding $1.37 < \lvert\eta\rvert < 1.52$ $\lvert z_0 \sin\theta\rvert < 0.5$ mm		$p_{\mathrm{T}} > 15$ GeV $\lvert\eta\rvert < 2.45$ $\lvert z_0 \sin\theta\rvert < 0.5$ mm	
Pass LHTight if $p_{\mathrm{T}} < 25$GeV Pass LHMedium if $p_{\mathrm{T}} > 25$GeV	Pass LHLoose	Pass Quality Tight	Pass Quality Medium
$\lvert d_0\rvert/\sigma(d_0) < 5$		$\lvert d_0\rvert/\sigma(d_0) < 3$	$\lvert d_0\rvert/\sigma(d_0) < 15$
Pass FixedCutTrackCone40 isolation if $p_{\mathrm{T}} < 25$GeV		$E_T^{cone20}/p_{\mathrm{T}} < 0.09$	
Pass Gradient isolation if $p_{\mathrm{T}} > 25$GeV		$p_{\mathrm{T}}^{varcone30}/p_{\mathrm{T}} < 0.06$	
Author $= 1$			
	Veto against identified electron		Veto against identified muon

Table 6.1: Summary of the requirements for fully identified (ID) and anti-identified (Anti-ID) electrons (left) and muons (right).

6.3 W+jets Control Region

The sample to which the fake factor is applied is referred to as the W+jets control region. It is defined using the full selection criteria as the signal region, except requiring ID + Anti-ID leptons rather than ID + ID leptons. The loosening of the identification for one of the leptons enhances the contribution in which one lepton comes from a fake or non-prompt source. However, the sample still contains electroweak backgrounds (e.g. Z+jets, diboson and top processes) that must be subtracted[2].

The final W+jets yield is obtained only once the fake factor is applied (and in doing so, extrapolating to the ID + ID phase space)[3]. A distinction is often made based on the flavor of Anti-ID lepton - events with an Anti-ID electron are also referred to as "e-fake", while events with an Anti-ID muon are also referred to as "μ-fake". When they are combined, however, they are referred to as either "fakes" or "Mis-Id". The W+jets control region yields before the fake factor is applied can be seen in Table 6.2, while m_{T} distributions of the W+jets control region separated for each jet category and fake flavor are shown in Figure 6.1.

[2]The $V + \gamma$ process is also subtracted in order not to double count since it is estimated directly in the analysis by Monte Carlo.

[3]A correction factor must also be applied to account for flavor composition differences, which is introduced in section 6.6.

CR for e-fakes	Z+jets	$V\gamma$	Diboson	Top	Total Bkg	Data	Subtracted
$e\mu$, $N_{\text{jet}}=0$	1 ± 2	139 ± 13	152 ± 5	32 ± 2	324 ± 14	861 ± 29	537 ± 33
μe, $N_{\text{jet}}=0$	72 ± 28	241 ± 17	361 ± 7	77 ± 4	752 ± 34	2823 ± 53	2071 ± 63
$e\mu$, $N_{\text{jet}}=1$	8 ± 2	51 ± 8	54 ± 3	62 ± 3	175 ± 9	400 ± 20	225 ± 22
μe, $N_{\text{jet}}=1$	21 ± 6	155 ± 16	158 ± 5	167 ± 5	500 ± 18	1482 ± 38	982 ± 43
$e\mu$, VBF	10 ± 2	13 ± 3	13 ± 1	38 ± 3	74 ± 5	174 ± 13	100 ± 14
μe, VBF	36 ± 5	28 ± 7	42 ± 2	119 ± 5	225 ± 10	470 ± 22	245 ± 24
CR for μ-fakes	Z+jets	$V\gamma$	Diboson	Top	Total Bkg	Data	Subtracted
$e\mu$, $N_{\text{jet}}=0$	19 ± 5	19 ± 4	333 ± 7	80 ± 4	451 ± 11	1694 ± 41	1243 ± 43
μe, $N_{\text{jet}}=0$	4 ± 1	6 ± 3	105 ± 4	20 ± 2	135 ± 5	293 ± 17	158 ± 18
$e\mu$, $N_{\text{jet}}=1$	45 ± 10	8 ± 3	162 ± 6	174 ± 6	389 ± 13	983 ± 31	594 ± 34
μe, $N_{\text{jet}}=1$	12 ± 3	5 ± 2	47 ± 3	40 ± 2	105 ± 5	219 ± 15	114 ± 16
$e\mu$, VBF	66 ± 10	1 ± 1	44 ± 2	129 ± 5	239 ± 11	521 ± 23	282 ± 25
μe, VBF	17 ± 4	4 ± 2	21 ± 4	41 ± 3	83 ± 6	133 ± 12	50 ± 13

Table 6.2: W+jets control region event yields, separated by different fake flavors for the $N_{\text{jet}}=0$ and $N_{\text{jet}}=1$ ggF signal regions as well as the VBF signal region and shown for $e\mu$ and μe channels separately. The fake factors and correction factors have not been applied to these yields and the uncertainties are statistical only.

6.4 Z+jets Fake Factor

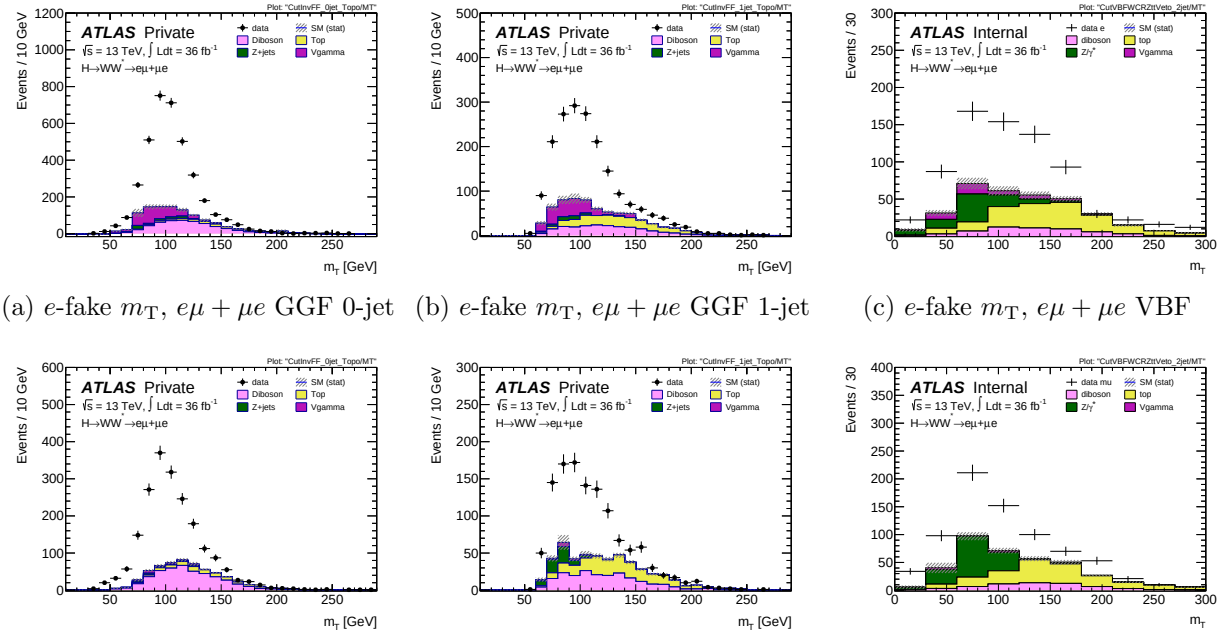
6.4.1 Sample Selection

A sample enriched in Z+jets containing one additional fake lepton is obtained by starting with events that have exactly three loosely defined reconstructed leptons, each with $p_{\text{T}} > 15\,\text{GeV}$. Then, two of the leptons are required to be tagged as Z boson candidates according to the following criteria:

- The pair must be same-flavor opposite-sign (ee or $\mu\mu$)
- Both must fulfill the "ID" criteria as listed in Table 6.1
- Their invariant mass must be within the Z mass window $80\,(70)\,\text{GeV} < m_{\ell\ell} < 110\,\text{GeV}$ for $ee\,(\mu\mu)$
- At least one must be matched to one of the single lepton triggers used in the analysis

The Z mass window is narrowed for electrons in order to reduce the $Z+\gamma$ background. If more than one Z candidate pair is identified, then the pair with invariant mass closer to the Z mass pole is chosen. The third lepton is subsequently classified as the fake candidate. Finally, the event is vetoed if its fake candidate has $m_{\text{T}}^{\ell} > 50\,\text{GeV}$ as defined from Equation 4.1 in order to reduce the amount of electroweak background from WZ events.

After the cuts above have been applied, the sample is divided based on the flavor of the fake candidate and whether it passes the "ID" or "Anti-ID" definitions from Table 6.1. Before the fake factor can be calculated, contamination from electroweak processes (which can be substantial, particularly in the case that the fake candidate fulfills the "ID" criteria) must be subtracted. The set of backgrounds considered include $V+\gamma$, diboson processes such as WW, WZ and ZZ, as well as top processes and are estimated using Monte Carlo. A full cutflow for the Z+jets fake factor estimate is provided in Table 6.3. Distributions of



(a) e-fake m_{T}, $e\mu + \mu e$ GGF 0-jet (b) e-fake m_{T}, $e\mu + \mu e$ GGF 1-jet (c) e-fake m_{T}, $e\mu + \mu e$ VBF

(d) μ-fake m_{T}, $e\mu + \mu e$ GGF 0-jet (e) μ-fake m_{T}, $e\mu + \mu e$ GGF 1-jet (f) μ-fake m_{T}, $e\mu + \mu e$ VBF

Figure 6.1: m_{T} distributions of the W+jets control region, separated by different fake flavors for the $N_{\mathrm{jet}} = 0$ and $N_{\mathrm{jet}} = 1$ ggF signal regions as well as the VBF signal region and shown for $e\mu$ and μe channels combined. The backgrounds are subtracted as electroweak processes, while the excess data events are taken to be from fake sources. Here, the fake factors have not yet been applied.

the ID/Anti-ID fake candidate p_{T} and η are shown in Figure 6.2 for electrons and Figure 6.3 for muons, which are the variables in which the fake factor is binned.

6.4.2 WZ Control Region

Due to the fact that the WZ process is the largest component of the electroweak background that must be subtracted for the Z+jets fake factor estimate and because other literature has reported the need for a WZ normalization factor with significant deviation from unity [45, 42, 73], a control region is established in order to provide a better handle on the WZ background prediction. The WZ control region is defined using an inversion of the WZ veto (i.e. the fake candidate must have $m_{\mathrm{T}}^{\ell} > 50\,\mathrm{GeV}$) and requiring the fake candidate to pass the full "ID" definition. It is 86% pure in WZ, with backgrounds including ZZ and Z+jets among others with smaller contributions.

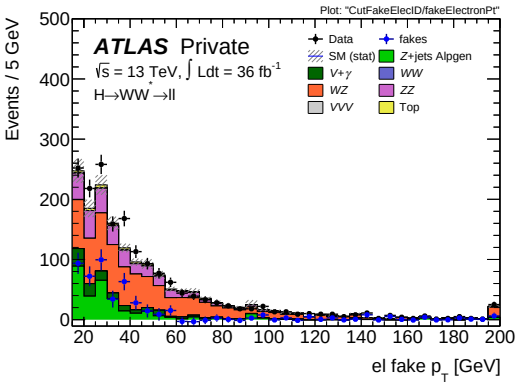
A WZ normalization factor is estimated by performing a χ^2 fit of Monte Carlo to data in the distribution of fake candidate m_{T}^{ℓ}. The χ^2 function is defined as

$$\chi^2 = \sum_k \frac{(x_k - \phi_k(\alpha))^2}{\sigma_k^2}; \quad \phi_k(\alpha) = B_k + \alpha S_k, \tag{6.5}$$

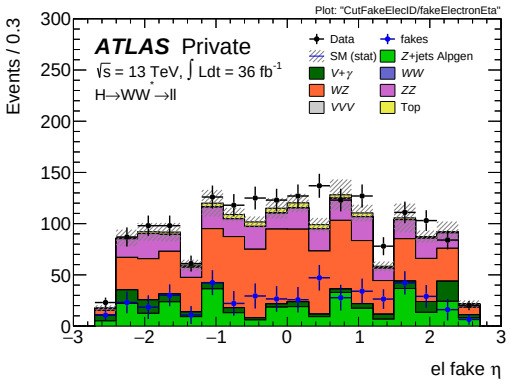
where the index k sums over the bins in the m_{T}^{ℓ} distribution. Here B_k represents the total number of background events in bin k, while S_k represents the total number of WZ events in bin k and α is the global normalization factor that is being measured. In addition, x_k denotes the number of data events in bin k, with σ_k being the sum in quadrature of the data

$\sqrt{s} = 13TeV$, $\mathcal{L} = 36fb^{-1}$,	Z+jets Alpgen	$V+\gamma$	WW	WZ	ZZ	Top	VVV	Total Bkg	Data	fakes	ZFake purity(%)
Channel Selection	44919007.54 ± 17169.55	1208250.57 ± 874.13	56405.60 ± 94.00	40014.49 ± 104.57	57917.72 ± 312.23	501174.99 ± 311.76	172.63 ± 0.83	1863936.01 ± 989.22	59879870	58015933.99 ± 7801.18	96.89 ± 0.02
Overlap: Vgamma/Vjets	44919007.54 ± 17169.55	1208250.57 ± 874.13	56405.60 ± 94.00	40014.49 ± 104.57	57917.72 ± 312.23	501174.99 ± 311.76	172.63 ± 0.83	1863936.01 ± 989.22	59879870	58015933.99 ± 7801.18	96.89 ± 0.02
lll final state	130752.35 ± 822.79	62970.78 ± 173.92	234.83 ± 5.20	14107.83 ± 59.72	9156.11 ± 161.47	17182.66 ± 56.48	70.93 ± 0.55	103723.15 ± 251.21	226744	123020.85 ± 538.38	54.26 ± 0.26
lep $p_T > 15$ GeV	59392.75 ± 539.66	31939.83 ± 121.54	118.79 ± 3.54	11411.27 ± 52.98	4555.57 ± 51.66	8706.88 ± 39.69	60.51 ± 0.51	56792.86 ± 147.77	100268	43475.14 ± 349.43	43.36 ± 0.37
Z-tagging	17610.35 ± 251.24	4176.02 ± 41.92	9.33 ± 0.86	7053.82 ± 41.41	1344.87 ± 28.01	968.91 ± 11.78	19.01 ± 0.22	13571.96 ± 66.31	29477	15905.04 ± 184.05	53.96 ± 0.70
WZ veto	16218.57 ± 243.60	3501.67 ± 39.07	4.08 ± 0.54	2315.85 ± 24.40	985.75 ± 25.03	447.10 ± 8.53	5.02 ± 0.12	7259.46 ± 53.12	21857	14597.54 ± 157.09	66.79 ± 0.85
fake type: electron	13174.16 ± 218.08	3472.56 ± 38.92	3.38 ± 0.46	1111.11 ± 16.77	497.32 ± 15.65	193.82 ± 5.17	2.35 ± 0.07	5280.54 ± 45.47	15916	10635.46 ± 134.10	66.82 ± 1.00
Fake el eta cut	11269.99 ± 202.07	3261.99 ± 37.69	2.50 ± 0.42	1049.12 ± 16.28	464.69 ± 14.99	180.86 ± 5.00	2.22 ± 0.07	4961.39 ± 44.00	13943	8981.61 ± 126.01	64.42 ± 1.06
ID cuts	313.29 ± 31.28	120.65 ± 7.31	0.02 ± 0.01	831.49 ± 14.25	302.56 ± 11.99	44.50 ± 1.31	1.75 ± 0.07	1300.97 ± 20.05	1769	468.03 ± 46.59	26.46 ± 2.71
pT < 20	88.78 ± 18.35	29.24 ± 3.44	0	81.31 ± 5.00	44.14 ± 4.34	3.55 ± 0.48	0.15 ± 0.02	158.39 ± 7.48	252	93.61 ± 17.55	37.15 ± 7.35
20 < pT < 25	38.99 ± 12.62	20.80 ± 2.98	0	75.47 ± 4.34	45.95 ± 4.92	3.53 ± 0.41	0.16 ± 0.02	145.91 ± 7.21	218	72.09 ± 16.43	33.07 ± 7.86
25 < pT < 35	100.40 ± 17.59	24.42 ± 3.05	0.01 ± 0.01	177.23 ± 6.63	72.49 ± 5.35	8.97 ± 0.68	0.34 ± 0.03	283.47 ± 9.07	417	133.53 ± 22.35	32.02 ± 5.58
pT > 35	85.13 ± 13.17	46.18 ± 4.84	0.01 ± 0.01	497.48 ± 10.73	139.98 ± 8.50	28.45 ± 0.92	1.10 ± 0.05	713.20 ± 14.54	882	168.80 ± 33.07	19.14 ± 3.80
Central eta	167.89 ± 23.62	40.58 ± 3.91	0.02 ± 0.01	590.14 ± 11.98	185.27 ± 9.34	35.50 ± 1.18	1.26 ± 0.06	852.77 ± 15.73	1145	292.23 ± 37.32	25.52 ± 3.35
Forward eta	145.40 ± 20.51	80.06 ± 6.17	0	241.35 ± 7.70	117.29 ± 7.52	9.00 ± 0.57	0.49 ± 0.03	448.20 ± 12.42	624	175.80 ± 27.90	28.17 ± 4.61
Anti-ID cuts	4551.26 ± 136.92	439.32 ± 14.15	0.93 ± 0.20	176.54 ± 7.23	141.22 ± 8.54	113.80 ± 4.47	0.37 ± 0.03	872.19 ± 18.59	5783	4910.81 ± 78.28	84.92 ± 1.75
pT < 20	2301.33 ± 100.99	206.46 ± 9.33	0.54 ± 0.14	59.11 ± 4.13	68.23 ± 5.85	46.37 ± 2.84	0.13 ± 0.02	380.84 ± 12.10	2975	2594.16 ± 55.87	87.20 ± 2.47
20 < pT < 25	946.01 ± 56.71	93.92 ± 7.14	0.16 ± 0.10	46.89 ± 4.03	36.00 ± 3.95	29.15 ± 2.47	0.08 ± 0.01	206.20 ± 9.43	1199	992.80 ± 35.89	82.80 ± 3.83
25 < pT < 35	668.60 ± 59.71	68.13 ± 5.36	0.12 ± 0.05	25.61 ± 2.87	21.84 ± 2.48	22.57 ± 1.79	0.06 ± 0.01	138.34 ± 6.80	848	709.66 ± 29.90	83.69 ± 4.55
pT > 35	635.33 ± 42.02	70.80 ± 5.79	0.12 ± 0.08	44.94 ± 3.29	15.15 ± 4.12	15.71 ± 1.63	0.10 ± 0.01	146.81 ± 8.00	761	614.19 ± 28.72	80.71 ± 4.78
Central eta	2844.78 ± 110.72	189.08 ± 8.83	0.49 ± 0.15	116.95 ± 5.71	101.58 ± 7.47	86.61 ± 3.86	0.25 ± 0.02	494.97 ± 13.46	3682	3187.03 ± 62.16	86.56 ± 2.21
Forward eta	1706.48 ± 80.55	250.24 ± 11.05	0.44 ± 0.13	59.59 ± 4.44	39.64 ± 4.15	27.19 ± 2.25	0.12 ± 0.02	377.22 ± 12.81	2101	1723.78 ± 47.59	82.05 ± 2.89
fake type: muon	3044.41 ± 108.55	29.11 ± 3.44	0.69 ± 0.28	1204.74 ± 17.73	488.43 ± 19.53	253.28 ± 6.78	2.66 ± 0.09	1978.92 ± 27.46	5941	3962.08 ± 81.82	66.69 ± 1.63
Fake mu —eta—¡2.5	2844.89 ± 104.79	26.58 ± 3.30	0.65 ± 0.28	1170.46 ± 17.45	478.90 ± 19.45	240.41 ± 6.64	2.60 ± 0.09	1919.60 ± 27.16	5669	3749.40 ± 80.04	66.14 ± 1.66
ID cuts	243.62 ± 31.65	2.25 ± 0.90	-0.13 ± 0.14	967.61 ± 16.00	292.00 ± 17.62	54.59 ± 1.90	2.10 ± 0.08	1318.41 ± 23.89	1643	324.59 ± 47.05	19.76 ± 2.90
pT < 20	147.32 ± 25.21	0.76 ± 0.49	-0.07 ± 0.08	117.82 ± 5.79	94.99 ± 5.63	10.15 ± 1.05	0.23 ± 0.02	223.87 ± 8.16	424	200.13 ± 22.15	47.20 ± 5.70
20 < pT < 25	43.15 ± 13.11	0	0.03 ± 0.03	119.50 ± 6.69	45.87 ± 4.72	6.28 ± 0.78	0.26 ± 0.03	171.94 ± 8.22	247	75.06 ± 17.74	30.39 ± 7.44
pT > 25	53.16 ± 13.93	1.49 ± 0.76	-0.10 ± 0.11	730.29 ± 13.33	151.14 ± 16.02	38.17 ± 1.37	1.62 ± 0.07	922.60 ± 20.90	972	49.40 ± 37.53	5.08 ± 3.86
Central eta	116.08 ± 21.31	0.31 ± 0.31	0.01 ± 0.01	463.63 ± 10.31	148.06 ± 16.07	30.95 ± 1.30	1.03 ± 0.06	644.00 ± 19.14	776	132.00 ± 33.80	17.01 ± 4.40
Forward eta	127.54 ± 23.40	1.94 ± 0.85	-0.14 ± 0.14	503.98 ± 12.23	143.94 ± 7.23	23.63 ± 1.38	1.07 ± 0.05	674.42 ± 14.30	867	192.58 ± 32.73	22.21 ± 3.85
Anti-ID cuts	2117.63 ± 94.68	20.31 ± 2.92	0.65 ± 0.23	177.62 ± 6.58	178.98 ± 8.08	173.65 ± 5.97	0.42 ± 0.04	551.63 ± 12.36	3439	2887.37 ± 59.93	83.96 ± 2.26
pT < 20	1288.31 ± 75.03	15.32 ± 2.49	0.45 ± 0.18	56.40 ± 3.73	89.88 ± 5.52	89.05 ± 4.43	0.09 ± 0.01	251.19 ± 8.38	2096	1844.81 ± 46.54	88.02 ± 2.94
20 < pT < 25	453.56 ± 35.64	3.06 ± 1.30	0.05 ± 0.11	30.55 ± 2.87	41.59 ± 3.67	40.89 ± 2.77	0.06 ± 0.01	116.21 ± 5.58	717	600.79 ± 27.35	83.79 ± 4.93
pT > 25	375.75 ± 45.44	1.92 ± 0.80	0.14 ± 0.07	90.67 ± 4.60	47.50 ± 4.62	43.72 ± 2.88	0.27 ± 0.03	184.22 ± 7.17	626	441.78 ± 26.03	70.57 ± 5.02
Central eta	1038.64 ± 66.00	8.24 ± 1.80	0.24 ± 0.10	89.39 ± 4.62	92.10 ± 5.58	82.90 ± 3.93	0.22 ± 0.02	273.10 ± 8.44	1575	1301.90 ± 40.57	82.66 ± 3.31
Forward eta	1078.99 ± 67.89	12.07 ± 2.30	0.41 ± 0.20	88.23 ± 4.68	86.87 ± 5.84	90.75 ± 4.49	0.20 ± 0.03	278.53 ± 9.03	1864	1585.47 ± 44.11	85.06 ± 3.08

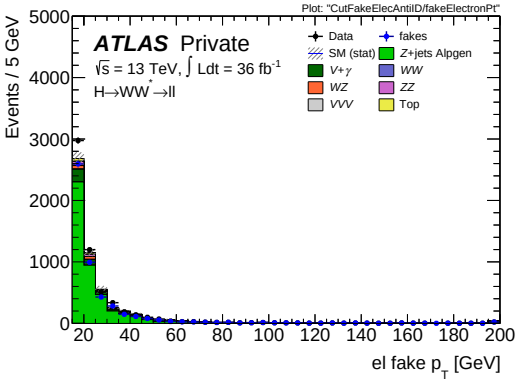
Table 6.3: Cutflow for the Z+jets fake factor estimate with $ee + \mu\mu$ channels combined. The top section shows the common selection, while the fake electron and fake muon yields are displayed in the middle and the bottom, respectively. The "fakes" column corresponds to the data subtracted by the total background and represents the yields that are later used to derive the fake factors. The Z+jets Alpgen column is meant only as a comparison. The WZ normalization factor as described in subsection 6.4.2 has been applied.



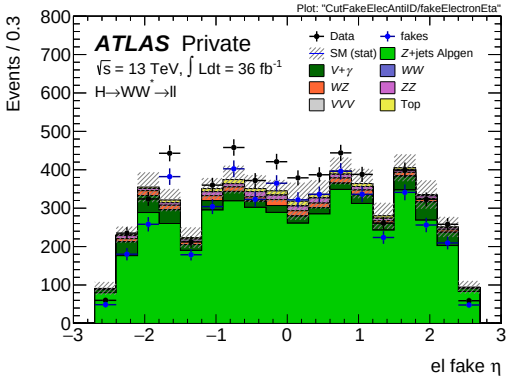
(a) "ID" electron p_{T}



(b) "ID" electron η

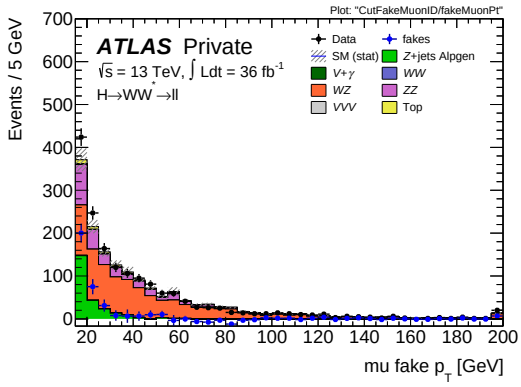


(c) "Anti-ID" electron p_{T}

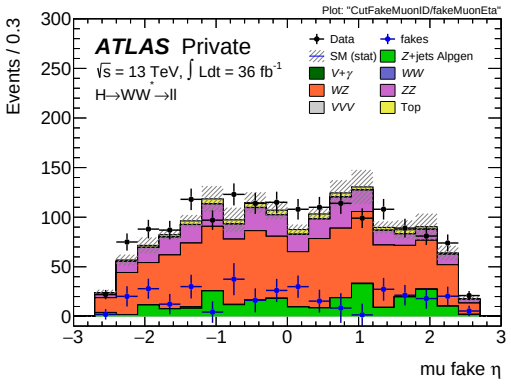


(d) "Anti-ID" electron η

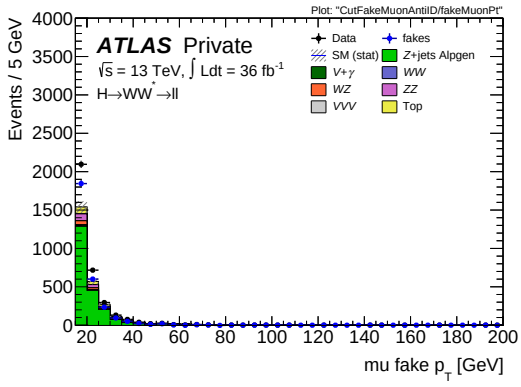
Figure 6.2: Electron fake candidate p_{T} (left) and η (right) distributions for "ID" (top) and "Anti-ID" (bottom) categories. The "fakes" contribution shown in blue is computed by subtracting the EW background processes (excluding the green Z+jets prediction, included for comparison) from the data. The WZ normalization factor as described in subsection 6.4.2 has been applied.



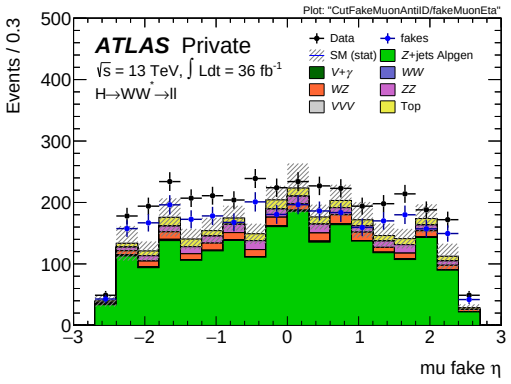
(a) "ID" muon p_{T}



(b) "ID" muon η



(c) "Anti-ID" muon p_{T}



(d) "Anti-ID" muon η

Figure 6.3: Muon fake candidate p_{T} (left) and η (right) distributions for "ID" (top) and "Anti-ID" (bottom) categories. The "fakes" contribution shown in blue is computed by subtracting the EW background processes (excluding the green Z+jets prediction, included for comparison) from the data. The WZ normalization factor as described in subsection 6.4.2 has been applied.

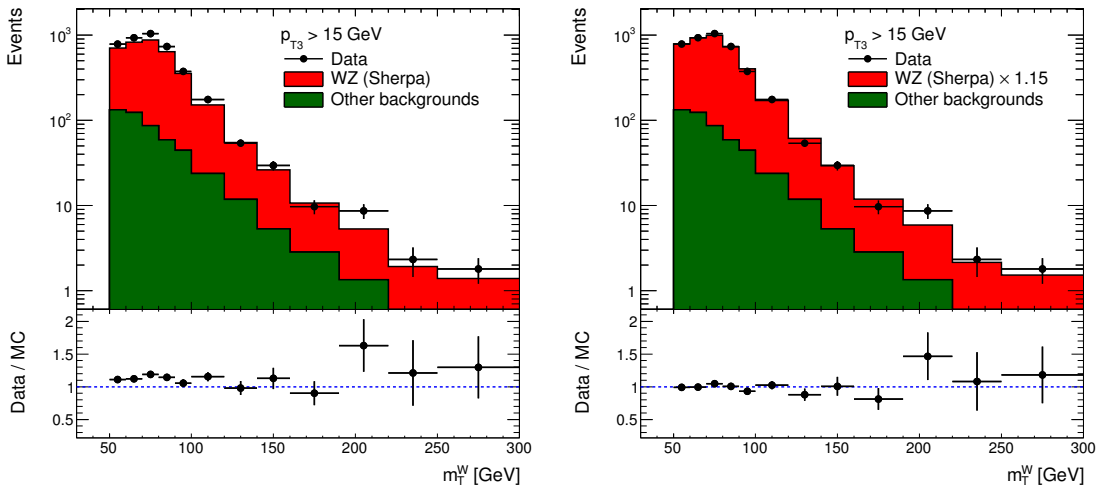


Figure 6.4: Fake candidate m_{T}^{ℓ} before (left) and after (right) the normalization factor is applied to the WZ control region, with WZ in red and other backgrounds in green. The WZ normalization factor is found to be $\alpha = 1.15 \pm 0.02$ (stat.).

and Monte Carlo statistical uncertainties. A normalization factor of $\alpha = 1.15 \pm 0.02$ (stat.) is returned by the fit, with a minimum χ^2 of 8.9 for 11 degrees of freedom. Figure 6.4 shows the fake candidate m_{T}^{ℓ} both before and after the normalization factor is applied in the WZ control region.

The p_{T} distributions of the fake candidate in the WZ control region under the scenarios that it is the leading, subleading, and third leading lepton are shown in Figure 6.5, where the WZ normalization factor has been applied and after which reasonable agreement between data and Monte Carlo is observed.

6.4.3 Results

The Z+jets fake factor is finally computed according to Equation 6.3 as a binned ratio in p_{T} using four bins for electrons ($[15, 20, 25, 35, 1000]$) and three bins for muons ($[15, 20, 25, 1000]$). For electrons, the fake factor is also divided between central and forward regions with two bins in $|\eta|$ ($[0, 1.5, 2.5]$), excluding the EM calorimeter crack region $1.327 < |\eta| < 1.52$. For muons, no statistically significant difference is observed between central and forward regions and therefore the fake factor is integrated in $|\eta|$ so as to gain statistical precision.

The central values as well as the statistical errors of the fake factor for each p_{T} and η bin is presented in Table 6.4. The p_{T}-differential distributions of the fake factors for each η bin are shown in Figure 6.6, where a comparison is also made with Monte Carlo predictions using POWHEG, ALPGEN and SHERPA Z+jets. It can be seen that for fake factors derived from Z+jets events are dominated by statistical uncertainties, particularly for muon fake candidates.

6.5 Dijets Fake Factor

While they are not applied nominally in the analysis, it is useful in addition to derive fake factors from a dijets sample not only as a cross-check with the Z+jets fake factors, but also to be used for a special case in which a trigger bias appears in the estimate. The details for

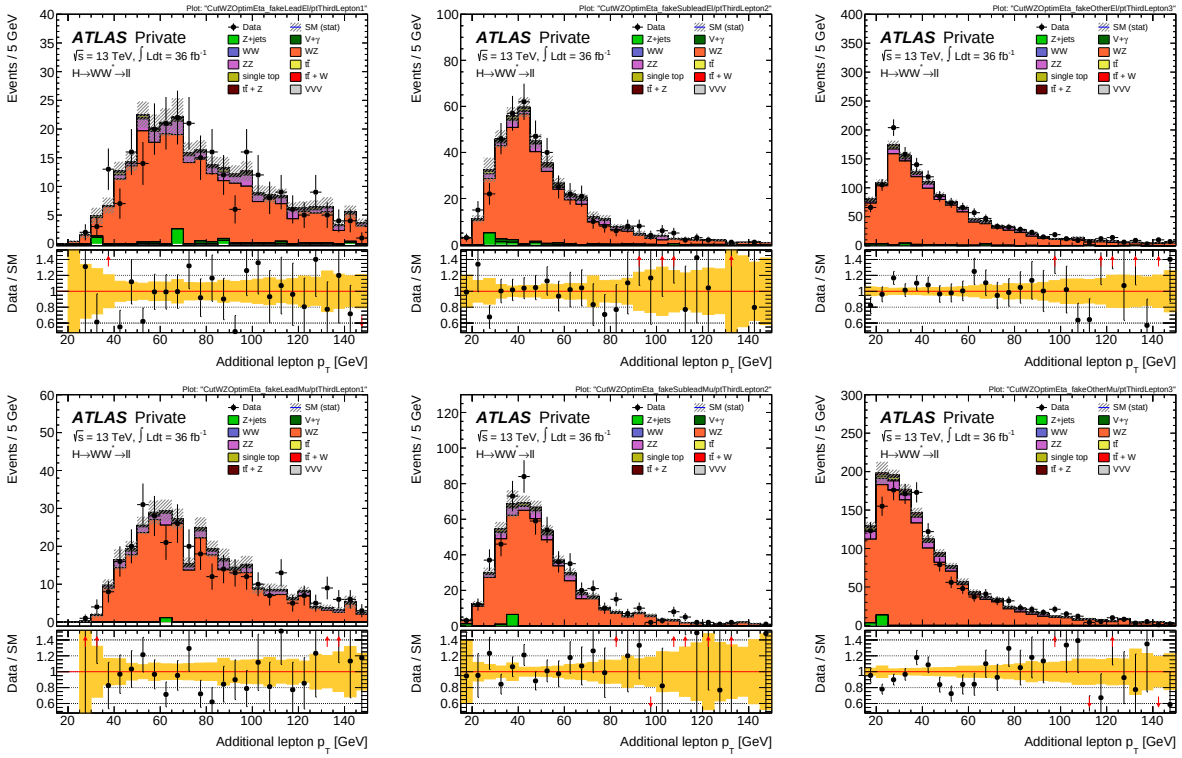
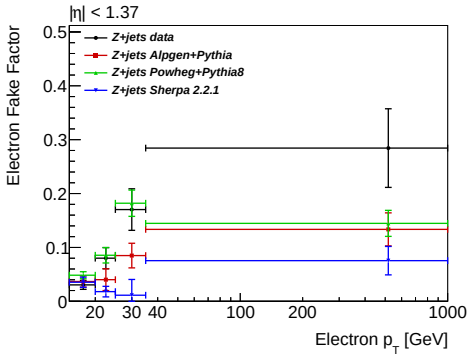


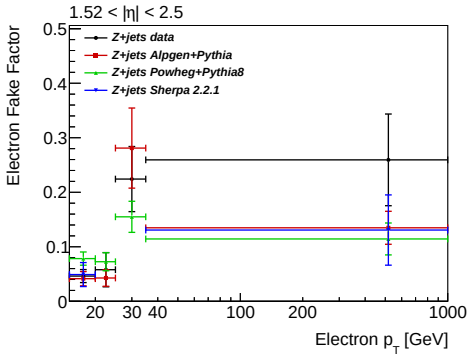
Figure 6.5: p_{T} distributions of the fake lepton candidate in the WZ control region for electron flavor (top) and muon flavor (bottom) under the scenarios that it is the leading (left) subleading (middle) or third leading (right) lepton in the event. The WZ normalization factor has been applied.

p_T range [GeV]	electron $\lvert\eta\rvert < 1.5$	electron $\lvert\eta\rvert > 1.5$	muon
15.0 – 20.0	0.030 ± 0.008	0.046 ± 0.012	0.108 ± 0.012
20.0 – 25.0	0.080 ± 0.020	0.058 ± 0.031	0.125 ± 0.030
25.0 – 35.0	0.170 ± 0.039	0.224 ± 0.060	0.112 ± 0.085
35.0 – 1000.0	0.284 ± 0.073	0.259 ± 0.084	

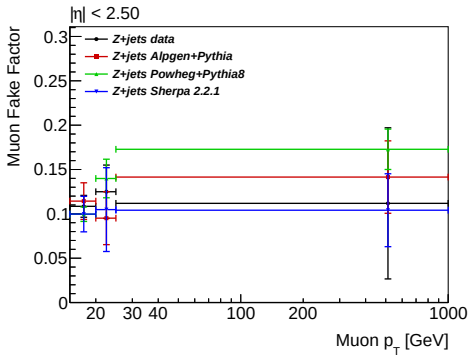
Table 6.4: Summary of fake factors derived from the Z+jets sample. The uncertainties are statistical only and the EM calorimeter crack region is excluded in the case of electron fake candidates.



(a) e: $|\eta| < 1.37$



(b) e: $1.52 < |\eta| < 2.5$



(c) μ: $|\eta| < 2.5$

Figure 6.6: Differential distributions of electron (top) and muon (bottom) fake factors as a function of fake candidate p_{T} for each $|\eta|$ bin. Monte Carlo predictions from POWHEG, ALPGEN and SHERPA Z+jets are also shown. The uncertainties are statistical only.

Sample	Trigger	Scope
Z+jets	e: HLT_e24(26) OR HLT_60M OR HLT_120L μ: HLT_m20(24/26) OR HLT_50	nominal F.F.
Dijet	e: HLT_e12_lhvloose_nod0_L1EM10VH μ: HLT_mu14_L1_MU10	nominal F.F.(for cross check)
Dijet	e: HLT_e24(26) OR HLT_60M OR HLT_120L μ: HLT_m20(24/26) OR HLT_50	"triggered" F.F.

Table 6.5: Summary of the strategy for which samples are used to derive fake factors in different scopes, along with the corresponding trigger selection.

both of these applications are described in this section, with a summary of the scope for each fake factor variant provided in Table 6.5.

6.5.1 Nominal

Nominal fake factors are also derived from a dijet-like, lepton-plus-jet, fake-enriched sample using the prescaled triggers HLT_e12_lhvloose_nod0_L1EM10VH and HLT_mu14_L1_MU10 for electrons and muons, respectively. These triggers are specifically chosen so as not to introduce a trigger bias since they are looser than the Anti-ID lepton definition. In order to enhance the contribution from dijet events and suppress the background contributions from electroweak processes, the following requirements are made:

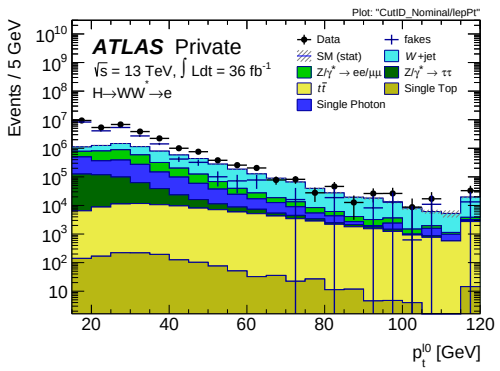
- Exactly one loosely defined lepton with $p_{\mathrm{T}} > 15\,\mathrm{GeV}$, considered the fake candidate
- At least one jet with $p_{\mathrm{T}} > 22\,\mathrm{GeV}$
- Angular separation of $\Delta\phi > 2.5$ between the leading p_{T} jet and the fake candidate
- $p_{\mathrm{T}}^{\mathrm{miss}} < 30\,\mathrm{GeV}$
- $m_{\mathrm{T}}^{\ell} < 60\,\mathrm{GeV}$

where the angular separation requirement between the leading jet and the fake candidate is meant to target the topology of a dijet event and the last two cuts are made to reduce the contribution from electroweak backgrounds (particularly W+jets).

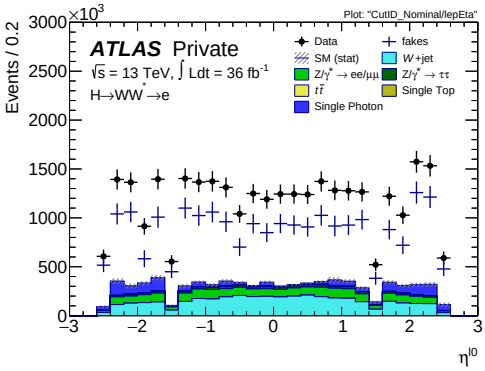
The data is also "un-prescaled" by applying a weight according to the prescale evolution over time so as to recover the yields that were present before prescaling. After the remaining electroweak backgrounds are subtracted (W+jets, Z+jets, $t\bar{t}$ and γ+jets estimated from Monte Carlo), the events are split into the lepton "ID" and "Anti-ID" categories defined in Table 6.1 which are then used to derive the fake factor. Distributions of the ID/Anti-ID fake candidate p_{T} and η are shown in Figure 6.7 for electrons and Figure 6.8 for muons, where it can be seen that the contamination of electroweak background is reduced relative to the Z+jets sample.

6.5.2 Triggered

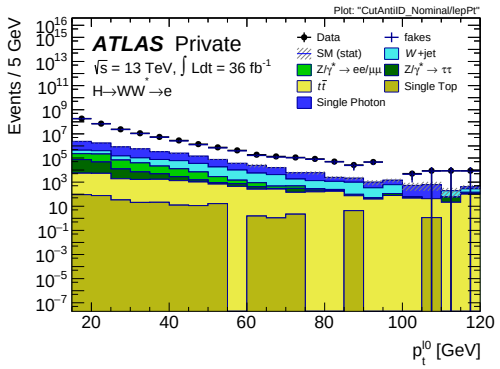
In situations where the Anti-ID lepton in the W+jets control sample is solely responsible for firing a single lepton trigger (i.e. it is the only lepton matched to one of the single lepton triggers used in the analysis from Table 4.1), a bias in the estimation is introduced if the



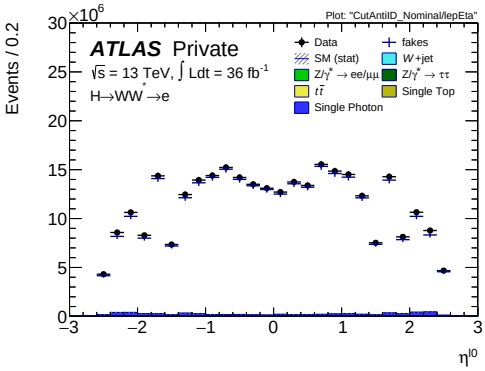
(a) "ID" electron p_{T} (log)



(b) "ID" electron η (lin)

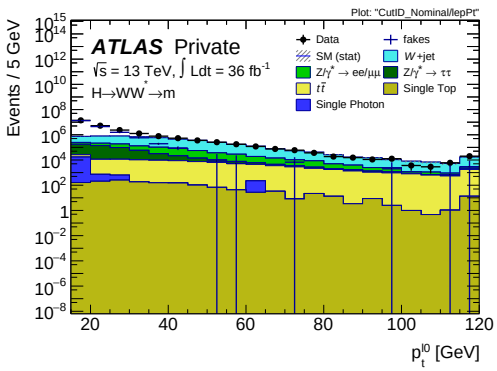


(c) "Anti-ID" electron p_{T} (log)

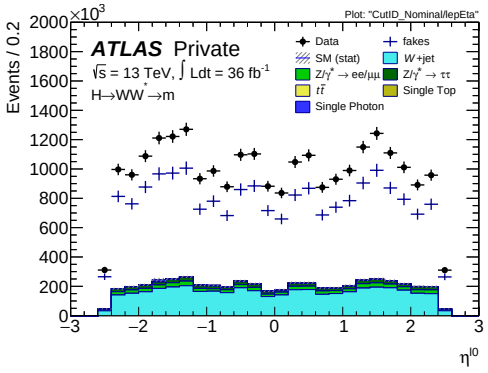


(d) "Anti-ID" electron η (lin)

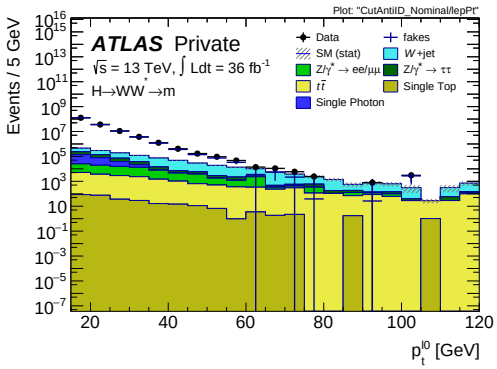
Figure 6.7: Electron fake candidate p_{T} (left) and η (right) distributions for "ID" (top) and "Anti-ID" (bottom) categories. The "fakes" contribution shown in blue is computed by subtracting the EW background processes from the data.



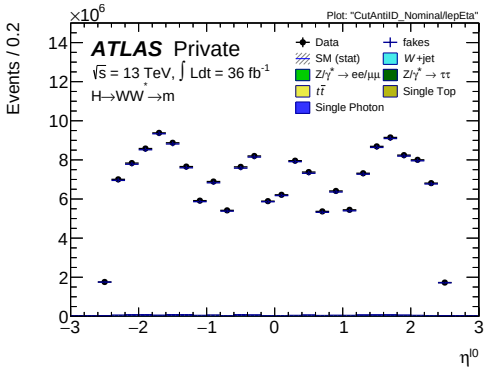
(a) "ID" muon p_{T} (log)



(b) "ID" muon η (lin)



(c) "Anti-ID" muon p_{T} (log)



(d) "Anti-ID" muon η (lin)

Figure 6.8: Muon fake candidate p_{T} (left) and η (right) distributions for "ID" (top) and "Anti-ID" (bottom) categories. The "fakes" contribution shown in blue is computed by subtracting the EW background processes from the data.

nominal fake factors are applied since the online lepton identification used in these triggers is tighter than the Anti-ID definition[4].

To avoid a bias in this case, dedicated "triggered" fake factors are applied which are derived using the same dijet sample as described in subsection 6.5.1 but using the analysis triggers rather than the prescaled ones. The reason for using the dijet sample as opposed to the Z+jets sample is because of the low statistics available after requiring the fake lepton candidate to fire one of the single lepton triggers in the Z+jets fake factor estimate[5].

The fraction of events for different categories in the W+jets control sample that require the triggered fake factors are shown in Table 6.6, where it can be seen that the category with the highest percentage is for Anti-ID muons that happen to be the leading lepton in the event. Separate sets of triggered fake factors are derived for each of the three trigger configurations in the analysis, as are defined in Table 6.7.

	$e\mu$, $N_{\text{jet}}=0$	μe, $N_{\text{jet}}=0$	$e\mu$, $N_{\text{jet}}=1$	μe, $N_{\text{jet}}=1$	$e\mu$, VBF	μe, VBF
e-fakes	3.6%	0.0%	6.0%	0.6%	6.5%	2.2%
μ-fakes	0.1%	17.2 %	0.1%	19.6 %	0.5%	18.5%

Table 6.6: Fraction of events for different categories in the W+jets control sample that require a triggered fake factor (i.e. the Anti-ID lepton alone fired one of the analysis single lepton triggers.)

6.5.3 Results

The fake factors derived from the dijet samples are also computed from Equation 6.3 and with the same binning as for the Z+jets fake factors except that the muons are also split into central and forward regions, with a binning of $[0, 1.05, 2.5]$. The central values and statistical uncertainties for the nominal and triggered dijet fake factors are reported in Table 6.8 + Table 6.9 for electrons and in Table 6.10 + Table 6.11 for muons.

*Trig*2015	*Trig*2016	*Trig*2016*D*
2015	Pre-D4 Period 2016	Post-D4 period 2016
e		
e24_lhmedium_L1EM20VH	e24_lhtight_nod0_ivarloose	e26_lhtight_nod0_ ivarloose
e60_lhmedium	e60_lhmedium_nod0	e60_lhmedium_ nod0
e120_lhloose	e140_lhloose_nod0	e140_lhloose_nod0
μ		
mu20_iloose_L1MU15	mu24_ivarmedium	mu26_ivarmedium
mu50	mu50	mu50

Table 6.7: The three distinct trigger configurations in the analysis that are used in order to collect samples for the sets of triggered fake factors.

[4]No such bias occurs in the case of the dilepton triggers since their selection is not tighter than the Anti-ID definition.

[5]Recall that in the Z+jets fake factor estimate, one of the Z candidate leptons is instead required to fire a single lepton trigger, allowing the fake candidate to initially remain potentially looser than the Anti-ID definition.

	Nominal	*Trig*2015	*Trig*2016	*Trig*2016*D*
$15 < p_{\mathrm{T}} < 20$	0.042 ± 0.002	0 ± 0	0 ± 0	0 ± 0
$20 < p_{\mathrm{T}} < 25$	0.054 ± 0.004	0 ± 0	0 ± 0	0 ± 0
$25 < p_{\mathrm{T}} < 35$	0.212 ± 0.011	0.420 ± 0.006	1.259 ± 0.030	1.162 ± 0.033
$p_{\mathrm{T}} > 35$	0.203 ± 0.026	0.537 ± 0.006	1.196 ± 0.074	1.164 ± 0.045

Table 6.8: Electron dijet fake factors for nominal and triggered events in the central region with $|\eta| < 1.37$.

	Nominal	*Trig*2015	*Trig*2016	*Trig*2016*D*
$15 < p_{\mathrm{T}} < 20$	0.05 ± 0.003	0 ± 0	0 ± 0	0 ± 0
$20 < p_{\mathrm{T}} < 25$	0.069 ± 0.006	0 ± 0	0 ± 0	0 ± 0
$25 < p_{\mathrm{T}} < 35$	0.285 ± 0.018	0.508 ± 0.029	1.190 ± 0.058	1.093 ± 0.031
$p_{\mathrm{T}} > 35$	0.261 ± 0.031	0.598 ± 0.030	1.217 ± 0.077	1.080 ± 0.033

Table 6.9: Electron dijet fake factors for nominal and triggered events in the forward region with $1.52 < |\eta| < 2.5$.

	Nominal	*Trig*2015	*Trig*2016	*Trig*2016*D*
$15 < p_{\mathrm{T}} < 20$	0.107 ± 0.002	0 ± 0	0 ± 0	0 ± 0
$20 < p_{\mathrm{T}} < 25$	0.130 ± 0.004	0.324 ± 0.006	0 ± 0	0 ± 0
$p_{\mathrm{T}} > 25$	0.151 ± 0.010	0.383 ± 0.021	0.648 ± 0.025	0.549 ± 0.024

Table 6.10: Muon dijet fake factors for nominal and triggered events in the central region with $|\eta| < 1.05$.

	Nominal	*Trig*2015	*Trig*2016	*Trig*2016*D*
$15 < p_{\mathrm{T}} < 20$	0.104 ± 0.002	0 ± 0	0 ± 0	0 ± 0
$20 < p_{\mathrm{T}} < 25$	0.121 ± 0.003	0.292 ± 0.004	0 ± 0	0 ± 0
$p_{\mathrm{T}} > 25$	0.147 ± 0.008	0.368 ± 0.013	0.626 ± 0.018	0.529 ± 0.017

Table 6.11: Muon dijet fake factors for nominal and triggered events in the forward region with $1.05 < |\eta| < 2.5$.

6.6 Flavor Composition and Correction Factor

As described so far in this chapter, the background from one misidentified lepton is estimated nominally with fake factors derived in a Z+jets control sample and applied to a W+jets control sample. One natural complication with this method is that misidentified leptons can originate from a number of different sources whose relative abundance will vary between samples. If these sources also have separate rates for passing ID and Anti-ID requirements (due to differences in jet kinematics, impact parameters, etc.), then this will in turn reflect as a discrepancy in fake factors derived from separate samples.

In order to account for this discrepancy, the flavor composition of the fake leptons for opposite sign (OS) W+jets and Z+jets samples is first investigated using Monte Carlo truth information[6]. A correction factor (CF) is then derived from this sample which is applied nominally in the analysis on top of the fake factor - that is, the W+jets background in the signal region is

$$N_{\text{id+id}}^{\text{W+jets(OS)}} = f_{W\text{+jets}}^{\text{OS}} \cdot N_{\text{id+anti-id}}^{\text{W+jets(OS)}} = f_{Z\text{+jets}}^{\text{incl.}} \cdot \frac{f_{W\text{+jets}}^{\text{OS}}}{f_{Z\text{+jets}}^{\text{incl.}}} N_{\text{id+anti-id}}^{\text{W+jets(OS)}} \tag{6.6}$$

where $f_{Z\text{+jets}}^{\text{incl.}}$ is the nominal fake factor derived from the Z+jets data-driven sample as described in section 6.4 and $f_{W\text{+jets}}^{\text{OS}}/f_{Z\text{+jets}}^{\text{incl.}}$ is the correction factor derived exclusively from W+jets and Z+jets Monte Carlo.

For the flavor composition study and correction factor derivation, the fake leptons are selected in the OS W+jets sample by performing an event selection similar to the analysis selection, except that the reconstructed lepton originating from the leptonically decaying W is identified by being matched to the truth prompt lepton. For the Z+jets sample an event selection similar to the Z+jets fake factor estimate presented in subsection 6.4.1 is performed, except that the two Z candidate leptons are also identified by being matched to the truth leptons from the Z decay and allowing the additional lepton to be classified as fake.

The flavor of the fake lepton is assigned according to the following truth matching scheme based on truth objects found to be in close proximity:

- The algorithm begins by searching for a bottom truth object (either a quark or meson/baryon state) within $\Delta R < 0.4$ to the fake lepton. If one is found, the fake lepton flavor is classified as "bottom".
- If none are found, the process repeats for charm, strange, and light objects, in that order.
- If still nothing is found, then the fake lepton is classified as "other".
- At each stage in the process, the fake lepton is classified also as "leptonic" if there is a truth lepton within $\Delta R < 0.03$. Otherwise, it is classified as "hadronic".

Table 6.12 and Table 6.13 report the flavor fractions of ID and Anti-ID leptons using Powheg+Pythia 8 for electron fakes and muon fakes, respectively. Distributions of fake lepton p_{T} illustrating the same split into flavor are provided in Figure 6.9 and Figure 6.10. It can be seen that for electron fakes, contributions from the "light" and "other" categories

[6]Truth information here simply refers to the details of the particles that were produced by the Monte Carlo generator along with the subsequent parton shower and hadronization, i.e. before detector simulation.

e-fake	% Bottom	% Charm	% Strange	% Light	% Other
ID electron					
OS W+jets	2.5 ± 0.8	36.5 ± 4.6	5.5 ± 1.7	33.2 ± 4.1	22.3 ± 3.5
Z+jets	22.7 ± 3.3	17.7 ± 3.0	6.1 ± 1.3	32.5 ± 3.7	21.0 ± 2.8
Anti-ID electron					
OS W+jets	1.9 ± 0.2	43.1 ± 1.5	9.4 ± 0.6	39.6 ± 1.3	5.9 ± 0.5
Z+jets	21.9 ± 1.0	17.9 ± 0.8	13.1 ± 0.7	42.5 ± 1.4	4.6 ± 0.4

Table 6.12: ID and Anti-ID fake electron flavor percentages for OS W+jets and Z+jets samples using POWHEG+PYTHIA 8.

μ-fake	% Bottom	% Charm	% Strange	% Light	% Other
ID muon					
OS W+jets	4.4 ± 1.3	82.6 ± 9.6	3.8 ± 1.2	6.7 ± 1.7	2.6 ± 1.0
Z+jets	59.2 ± 6.3	29.4 ± 3.5	2.2 ± 0.8	4.8 ± 1.2	4.3 ± 1.3
Anti-ID muon					
OS W+jets	7.7 ± 1.1	77.2 ± 3.6	6.5 ± 0.7	6.0 ± 0.8	2.7 ± 0.6
Z+jets	65.0 ± 2.6	22.7 ± 1.2	5.6 ± 0.5	4.7 ± 0.6	2.0 ± 0.4

Table 6.13: ID and Anti-ID fake muon flavor percentages for OS W+jets and Z+jets samples using POWHEG+PYTHIA 8.

are present in addition to a heavy flavor component which is made up largely of the "charm" category in the case of OS W+jets and more of the "bottom" category in the case of Z+jets. For muon fakes the source is almost exclusively heavy flavor, being composed of similar ratios for "charm" and "bottom" as with the electron fakes. The asymmetry observed in the heavy flavor component between OS W+jets and Z+jets is expected on a theoretical level due to an additional W+charm channel that is not available to Z+jets.

Fake factors are also calculated separately for each flavor category in both samples and are reported in Table 6.14 and Table 6.15 for electron and muon fakes, respectively. However due to a lack of statistics, they are fully integrated in p_{T} and η providing only a global view.

A comparison of the total (flavor combined) fake factors between OS W+jets and Z+jets Monte Carlo is shown in Figure 6.11, also with the deviation of the correction factor from unity, $(f^{\mathrm{OS}}_{W+\mathrm{jets}}/f^{\mathrm{incl.}}_{Z+\mathrm{jets}}) - 1$. The final correction factors applied in the analysis are provided in Table 6.19, along with their systematic uncertainty as described in subsection 6.7.2.

e-fake	Bottom FF	Charm FF	Strange FF	Light FF	Other FF
OS W+jets	0.129 ± 0.044	0.082 ± 0.009	0.056 ± 0.018	0.081 ± 0.009	0.362 ± 0.060
Z+jets	0.094 ± 0.013	0.090 ± 0.015	0.042 ± 0.009	0.069 ± 0.007	0.412 ± 0.060

Table 6.14: Electron fake factors fully integrated in p_{T} and η for each flavor component. OS W+jets and Z+jets samples are compared using POWHEG+PYTHIA 8.

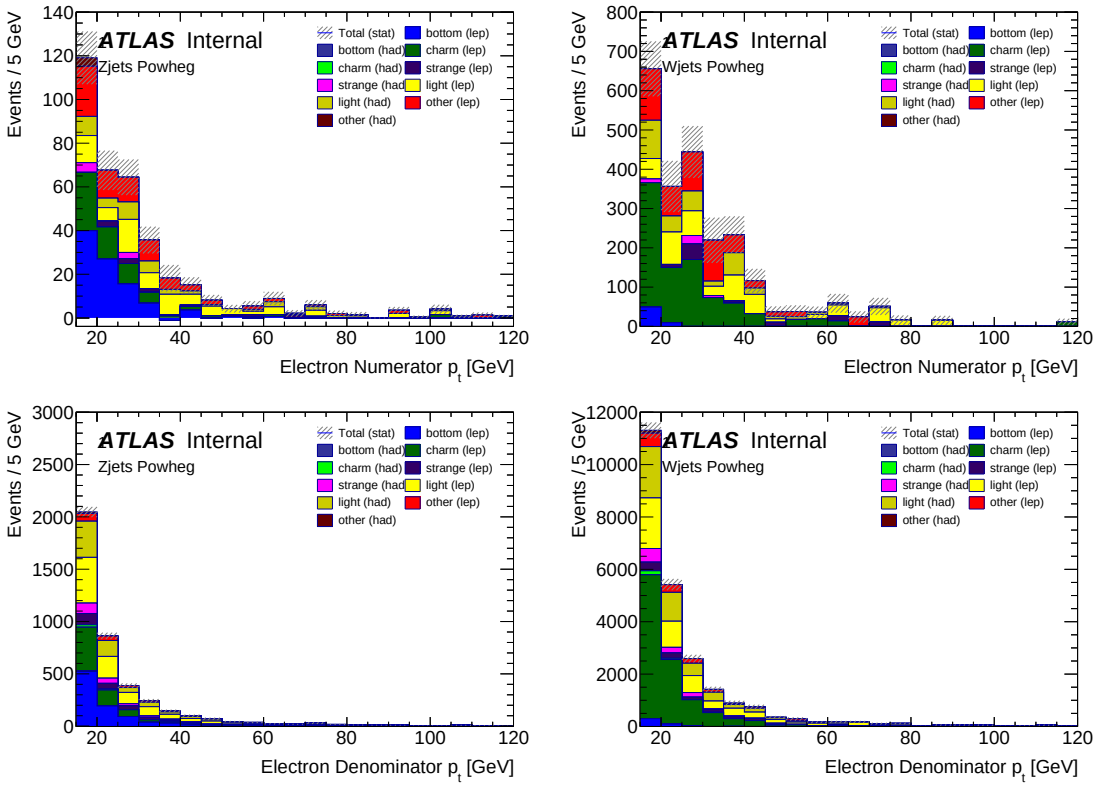


Figure 6.9: Comparison of the flavor composition in p_{T} distributions for fake electrons between the Z+jets (left) and OS W+jets (right) samples. Both ID (top) and Anti-ID (bottom) populations are shown. The events shown are generated using Powheg+PYTHIA 8.

μ-fake	Bottom FF	Charm FF	Strange FF	Light FF	Other FF
OS W+jets	0.102 ± 0.033	0.191 ± 0.018	0.103 ± 0.035	0.198 ± 0.054	0.172 ± 0.079
Z+jets	0.115 ± 0.011	0.163 ± 0.018	0.050 ± 0.018	0.128 ± 0.033	0.270 ± 0.098

Table 6.15: Muon fake factors fully integrated in p_{T} and η for each flavor component. OS W+jets and Z+jets samples are compared using Powheg+PYTHIA 8.

6.7 Fake Factor Systematics

The systematic uncertainties on the Z+jets fake factor estimate of the W+jets background in the analysis can be divided into three sources:

- Statistical uncertainties on the fake factors themselves that are applied as systematic variations to the fake estimate by varying all bins independently, i.e. the uncertainties are treated as uncorrelated across all bins
- Uncertainties associated with the prompt lepton contamination from electroweak processes in the Z+jets sample, estimated by varying the amount of background subtracted
- Difference between Z+jets and W+jets fake factors due to sample composition, estimated using Monte Carlo

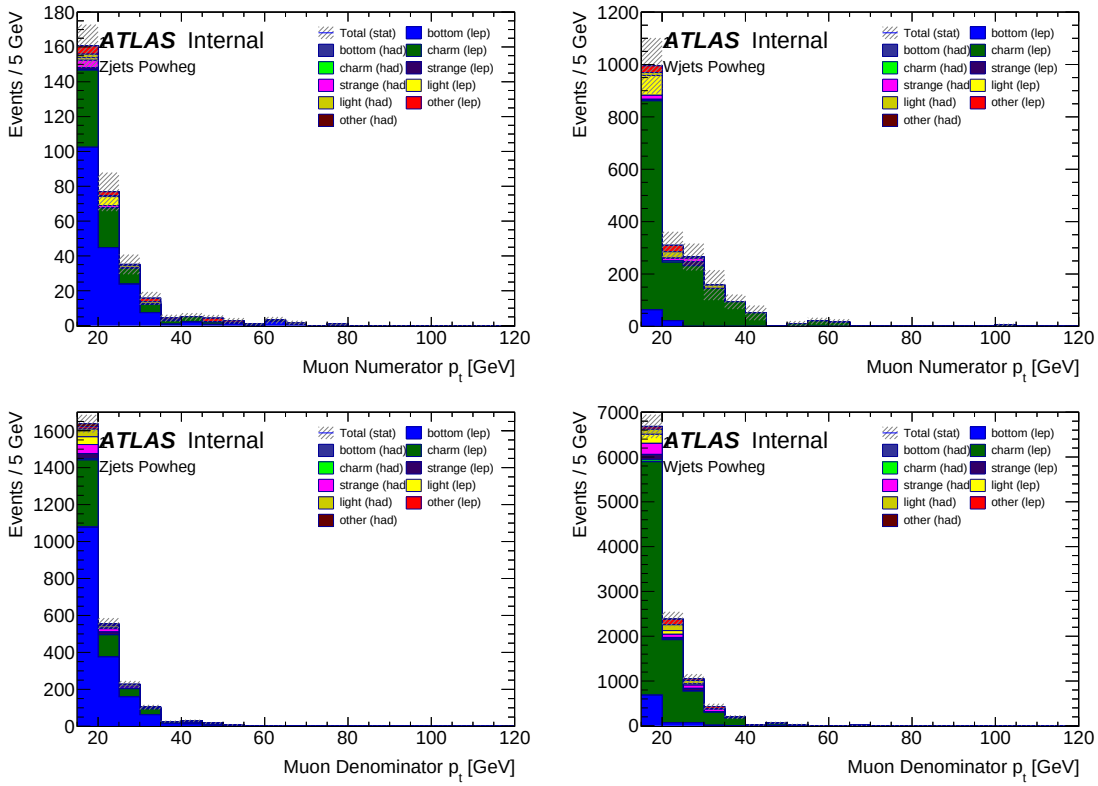


Figure 6.10: Comparison of the flavor composition in p_{T} distributions for fake muons between the Z+jets (left) and OS W+jets (right) samples. Both ID (top) and Anti-ID (bottom) populations are shown. The events shown are generated using POWHEG+PYTHIA 8.

A summary of all fake factor related systematics is provided in Table 6.16. The contribution from each source varies depending on the p_{T} and η of the fake candidate, except in some cases such as for muon flavor at high p_{T} where the electroweak subtraction uncertainty dominates. The determination of electroweak subtraction and sample composition uncertainties are described in more detail below.

6.7.1 Electroweak Subtraction Uncertainty

The samples used to derive the fake factor contain prompt backgrounds that are estimated by Monte Carlo and whose yields are also subject to uncertainty. While the Anti-ID population naturally has a high fake purity, the ID population is contaminated to a larger degree with electroweak processes, particularly at high p_{T} as can be seen in Figure 6.2 and Figure 6.3. In order to assess the uncertainty on the fake factors due to electroweak subtraction, each process is varied by its corresponding uncertainties described below. The fake factors are then recomputed and the difference between the variations and the central values are taken as the result.

The main backgrounds in the Z+jets fake factor sample are WZ, ZZ and $V+\gamma$ (mostly $Z+\gamma$). The uncertainty on WZ is taken from systematic variations of the normalization factor described in section 6.3 which include WZ theory uncertainties and a p_{T} dependence for the result of the χ^2 fit, together amounting to 7.7% [48]. The uncertainties on ZZ and $V+\gamma$ backgrounds are estimated based on theory uncertainties which include the choice of generator, QCD scale variations, and PDF, amounting to 7.7% and 10%, respectively.

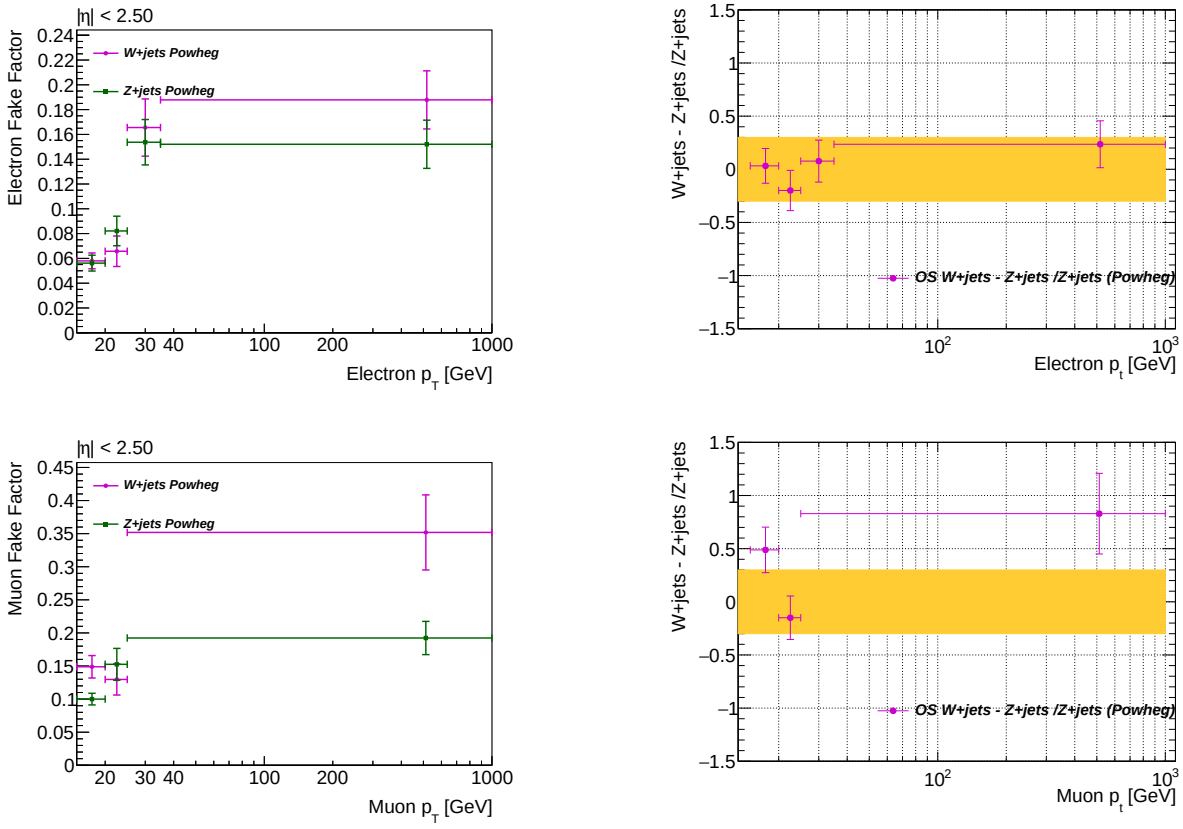


Figure 6.11: Electron (top) and muon (bottom) fake factor comparison between OS W+jets and Z+jets Monte Carlo. Plots for the deviation of the correction factor from unity, $(f^{\mathrm{OS}}_{W+\mathrm{jets}}/f^{\mathrm{incl.}}_{Z+\mathrm{jets}}) - 1$, are also shown where the yellow band is drawn simply as a point of reference to ± 30%. POWHEG+PYTHIA 8 is used to generate the samples shown.

Kinematic region ($\|\eta\|$ and p_T range)	Statistical	EW Subtraction	Sample Composition	Total
Electron:				
$0.0 < \|\eta\| < 1.5$				
$15 - 20$ GeV	27	13	32	44
$20 - 25$ GeV	25	16	32	44
$25 - 35$ GeV	23	16	13	31
$35 - 1000$ GeV	26	33	13	44
$1.5 < \|\eta\| < 2.5$				
$15 - 20$ GeV	26	13	32	43
$20 - 25$ GeV	54	16	32	65
$25 - 35$ GeV	27	16	13	34
$35 - 1000$ GeV	32	33	13	47
Muon:				
$0.0 < \|\eta\| < 2.5$				
$15 - 20$ GeV	11	9	23	27
$20 - 25$ GeV	24	17	23	37
$25 - 1000$ GeV	76	143	23	163

Table 6.16: Summary of systematic uncertainties on the Z+jets fake factor estimate in percentage, separated based on the flavor and kinematic phase space of the fake lepton candidate.

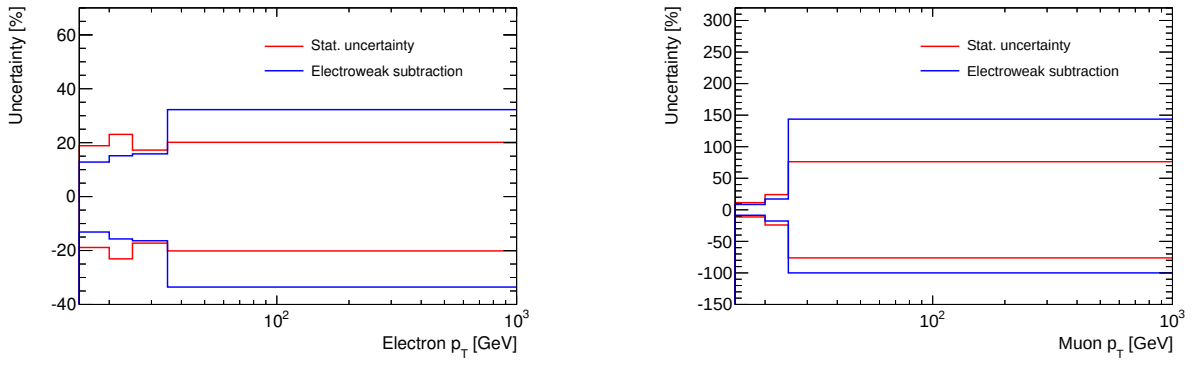
	POWHEG+PYTHIA 8	ALPGEN+PYTHIA 6
$15 < p_T < 20$	1.03 ± 0.16	1.19 ± 0.32
$20 < p_T < 25$	0.80 ± 0.19	1.16 ± 0.43
$25 < p_T < 35$	1.08 ± 0.20	0.92 ± 0.21
$p_T > 35$	1.24 ± 0.22	1.28 ± 0.23
p_T average	1.02 ± 0.09	1.11 ± 0.13

Table 6.17: Comparison of electron $f^{\mathrm{OS}}_{W+\mathrm{jets}}/f^{\mathrm{incl.}}_{Z+\mathrm{jets}}$ correction factors derived using nominal (POWHEG+PYTHIA 8) and alternative (ALPGEN+PYTHIA 6) generators.

The rest of the backgrounds are also varied by 10%. The relative uncertainties on the fake factors from varying the background by these percentages collectively up and down are shown in Figure 6.12, after integrating the fake factors in η.

6.7.2 Flavor Composition Uncertainty

The uncertainty due to differences in flavor composition between the Z+jets sample used to derive the nominal fake factors and the OS W+jets sample which they're applied is determined by comparing the correction factors obtained using different Monte Carlo generators. Table 6.17 and Table 6.18 show the correction factors side-by-side for POWHEG and ALPGEN. POWHEG is chosen as the nominal generator, providing the central values of the correction factors due to its superior statistical precision. The systematic uncertainties are then evaluated by comparing with the central values of ALPGEN.



(a) electron (b) muon

Figure 6.12: Electroweak subtraction uncertainties on the Z+jets fake factor estimate in blue for electrons (left) and muons (right). The statistical uncertainties are also shown for comparison. The fake factors are integrated in $|\eta|$ due to lack of statistics. In the highest-p_{T} bin for muons, the variation increasing the amount of electroweak background being subtracted results in the fake factor being negative. Therefore, the fake factor is set to zero for this bin.

	POWHEG+PYTHIA 8	ALPGEN+PYTHIA 6
$15 < p_T < 20$	1.49 ± 0.21	0.99 ± 0.22
$20 < p_T < 25$	0.85 ± 0.20	1.31 ± 0.48
$p_T > 25$	1.83 ± 0.38	1.63 ± 0.53
p_T average	1.24 ± 0.14	1.12 ± 0.18

Table 6.18: Comparison of electron $f^{\mathrm{OS}}_{W+\mathrm{jets}}/f^{\mathrm{incl.}}_{Z+\mathrm{jets}}$ correction factors derived using nominal (POWHEG+PYTHIA 8) and alternative (ALPGEN+PYTHIA 6) generators.

	$f^{OS}_{W+jets}/f^{incl.}_{Z+jets}$	
	$p_{\mathrm{T}} < 25$ GeV	$p_{\mathrm{T}} > 25$ GeV
electrons	0.96 ± 0.13(stat) ± 0.28(syst)	1.15 ± 0.15(stat) ± 0.02(syst)
muons	1.34 ± 0.17(stat) ± 0.25(syst)	1.83 ± 0.38(stat) ± 0.20(syst)

Table 6.19: Final correction factors applied in the analysis and their corresponding uncertainties. POWHEG+PYTHIA 8 is used to derive the central values, while the systematic uncertainty is evaluated by comparing with ALPGEN+PYTHIA 6.

The final correction factors applied in the analysis are summarized in Table 6.19. They are split into bins of low and high p_{T} (below and above 25 GeV, respectively) due to their observed p_{T} dependence and are calculated by first taking the p_{T} average of the relevant fake factor bins and then dividing the resulting W+jets fake factor by the resulting Z+jets fake factor.

6.8 QCD Double Fakes

The previous sections of this chapter outline the procedure adopted for estimating the W+jets background which appears in the signal region due to a single misidentified lepton in the event. However, it is also possible for QCD multijet processes to be selected as well via a double lepton misidentification. In fact, this contribution is already included by design of the fake factor method as outlined above - although it is overestimated, as is demonstrated in Equation 6.7

$$\begin{aligned} N^{\text{FF estimate}}_{\text{id+id}} &= f^Z_e N_{\mu,\not{e}} + f^Z_\mu N_{e,\not{\mu}} \\ &= f^Z_e \times (N^{\text{data}}_{\mu,\not{e}} - N^{\text{EW MC}}_{\mu,\not{e}}) + f^Z_\mu \times (N^{\text{data}}_{e,\not{\mu}} - N^{\text{EW MC}}_{e,\not{\mu}}) \\ &= f^Z_e \times (N^{W+\text{jets}}_{\mu,\not{e}} + N^{\text{QCD}}_{\mu,\not{e}}) + f^Z_\mu \times (N^{W+\text{jets}}_{e,\not{\mu}} + N^{\text{QCD}}_{e,\not{\mu}}) \\ &= f^Z_e \times N^{W+\text{jets}}_{\mu,\not{e}} + f^Z_e \cdot f^D_\mu N^{\text{QCD}}_{\not{\mu},\not{e}} + f^Z_\mu \times N^{W+\text{jets}}_{e,\not{\mu}} + f^Z_\mu \cdot f^D_e N^{\text{QCD}}_{\not{e},\not{\mu}} \end{aligned} \tag{6.7}$$

where f^Z_e and f^Z_μ denote the Z+jets fake factors for electrons and muons respectively, f^D_e and f^D_μ denote the dijet fake factors, the leptons are not p_{T} ordered, and a strikethrough represents an anti-identification. In the last line the QCD terms have been expanded using the fake factor method, where it is clear that if the Z+jets and dijet fake factors are the same, the QCD contribution is exactly double counted. Therefore, the overestimation can be accounted for by adding to $N^{\text{FF estimate}}_{\text{id+id}}$ a correction term

$$N^{\text{QCDcorr}}_{\text{id+id}} = N^{QCD}_{\not{e},\not{\mu}} \cdot \text{FF}^{\text{QCD}} = (N^{data}_{\not{e},\not{\mu}} - N^{EWMC}_{\not{e},\not{\mu}}) \cdot (f^D_e f^D_\mu - f^Z_e \cdot f^D_\mu - f^Z_\mu \cdot f^D_e) \tag{6.8}$$

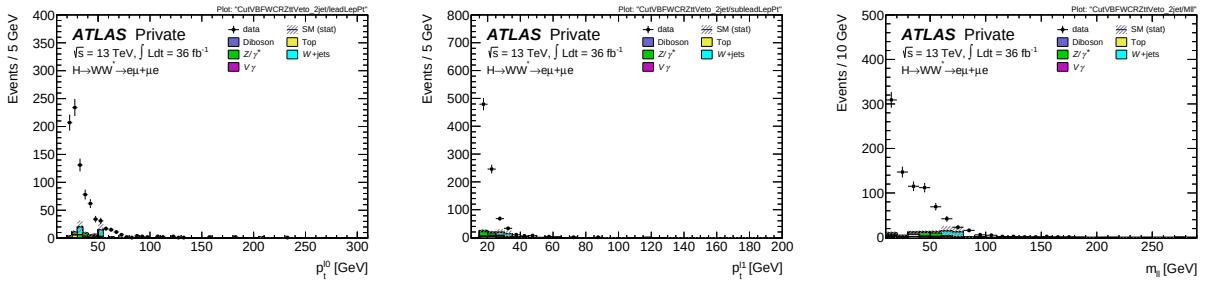
which is derived from a region with two anti-identified leptons where $N^{EWMC}_{\not{e},\not{\mu}}$ is the contamination of this region from electroweak processes containing at least one prompt lepton and FF^{QCD} is a negative number. In Table 6.20, the overestimation of the total misidentified background without the correction is provided.

In the ggF categories, the overestimation is no more than 7.3%. Considering that this is a small bias compared with the uncertainties on the overall estimation found in Table 6.16,

$\sqrt{s} = 13TeV$, $\mathcal{L} = 36fb^{-1}$	e W+jets	μ W+jets	QCDCorr	Overestimation (%)	e+μ W+jets	Misid (corrected)
0-jet SR with b-veto ($e\mu$)	129.9 ± 10.4	192.8 ± 6.9	-14.0 ± 2.3	4.5 ± 0.8	322.8 ± 12.5	308.8 ± 12.7
0-jet SR with b-veto (μe)	130.1 ± 6.5	48.0 ± 6.3	-4.7 ± 1.9	2.7 ± 1.1	178.0 ± 9.1	173.3 ± 9.3
1-jet SR (μe)	73.1 ± 6.0	35.2 ± 5.5	-6.0 ± 0.8	5.8 ± 0.9	108.4 ± 8.1	102.4 ± 8.2
1-jet SR ($e\mu$)	69.8 ± 8.2	96.1 ± 5.7	-11.3 ± 2.0	7.3 ± 1.4	165.9 ± 10.0	154.6 ± 10.2
VBF $Z \to \tau\tau$ veto ($e\mu$)	36.2 ± 6.3	51.3 ± 4.9	-18.6 ± 1.7	26.9 ± 4.1	87.4 ± 7.9	68.9 ± 8.1
VBF $Z \to \tau\tau$ veto (μe)	32.4 ± 5.3	18.9 ± 4.9	-7.9 ± 1.0	18.1 ± 3.8	51.3 ± 7.2	43.4 ± 7.3

Table 6.20: Misidentified background yields before and after applying the QCD correction. The "e W+jets" and "μ W+jets" columns correspond to events with Anti-ID e and μ respectively, with the overestimation being computed as $N_{\text{id+id}}^{\text{FF estimate}}/(N_{\text{id+id}}^{\text{FF estimate}}+N_{\text{id+id}}^{\text{QCDcorr}})$.

no QCD correction is applied. However in the VBF $N_{\text{jet}} \geq 2$ category, the overestimation is larger than 20% and so the QCD correction is applied. Distributions of select variables in the Anti-ID + Anti-ID control sample for the VBF signal region are shown in Figure 6.13. Analogous to the W+jets control region, it is extrapolated to the signal region by applying FF^{QCD} from Equation 6.8.



(a) $p_{\text{T}}^{\text{lead}}$ (b) $p_{\text{T}}^{\text{sublead}}$ (c) $m_{\ell\ell}$

Figure 6.13: Kinematic variable distributions in the Anti-ID + Anti-ID control sample for the VBF signal region before applying fake factor weights. The QCD correction for the VBF category is taken from this sample after EW components are subtracted from data and fake factor weights have been applied.

Chapter 7

Systematic Uncertainties

This chapter contains a general summary of the systematic uncertainties that are considered in the analysis. Uncertainties related to detector and reconstruction effects are described in section 7.1, while uncertainties related to theoretical predictions of the signal and main backgrounds are described in section 7.2.

7.1 Experimental Uncertainties

Systematic uncertainties related to detector and reconstruction effects (also known as experimental uncertainties) can be sub-divided into two distinct types:

- **four-vector (P4) systematics** - those that are applied as $\pm 1\sigma$ variations to the four-momentum of an object
- **scale factor (SF) systematics** - those that are applied as $\pm 1\sigma$ variations to the weight of the event or an individual particle

The complete set of experimental uncertainties that are considered in the analysis is listed in Table 7.1, along with their types. The electron-related systematics include uncertainties on their reconstruction, identification, isolation and trigger efficiencies, as well as on their associated four-momentum scale and resolution. The muon-related systematics are similar, except with an additional uncertainty on the efficiency of track-to-vertex association (TTVA) impact parameter cuts. The lepton uncertainties are derived from studying $J/\psi \to \ell^+\ell^-$, $W^\pm \to \ell^\pm\nu$ and $Z \to \ell^+\ell^-$ decays [28, 27, 34].

The uncertainties on the jet energy scale are derived as a function of p_{T} and η, containing terms that account also for pileup conditions as well as flavor composition of the jet. The uncertainty on the jet energy resolution on the other hand, is modeled with a single parameter. Both are derived based on *in-situ* studies of dijet, Z+jet, and γ+jet events [16]. A scale factor uncertainty is added to account for the JVT efficiency. Uncertainties related to jet flavor tagging are modeled by parameters that are the result of eigen-vector decomposition [21] as well as a couple of additional parameters for charm quark and Run1-to-Run2 extrapolation.

Uncertainties on the TST missing transverse energy $E_{\mathrm{T}}^{\mathrm{miss}}$ are measured in $Z \to \mu\mu$ events using the variable $\vec{p}_{\mathrm{T}}^{\,\mathrm{hard}}$ defined as the p_{T} sum of the hard terms to discriminate between soft term scale and resolution effects, taking into account detector material uncertainties [17]. Other experimental systematics include uncertainties on the pileup reweighting data scale factor introduced in subsection 3.4.2 and an uncertainty on the integrated luminosity, which has been estimated to be $\pm 2.1\%$ for the dataset using in this analysis [44].

Systematic uncertainty	Short description	syst. type
	Event	
Luminosity	uncertainty on total integrated luminosity	SF
Pileup Reweighting	uncertainty on pileup reweighting	SF
	Electrons	
EL_EFF_Reco_Total_1NPCOR_PLUS_UNCOR	reconstruction efficiency uncertainty	SF
EL_EFF_ID_CorrUncertaintyNP (0 to 14)	ID efficiency uncertainty split in 15 components	SF
EL_EFF_ID_SIMPLIFIED_UncorrUncertaintyNP (0 to 15)	ID efficiency uncertainty split in 16 components	SF
EL_EFF_Iso_Total_1NPCOR_PLUS_UNCOR	isolation efficiency uncertainty	SF
EL_EFF_Trigger_Total_1NPCOR_PLUS_UNCOR	trigger efficiency uncertainty	SF
EG_SCALE_ALLCORR		P4
EG_SCALE_E4SCINTILLATOR		P4
EG_SCALE_LARTEMPERATURE_EXTRA2015PRE	energy scale uncertainty	P4
EG_SCALE_LARTEMPERATURE_EXTRA2016PRE		P4
EG_SCALE_LARCALIB_EXTRA2015PRE		P4
EG_RESOLUTION_ALL	energy resolution uncertainty	P4
	Muons	
MUON_EFF_STAT	reconstruction and ID efficiency uncertainty for muons with $p_{\mathrm{T}} > 15$ GeV	SF
MUON_EFF_SYS		SF
MUON_ISO_STAT	isolation efficiency uncertainty	SF
MUON_ISO_SYS		SF
MUON_TTVA_STAT	track-to-vertex association efficiency uncertainty	SF
MUON_TTVA_SYS		SF
MUON_EFF_TrigStatUncertainty	trigger efficiency uncertainty	SF
MUON_EFF_TrigSystUncertainty		SF
MUON_SCALE	momentum scale uncertainty	P4
MUON_ID	momentum resolution uncertainty from inner detector	P4
MUON_MS	momentum resolution uncertainty from muon system	P4
	Jets	
JET_21NP_JET_EffectiveNP_1		P4
JET_21NP_JET_EffectiveNP_2		P4
JET_21NP_JET_EffectiveNP_3		P4
JET_21NP_JET_EffectiveNP_4	energy scale uncertainty from the in situ analyses split into 8 components	P4
JET_21NP_JET_EffectiveNP_5		P4
JET_21NP_JET_EffectiveNP_6		P4
JET_21NP_JET_EffectiveNP_7		P4
JET_21NP_JET_EffectiveNP_8restTerm		P4
JET_21NP_JET_EtaIntercalibration_Modeling	energy scale uncertainty on eta-intercalibration (modeling)	P4
JET_21NP_JET_EtaIntercalibration_TotalStat	energy scale uncertainty on eta-intercalibrations (statistics/method)	P4
JET_21NP_JET_EtaIntercalibration_NonClosure	energy scale uncertainty on eta-intercalibrations (non-closure)	P4
JET_21NP_JET_Pileup_OffsetMu	energy scale uncertainty on pileup (mu dependent)	P4
JET_21NP_JET_Pileup_OffsetNPV	energy scale uncertainty on pileup (NPV dependent)	P4
JET_21NP_JET_Pileup_PtTerm	energy scale uncertainty on pileup (pt term)	P4
JET_21NP_JET_Pileup_RhoTopology	energy scale uncertainty on pileup (density ρ)	P4
JET_21NP_JET_Flavor_Composition	energy scale uncertainty on flavor composition	P4
JET_21NP_JET_Flavor_Response	energy scale uncertainty on sample flavor response	P4
JET_21NP_JET_BJES_Response	energy scale uncertainty on b-jets	P4
JET_21NP_JET_PunchThrough_MC15	energy scale uncertainty for punch-through jets	P4
JET_21NP_JET_SingleParticle_HighPt	energy scale uncertainty from the behaviour of high-p_{T} jets	P4
JET_JER_SINGLE_NP	energy resolution uncertainty	P4
JET_JvtEfficiency	JVT efficiency uncertainty	SF
FT_EFF_Eigen_B	b-tagging efficiency uncertainties ("BTAG_MEDIUM"): 3	SF
FT_EFF_Eigen_C	components for b jets, 4 for c jets and 5 for light jets	SF
FT_EFF_Eigen_L		SF
FT_EFF_Eigen_extrapolation	b-tagging efficiency uncertainty on the extrapolation to high-p_{T} jets	SF
FT_EFF_Eigen_extrapolation_from_charm	b-tagging efficiency uncertainty on tau jets	SF
	MET	
MET_SoftTrk_ResoPara	track-based soft term related longitudinal resolution uncertainty	P4
MET_SoftTrk_ResoPerp	track-based soft term related transverse resolution uncertainty	P4
MET_SoftTrk_Scale	track-based soft term related longitudinal scale uncertainty	P4

Table 7.1: Summary of the experimental systematic uncertainties considered in the analysis. The last column indicates whether they are applied as a scale factor (SF) systematic or a four-vector (P4) systematic.

7.2 Theory Uncertainties

Theoretical uncertainties are applied based on how the sample in question is normalized. For the signal and the background processes normalized directly from theory predictions, uncertainties on the absolute expected yields in each signal and control region are taken into account. In the case of signal, these uncertainties also encompass migrations between the signal regions. For the background processes that are normalized using dedicated control regions, the theory uncertainties are instead derived based on the resulting variations of the extrapolation from the control to signal regions.

The theory uncertainties considered for each process include QCD scale variations, modeling uncertainties of the parton shower and underlying event (PS/UE), as well as variations of the PDF set. In some cases, additional process specific theory uncertainties are also included. In other cases, theory uncertainties are excluded if the systematic variation is smaller than the statistical uncertainty of the Monte Carlo sample. An overview of the theory uncertainties included in the analysis is provided in Table 7.2.

For the ggF Higgs production process, cross sections are calculated separately for the $N_{\text{jet}} = 0$ and $N_{\text{jet}} = 1$ categories. QCD scale uncertainties are evaluated by independently varying the the renormalization and factorization scales by 2.0 and 0.5 relative to the nominal value, while PDF uncertainties are evaluated using an envelope of the 68% confidence level PDF4LHC Hessian PDF eigenvectors added in quadrature as well as the differences compared with the CT10, MMHT14 and NNPDF3.0 PDF sets. An uncertainty is assigned on the generator and matching scheme by comparing the NLO matching of Powheg NNLOPS + Pythia8 with the NLO matching of MG5_aMC+Pythia8 up to $H+2$ jet production. Uncertainties on the modeling of the parton shower are also evaluated by comparing the nominal Powheg + Pythia8 with Powheg + Herwig7. This is performed using truth level information due to limited availability of the latter sample, so a folding matrix is applied in order to correct the acceptance for resolution effects that appear only for fully reconstructed events. In addition, the perturbative uncertainty on the ggF Higgs production process in the VBF $N_{\text{jet}} \geq 2$ category is estimated using the Stewart-Tackmann (ST) method [108] with MG5_aMC+Pythia8 inclusive 2-jet and inclusive 3-jet cross sections ($\sigma_{\geq 2}$ and $\sigma_{\geq 3}$), consisting of both a normalization uncertainty on the VBF signal region and a shape uncertainty on the BDT distribution. For the VBF Higgs production process, QCD scale, PDF, PS/UE, and matching uncertainties are similarly derived.

For the WW background, theory uncertainties are applied differently in the ggF and VBF categories since only dedicated control regions exist for the former. In the case of the $N_{\text{jet}} = 0$ and $N_{\text{jet}} = 1$ categories, they are applied to an extrapolation factor $\alpha_{i,j-\text{jet}}$ which relates the number of WW events in the j-jet WW control region to the number of WW events in the signal region category i in jet-bin j through $N^{WW}_{\text{SR},i,j-\text{jet}} = \alpha_{i,j-\text{jet}} \cdot N^{WW}_{\text{CR},j-\text{jet}}$. In the case of the $N_{\text{jet}} \geq 2$ category, they are applied directly to the nominal SHERPA prediction for normalization and BDT shape in the VBF signal region. Due to limited Monte Carlo statistics, the WW theory uncertainties are derived at truth level and the different flavor channels $e\mu$ and μe are combined. The QCD scale and PDF uncertainties are evaluated similarly to the strategy used for the signal processes. For PS/UE uncertainties, a comparison is made between Powheg+PYTHIA 8 and Powheg+Herwig++, while a matrix element matching scale (CKKW) uncertainty is applied to account for the choice of SHERPA showering parameters. Also, uncertainties on additional electroweak corrections as well as on the normalization of the $gg \to WW$ process are considered, which have been

extrapolated from a previous publication [43].

For the top background, theory uncertainties are applied to the extrapolation factor between control and signal regions. The QCD scale and PDF uncertainties are evaluated similarly to the strategy used for the signal processes. For PS/UE uncertainties, POWHEG+PYTHIA 8 is compared with POWHEG+HERWIG7 for $t\bar{t}$ and POWHEG+PYTHIA 6 is compared with POWHEG+HERWIG++ for Wt. For a generator matching uncertainty, POWHEG+HERWIG++ is compared with MG5_aMC+HERWIG++. Variations in shower radiation are also evaluated using POWHEG+PYTHIA 8. An uncertainty on the treatment of the interference between $t\bar{t}$ and Wt is derived by comparing samples with different overlap subtraction schemes [70]. Due to limited Monte Carlo statistics, the $N_{\text{jet}} = 0$ and $N_{\text{jet}} = 1$ signal regions are combined for evaluating each of the top theory uncertainties.

For the $Z \to \tau\tau$ background, an uncertainty on the generator modeling is evaluated on the extrapolation factor between control and signal regions by comparing the nominal SHERPA prediction to the alternative MG5_aMC+PYTHIA 8.

For the Non-WW diboson backgrounds, theory uncertainties are applied on the absolute normalization from Monte Carlo prediction. The QCD renormalization and factorization scales are varied for both WZ/γ^* and $W\gamma$, while an uncertainty in the merging scale for WZ/γ^* is determined by varying the choice of SHERPA shower parameters and a generator uncertainty is applied to $W\gamma$ by comparing the predictions of SHERPA v2.2 with those of MG5_aMC@NLO v2.3.3.

process	uncertainty	source	included
ggF	jet veto uncertainty		Yes
	matching		Yes
	PS/UE		Yes
	PDF		Yes
	QCD scale		Yes
VBF	matching		Yes
	PS/UE		Yes
	PDF		Yes
	QCD scale		Yes
WW	ckkw matching		Yes
	PS/UE (for VBF full generator)		Yes
	PDF		Yes
	QCD scale		Yes
	EW correction	[43]	Yes
	$gg \to WW$ fraction	[43]	Yes
top	radiation		Yes
	PS/UE		Yes
	matching		Yes
	Wt diagram removal		Yes
	single top cross-section	[43]	No
Z/DY	alternative generators		Yes
$WZ/W\gamma^*$	QCD and merging scales		Yes
$W\gamma$	QCD scale		Yes
$W\gamma$	generator/matching		Yes

Table 7.2: Summary of the theoretical systematic uncertainties considered in the analysis. The "source" column indicates where the uncertainty is taken from, with an empty entry signifying that the uncertainty was rederived for this analysis. The "included" column indicates whether or not the uncertainty is used in the final likelihood fit.

Chapter 8

Statistical Treatment

This chapter demonstrates how the final analysis results are obtained in a statistically robust way. The general methodology is first described in section 8.1, while the specific details for the ggF and VBF production modes are provided in section 8.2 and section 8.3, respectively.

8.1 The Profile Likelihood Method

When measuring the coupling of the Higgs boson, it is common to frame the task in terms of *hypothesis testing* in which two hypotheses are compared with the observed data. The *null hypothesis*, H_0, typically refers to the hypothesis that the prediction of the current Standard Model is correct and that any deviation in the observed data is the result of chance alone. The *alternative hypothesis*, H_1, on the other hand often refers to the hypothesis that some physical effect which is not accounted for by the current Standard Model is responsible for producing the observed deviations in data.

In order to quantify the agreement between the observed data and either of the hypotheses, a test statistic t is used. Due to the fact that the hypotheses in this case are parametric, the test statistic is constructed using *likelihood functions*. A likelihood function

$$\mathcal{L}(\mu, \vec{\theta}|x)$$

is a function of the parameters of the underlying model which expresses the likelihood that a particular set of parameter values correspond to the true values of the model, given an observed dataset x that has been sampled from that model. Here, a conceptual distinction is made between two types of parameters. A *parameter of interest* (denoted as μ) is a parameter which is typically unconstrained and is either not present or fixed to a particular value in H_0. On the other hand, a *nuisance parameter* (denoted as θ) is a parameter which is not directly being measured but nevertheless must be modeled and is typically constrained within certain bounds using prior knowledge. The likelihood function can be computed for a particular set of parameters by evaluating the corresponding probability distribution function (pdf) of the model using the given observed dataset.

In this way, the most likely parameters that describe the underlying model given the observed dataset can also be viewed as those which maximize the likelihood function. Also called maximum likelihood estimators (MLEs), they are obtained through differentiation of the likelihood function in a process also known as a *likelihood fit*. Furthermore, the *likelihood ratio* $\Lambda(\mu)$ is defined as

$$\Lambda(\mu) = \frac{\mathcal{L}(\mu, \hat{\hat{\vec{\theta}}}(\mu))}{\mathcal{L}(\hat{\mu}, \hat{\hat{\vec{\theta}}})} \tag{8.1}$$

where $\hat{\mu}$ and $\hat{\hat{\vec{\theta}}}$ denote the unconditional MLEs which maximize the likelihood function in absolute terms, while $\hat{\hat{\vec{\theta}}}(\mu)$ denotes the conditional MLEs of $\vec{\theta}$ that maximize the likelihood for a particular value of μ. It is common to refer to the numerator of the likelihood ratio as the *profile likelihood* since it 'profiles' only a slice of likelihood surface for a given μ after having replaced the nuisance parameters by their conditional MLEs. Most often, the likelihood maximization must be solved numerically rather than analytically. In order to facilitate the calculation, the *negative log likelihood ratio*

$$t_\mu = -\log \Lambda(\mu) \tag{8.2}$$

is taken as the final test statistic. Using this definition, the test statistic will result in values most often close to 0[1] with only a single parameter of interest if the difference between numerator and denominator of the likelihood ratio is due only to sampling error. In fact, this test statistic approaches a χ^2 distribution with n degrees of freedom where n is the number of parameters of interest in the asymptotic limit.

For the analysis presented in this thesis, the parameters of interest are the signal strengths for the ggF and VBF Higgs production modes. The statistical significance Z which is the number of standard deviations the measurement is away from $\mu = 0$ can be calculated from the corresponding test statistic t_0 [74] and is quoted as part of each fit result. Since the target N_{jet} categories are orthogonal to one another, two separate fits described in the rest of this chapter are performed to extract both parameters independently. However, the results of a combined fit are also provided in chapter 9.

8.2 ggF Statistical Treatment

After the signal selection in the $N_{\text{jet}} = 0$ and $N_{\text{jet}} = 1$ categories described in subsection 4.5.2, events are further split into $e\mu$ and μe channels in addition to two bins each of the variables $p_{\text{T}}^{\text{sublead}}$ and $m_{\ell\ell}$ for a total of 16 final signal regions. Table 8.1 provides an overview of the definitions for each signal region category.

N_j	$m_{\ell\ell}$	$p_{\text{T}}^{\text{sublead}}$	$e\mu/\mu e$-channel
$N_{\text{jet}} = 0$	[10-30, 30-55]	[15-20, 20-∞]	$[e\mu, \mu e]$
$N_{\text{jet}} = 1$	[10-30, 30-55]	[15-20, 20-∞]	$[e\mu, \mu e]$

Table 8.1: Signal region categories in the ggF analysis, with bin boundaries for $m_{\ell\ell}$ and $p_{\text{T}}^{\text{sublead}}$ given in GeV. In total there are 16 categories, 8 for each jet bin.

In addition to the signal regions, 6 control regions for WW, top, and $Z \to \tau\tau$ ($N_{\text{jet}} = 0$ and $N_{\text{jet}} = 1$ categories each) are included in the fit that are described in more detail in section 5.2 and section 5.3. For each control region, only the total event yields are considered.

[1] corresponding to the likelihood ratio most often being close to 1

For the signal regions, m_{T} is used as a discriminating variable, with the original histograms having 50 uniform bins in the range 80 GeV $< m_{\mathrm{T}} <$ 130 GeV as well as overflow and underflow bins covering the ranges $m_{\mathrm{T}} <$ 80 GeV and $m_{\mathrm{T}} >$ 130 GeV. The bins are then combined using a heuristic tree-search algorithm until there are 8 (6) bins left for the $N_{\mathrm{jet}} = 0$ ($N_{\mathrm{jet}} = 1$) category, each containing as similar a signal yield as possible. The particular choice of 6 bins for the $N_{\mathrm{jet}} = 1$ category is to ensure that each bin is sufficiently populated so as to reduce the chance of the likelihood fit becoming unstable. The final signal region bin boundaries are displayed in Table 8.2, while an overview of all the regions considered in the fit with pre-fit background yields is shown in Figure 8.1.

Region	Bin Boundaries								
SR_0j_DF_Mll1_PtSubLead2_e	0	84.0	92.0	98.0	103.0	108.0	114.0	121.0	inf
SR_0j_DF_Mll1_PtSubLead2_m	0	82.0	90.0	96.0	102.0	108.0	115.0	124.0	inf
SR_0j_DF_Mll1_PtSubLead3_e	0	93.0	101.0	107.0	113.0	118.0	124.0	130.0	inf
SR_0j_DF_Mll1_PtSubLead3_m	0	93.0	101.0	107.0	112.0	117.0	123.0	130.0	inf
SR_0j_DF_Mll2_PtSubLead2_e	0	87.0	95.0	101.0	106.0	111.0	117.0	124.0	inf
SR_0j_DF_Mll2_PtSubLead2_m	0	86.0	94.0	100.0	106.0	111.0	117.0	125.0	inf
SR_0j_DF_Mll2_PtSubLead3_e	0	94.0	102.0	108.0	113.0	118.0	124.0	130.0	inf
SR_0j_DF_Mll2_PtSubLead3_m	0	93.0	101.0	107.0	112.0	117.0	123.0	130.0	inf
SR_1j_DF_Mll1_PtSubLead2_e	0	80.0	91.0	101.0	111.0	123.0	inf		
SR_1j_DF_Mll1_PtSubLead2_m	0	80.0	91.0	101.0	110.0	121.0	inf		
SR_1j_DF_Mll1_PtSubLead3_e	0	85.0	96.0	107.0	116.0	129.0	inf		
SR_1j_DF_Mll1_PtSubLead3_m	0	85.0	97.0	107.0	117.0	130.0	inf		
SR_1j_DF_Mll2_PtSubLead2_e	0	86.0	97.0	106.0	115.0	125.0	inf		
SR_1j_DF_Mll2_PtSubLead2_m	0	88.0	98.0	106.0	115.0	125.0	inf		
SR_1j_DF_Mll2_PtSubLead3_e	0	89.0	99.0	108.0	118.0	130.0	inf		
SR_1j_DF_Mll2_PtSubLead3_m	0	89.0	100.0	109.0	119.0	130.0	inf		

Table 8.2: Boundaries of the signal region bins in GeV after the heuristic tree-search algorithm remapping procedure. The convention for the signal region names is as follows: SR_0j_DF (SR_1j_DF) signifies a $N_{\mathrm{jet}} = 0$ ($N_{\mathrm{jet}} = 1$) different flavor signal region category, $Mll1$ ($Mll2$) denotes the regions with $10 < m_{\ell\ell} < 30$ GeV ($30 < m_{\ell\ell} < 55$ GeV), while PtSubLead2 (PtSubLead3) denotes the regions with $15 < p_{\mathrm{T}}^{\mathrm{sublead}} < 20$ GeV ($p_{\mathrm{T}}^{\mathrm{sublead}} > 20$ GeV) and the final suffix e (μ) denotes that the subleading lepton is an electron (muon).

The likelihood function for the fit is constructed primarily as the product of Poisson terms, one for each bin in all signal regions and one for each control region. Additionally, systematic uncertainties are taken into account using $\pm 1\sigma$ variations when possible and are modeled as nuisance parameters, each with a dedicated constraint term. The full likelihood can be written as

$$\mathcal{L}\left(\mu, \vec{\theta}\right) = \{\prod_{l=e,\mu}^{\ell_2} \prod_{p=1,2}^{p_{\mathrm{T}}} \prod_{m=1,2}^{m_{\ell\ell}} \prod_{j=0,1}^{N_{\mathrm{jets}}} \prod_{i=1}^{N_{m_{\mathrm{T}}\mathrm{bins}}} P(N_{lpmji}|\mu \cdot s_{lpmji}(\vec{\theta}) + \sum_{\Phi} \beta_{\Phi} b^{\Phi}_{lpmji}(\vec{\theta}))\} \times \{\prod_{c=1}^{N_{\mathrm{CR}}} P(N_c|\mu \cdot s_c(\vec{\theta}) + \sum_{\Phi} \beta_{\Phi} b^{\Phi}_{c}(\vec{\theta}))\} \times \{\prod_{\theta \in \vec{\theta}} C(\theta)\} \tag{8.3}$$

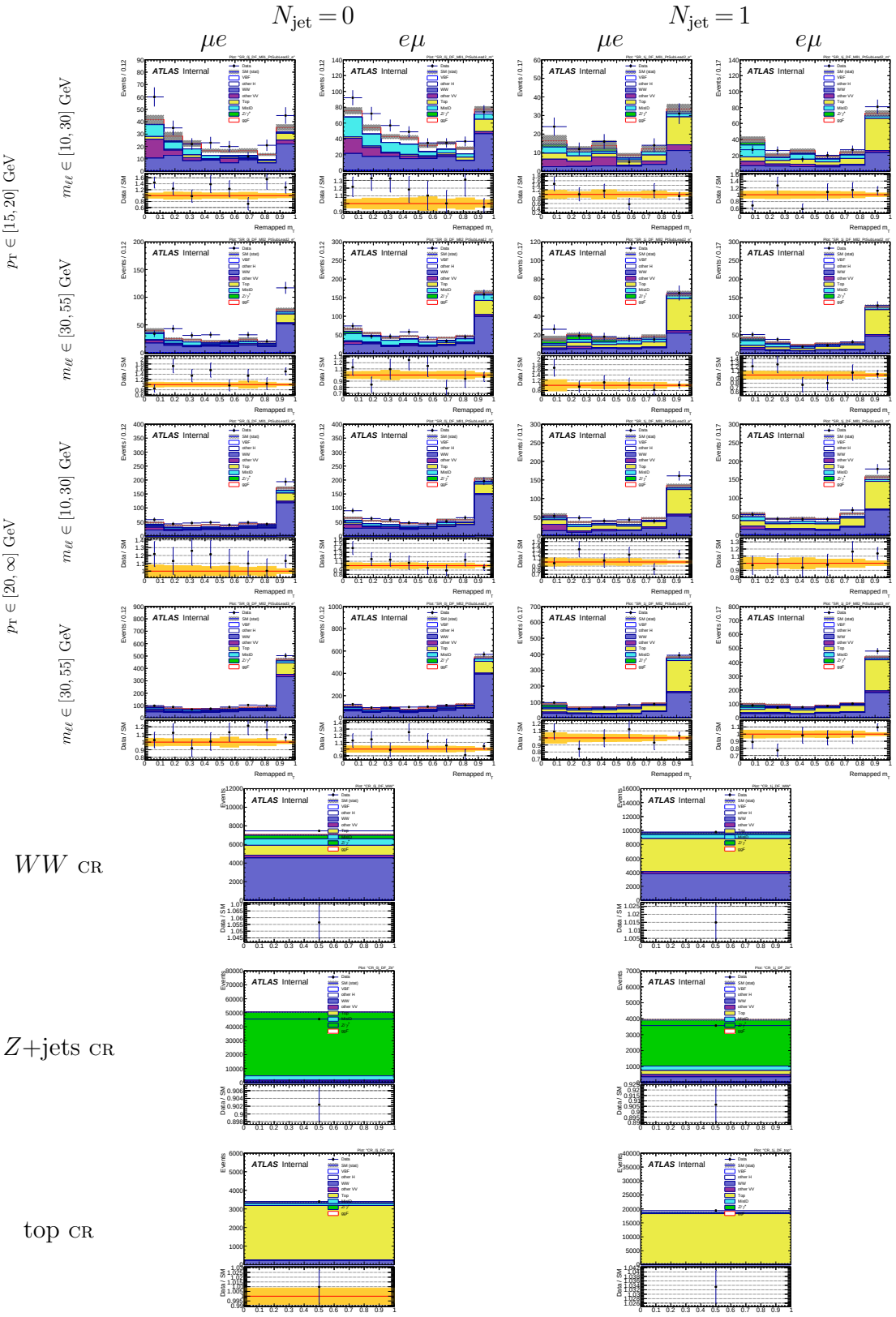


Figure 8.1: Overview of all regions included in the ggF fit. In each bin, the data is compared with the total pre-fit expected background yields. The uncertainty bands are statistical only.

where s and b denote the expected signal and background yields in a particular bin, while μ is the signal strength parameter of interest and Φ denotes a particular background process with normalization factor β_Φ. Only the normalization factors for the WW, top, and $Z \to \tau\tau$ processes are allowed to float, however, with the rest being fixed to unity. In addition, $P(x|y)$ represents the Poisson probability to observe x events when y are expected and $C(\theta)$ represents a constraint term for the nuisance parameter θ.

The signal and background yields are allowed to vary as functions of the nuisance parameters $\vec{\theta}$. Specifically, they are parameterized as $s = s_0 \times \prod \nu(\theta)$, in which the response to each nuisance parameter is factorized from the nominally expected rate. Four general uncertainty sources are distinguished, each with a different form for $\nu(\theta)$: a flat rate source, a shape source, a statistical source, and a normalization source. Flat rate sources take the form $\nu_{\text{flat}(\theta)} = \kappa^\theta$ with κ being determined by measuring ν_{flat} at $\theta = \pm 1$. In this case, the constraint term $C(\theta)$ is a unit gaussian. Systematics which can affect both rate and shape are first split into a separate flat component and a pure shape component in such a way that varying the pure shape component has no effect on the overall expected rate. The pure shape component uses vertical linear interpolation to estimate the variation, with its constraint term being a truncated gaussian. The statistical uncertainties are modeled with a Poisson constraint $P(\tilde{\theta}|\theta\lambda)$, which represents an auxiliary measured number of events $\tilde{\theta}$ with an expected number $\theta\lambda$. The constraint term for a normalization source is analogous to the one for a statistical source, except that the observed and expected number of events in the relevant control region are instead used.

The two types of experimental systematics described in section 7.1 can affect the fit in fundamentally different ways. While efficiency scale factor systematics only change the weight of each event and are therefore fully correlated to the nominal yields, the four-momentum systematics can cause events to migrate into or out of a signal or control region[2]. For Monte Carlo events with low statistics, this can also lead to artificially large and unphysical variations. In order to improve the fit stability and reduce convergence time, a procedure referred to as "pruning" is implemented in which certain uncertainties found to be sufficiently small are neglected. First, rate uncertainties are removed from the fit for a particular sample in a region if its variation is found to be less than 0.5%. Afterwards, any shape uncertainty for which no single bin in a region is over 0.5% is also removed from the fit.

A likelihood fit is first performed using an Asimov dataset that is generated by setting $\mu = 1$ in order to measure the expected ggF signal significance and the expected uncertainties on the ggF signal strength. The expected ggF signal significance is found to be $Z = 5.3$, while the expected uncertainties on μ are shown in Table 8.3 along with a breakdown of the uncertainties shown in Figure 8.2. A likelihood fit is then performed using the full observed dataset. The ranking of nuisance parameter impact on the signal strength is given in Figure 8.3, while correlations between the nuisance parameters are displayed in Figure 8.4. The final post-fit event yields for each signal region bin and control region are shown in Table 8.4.

[2]For example in an event with nominally one jet, a jet energy scale uncertainty might fluctuate its four-momentum such that it becomes subthreshold and the event migrates from an $N_{\text{jet}} = 1$ category into an $N_{\text{jet}} = 0$ category.

Source	$\Delta\mu/\mu$ [%]
Data statistics	± 9.7
MC statistics	± 6.3
Theoretical uncertainties	± 12.0
ggF signal	± 9.1
WW	± 6.5
Top-quark	± 4.6
Other VV	± 2.9
Fake factor uncertainties	± 6.8
Electron sample composition	± 0.9
Electron EW subtraction	± 0.9
Muon sample composition	± 3.7
Muon EW subtraction	± 3.7
Experimental uncertainties	± 11.2
B-Tagging	± 6.7
Pile-up	± 6.1
JER	± 3.5
Electron eff.	± 3.0
Background normalization	± 9.4
Z+jets $N_{\text{jet}} = 0$	± 5.4
top $N_{\text{jet}} = 0$	± 5.9
TOTAL systematics	± 19.8
TOTAL	$+22.0/-21.0$

Table 8.3: Summary of the expected uncertainties on the ggF signal strength μ. Only the most important sources are listed for each category.

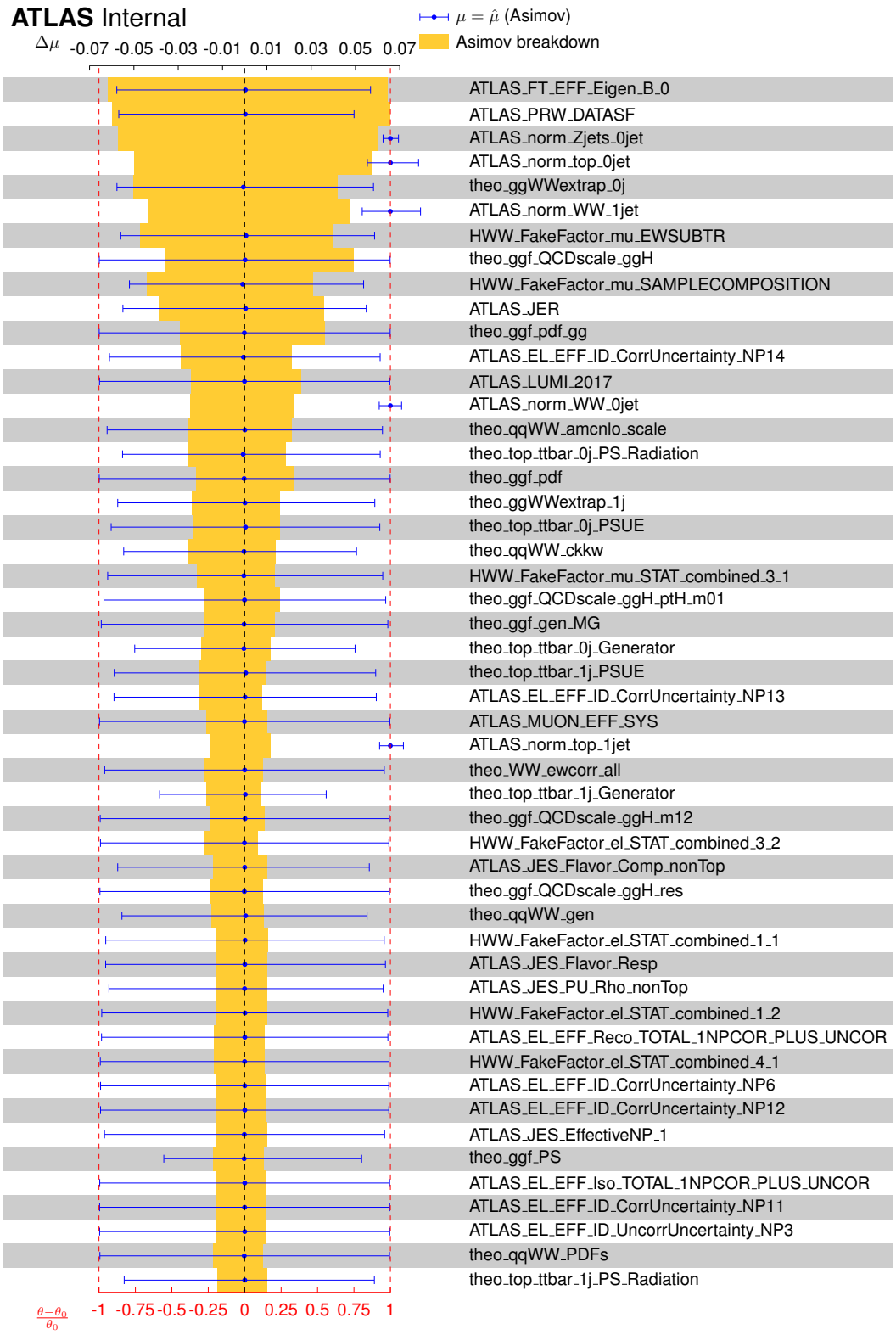


Figure 8.2: The ranking distribution of the nuisance parameters participating in the ggF fit to an Asimov dataset. Their pull and post-fit uncertainties are indicated by the blue dot and associated error bar, respectively, while the yellow bands represent their contribution to the total uncertainty in the analysis and is computed as the quadratic difference between the uncertainty on μ in the main fit with all nuisance parameters and a fit for which the nuisance parameter in question has been fixed to its best-fit value.

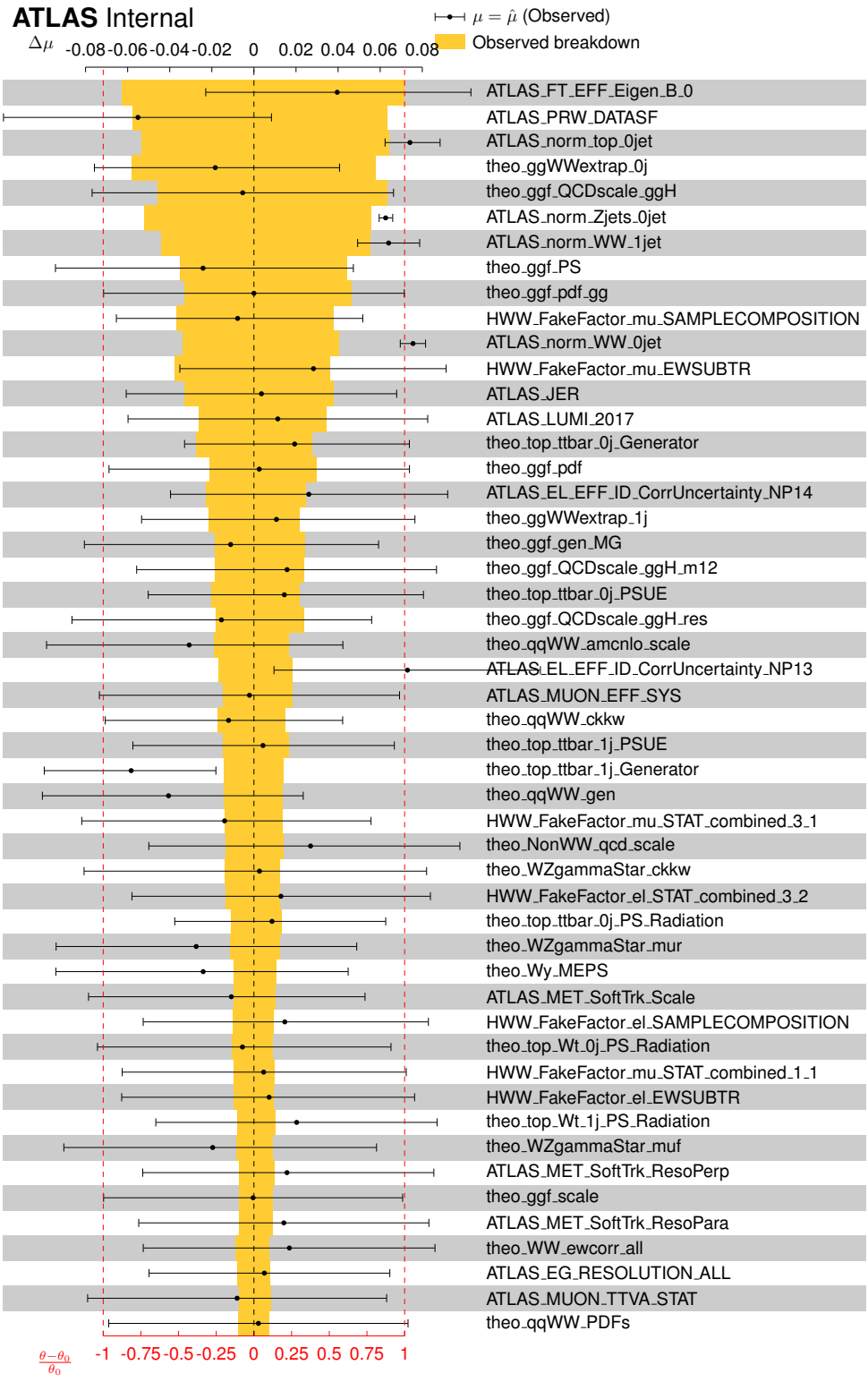


Figure 8.3: The ranking distribution of the nuisance parameters participating in the full ggF fit to the observed data. Their pull and post-fit uncertainties are indicated by the black dot and associated error bar, respectively, while the yellow bands represent their contribution to the total uncertainty in the analysis and is computed as the quadratic difference between the uncertainty on μ in the main fit with all nuisance parameters and a fit for which the nuisance parameter in question has been fixed to its best-fit value.

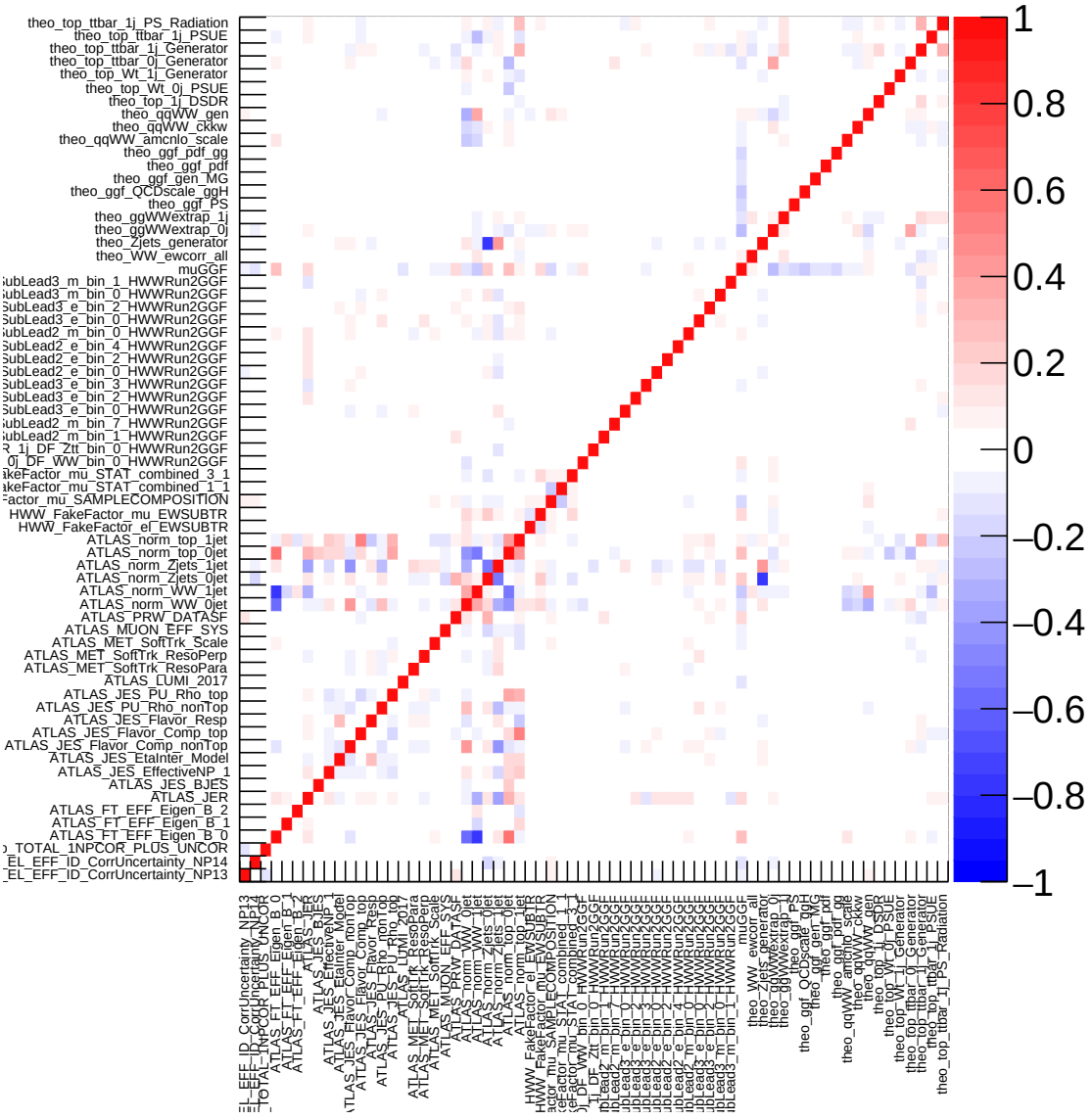


Figure 8.4: Post-fit correlations of the nuisance parameters participating in the full ggF fit to the observed data. Only nuisance parameters with correlations greater than 10% are shown for visibility.

	Data	$V\gamma$	Fakes	Z/γ^*	$gg \to WW$	other H	Other $qq \to VV$	$qq \to WW$	ggF [125 GeV]	VBF [125 GeV]	$t\bar{t}$	Single Top	Total
$N_{\text{jet}} = 0$													
WW CR	7461	86.28 ± 12.36	622.58 ± 130.33	361.40 ± 51.73	521.93 ± 36.60	4.64 ± 0.44	224.95 ± 30.27	4437.01 ± 273.41	112.28 ± 18.75	1.08 ± 0.12	752.27 ± 171.68	429.22 ± 97.92	7328.67 ± 82.69
$Z \to \tau\tau$ CR	45463	765.83 ± 72.17	2467.58 ± 637.68	41330.27 ± 677.77	40.15 ± 2.88	102.17 ± 4.03	175.25 ± 24.41	860.67 ± 55.09	56.43 ± 9.39	0.53 ± 0.06	42.82 ± 12.01	27.24 ± 7.38	45693.68 ± 221.03
Top CR	3399	8.80 ± 3.43	93.76 ± 22.68	48.27 ± 16.51	39.63 ± 7.33	0.31 ± 0.06	24.21 ± 5.98	207.50 ± 38.35	22.76 ± 4.63	0.44 ± 0.08	2303.96 ± 130.63	652.53 ± 102.08	3377.97 ± 59.76
SR	5089	114.16 ± 16.01	467.70 ± 89.95	26.76 ± 8.86	425.39 ± 99.08	0.01 ± 0.00	204.51 ± 36.17	2610.67 ± 246.05	679.98 ± 106.48	6.81 ± 0.81	347.07 ± 112.99	234.50 ± 59.17	4913.06 ± 203.79
$m_{\ell\ell} < 30$GeV, $p_{\text{T}}^{\text{sublead}} < 20$GeV: μe	238	22.37 ± 3.14	27.59 ± 5.31	0.92 ± 0.31	11.43 ± 2.66	0.00	14.79 ± 2.62	90.47 ± 8.53	44.53 ± 6.97	0.43 ± 0.05	9.35 ± 3.05	6.31 ± 1.59	213.42 ± 8.85
$m_{\ell\ell} < 30$GeV, $p_{\text{T}}^{\text{sublead}} < 20$GeV: $e\mu$	451	16.48 ± 2.16	84.51 ± 12.12	-0.35 ± 0.11	20.85 ± 4.78	0.00	35.76 ± 5.37	162.12 ± 12.93	80.55 ± 12.79	0.70 ± 0.09	20.53 ± 5.74	12.53 ± 3.21	397.91 ± 12.21
$m_{\ell\ell} < 30$GeV, $p_{\text{T}}^{\text{sublead}} \geq 20$GeV: μe	518	9.65 ± 1.40	31.69 ± 8.09	1.00 ± 0.31	51.18 ± 11.57	0.00	21.26 ± 3.32	253.64 ± 17.57	82.54 ± 13.27	0.91 ± 0.11	33.16 ± 9.60	23.31 ± 5.83	487.09 ± 12.92
$m_{\ell\ell} < 30$GeV, $p_{\text{T}}^{\text{sublead}} \geq 20$GeV: $e\mu$	612	16.25 ± 3.49	42.58 ± 9.79	0.46 ± 0.14	59.92 ± 13.54	0.00	23.92 ± 4.89	315.63 ± 22.68	107.25 ± 17.75	1.09 ± 0.16	34.93 ± 9.77	31.02 ± 7.31	609.12 ± 14.78
$m_{\ell\ell} \geq 30$GeV, $p_{\text{T}}^{\text{sublead}} < 20$GeV: μe	329	6.30 ± 1.42	38.58 ± 7.69	0.02 ± 0.01	17.98 ± 4.16	0.01 ± 0.00	9.63 ± 1.73	145.54 ± 12.01	43.25 ± 6.94	0.35 ± 0.04	20.26 ± 6.29	11.16 ± 2.69	283.45 ± 10.48
$m_{\ell\ell} \geq 30$GeV, $p_{\text{T}}^{\text{sublead}} < 20$GeV: $e\mu$	503	2.94 ± 0.82	90.05 ± 14.04	9.31 ± 4.13	30.85 ± 7.03	0.01 ± 0.00	24.69 ± 4.00	241.10 ± 17.94	74.95 ± 12.52	0.64 ± 0.06	36.86 ± 9.45	23.93 ± 5.59	510.65 ± 14.11
$m_{\ell\ell} \geq 30$GeV, $p_{\text{T}}^{\text{sublead}} \geq 20$GeV: μe	1130	29.49 ± 3.25	71.34 ± 18.37	9.45 ± 3.23	107.88 ± 24.32	0.00 ± 0.00	37.57 ± 5.89	638.85 ± 41.68	109.07 ± 17.54	1.18 ± 0.16	89.69 ± 23.63	57.28 ± 13.75	1114.23 ± 21.50
$m_{\ell\ell} \geq 30$GeV, $p_{\text{T}}^{\text{sublead}} \geq 20$GeV: $e\mu$	1308	10.74 ± 1.71	81.45 ± 19.47	5.53 ± 2.47	125.30 ± 28.28	0.00 ± 0.00	36.88 ± 6.42	763.32 ± 50.40	137.82 ± 22.48	1.50 ± 0.16	102.28 ± 27.59	68.97 ± 16.53	1296.91 ± 23.26
$N_{\text{jet}} = 1$													
WW CR	9784	56.88 ± 10.00	501.49 ± 124.05	206.85 ± 51.28	266.28 ± 59.99	5.04 ± 0.23	285.92 ± 58.74	3314.64 ± 680.17	1.86 ± 0.34	0.19 ± 0.02	3972.37 ± 611.32	1200.67 ± 168.47	9526.27 ± 98.30
$Z \to \tau\tau$ CR	3571	85.96 ± 17.27	233.19 ± 59.85	2565.76 ± 88.66	14.95 ± 3.47	35.75 ± 1.41	56.46 ± 12.40	277.94 ± 58.91	28.87 ± 5.16	3.54 ± 0.19	231.85 ± 31.21	69.86 ± 10.77	3547.68 ± 61.94
Top CR	19428	6.95 ± 1.95	344.32 ± 87.18	77.37 ± 14.12	22.21 ± 5.79	0.45 ± 0.06	24.57 ± 5.70	194.52 ± 52.65	23.97 ± 4.71	2.14 ± 0.30	14710.54 ± 308.09	4029.01 ± 264.05	19411.48 ± 144.89
SR	3264	69.95 ± 13.14	245.67 ± 48.57	84.21 ± 42.50	141.61 ± 55.80	0.12 ± 0.01	130.61 ± 29.23	901.83 ± 194.66	303.16 ± 57.89	29.98 ± 1.81	1047.40 ± 170.53	351.65 ± 59.49	3175.58 ± 183.06
$m_{\ell\ell} < 30$GeV, $p_{\text{T}}^{\text{sublead}} < 20$GeV: μe	102	12.34 ± 2.32	12.15 ± 2.40	3.06 ± 1.54	3.29 ± 1.30	0.00	6.44 ± 1.44	20.59 ± 4.45	12.15 ± 2.32	0.96 ± 0.06	27.58 ± 4.49	7.21 ± 1.22	99.34 ± 5.73
$m_{\ell\ell} < 30$GeV, $p_{\text{T}}^{\text{sublead}} < 20$GeV: $e\mu$	196	4.09 ± 1.05	32.36 ± 4.59	3.73 ± 1.69	6.60 ± 2.58	0.00 ± 0.00	12.76 ± 2.95	42.47 ± 9.42	23.25 ± 4.12	1.99 ± 0.14	57.69 ± 8.15	18.30 ± 3.44	190.49 ± 7.46
$m_{\ell\ell} < 30$GeV, $p_{\text{T}}^{\text{sublead}} \geq 20$GeV: μe	384	18.52 ± 3.46	21.07 ± 6.30	1.52 ± 0.87	16.74 ± 6.48	0.00	18.11 ± 4.07	91.63 ± 19.10	47.96 ± 8.25	5.00 ± 0.33	107.45 ± 14.79	38.25 ± 5.89	348.14 ± 10.80
$m_{\ell\ell} < 30$GeV, $p_{\text{T}}^{\text{sublead}} \geq 20$GeV: $e\mu$	433	8.59 ± 4.09	28.51 ± 8.98	0.71 ± 0.22	19.73 ± 7.62	0.00	16.09 ± 3.55	114.01 ± 23.43	56.15 ± 9.63	5.67 ± 0.36	125.00 ± 18.79	48.87 ± 6.94	407.23 ± 12.30
$m_{\ell\ell} \geq 30$GeV, $p_{\text{T}}^{\text{sublead}} < 20$GeV: μe	159	0.66 ± 0.34	19.62 ± 3.77	12.74 ± 3.66	5.47 ± 2.15	0.00 ± 0.00	9.62 ± 2.43	40.79 ± 8.83	12.09 ± 2.29	1.04 ± 0.07	52.37 ± 7.66	12.33 ± 2.46	157.11 ± 7.51
$m_{\ell\ell} \geq 30$GeV, $p_{\text{T}}^{\text{sublead}} < 20$GeV: $e\mu$	289	5.84 ± 1.97	37.78 ± 5.77	12.14 ± 5.73	10.09 ± 3.96	0.06 ± 0.00	10.57 ± 3.84	73.60 ± 15.65	21.67 ± 4.22	1.84 ± 0.11	93.17 ± 13.90	28.97 ± 4.20	285.17 ± 10.24
$m_{\ell\ell} \geq 30$GeV, $p_{\text{T}}^{\text{sublead}} \geq 20$GeV: μe	788	10.40 ± 3.76	48.02 ± 12.18	23.27 ± 7.45	36.50 ± 14.10	0.03 ± 0.00	25.34 ± 5.29	242.25 ± 48.80	57.96 ± 10.10	6.28 ± 0.43	266.51 ± 38.05	88.29 ± 12.06	779.51 ± 18.27
$m_{\ell\ell} \geq 30$GeV, $p_{\text{T}}^{\text{sublead}} \geq 20$GeV: $e\mu$	913	9.73 ± 1.69	46.17 ± 14.07	27.12 ± 8.00	43.19 ± 16.74	0.03 ± 0.00	31.67 ± 6.75	276.49 ± 56.01	71.91 ± 12.42	7.20 ± 0.46	317.62 ± 43.06	109.42 ± 15.44	908.90 ± 18.73

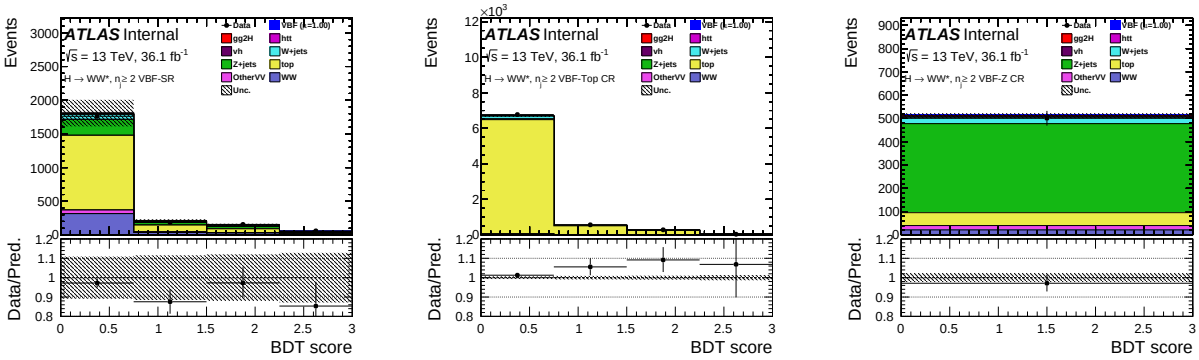
Table 8.4: Post-fit event yields for each signal region bin and control region for the ggF $N_{\text{jet}} = 0$ and $N_{\text{jet}} = 1$ categories. Both statistical and systematic uncertainties are included.

8.3 VBF Statistical Treatment

The setup for the likelihood fit in the $N_{\text{jet}} \geq 2$ category to extract the VBF signal strength parameter shares many similarities with the ggF likelihood fit setup. Therefore, this section focuses mainly on the differences between the two. After the signal selection in the $N_{\text{jet}} \geq 2$ category described in subsection 4.5.3, one signal region is included in the fit by combining the $e\mu$ and μe channels together into a single region. The BDT output distribution is used as the discriminating variable in the final fit, with the BDT shape being considered in the top control region and the $Z \to \tau\tau$ control region being considered only as a single bin. The BDT distribution is rebinned by scanning for the boundaries which provide the best signal significance. Due to the analysis being affected by large Monte Carlo statistical uncertainties, they are also considered in the significance which is defined as

$$\text{Sig} = \frac{N_S}{\sqrt{N_S + N_B + \Delta N_B^2}} \tag{8.4}$$

where N_S is the signal yield, N_B is the total background yield and ΔN_B is the total background statistical uncertainty. The results of the BDT bin optimization are the following bin boundaries: [-1,0.26,0.61,0.86,1]. An overview of all the regions considered in the fit with pre-fit background yields is shown in Figure 8.5.



(a) VBF Signal Region (b) Top Control Region (c) $Z \to \tau\tau$ Control Region

Figure 8.5: Overview of all regions included in the VBF fit. In each bin, the data is compared with the total pre-fit expected background yields. The uncertainty bands include both statistical and systematic uncertainties.

The Monte Carlo statistical uncertainties tend to be relatively large in the VBF phase space because of its unique event topology. A typical treatment (which is also utilized for the ggF fit) is to group them together in a single bin for all backgrounds combined. However, the assumption behind this simplification breaks down when the background templates suffer from low statistics. Therefore, the Monte Carlo statistical uncertainties are decorrelated between the separate processes for each bin in the VBF fit.

Each of the systematic uncertainties included in the VBF fit are listed in Table 8.5. Uncertainties with negligible contributions are removed following the same pruning procedure described in section 8.2. In addition, a so-called "smoothing" algorithm is implemented in order to reduce the impact of Monte Carlo statistical uncertainties due to bin migration effects that are present when considering four-vector systematics. Here, approximate uncertainties for problematic bins are applied which are derived after combining bins until

the statistical uncertainty in each of the merged bins (calculated in the nominal template) is less than 5%.

Similarly to the ggF categories, a likelihood fit is first performed using an Asimov dataset that is generated by setting $\mu = 1$. The expected VBF signal significance is found to be $Z = 2.6$, while the expected uncertainties on μ are shown in Table 8.6 along with a breakdown of the uncertainties shown in Figure 8.6. A likelihood fit is then performed using the full observed dataset. The ranking of nuisance parameter impact on the signal strength is given in Figure 8.7, while correlations between the nuisance parameters are displayed in Figure 8.8. The final post-fit event yields for the VBF signal region and control regions as well as for each VBF signal region BDT bin are shown in Table 8.7 and Table 8.8, respectively.

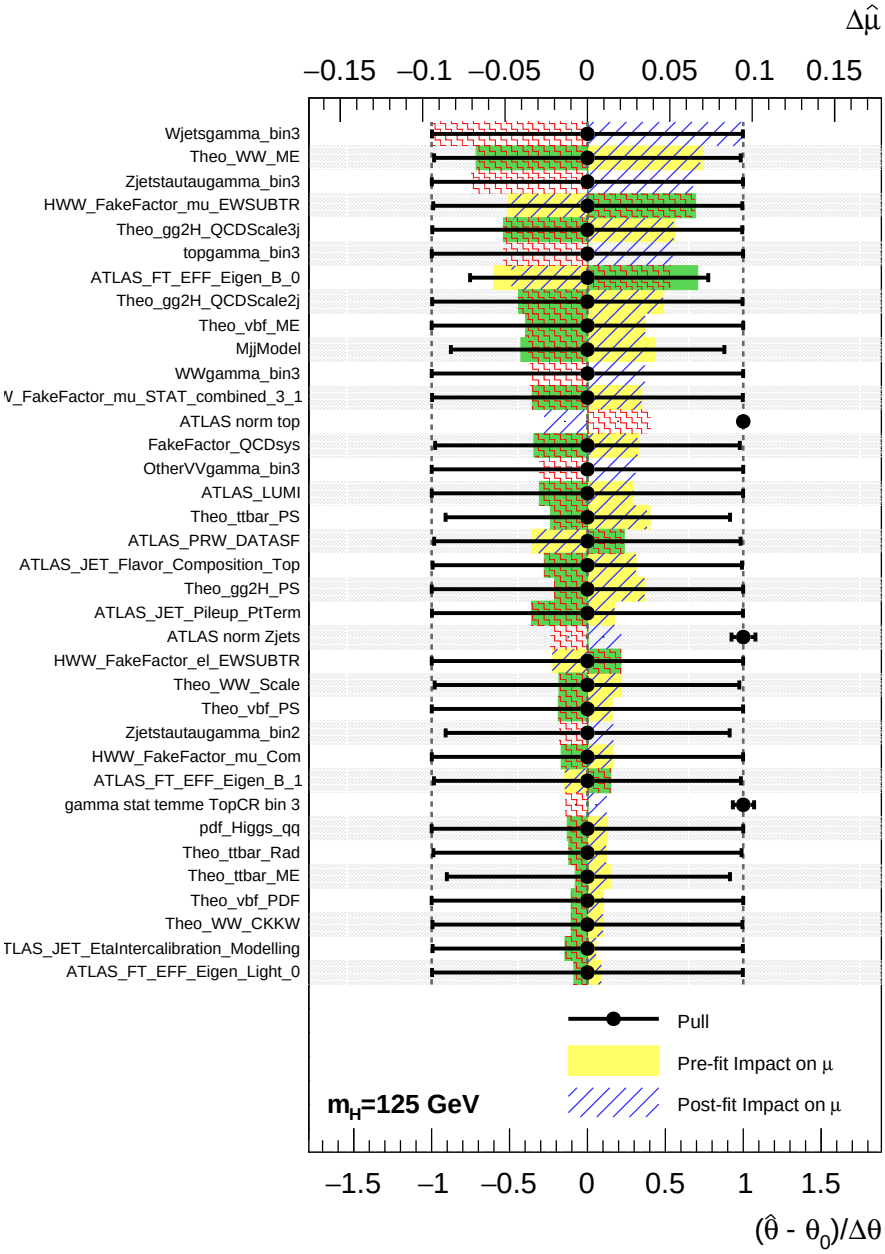


Figure 8.6: The ranking distribution of the nuisance parameters participating in the VBF fit to an Asimov dataset. Their pull and post-fit uncertainties are indicated by the black dot and associated error bar, respectively. The dashed lines represent their post-fit impacts, which are evaluated by changing them by their profiled error at the maximum likelihood. The yellow and green bands represent their pre-fit impacts, which are evaluated by changing their value by their pre-fit uncertainty.

Systematic uncertainty	Type	Comp.
Luminosity	N	1
Physics Objects		
Jet energy scale	SN	21(23)
Jet energy resolution	SN	1
b-tagging efficiency	SN	3
c-tagging efficiency	SN	4
Light jet-tagging efficiency	SN	5
Muon resolution	SN	1
Muon momentum scale	SN	1
Muon efficiency	SN	6
Electron resolution	SN	1
Electron momentum scale	SN	6
Electron efficiency	SN	32
MET uncertainty	SN	3
Pile up Reweighing	SN	1
JVT	SN	1
FJVT	SN	1
Background MC Model x		
ggF+2j- QCD scale	NS	1
ggF+3j- QCD scale	NS	1
ggF+2j- UEPS	N	1
ggF+2j- Matching	N	1
Top generator	NS	2
Top PS	NS	2
Top Radiation	NS	2
Top Interfeence	NS	1
QCD WW generator	NS	1
QCD WW- QCD Scale	NS	1
QCD WW- PDF	NS	1
QCD WW- CKKW	N	1
Data driven background		
Fake systematics	N	15
Fake systematics on QCD Corr.	S	1
Signal Model		
vbf- Generator	N	1
vbf- QCD scale	NS	1
vbf- PS	NS	1
vbf- PDF	NS	1

Table 8.5: The list of systematic uncertainties considered in the VBF fit. An "N" denotes a normalization or rate only systematic, while an "S" denotes a shape only systematic. An "NS" means that both shape and rate systematics are taken into account. The number of components in a given systematic uncertainty are also reported. For the jet energy scale uncertainty, the effective number of nuisance parameters included in the fit is 23 due to the jet flavor composition systematics being decorrelated between WW, top, and other signal/background processes.

Component	Error on μ
CTRL	+0.261 / -0.249
SYS	+0.251 / -0.245
SR STAT	+0.332 / -0.31
STAT	+0.336 / -0.313
TOT	+0.417 / -0.397
$Norm_{Z\to\tau\tau}$	+0.0493 / -0.0442
$Norm_{top}$	+0.0217 / -0.0199
μ-Fake EWSUBTR	+0.0777 / -0.0841
Theo: WW ME	+0.0798 / -0.0804
Theo: ggf QCDScale3j	+0.0716 / -0.0711
Fake QCD Corr.	+0.0714 / -0.0687
Theo: $t\bar{t}$ PS	+0.0654 / -0.0648
Theo: ggf QCDScale2j	+0.0608 / -0.0617
Mjj Model	+0.062 / -0.0564
Theo: $t\bar{t}$ ME	+0.0529 / -0.0518
Theo: WW Scale	+0.0519 / -0.0515
μ-Fake STAT Bin3	+0.0495 / -0.0486
JER	+0.0457 / -0.0489
e-Fake EWSUBTR	+0.0497 / -0.0484

Table 8.6: Summary of the expected uncertainties on the VBF signal strength μ for data and Monte Carlo statistics as well as the most highly ranking experimental and theory systematics.

Process	SR	Top CR	Ztt CR
OtherVV	70.55 ± 13.72	9.71 ± 4.93	18.47 ± 5.06
WW	384.48 ± 59.27	51.75 ± 6.40	21.21 ± 2.98
W+jets	108.49 ± 38.19	181.81 ± 51.86	25.55 ± 2.94
$Z \to \tau\tau$	297.64 ± 41.83	50.50 ± 9.82	369.39 ± 24.71
top	1234.01 ± 89.38	7361.64 ± 102.59	54.80 ± 6.16
htt	2.26 ± 0.32	0.30 ± 0.06	5.32 ± 0.69
vh	3.63 ± 0.33	1.08 ± 0.13	0.22 ± 0.09
gg2H	36.67 ± 13.37	5.32 ± 0.75	2.81 ± 0.18
Bkg	2137.74 ± 47.43	7662.31 ± 87.39	497.83 ± 22.33
Signal	29.90 ± 16.36	3.48 ± 1.97	2.77 ± 1.52
SignalExpected	43.30 ± 23.70	5.04 ± 2.85	4.02 ± 2.21
S/B	1.41e-02	4.55e-04	5.57e-03
$S/\sqrt{(S+B)}$	6.42e-01	3.98e-02	1.24e-01
data	2164	7668	501

Table 8.7: Post-fit event yields for the VBF signal region and control regions. Both statistical and systematic uncertainties are included.

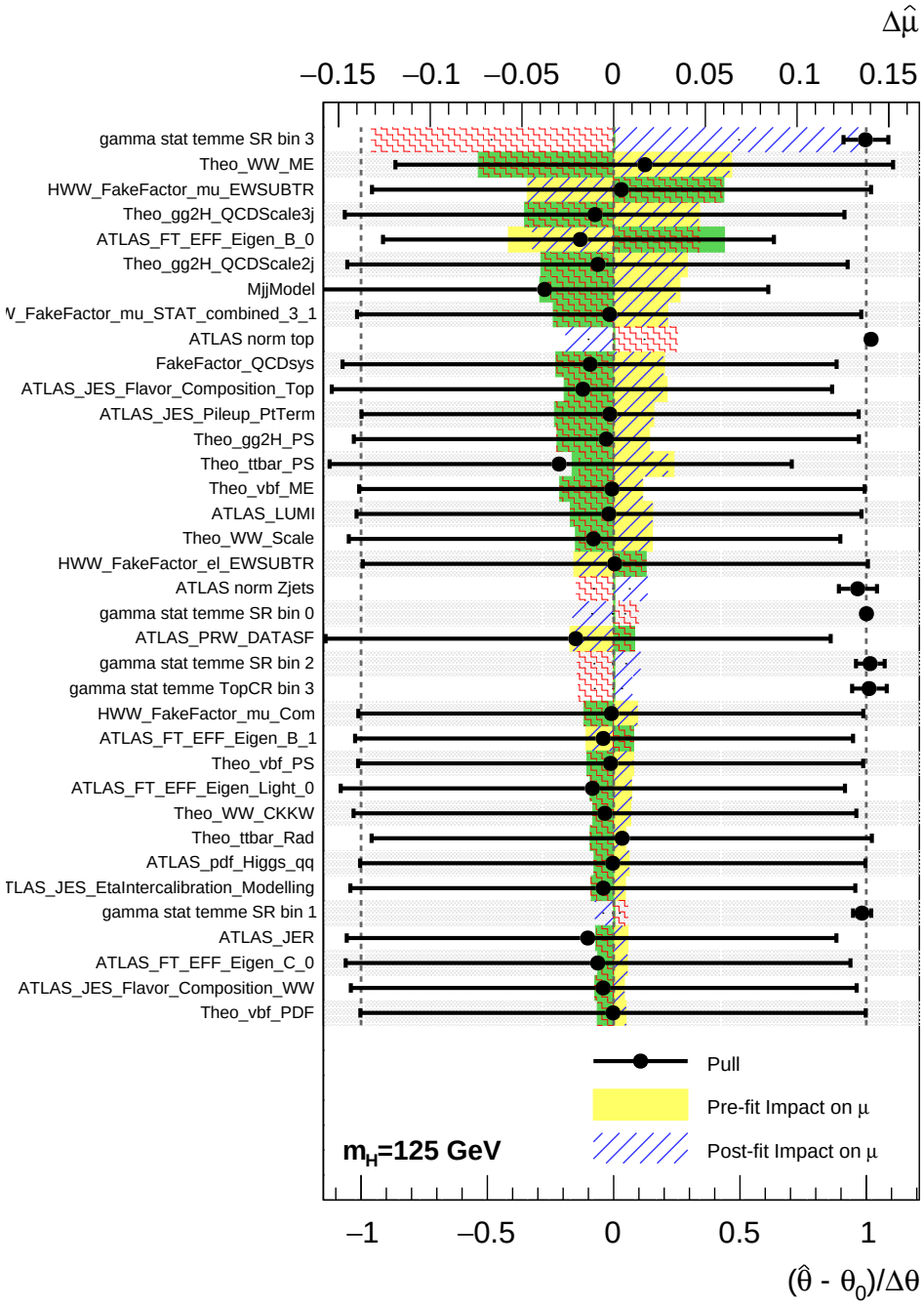


Figure 8.7: The ranking distribution of the nuisance parameters participating in the full VBF fit to the observed data. Their pull and post-fit uncertainties are indicated by the black dot and associated error bar, respectively. The dashed lines represent their post-fit impacts, which are evaluated by changing them by their profiled error at the maximum likelihood. The yellow and green bands represent their pre-fit impacts, which are evaluated by changing their value by their pre-fit uncertainty.

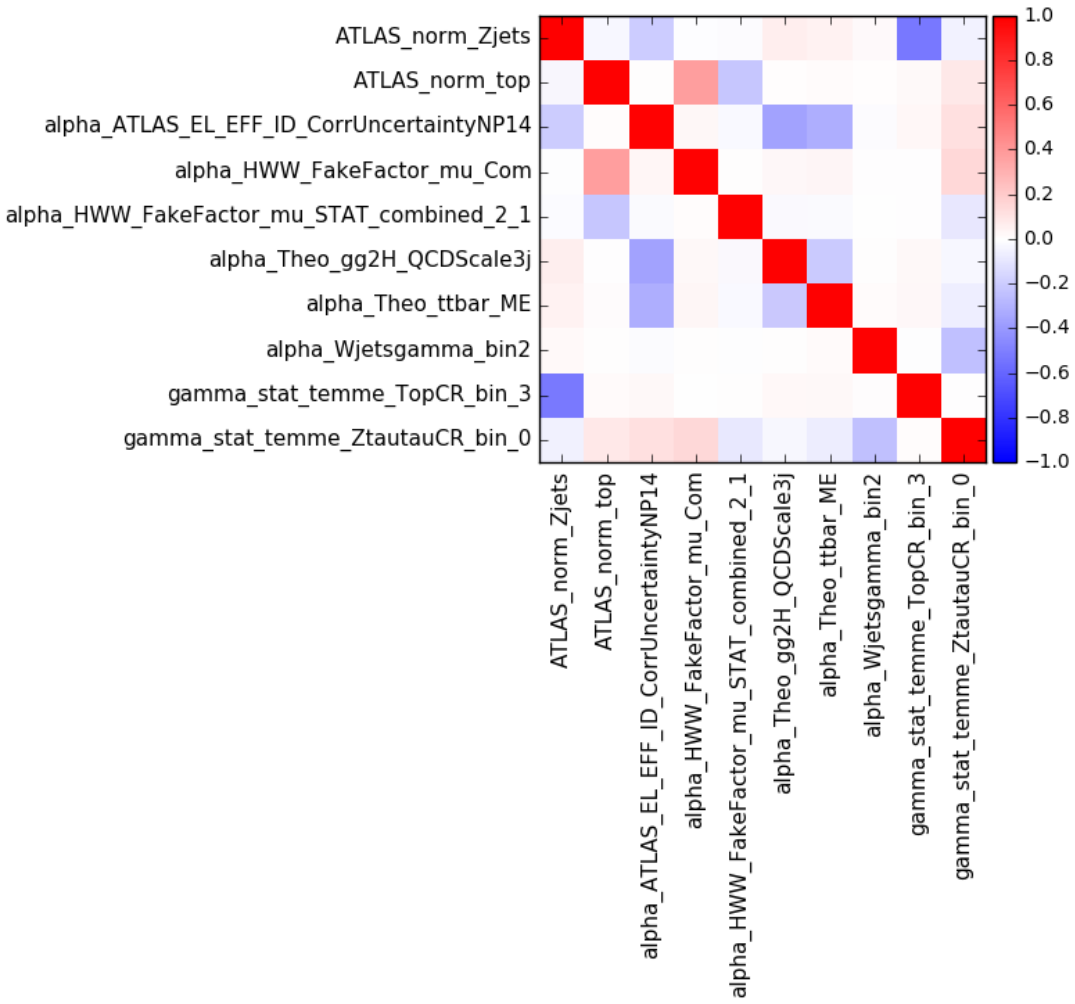


Figure 8.8: Post-fit correlations of the nuisance parameters participating in the full VBF fit to the observed data.

Process	Bin1	Bin2	Bin3	Bin4
OtherVV	52 ± 10	10.7 ± 2.1	5.0 ± 1.0	2.7 ± 0.6
WW	309 ± 48	34.6 ± 5.6	28.9 ± 5.7	10.2 ± 2.4
W+jets	73 ± 32	14.9 ± 5.7	13.8 ± 5.8	7.54 ± 2.5
$Z \to \tau\tau$	226 ± 32	34.7 ± 5.0	31.8 ± 4.7	4.36 ± 0.7
top	1075 ± 74	91.6 ± 9.6	54.5 ± 7.3	13.4 ± 2.3
htt	1.3 ± 0.2	0.41 ± 0.06	0.38 ± 0.06	0.19 ± 0.03
vh	3.0 ± 0.3	0.32 ± 0.03	0.27 ± 0.03	0.05 ± 0.01
gg2H	14.1 ± 5.2	8.2 ± 3.0	9.1 ± 3.1	5.16 ± 2.07
Bkg	1755 ± 40	195.5 ± 9.6	143.7 ± 9.5	43.660 ± 5.04
Signal	1.8 ± 1.0	3.0 ± 1.7	8.1 ± 4.4	17.20 ± 9.34
SignalExpected	2.6 ± 1.4	4.3 ± 2.4	11.6 ± 6.4	24.660 ± 13.44
data	1761	187	156	60

Table 8.8: Post-fit event yields in the VBF signal region for each BDT bin. Both statistical and systematic uncertainties are included.

Chapter 9

Results

This chapter summarizes the results of the analysis presented in this thesis. The expected ggF and VBF signal strengths are first obtained from fits using Asimov datasets and are found to be

$$\begin{aligned} \mu_{\text{ggF}}^{\text{exp}} &= 1.00 \pm 0.10(\text{stat.}) \pm +0.19_{-0.18}(\text{syst.}) = 1.00^{+0.22}_{-0.21} \\ \mu_{\text{VBF}}^{\text{exp}} &= 1.00^{+0.33}_{-0.31}(\text{stat.}) \pm 0.25(\text{syst.}) = 1.00^{+0.42}_{-0.40} \end{aligned}$$

where statistical and systematic uncertainties are also reported separately. A combined fit using the full observed dataset is then performed, resulting in observed ggF and VBF signal strengths that are simultaneously determined to be

$$\begin{aligned} \mu_{\text{ggF}}^{\text{obs}} &= 1.10^{+0.10}_{-0.09}(\text{stat.})^{+0.13}_{-0.11}(\text{theo syst.})^{+0.14}_{-0.13}(\text{exp syst.}) = 1.10^{+0.21}_{-0.20} \\ \mu_{\text{VBF}}^{\text{obs}} &= 0.62^{+0.29}_{-0.27}(\text{stat.})^{+0.12}_{-0.13}(\text{theo syst.}) \pm 0.15(\text{exp syst.}) = 0.62^{+0.36}_{-0.35}. \end{aligned}$$

The observed (expected) ggF and VBF signals have significances of 6.0 (5.3) and 1.8 (2.6) standard deviations, respectively.

The post-fit m_{T} distribution for the combination of $N_{\text{jet}} = 0$ and $N_{\text{jet}} = 1$ signal regions is shown in Figure 9.1, while the post-fit BDT score distribution in the VBF signal region is shown in Figure 9.2. Both of these figures include the final background normalization factors that are extracted from the fit and displayed in Table 9.1, as well as the measured signal strengths. The normalization factors obtained from the fit are compatible with those shown in Table 5.9 which use the simple matrix inversion method, where differences are attributable to correlations with the constrained nuisance parameters.

Category	WW	$t\bar{t}/Wt$	Z/γ^*
$N_{\text{jet},(p_{\text{T}}>30\text{ GeV})} = 0$ ggF	1.06 ± 0.09	0.99 ± 0.17	0.84 ± 0.04
$N_{\text{jet},(p_{\text{T}}>30\text{ GeV})} = 1$ ggF	0.97 ± 0.17	0.98 ± 0.08	0.90 ± 0.12
$N_{\text{jet},(p_{\text{T}}>30\text{ GeV})} \geq 2$ VBF	–	1.01 ± 0.01	0.93 ± 0.07

Table 9.1: Post-fit normalization factors which are applied to the corresponding background estimates in the signal regions. The errors include statistical and systematic uncertainties.

The predicted cross-section times branching fraction values are 10.4 ± 0.6 pb and 0.81 ± 0.02 pb for ggF and VBF production [68], respectively. With the signal strength μ

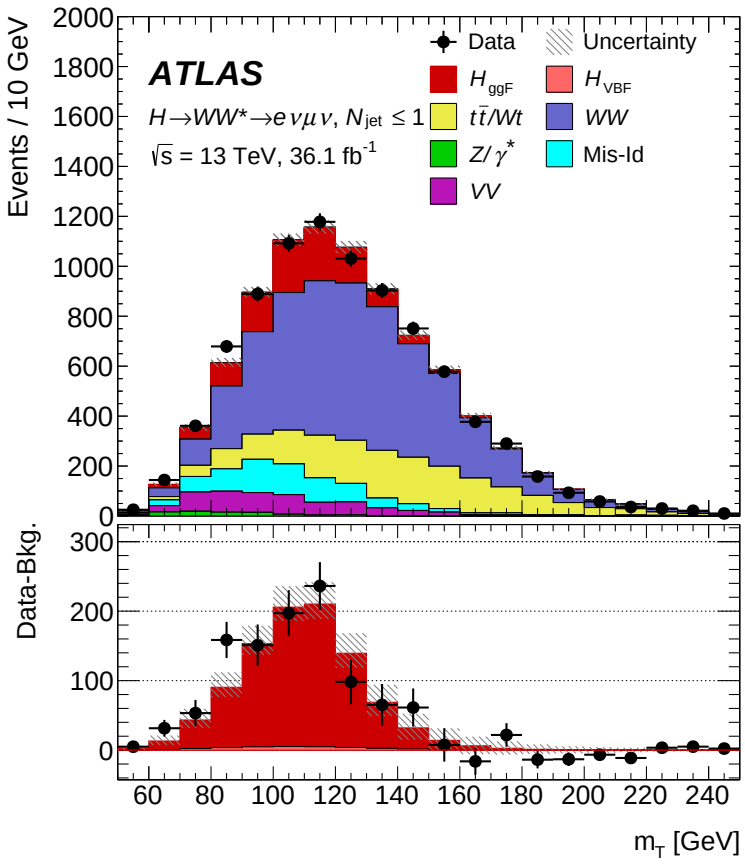


Figure 9.1: Post-fit m_{T} distribution for the $N_{\mathrm{jet}} = 0$ and $N_{\mathrm{jet}} = 1$ signal regions combined. The difference between the data and the total background is compared with the distribution for the Standard Model Higgs boson. The uncertainty band includes the total uncertainty of the signal and background modeling contributions.

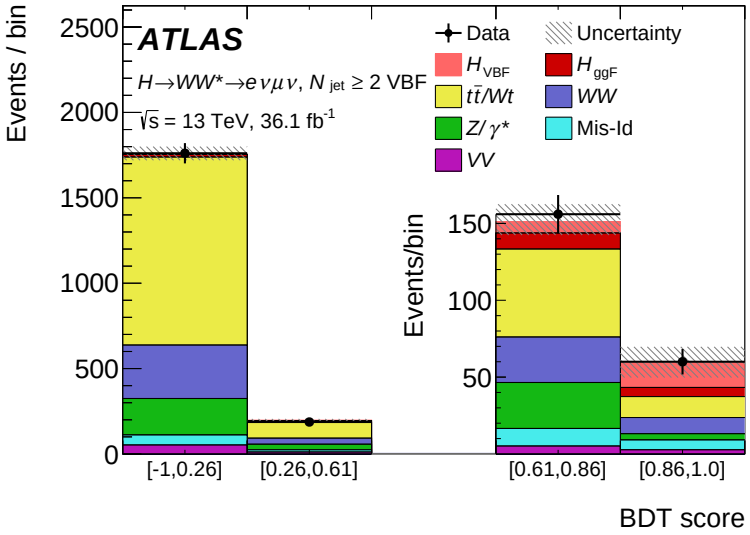


Figure 9.2: Post-fit distribution of BDT score in the VBF signal region. The uncertainty band includes the total uncertainty of the signal and background modeling contributions.

being defined as the ratio of the measured signal yield to that predicted by the Standard Model, the final observed cross-sections can be calculated in a straight-forward way as $\hat{\sigma} = \hat{\mu} \cdot \sigma_{\text{pred.}}$ and are determined to be

$$\begin{aligned}
\sigma_{\text{ggF}} \cdot \mathcal{B}_{H\to WW^*} &= 11.4^{+1.2}_{-1.1}(\text{stat.})^{+1.2}_{-1.1}(\text{theo syst.})^{+1.4}_{-1.3}(\text{exp syst.})\ \text{pb} = 11.4^{+2.2}_{-2.1}\ \text{pb} \\
\sigma_{\text{VBF}} \cdot \mathcal{B}_{H\to WW^*} &= 0.50^{+0.24}_{-0.22}(\text{stat.}) \pm 0.10(\text{theo syst.})^{+0.12}_{-0.13}(\text{exp syst.})\ \text{pb} = 0.50^{+0.29}_{-0.28}\ \text{pb}.
\end{aligned}$$

A breakdown of the largest contributions to the total uncertainty in the cross-sections are shown in Table 9.2. Finally, the 68% and 95% confidence level two-dimensional likelihood contours of $\sigma_{\text{ggF}} \cdot \mathcal{B}_{H\to WW^*}$ vs. $\sigma_{\text{VBF}} \cdot \mathcal{B}_{H\to WW^*}$ are displayed in Figure 9.3 which are in agreement with the Standard Model prediction.

Source	$\Delta\sigma_{\text{ggF}} \cdot \mathcal{B}_{H\to WW^*}$ [%]	$\Delta\sigma_{\text{VBF}} \cdot \mathcal{B}_{H\to WW^*}$ [%]
Data statistics	10	46
CR statistics	7	9
MC statistics	6	21
Theoretical uncertainties	10	19
ggF signal	5	13
VBF signal	<1	4
WW	6	12
Top-quark	5	5
Experimental uncertainties	8	9
b-tagging	4	6
Modeling of pileup	5	2
Jet	2	2
Lepton	3	<1
Misidentified leptons	6	9
Luminosity	3	3
TOTAL	18	57

Table 9.2: Breakdown of the largest contributions to the total uncertainty in $\sigma_{\text{ggF}} \cdot \mathcal{B}_{H\to WW^*}$ and $\sigma_{\text{VBF}} \cdot \mathcal{B}_{H\to WW^*}$, with the individual uncertainties being grouped together. The sum in quadrature of individual components differs from the total uncertainty due to correlations between components.

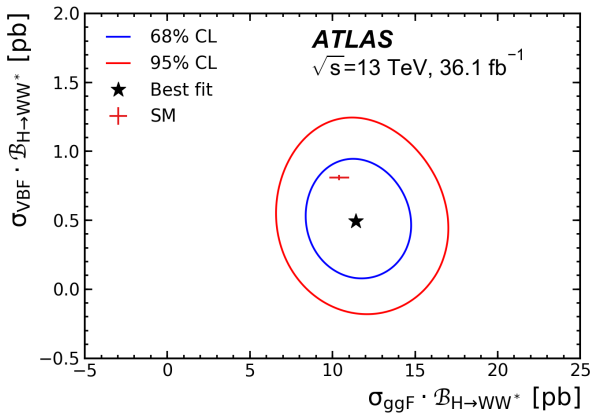


Figure 9.3: Two-dimensional likelihood contours at the 68% (blue) and 95% (red) confidence levels of $\sigma_{\text{ggF}} \cdot \mathcal{B}_{H\to WW^*}$ vs. $\sigma_{\text{VBF}} \cdot \mathcal{B}_{H\to WW^*}$. The Standard Model prediction is shown with a red marker, with error bars representing the respective ggF and VBF theory uncertainties [68].

Chapter 10

Conclusions

The analysis of the $H \to WW^*$ decay mode to two leptons, offering together one of the most challenging and the most sensitive measurements of Higgs boson production, is presented in this thesis. The product of the $H \to WW^*$ branching ratio times the ggF and VBF cross-sections are measured to be $11.4^{+1.2}_{-1.1}(\text{stat.})^{+1.8}_{-1.7}(\text{syst.})$ pb and $0.50^{+0.24}_{-0.22}$ (stat.) $\pm$ 0.17(syst.) pb, respectively, which correspond to observed signal strengths of $1.10^{+0.21}_{-0.20}$ and $0.62^{+0.36}_{-0.35}$, respectively, and are compatible with the Standard Model prediction of $\mu = 1$. These results are competitive with the analysis published previously on Run 1 data in which there was already an observation in the ggF production mode with a 6.1 sigma excess [18]. Although the present analysis had access to a larger integrated luminosity (36.1 fb^{-1} instead of 20.3 fb^{-1}) and a higher Higgs cross-section (by a factor of 2.3 for the ggF production mode), a number of factors prevent it from achieving significantly better results than the Run 1 analysis, including the exclusion of the same flavor channels and leptons between $10 < p_{\text{T}} < 15$ GeV in addition to larger background rates. For example, the top cross-section increased by a factor of 3.3, while significantly more mis-identified leptons are observed as well - which is likely attributable to higher pileup conditions, leading to less-performant lepton isolation.

Further studies will be performed in the future in addition to inclusive cross-section measurements in order to probe for any deviations from the Standard Model. For example, differential and fiducial cross-sections will be reported in bins of key variables such as number of jets and Higgs boson p_{T} as was done with the Run 1 dataset [32]. In addition, the results will be reported in the context of a simplified template cross-section (STXS) framework which is by now common across all Higgs analyses, simultaneously representing a way to more cleanly separate measurement from interpretation so as to reduce theory dependencies and providing more finely-graind measurements in a way which still allows for the global combination of the measurements in all decay channels. Efforts are also underway to incorporate future analyses into global effective field theory (EFT) parameters which interpret the current Standard Model as the low energy limit of some more fundamental theory.

The sensitivity of the VBF measurement is still driven by its statistical uncertainty, which will be improved once the full Run 2 dataset (approximately 139 fb^{-1} of integrated luminosity) is included with additional collision events recorded in the years 2017 and 2018. The sensitivity of the ggF measurement, on the other hand, is already driven by systematic uncertainties. Therefore, its largest improvements in the future will need to come from a reduction in experimental and theoretical uncertainties.

The reliance on theory predictions can be reduced by employing more data-driven

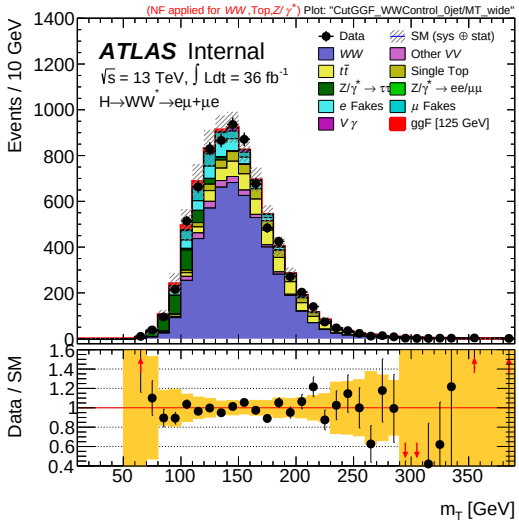
methods, which will benefit from an increase in data statistics. Experimental uncertainties that are constrained by external measurements will improve as the dataset grows, since these measurements will also profit from the increase in statistical power. In addition, new ATLAS algorithms will be used for the next $H \to WW^*$ measurements which are expected to improve the performance across multiple key areas of the analysis. For example, jets will be reconstructed using so-called "particle flow" objects which make use of both calorimeter and tracking information [25] and will be b-tagged with a deep neural network for better b-jet discriminatory power [24].

Appendix A

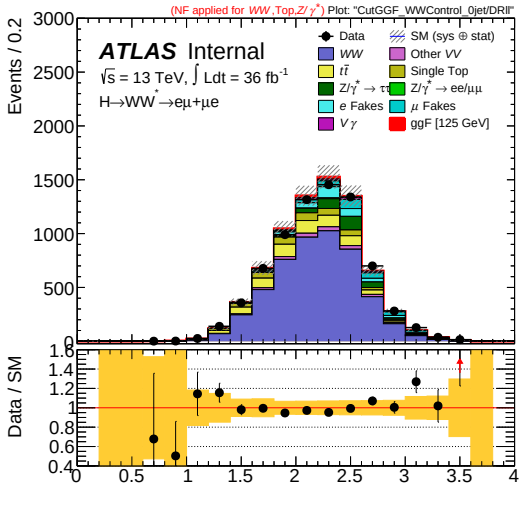
Additional Control Region Distributions

A.1 WW CRs

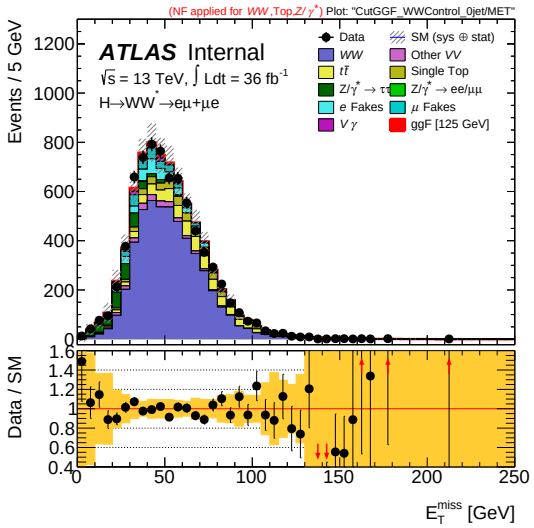
Additional kinematic distributions in the 0 and 1 jet WW control regions can be seen in Figure A.1 and Figure A.2, respectively, where reasonable data and MC agreement is consistently observed.



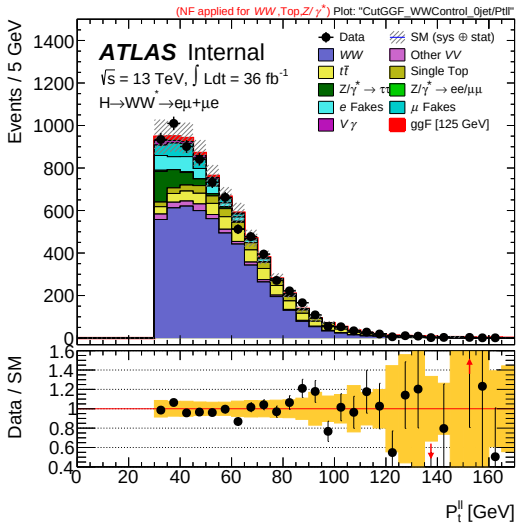
(a) m_{T}



(b) $\Delta R(\ell\ell)$

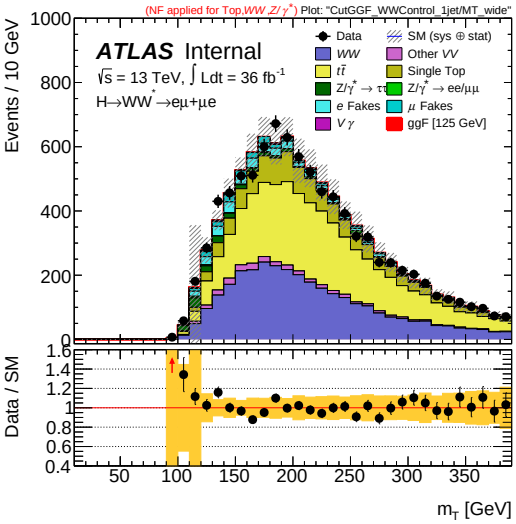


(c) $E_{\mathrm{T}}^{\mathrm{miss}}$

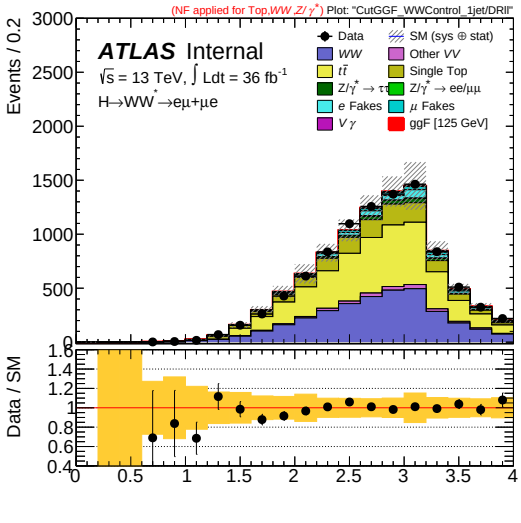


(d) $p_{\mathrm{T}}^{\ell\ell}$

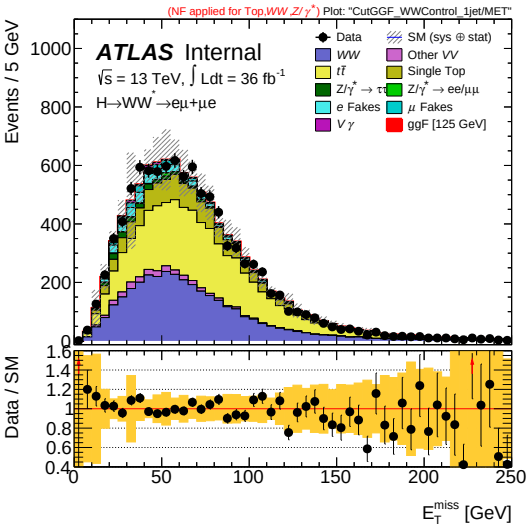
Figure A.1: Additional kinematic distributions in the 0 jet WW control region. The normalization factors from Table 5.9 have been applied. The uncertainty bands represent the quadratic sum of the statistical and experimental systematic rate uncertainties.



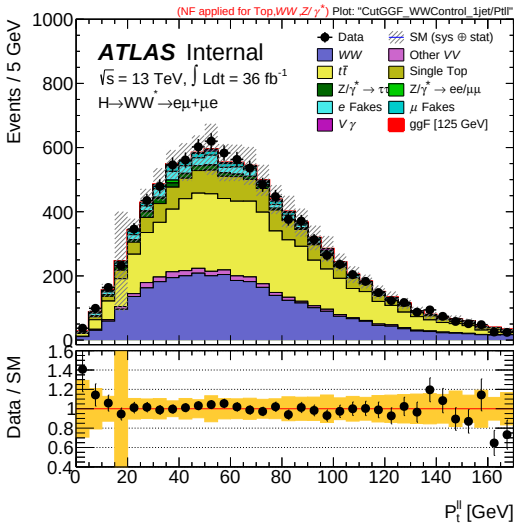
(a) m_{T}



(b) $\Delta R(\ell\ell)$



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

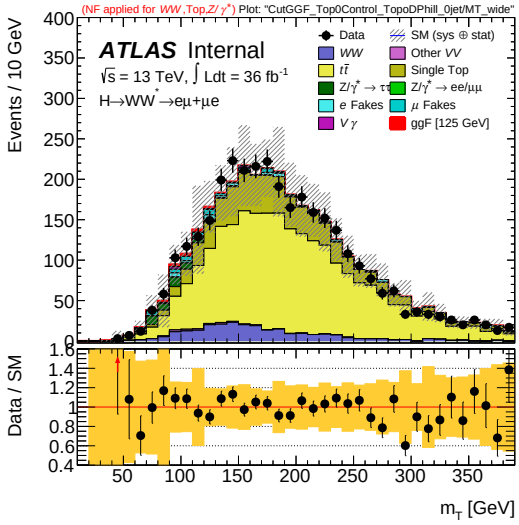


(d) $p_{\mathrm{T}}^{\ell\ell}$

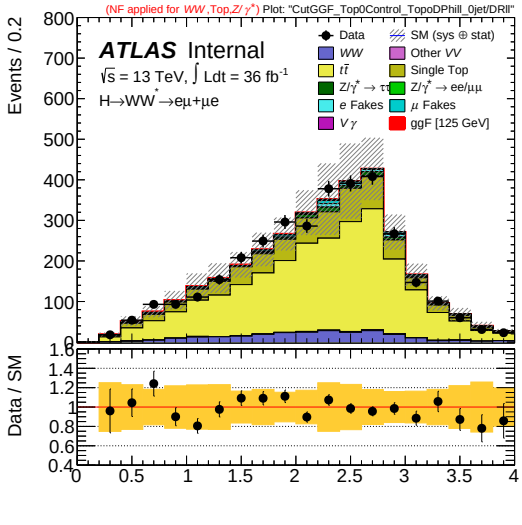
Figure A.2: Additional kinematic distributions in the 1 jet WW control region. The normalization factors from Table 5.9 have been applied. The uncertainty bands represent the quadratic sum of the statistical and experimental systematic rate uncertainties.

A.2 Top CRs

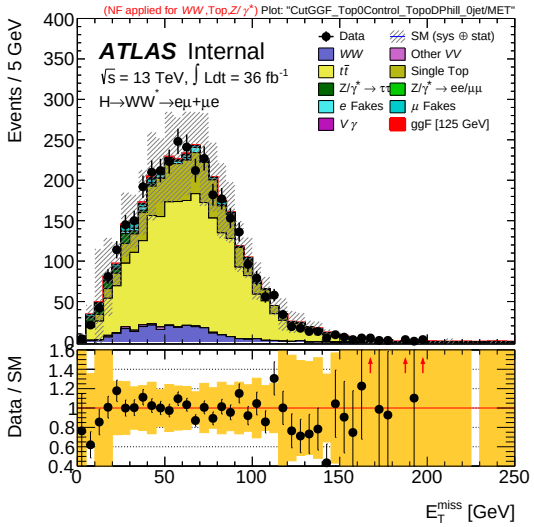
Additional kinematic distributions in the 0 and 1 jet Top control regions can be seen in Figure A.3 and Figure A.4, respectively, where reasonable data and MC agreement is consistently observed.



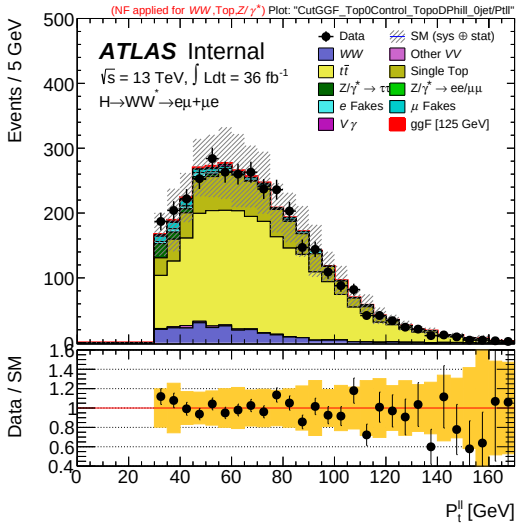
(a) m_{T}



(b) $\Delta R(\ell\ell)$

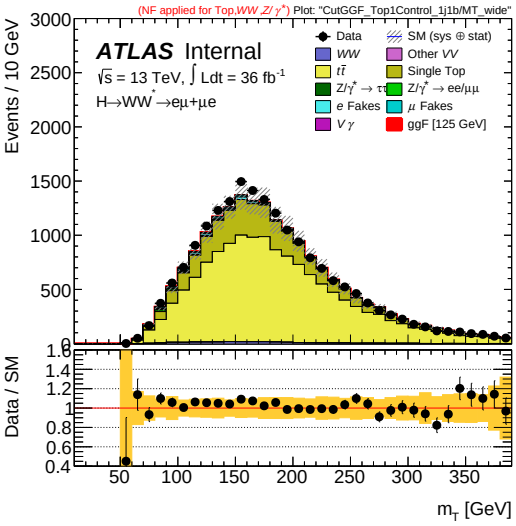


(c) $E_{\mathrm{T}}^{\mathrm{miss}}$

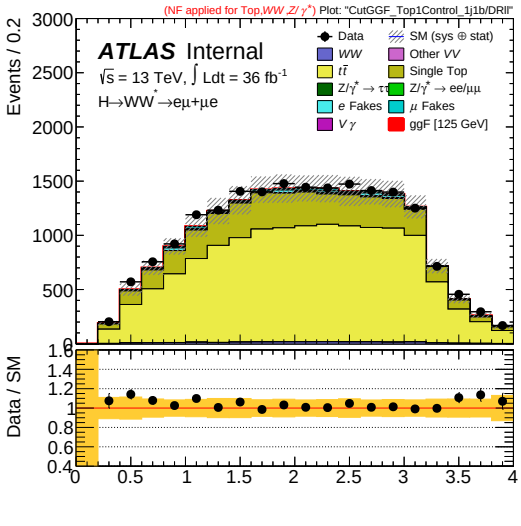


(d) $p_{\mathrm{T}}^{\ell\ell}$

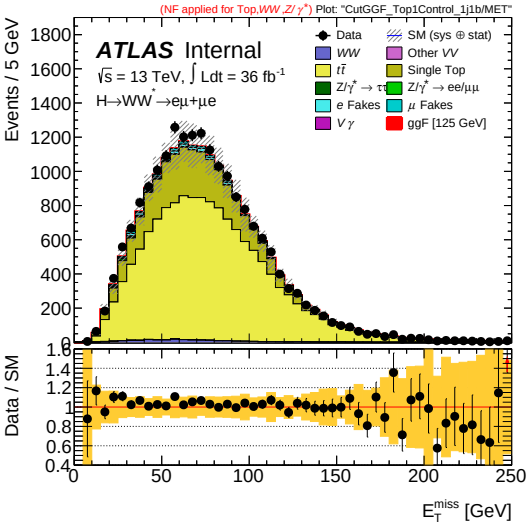
Figure A.3: Additional kinematic distributions in the 0 jet Top control region. The normalization factors from Table 5.9 have been applied. The uncertainty bands represent the quadratic sum of the statistical and experimental systematic rate uncertainties.



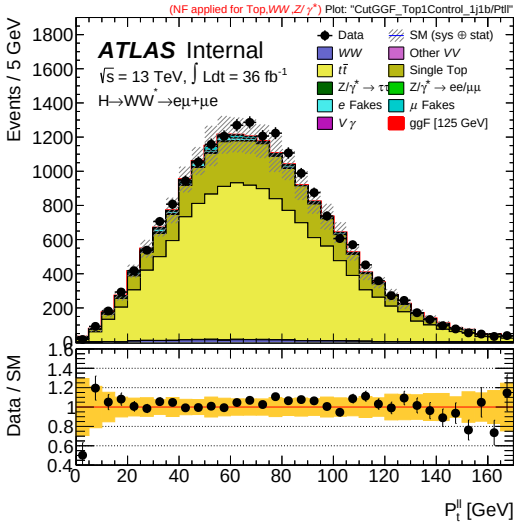
(a) m_{T}



(b) $\Delta R(\ell\ell)$



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

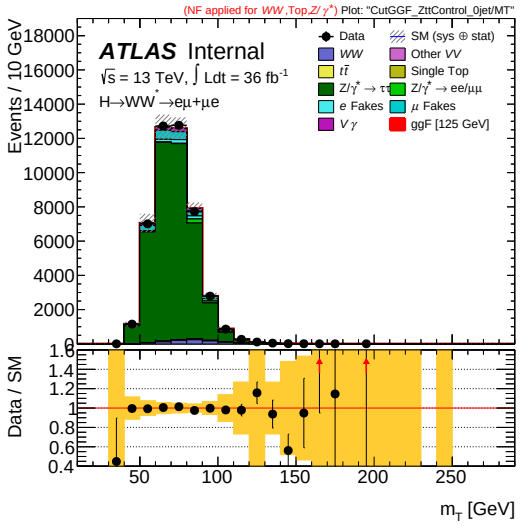


(d) $p_{\mathrm{T}}^{\ell\ell}$

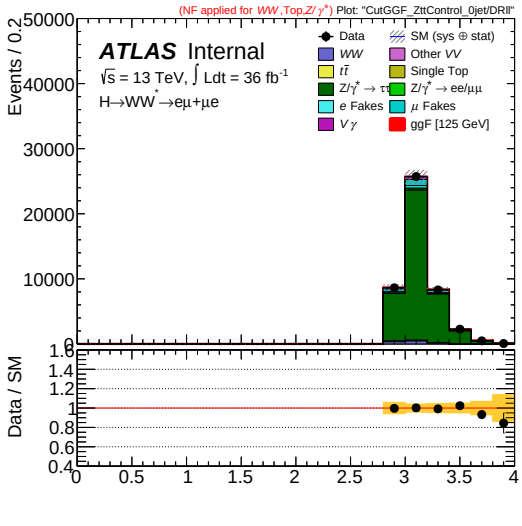
Figure A.4: Additional kinematic distributions in the 1 jet Top control region. The normalization factors from Table 5.9 have been applied. The uncertainty bands represent the quadratic sum of the statistical and experimental systematic rate uncertainties.

A.3 $Z \to \tau\tau$ CRs

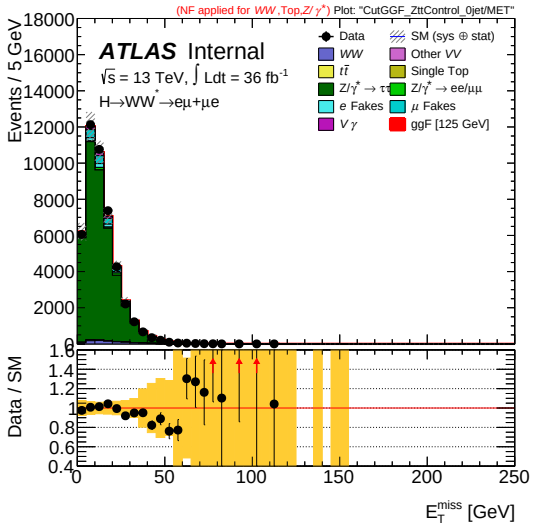
Additional kinematic distributions in the 0 and 1 jet $Z \to \tau\tau$ control regions can be seen in Figure A.5 and Figure A.6, respectively, where reasonable data and MC agreement is consistently observed.



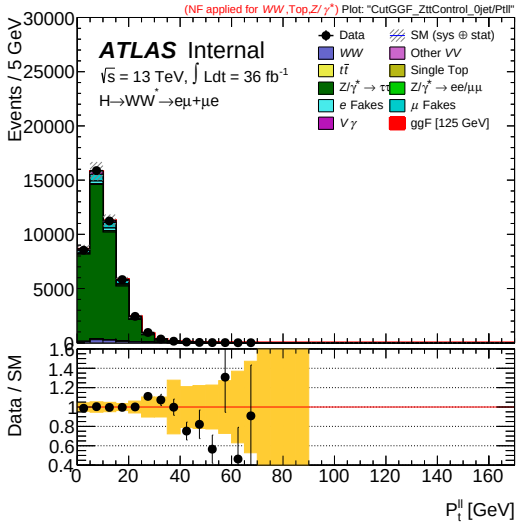
(a) m_{T}



(b) $\Delta R(\ell\ell)$

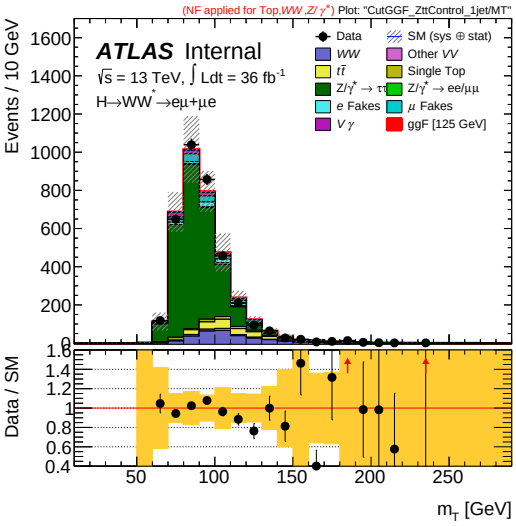


(c) $E_{\mathrm{T}}^{\mathrm{miss}}$

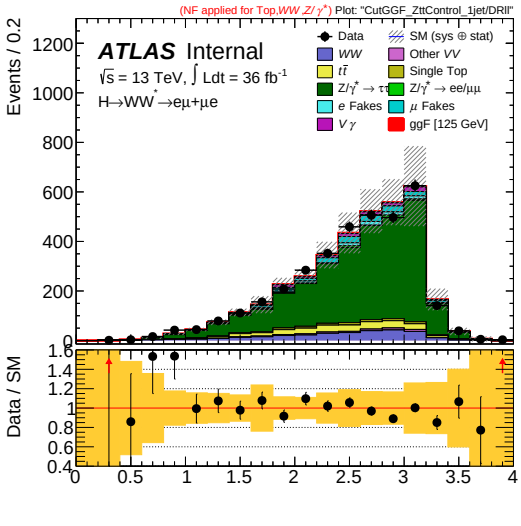


(d) $p_{\mathrm{T}}^{\ell\ell}$

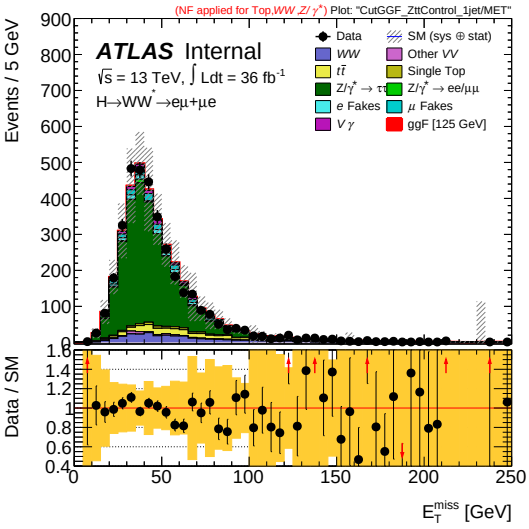
Figure A.5: Additional kinematic distributions in the 0 jet $Z \to \tau\tau$ control region. The normalization factors from Table 5.9 have been applied. The uncertainty bands represent the quadratic sum of the statistical and experimental systematic rate uncertainties.



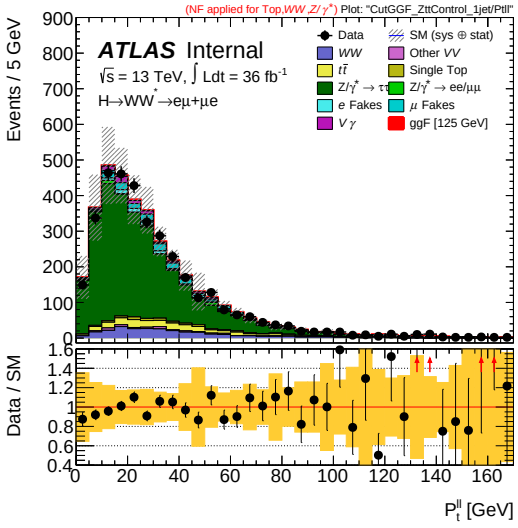
(a) m_{T}



(b) $\Delta R(\ell\ell)$



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



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

Figure A.6: Additional kinematic distributions in the 1 jet $Z \to \tau\tau$ control region. The normalization factors from Table 5.9 have been applied. The uncertainty bands represent the quadratic sum of the statistical and experimental systematic rate uncertainties.

Bibliography

[1] URL https://twiki.cern.ch/twiki/bin/view/AtlasPublic/LuminosityPublicResultsRun2.

[2] A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119 – 139, 1997. ISSN 0022-0000. doi: http://dx.doi.org/10.1006/jcss.1997.1504.

[3] *ATLAS detector and physics performance: Technical Design Report, 1.* Technical Design Report ATLAS. CERN, Geneva, 1999.

[4] *ATLAS detector and physics performance: Technical Design Report, 2.* Technical Design Report ATLAS. CERN, Geneva, 1999.

[5] The ATLAS experiment at the CERN large hadron collider. 3(08):S08003, 2008.

[6] Improved electron reconstruction in ATLAS using the Gaussian Sum Filter-based model for bremsstrahlung. Technical Report ATLAS-CONF-2012-047, CERN, Geneva, May 2012. URL http://cds.cern.ch/record/1449796.

[7] Summary of ATLAS Pythia 8 tunes. Technical Report ATL-PHYS-PUB-2012-003, CERN, Geneva, Aug 2012. URL https://cds.cern.ch/record/1474107.

[8] Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC. *Phys. Lett.*, B716:1–29, 2012. doi: 10.1016/j.physletb.2012.08.020.

[9] Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC. *Phys. Lett.*, B716:30–61, 2012. doi: 10.1016/j.physletb.2012.08.021.

[10] Search for the Standard Model Higgs boson in the $H \to Z\gamma$ decay mode with pp collisions at $\sqrt{s} = 7$ and 8 TeV. Technical Report ATLAS-CONF-2013-009, CERN, Geneva, Mar 2013.

[11] Search for a Higgs boson decaying into a Z and a photon in pp collisions at $\sqrt{s} = 7$ and 8 TeV. *Phys. Lett.*, B726:587–609, 2013. doi: 10.1016/j.physletb.2013.09.057.

[12] ATLAS Run 1 Pythia8 tunes. Technical Report ATL-PHYS-PUB-2014-021, CERN, Geneva, Nov 2014. URL https://cds.cern.ch/record/1966419.

[13] Tagging and suppression of pileup jets with the ATLAS detector. Technical Report ATLAS-CONF-2014-018, CERN, Geneva, May 2014. URL https://cds.cern.ch/record/1700870.

[14] Evidence for the 125 GeV Higgs boson decaying to a pair of τ leptons. *JHEP*, 05:104, 2014. doi: 10.1007/JHEP05(2014)104.

[15] Search for the standard model Higgs boson produced in association with a W or a Z boson and decaying to bottom quarks. *Phys. Rev.*, D89(1):012003, 2014. doi: 10.1103/PhysRevD.89.012003.

[16] Jet Calibration and Systematic Uncertainties for Jets Reconstructed in the ATLAS Detector at $\sqrt{s} = 13$ TeV. Technical Report ATL-PHYS-PUB-2015-015, CERN, Geneva, Jul 2015. URL http://cds.cern.ch/record/2037613.

[17] Expected performance of missing transverse momentum reconstruction for the ATLAS detector at $\sqrt{s} = 13$ TeV. Technical Report ATL-PHYS-PUB-2015-023, CERN, Geneva, Jul 2015. URL http://cds.cern.ch/record/2037700.

[18] Observation and measurement of Higgs boson decays to WW^* with the ATLAS detector. *Phys. Rev.*, D92(1):012006, 2015. doi: 10.1103/PhysRevD.92.012006.

[19] Search for a standard model-like Higgs boson in the $\mu^+\mu^-$ and e^+e^- decay channels at the LHC. *Phys. Lett.*, B744:184–207, 2015. doi: 10.1016/j.physletb.2015.03.048.

[20] Evidence for the Higgs-boson Yukawa coupling to tau leptons with the ATLAS detector. *JHEP*, 04:117, 2015. doi: 10.1007/JHEP04(2015)117.

[21] Optimisation of the ATLAS b-tagging performance for the 2016 LHC Run. Technical Report ATL-PHYS-PUB-2016-012, CERN, Geneva, Jun 2016. URL http://cds.cern.ch/record/2160731.

[22] Search for Higgs bosons decaying into di-muon in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector. Technical Report ATLAS-CONF-2016-041, CERN, Geneva, Aug 2016.

[23] Search for the Standard Model Higgs boson produced in association with a vector boson and decaying to a $b\bar{b}$ pair in pp collisions at 13 TeV using the ATLAS detector. Technical Report ATLAS-CONF-2016-091, CERN, Geneva, Aug 2016.

[24] Optimisation and performance studies of the ATLAS b-tagging algorithms for the 2017-18 LHC run. Technical Report ATL-PHYS-PUB-2017-013, CERN, Geneva, Jul 2017. URL http://cds.cern.ch/record/2273281.

[25] M. Aaboud. Jet reconstruction and performance using particle flow with the atlas detector. *The European Physical Journal C*, 77(7):466, Jul 2017. ISSN 1434-6052. doi: 10.1140/epjc/s10052-017-5031-2. URL https://doi.org/10.1140/epjc/s10052-017-5031-2.

[26] Morad Aaboud et al. Performance of the ATLAS Trigger System in 2015. *Eur. Phys. J.*, C77(5):317, 2017. doi: 10.1140/epjc/s10052-017-4852-3.

[27] Morad Aaboud et al. Electron and photon energy calibration with the ATLAS detector using 2015–2016 LHC proton-proton collision data. *JINST*, 14(03):P03017, 2019. doi: 10.1088/1748-0221/14/03/P03017.

[28] Morad Aaboud et al. 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.*, C79(8):639, 2019. doi: 10.1140/epjc/s10052-019-7140-6.

[29] G. Aad et al. The ATLAS Simulation Infrastructure. *Eur. Phys. J.*, C70:823–874, 2010. doi: 10.1140/epjc/s10052-010-1429-9.

[30] Georges Aad et al. Improved luminosity determination in pp collisions at sqrt(s) = 7 TeV using the ATLAS detector at the LHC. *Eur. Phys. J.*, C73(8):2518, 2013. doi: 10.1140/epjc/s10052-013-2518-3.

[31] Georges Aad et al. Measurement of the Z/γ^* boson transverse momentum distribution in pp collisions at $\sqrt{s} = 7$ TeV with the ATLAS detector. *JHEP*, 09:145, 2014. doi: 10.1007/JHEP09(2014)145.

[32] Georges Aad et al. Measurement of fiducial differential cross sections of gluon-fusion production of Higgs bosons decaying to $WW^* \to e\nu\mu\nu$ with the ATLAS detector at $\sqrt{s} = 8$ TeV. *JHEP*, 08:104, 2016. doi: 10.1007/JHEP08(2016)104.

[33] Georges Aad et al. Performance of b-Jet Identification in the ATLAS Experiment. *JINST*, 11(04):P04008, 2016. doi: 10.1088/1748-0221/11/04/P04008.

[34] Georges Aad et al. Muon reconstruction performance of the ATLAS detector in proton–proton collision data at $\sqrt{s}$ =13 TeV. *Eur. Phys. J.*, C76(5):292, 2016. doi: 10.1140/epjc/s10052-016-4120-y.

[35] Georges Aad et al. Topological cell clustering in the ATLAS calorimeters and its performance in LHC Run 1. *Eur. Phys. J.*, C77:490, 2017. doi: 10.1140/epjc/s10052-017-5004-5.

[36] LHC Higgs Cross Section Working Group. Picture gallery. URL https://twiki.cern.ch/twiki/bin/view/LHCPhysics/LHCHXSWGCrossSectionsFigures.

[37] S. Agostinelli et al. GEANT4: A Simulation toolkit. *Nucl. Instrum. Meth.*, A506: 250–303, 2003. doi: 10.1016/S0168-9002(03)01368-8.

[38] G. Altarelli and G. Parisi. Asymptotic freedom in parton language. *Nuclear Physics B*, 126(2):298 – 318, 1977. ISSN 0550-3213. doi: https://doi.org/10.1016/0550-3213(77)90384-4. URL http://www.sciencedirect.com/science/article/pii/0550321377903844.

[39] Charalampos Anastasiou, Claude Duhr, Falko Dulat, Elisabetta Furlan, Thomas Gehrmann, Franz Herzog, Achilleas Lazopoulos, and Bernhard Mistlberger. High precision determination of the gluon fusion Higgs boson cross-section at the LHC. *JHEP*, 05:058, 2016. doi: 10.1007/JHEP05(2016)058.

[40] B. Andersson, G. Gustafson, G. Ingelman, and T. Sjöstrand. Parton fragmentation and string dynamics. *Physics Reports*, 97(2):31 – 145, 1983. ISSN 0370-1573. doi: https://doi.org/10.1016/0370-1573(83)90080-7. URL http://www.sciencedirect.com/science/article/pii/0370157383900807.

[41] K. Arnold et al. VBFNLO: A Parton level Monte Carlo for processes with electroweak bosons. *Comput. Phys. Commun.*, 180:1661–1670, 2009. doi: 10.1016/j.cpc.2009.03.006.

[42] ATLAS Collaboration. Measurement of $W^{\pm}Z$ boson pair-production in pp collisions at $\sqrt{s} = 13$TeV with the ATLAS Detector and confidence intervals for anomalous triple gauge boson couplings. ATLAS-CONF-2016-043. URL https://cds.cern.ch/record/2206093.

[43] ATLAS Collaboration. Observation and measurement of Higgs boson decays to WW^* with the ATLAS detector. *Phys. Rev. D*, 92:012006, 2015. doi: 10.1103/PhysRevD.92.012006.

[44] ATLAS Collaboration. Luminosity determination in pp collisions at $\sqrt{s} = 8$ TeV using the ATLAS detector at the LHC. *Eur. Phys. J. C*, 76:653, 2016. doi: 10.1140/epjc/s10052-016-4466-1.

[45] ATLAS Collaboration. Measurement of the $W^{\pm}Z$ boson pair-production cross section in pp collisions at $\sqrt{s} = 13$TeV with the ATLAS Detector. *Phys. Lett. B 762 1*, 2016. doi: 10.1016/j.physletb.2016.08.052.

[46] Laura Barranco Navarro. Alignment of the ATLAS Inner Detector in the LHC Run II. Technical Report ATL-PHYS-PROC-2015-190, CERN, Geneva, Dec 2015. URL http://cds.cern.ch/record/2114708.

[47] Kathrin Becker, Claudia Bertella, Carsten Daniel Burgard, Lucrezia Stella Bruni, Remco Castelijn, João Barreiro Guimarães da Costa, Karri Folan Di Petrillo, Dongshuo Du, Dominik Duda, Pamela Ferrari, Frank Filthaut, Alexander Gavrilyuk, Marc Geisen, Ralf Gugel, Ruchi Gupta, Chris Hays, Pai-hsien Jennifer Hsu, Paul Jackson, Benjamin Paul Jaeger, Karsten Koeneke, Ashutosh Kotwal, Javier Llorente Merino, Ya-feng Lo, Yun-Ju Lu, Lianliang Ma, Jason Oliver, Sophio Pataraia, Sourav Sen, Per Edvin Sidebo, Weimin Song, Christian Schmitt, David Richard Shope, Jonas Strandberg, Mike Strauss, Ilya Tsukerman, Gabija Zemaityte, Yongke Zhao, Meng-ju Tsai, Alexander Naip Tuna, and Martin White. Optimisation note of the ggF+VBF analysis in $H \to WW^*$ using 36 fb^{-1} of data collected with the ATLAS detector at $\sqrt{s} = 13$ TeV. Technical Report ATL-COM-PHYS-2017-1089, CERN, Geneva, Jul 2017. URL https://cds.cern.ch/record/2276101.

[48] Kathrin Becker, Claudia Bertella, Carsten Daniel Burgard, Lucrezia Stella Bruni, Remco Castelijn, João Barreiro Guimarães da Costa, Karri Folan Di Petrillo, Dongshuo Du, Dominik Duda, Pamela Ferrari, Frank Filthaut, Alexander Gavrilyuk, Marc Geisen, Ralf Gugel, Ruchi Gupta, Chris Hays, Pai-hsien Jennifer Hsu, Paul Jackson, Benjamin Paul Jaeger, Karsten Koeneke, Ashutosh Kotwal, Javier Llorente Merino, Ya-feng Lo, Yun-Ju Lu, Lianliang Ma, Jason Oliver, Sophio Pataraia, Sourav Sen, Per Edvin Sidebo, Weimin Song, Christian Schmitt, David Richard Shope, Jonas Strandberg, Mike Strauss, Ilya Tsukerman, Gabija Zemaityte, Yongke Zhao, Meng-ju Tsai, Alexander Naip Tuna, and Martin White. Measurements of the Higgs boson production cross section via ggF and VBF in $H \to WW^* \to \ell\nu\ell\nu$ with 36.1 fb^{-1} of data collected with the ATLAS detector at $\sqrt{s} = 13$ TeV. Technical Report ATL-COM-PHYS-2017-1094, CERN, Geneva, Jul 2017. URL https://cds.cern.ch/record/2276143.

[49] J. Benecke, T. T. Chou, C. N. Yang, and E. Yen. Hypothesis of limiting fragmentation in high-energy collisions. *Phys. Rev.*, 188:2159–2169, Dec 1969. doi: 10.1103/PhysRev.188.2159. URL https://link.aps.org/doi/10.1103/PhysRev.188.2159.

[50] Paolo Bolzoni, Fabio Maltoni, Sven-Olaf Moch, and Marco Zaro. Higgs production via vector-boson fusion at NNLO in QCD. *Phys. Rev. Lett.*, 105:011801, 2010. doi: 10.1103/PhysRevLett.105.011801.

[51] L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification and Regression Trees*. Wadsworth and Brooks, Monterey, CA, 1984.

[52] Oliver Sim Brüning, Paul Collier, P Lebrun, Stephen Myers, Ranko Ostojic, John Poole, and Paul Proudlock. *LHC Design Report, 1*. Technical Design Report LHC. CERN, Geneva, 2004.

[53] Oliver Sim Brüning, Paul Collier, P Lebrun, Stephen Myers, Ranko Ostojic, John Poole, and Paul Proudlock. *LHC Design Report, 2*. Technical Design Report LHC. CERN, Geneva, 2004.

[54] Oliver Sim Brüning, Paul Collier, P Lebrun, Stephen Myers, Ranko Ostojic, John Poole, and Paul Proudlock. *LHC Design Report, 3*. Technical Design Report LHC. CERN, Geneva, 2004.

[55] Andy Buckley et al. General-purpose event generators for LHC physics. *Phys. Rept.*, 504:145–233, 2011. doi: 10.1016/j.physrep.2011.03.005.

[56] Matteo Cacciari, Gavin P. Salam, and Gregory Soyez. The anti-k_t jet clustering algorithm. *JHEP*, 04:063, 2008. doi: 10.1088/1126-6708/2008/04/063.

[57] Valentina Maria Cairo, Graham Richard Lee, Dave Casper, Izaac Gregory Sanderswood, Nora Emilia Pettersson, and Matthias Danninger. Development of ATLAS Primary Vertex Reconstruction for LHC Run 3. Technical Report ATL-COM-PHYS-2019-158, CERN, Geneva, Mar 2019. URL https://cds.cern.ch/record/2665238.

[58] John M. Campbell, R. Keith Ellis, and Ciaran Williams. Gluon-Gluon Contributions to W+ W- Production and Higgs Interference Effects. *JHEP*, 10:005, 2011. doi: 10.1007/JHEP10(2011)005.

[59] Fabrizio Caola, Kirill Melnikov, Raoul Röntsch, and Lorenzo Tancredi. QCD corrections to W^+W^- production through gluon fusion. *Phys. Lett.*, B754:275–280, 2016. doi: 10.1016/j.physletb.2016.01.046.

[60] M Capeans, G Darbo, K Einsweiler, M Elsing, T Flick, M Garcia-Sciveres, C Gemme, H Pernegger, O Rohne, and R Vuillermet. ATLAS Insertable B-Layer Technical Design Report. Technical Report CERN-LHCC-2010-013. ATLAS-TDR-19, CERN, Sep 2010.

[61] F. Cascioli, S. Höche, F. Krauss, P. Maierhöfer, S. Pozzorini, and F. Siegert. Precise Higgs-background predictions: merging NLO QCD and squared quark-loop corrections to four-lepton + 0,1 jet production. *JHEP*, 01:046, 2014. doi: 10.1007/JHEP01(2014)046.

[62] M. Ciccolini, Ansgar Denner, and S. Dittmaier. Strong and electroweak corrections to the production of Higgs + 2jets via weak interactions at the LHC. *Phys. Rev. Lett.*, 99:161803, 2007. doi: 10.1103/PhysRevLett.99.161803.

[63] T G Cornelissen, N Van Eldik, M Elsing, W Liebig, E Moyse, N Piacquadio, K Prokofiev, A Salzburger, and A Wildauer. Updates of the ATLAS Tracking Event Data Model (Release 13). Technical Report ATL-SOFT-PUB-2007-003. ATL-COM-SOFT-2007-008, CERN, Geneva, Jun 2007. URL https://cds.cern.ch/record/1038095.

[64] T G Cornelissen, M Elsing, I Gavrilenko, J-F Laporte, W Liebig, M Limper, K Nikolopoulos, A Poppleton, and A Salzburger. The global χ^2 track fitter in ATLAS. *Journal of Physics: Conference Series*, 119(3):032013, jul 2008. doi: 10.1088/1742-6596/119/3/032013. URL https://doi.org/10.1088%2F1742-6596%2F119%2F3%2F032013.

[65] Michał Czakon, Paul Fiedler, and Alexander Mitov. Total Top-Quark Pair-Production Cross Section at Hadron Colliders Through $O(\alpha\frac{4}{S})$. *Phys. Rev. Lett.*, 110:252004, 2013. doi: 10.1103/PhysRevLett.110.252004.

[66] François Baron Englert and Robert Brout. Broken Symmetry and the Mass of Gauge Vector Mesons. *Physical Review Letters*, 13:321–323, August 1964. doi: 10.1103/PhysRevLett.13.321.

[67] Lyndon Evans and Philip Bryant. LHC machine. *Journal of Instrumentation*, 3 (08):S08001–S08001, aug 2008. doi: 10.1088/1748-0221/3/08/s08001. URL https://doi.org/10.1088%2F1748-0221%2F3%2F08%2Fs08001.

[68] D Florian et al. Handbook of LHC Higgs Cross Sections: 4. Deciphering the Nature of the Higgs Sector. 2016. doi: 10.23731/CYRM-2017-002.

[69] Jerome H. Friedman. Stochastic gradient boosting. *Comput. Stat. Data Anal.*, 38(4): 367–378, 2002. doi: 10.1016/S0167-9473(01)00065-2.

[70] Stefano Frixione, Eric Laenen, Patrick Motylinski, Bryan R. Webber, and Chris D. White. Single-top hadroproduction in association with a W boson. *JHEP*, 07:029, 2008. doi: 10.1088/1126-6708/2008/07/029.

[71] R. Frühwirth. Application of kalman filtering to track and vertex fitting. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 262(2):444 – 450, 1987. ISSN 0168-9002. doi: https://doi.org/10.1016/0168-9002(87)90887-4. URL http://www.sciencedirect.com/science/article/pii/0168900287908874.

[72] R Frühwirth, Wolfgang Waltenberger, and Pascal Vanlaer. Adaptive Vertex Fitting. Technical Report CMS-NOTE-2007-008, CERN, Geneva, Mar 2007. URL http://cds.cern.ch/record/1027031.

[73] G. Artoni et al. Study of $\ell\ell$+MET with proton-proton collisions at $\sqrt{s} = 13$TeV in the HZZ group: introduction of common analysis strategies. ATL-COM-PHYS-2016-1600. URL https://cds.cern.ch/record/2231582.

[74] E. Gross G. Cowan, K. Cranmer and O. Vitells. Asymptotic formulae for likelihood-based tests of new physics. C71:1554, 2011.

[75] M. Gell-Mann. A schematic model of baryons and mesons. *Physics Letters*, 8(3):214 – 215, 1964. ISSN 0031-9163. doi: https://doi.org/10.1016/S0031-9163(64)92001-3. URL http://www.sciencedirect.com/science/article/pii/S0031916364920013.

[76] Diptimoy Ghosh, Rohini Godbole, Monoranjan Guchait, Kirtimaan Mohan, and Dipan Sengupta. Looking for an Invisible Higgs Signal at the LHC. *Phys. Lett.*, B725: 344–351, 2013. doi: 10.1016/j.physletb.2013.07.042.

[77] S. L. Glashow. Partial Symmetries of Weak Interactions. *Nucl. Phys.*, 22:579–588, 1961. doi: 10.1016/0029-5582(61)90469-2.

[78] T. Gleisberg, Stefan. Hoeche, F. Krauss, M. Schonherr, S. Schumann, F. Siegert, and J. Winter. Event generation with SHERPA 1.1. *JHEP*, 02:007, 2009. doi: 10.1088/1126-6708/2009/02/007.

[79] Tanju Gleisberg and Stefan Hoeche. Comix, a new matrix element generator. *JHEP*, 12:039, 2008. doi: 10.1088/1126-6708/2008/12/039.

[80] Jeffrey Goldstone, Abdus Salam, and Steven Weinberg. Broken symmetries. 127: 965–970, Aug 1962. doi: 10.1103/PhysRev.127.965.

[81] David Griffiths. *Introduction to Elementary Particles*. Wiley-VCH, 2004. ISBN 978-3-527-40601-2.

[82] Kathryn Grimm, S. Boutle, D. Casper, B. Hooberman, B. Gui, G. Lee, J. Maurer, A. Morley, S. Pagan Griso, B. Petersen, K. Prokofiev, L. Shan, D. Shope, A. Wharton, B. Whitmore, and M. Zhang. Primary vertex reconstruction at the ATLAS experiment. Technical Report ATL-SOFT-PROC-2017-051. 4, CERN, Geneva, Feb 2017. URL https://cds.cern.ch/record/2253428.

[83] Gerald Stanford Guralnik, Carl Richard Hagen, and Tom Walter Bannerman Kibble. Global Conservation Laws and Massless Particles. *Physical Review Letters*, 13:585–587, November 1964. doi: 10.1103/PhysRevLett.13.585.

[84] F. Halzen and Alan D. Martin. *QUARKS AND LEPTONS: AN INTRODUCTORY COURSE IN MODERN PARTICLE PHYSICS*. 1984. ISBN 0471887412, 9780471887416.

[85] Keith Hamilton, Paolo Nason, Emanuele Re, and Giulia Zanderighi. NNLOPS simulation of Higgs boson production. *JHEP*, 10:222, 2013. doi: 10.1007/JHEP10(2013)222.

[86] Lukas Heinrich. The ATLAS Trigger Core Configuration and Execution System in Light of the ATLAS Upgrade for LHC Run 2. Technical Report ATL-DAQ-PROC-2015-016. 8, CERN, Geneva, May 2015. URL http://cds.cern.ch/record/2016643.

[87] Werner Herr and B Muratori. Concept of luminosity. 2006. doi: 10.5170/CERN-2006-002.361. URL http://cds.cern.ch/record/941318.

[88] Peter Ware Higgs. Broken Symmetries and the Masses of Gauge Bosons. *Physical Review Letters*, 13:508–509, October 1964. doi: 10.1103/PhysRevLett.13.508.

[89] Peter Ware Higgs. Spontaneous symmetry breakdown without massless bosons. *Phys. Rev.*, 145:1156–1163, May 1966. doi: 10.1103/PhysRev.145.1156.

[90] Stefan Höche. Introduction to parton-shower event generators. In *Proceedings, Theoretical Advanced Study Institute in Elementary Particle Physics: Journeys Through the Precision Frontier: Amplitudes for Colliders (TASI 2014): Boulder, Colorado, June 2-27, 2014*, pages 235–295, 2015. doi: 10.1142/9789814678766˙0005.

[91] Bora Isildak. *Measurement of the differential dijet production cross section in proton-proton collisions at* $\sqrt{s} = 7$ *tev*. PhD thesis, Bogazici U., 2011.

[92] Matthew Henry Klein, Francesco Rubbo, and Ariel Schwartzman. Forward Jet Vertex Tagging: A new technique for the identification and rejection of forward pileup jets. Technical Report ATL-COM-PHYS-2015-723, CERN, Geneva, Jul 2015. URL https://cds.cern.ch/record/2034507.

[93] Hung-Liang Lai, Marco Guzzi, Joey Huston, Zhao Li, Pavel M. Nadolsky, Jon Pumplin, and C. P. Yuan. New parton distributions for collider physics. *Phys. Rev.*, D82:074024, 2010. doi: 10.1103/PhysRevD.82.074024.

[94] D. J. Lange. The EvtGen particle decay simulation package. *Nucl. Instrum. Meth.*, A462:152–155, 2001. doi: 10.1016/S0168-9002(01)00089-4.

[95] C. G. Lester and D. J. Summers. Measuring masses of semiinvisibly decaying particles pair produced at hadron colliders. *Phys. Lett.*, B463:99–103, 1999. doi: 10.1016/S0370-2693(99)00945-4.

[96] Michelangelo L. Mangano, Mauro Moretti, Fulvio Piccinini, Roberto Pittau, and Antonio D. Polosa. ALPGEN, a generator for hard multiparton processes in hadronic collisions. *JHEP*, 07:001, 2003. doi: 10.1088/1126-6708/2003/07/001.

[97] A. D. Martin, W. J. Stirling, R. S. Thorne, and G. Watt. Parton distributions for the lhc. *The European Physical Journal C*, 63(2):189–285, Sep 2009. ISSN 1434-6052. doi: 10.1140/epjc/s10052-009-1072-5. URL https://doi.org/10.1140/epjc/s10052-009-1072-5.

[98] Kirill Melnikov and Frank Petriello. Electroweak gauge boson production at hadron colliders through $\mathcal{O}(\alpha_s^2)$. *Phys. Rev. D*, 74:114017, Dec 2006. doi: 10.1103/PhysRevD.74.114017. URL https://link.aps.org/doi/10.1103/PhysRevD.74.114017.

[99] Esma Mobs. The CERN accelerator complex. Complexe des accélérateurs du CERN. Jul 2016. URL http://cds.cern.ch/record/2197559. General Photo.

[100] P. Nason and C. Oleari. NLO Higgs boson production via vector-boson fusion matched with shower in POWHEG. *JHEP*, 1002:037, 2010. doi: 10.1007/JHEP02(2010)037.

[101] K. A. Olive et al. Review of Particle Physics. *Chin. Phys.*, C38:090001, 2014. doi: 10.1088/1674-1137/38/9/090001. URL http://pdg.lbl.gov.

[102] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[103] Michael Edward Peskin and Daniel V. Schroeder. *An introduction to quantum field theory*. Westview Press Reading, 1995. ISBN 978-0-201-50397-5.

[104] T. Plehn, David L. Rainwater, and D. Zeppenfeld. A Method for identifying $H \to \tau^+\tau^- \to e^\pm\mu^\mp p_T$ at the CERN LHC. *Phys. Rev.*, D61:093005, 2000. doi: 10.1103/PhysRevD.61.093005.

[105] Steffen Schumann and Frank Krauss. A Parton shower algorithm based on Catani-Seymour dipole factorisation. *JHEP*, 03:038, 2008. doi: 10.1088/1126-6708/2008/03/038.

[106] Matthew D. Schwartz. *Quantum Field Theory and the Standard Model*. Cambridge University Press, 2013. ISBN 978-1-107-03473-0.

[107] Torbjorn Sjostrand, Stephen Mrenna, and Peter Z. Skands. PYTHIA 6.4 Physics and Manual. *JHEP*, 05:026, 2006. doi: 10.1088/1126-6708/2006/05/026.

[108] Iain W. Stewart and Frank J. Tackmann. Theory Uncertainties for Higgs and Other Searches Using Jet Bins. *Phys. Rev.*, D85:034011, 2012. doi: 10.1103/PhysRevD.85.034011.

[109] James Stirling. Parton luminosity and cross section plots. URL `http://www.hep.ph.ic.ac.uk/~wstirlin/plots/plots.html`.

[110] T. Sjöstrand, S. Mrenna, and P. Z. Skands. A brief introduction to PYTHIA 8.1. *Computer Physics Communications*, 178(11):852–867, 2008.

[111] AC Team. The four main LHC experiments. Jun 1999. URL `https://cds.cern.ch/record/40525`.

List of Figures

List of Tables