Measurement of $H \rightarrow b\bar{b}$ in Associated Production with the CMS Detector

by

Daniel Robert Abercrombie

B.S., Pennsylvania State University (2014)

Submitted to the Department of Physics
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2021

Author . . . .
Department of Physics
April 22, 2021

Certified by . . . .
Christoph M. E. Paus
Professor of Physics
Thesis Supervisor

Accepted by . . . .
Deepto Chakrabarty
Associate Department Head, Physics

Measurement of $H \rightarrow b\bar{b}$ in Associated Production with the CMS Detector

by

Daniel Robert Abercrombie

Submitted to the Department of Physics
on April 22, 2021, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

The differential cross section of $VH \rightarrow b\bar{b}$ is measured with the CMS Detector. The Simplified Template Cross Section framework is used. The inclusive strength of the measured signal relative to the Standard Model is $0.568^{+0.154}_{-0.147}(\text{stat})^{+0.134}_{-0.133}(\text{sys})$, which agrees with the Standard Model within 2.1 standard deviations. The measured spectrum of the recoiling vector boson transverse momentum has a p-value of 9.3%, assuming Standard Model predictions at the measured signal strength.

Thesis Supervisor: Christoph M. E. Paus
Title: Professor of Physics

Acknowledgments

I also would like to thank Christoph Paus, my thesis advisor, for giving me immense freedom to explore and educate myself during my time at MIT and for his patience while I eventually got around to finishing up this thesis.

I have the usual fear of leaving out individuals if I were to continue this section with names, so instead I will try to paint my positive experiences of the past seven years in broad strokes. Those who knew me will know where they fit in. Before arriving at MIT, I possessed a solid foundation in physics education. This was thanks to instructors and classmates that I had in high school, while taking classes at Lycoming College, and while obtaining my Bachelor degrees at Penn State. When I started at MIT, I met many exceptional students whom I explored Boston with at bars and concerts. Friends that I shared classes with continued to aid my education, and those I shared offices or floors with also aided in my research.

I also had a positive experience outside of work. I enjoyed my time in the Isshinryu Karate-do Club and in the IM sports of Air Pistol, Tennis, and Softball. I then found one of my favorite hobbies thanks to a welcoming group of powerlifters at the MIT gym. Despite being separated from this first group of "gym bros" due to the events of 2020, I was able to meet other motivational and kind people in weight rooms in Indiana and New Mexico. The climbers I met in Albuquerque have also encouraged me to experiment in one of the popular local hobbies. Outside of physical endeavors, a group of friends rescued me from creative torpidity by participating in games of Pathfinder. I also want to thank those friends from high school and Penn State that kept in touch through regular video calls and gave me a perspective of life outside of MIT at that time.

To end, I will return to individual names for those that I do and did consider immediate family. Though it is painful to remember how we grew apart as we learned

Contents

List of Figures

List of Tables

Chapter 1

Introduction

One of the most curious features of physics at small scales, which will likely frustrate students for the rest of time, is that certain events are not deterministic and only have a probability of happening. There is no guarantee that the initial state of an electron and a positron approaching each other will annihilate and produce the final state of a muon and an anti-muon, even if their energies in the center-of-mass frame are adequate for muon production. However, trying long enough with the same initial conditions will eventually produce a muon and anti-muon pair. Furthermore, the observation of resonances, where this is more likely to happen when the electron and positron approach each other at particular energies, does not mean that a Z boson was present in a given interaction. It just means that the weak component of the electroweak force significantly increases the probability of the muonic final state, when the total energy of the system is close to the mass of the Z boson. The sum of probabilities from different possible field interactions leading from a specific initial condition to a particular final state is the only thing that can be measured. This measurement is only done accurately when observing many events with the same initial conditions.

This point is difficult to convey concisely, so many laypeople, as well as some practicing physicists, are confused by the terminology adopted by the field. But this distinction is relevant to the topic of this work. This document presents a measurement of a cross section. Cross section is the name given to the probability of an

interaction occurring. Reported cross sections can be split up to describe different contributions to final states, and they can be collated into what are called "production cross sections" which describe the probabilities of particular intermediate states "occurring" (even though intermediate states never exist in reality). In the example above, the Z boson would be a possible intermediate state.

The main point of the example above is that physicists can learn about the Z boson by observing only electrons and muons. They never see the Z boson itself. This can be generalized to any interesting particle that interacts with other particles. Comparing only the initial and final particles is most common way measurements are done today in the field of particle physics. An increased probability of certain initial states resulting in certain final states can teach the observer much about the role of the intermediate particles without ever directly seeing them.

1.1   Measurement of the Higgs Cross Section

The purpose of the following document is to present the methods for and most recent results of measuring the strength of the coupling between the Higgs boson and bottom quarks. In this context, the Higgs boson makes up one of the previously mentioned intermediate states that cannot be shown as present in a given event. The coupling strength is directly related to the contribution the Higgs field has on the probability of having a final state with bottom quarks. However, the cross section measurement also relies on a number of other physics processes.

To measure this coupling, the Higgs boson must first be "produced" before measuring its coupling strength to bottom quarks. This analysis takes advantage of a process known as associated production, where a vector boson, one of the intermediate particles of the weak nuclear force, couples to and radiates a Higgs boson. This means the featured measurement includes information about the Higgs coupling to vector bosons. The vector boson itself is produced by the collision of high energy protons, so it too is an intermediate state that precedes the Higgs boson.

The measurement is not complete once the intermediate states are generated. The

Higgs boson can decay into a number of different particles, with bottom quarks being only one type, though it is predicted to be the most frequent Higgs decay. The vector bosons have multiple decay modes as well. In this analysis, we only use the leptonic decays because these give us the cleanest signature, where enough of the contribution to the final state probability is from associated production for it to be measured. The clean signal also gives a more accurate measurement of the overall energies involved in each event. This means that instead of looking for a Higgs and a vector boson, we must look for b quarks and leptons and measure these final state yields near the appropriate energy resonances.

There are also other physics processes that create the same final states, as well as processes that create final states that look similar enough to be practically indistinguishable. These processes must also be well understood through thorough study before a Higgs boson cross section measurement can be successfully completed.

1.2   Motivation for the Measurement

The measurement of a cross section of a known particle is "normal science," and that is the space in which this analysis operates [1]. Much of the community of physics researchers have been operating under the paradigm of the Standard Model [2] for the better part of a century. The Standard Model has known gaps, such as lack of explanation for neutrino mass, as well as the origin of Dark Matter and Dark Energy. However, none of these research fields have yielded any results that will trigger a paradigm shift. In fact, the most exciting discoveries of new particles, such as the weak W [3] and Z [4] bosons and the Higgs boson [5, 6] only confirmed predictions by the Standard Model.

In the meanwhile, precision measurements are performed on processes that we expect to already understand very well. Repeating measurements while the state of the art is improving is interesting, no matter the outcome. Over time, the uncertainty in the measurement outcome shrinks, leading to more precise knowledge of parameters of the Standard Model. If the parameters' precision eventually causes excessive

tension in that they cannot all exist at their measured value assuming the Standard Model is true, the discrepancies would need to be explained by a different or amended model.

The Higgs decaying to $b\bar{b}$, or a bottom quark and bottom anti-quark, was first observed in 2018 [7, 8]. The measurement outlined in this document goes further in that it measures the contribution of $H \rightarrow b\bar{b}$ in associated production to final states possessing different energies. This is called a differential cross section, and places greater constraints on the parameters of the Standard Model because the undetectable intermediate states play a large role in the energy distribution of the final state. Additional constraints lead to more precise measurements of the parameters, and have the potential to discover discrepancies that have hitherto been missed.

If discrepancies arise, not only in the frequency of events, but also the energy spectrum of the events, that means there are intermediate interactions that are not accounted for. The Standard Model describes all possible interactions between the particles we know of. If there is evidence of additional interactions, then additional particles must exist to allow them. Alternatively, the Higgs boson itself might not be the type of boson we expect it to be.

1.3 Using the CMS Detector at the LHC

This measurement is only possible due to massive efforts by the scientific community. Interactions near the Higgs boson resonance are not generated by typical conditions on Earth. There have been multiple colliders created over the years that attempted to find evidence of the Higgs boson. The highest energy hadron collider before the LHC, the Tevatron at Fermilab reached collision energies of nearly $2\,\mathrm{TeV}$ and ran for two decades [9], but was unable to discover the Higgs. At CERN, the Large Electron-Positron (LEP) collider [10] ran at energies up to $200\,\mathrm{GeV}$. In principle, this was enough energy to generate Higgs bosons, but the collider was decommissioned with only hints of the Higgs boson at LEP. The Large Hadron Collider (LHC) was built in the same tunnels as LEP, and required the efforts of thousands of scientists and

engineers, as well as funding from countries distributed all around the globe. The LHC performs proton-proton collisions at $13\,\mathrm{TeV}$, which we now know is more than enough energy to produce Higgs bosons.

To observe the final states of collisions at the LHC, multiple detectors have also been constructed, due to the efforts of hundreds or thousands of individuals. This analysis is done using data from the Compact Muon Solenoid (CMS). Other experiments are ATLAS, ALICE, LHCb, TOTEM, LHCf, MoDEL, and FASER. The CMS detector is a general purpose detector, which was used in the discovery of the Higgs boson, along with ATLAS. Though it is impossible to make a perfect detector, CMS utilizes a number of state-of-the-art technologies and analysis techniques to achieve unprecedented precision in this measurement of $VH b\bar{b}$, as well as other Standard Model measurements and searches for physics beyond the Standard Model.

Chapter 2

Theory

The Standard Model describes the interactions of all observable matter. There are many textbooks that cover the Standard Model, as there are many scientists and students who study it. For an in depth presentation of the Standard Model beyond what is presented in this chapter, please refer to Reference [11].

Matter is made up of 12 kinds of fermions. The forces between the fermions are mediated by the gauge fields created by the Standard Model's SU(3) $\times$ SU(2)$_L$ $\times$ U(1)$_Y$ symmetry. Fermions with the appropriate charge are affected by the gauge fields. Fermions are classified as quarks or leptons. There are six types of quarks, separated into three generations of two quarks each. Each quark has a color charge associated with the SU(3) symmetry. The interactions arising from this is called QCD. Each pair of quarks in a family also have an approximate SU(2) symmetry, which allows interactions via the weak force. Quarks also have a hypercharge, which is a relation of electromagnetic charge and weak isospin, meaning the gauge field from the U(1) symmetry affects them as well. Three charged leptons and three neutral leptons, called neutrinos, comprise the other six fermions. The leptons do not carry a color charge, so they are not affected by the SU(3) symmetry, but they do carry weak isospin and hypercharge. Table 2.1 displays the values of these charges for all fermions. Fermions also have anti-particles, which carry the opposite charges of their counterparts.

In the Standard Model, the gauge files from the SU(2)$_L$ $\times$ U(1)$_Y$ symmetries

Table 2.1: All of the fermions are listed below, along with their charges and weak isospin values. The three generations are listed from least to most massive, meaning only the first generation of quarks and charged leptons is stable. The masses and decays of neutrinos is beyond the scope of the Standard Model and this analysis.

	1st gen.	2nd gen.	3rd gen.	Color	Q	$I_W^{(3)}$	Y
down-type quarks	d	s	b	yes	$-\frac{1}{3}$	$-\frac{1}{2}$	$-\frac{1}{3}$
up-type quarks	u	c	t	yes	$+\frac{2}{3}$	$+\frac{1}{2}$	$+\frac{5}{3}$
charged leptons	e	μ	τ	no	-1	$-\frac{1}{2}$	-1
neutral leptons	ν_e	ν_μ	ν_τ	no	0	$+\frac{1}{2}$	-1

are mixed into what is known as the electroweak force. The electroweak force is mediated by the neutral photon and Z boson, and the charged W boson, which are also known as collectively as vector bosons. This mixing happens due to the Higgs boson, a scalar which grants all of the charged fermions and the Z and W bosons mass through interactions. The interactions allow associated production, where a Higgs boson is created through radiation from either a Z or W boson. The measurement of $H \to b\bar{b}$ in associated production, the focus of this work, is therefore a measurement of interactions between the Higgs boson and vector bosons and the coupling between the Higgs boson and bottom, or b, quarks. The generation and observable decay of the vector bosons are also important to make this measurement. The parameters for those interactions are measured more accurately by other analyses not involving an observation of the Higgs [12]. Since the coupling of the Higgs boson with the vector bosons and with fermions also gives rise to the masses of each in the Standard Model through what is called electroweak symmetry breaking or the Higgs mechanism, most discussions of Higgs couplings include an explanation of the Higgs mechanism.

The treatment of these topics in this chapter are arranged as follows. First, I will give a brief explanation of Higgs field's non-zero vacuum energy, a trait that makes the electroweak symmetry breaking possible. After that, the coupling of the Higgs boson to the W and Z vector bosons will be described. These interactions arise from a mixing of the weak force with the electromagnetic force. What results is collectively known as the electroweak force. The coupling of the electroweak force to fermions is

then discussed to understand both the generation of the vector boson intermediate states and the resulting final state that can be observed. Finally, the decay of the Higgs boson itself into bottom quarks is explained. This is allowed because the Higgs boson couples directly to massive fermions and gives them mass through the Higgs mechanism.

The Standard model is described by its Lagrangian, which is the difference between a system's kinetic and potential energies. Equations of motion are extracted from a Lagrangian $\mathcal{L}$ for a particle field ϕ_i using the Euler-Lagrange equations.

$$\delta_\mu \left(\frac{\delta\mathcal{L}}{\delta(\delta_\mu\phi_i)} \right) - \frac{\delta\mathcal{L}}{\delta\phi_i} = 0 \tag{2.1}$$

In order to make predictions of particle interactions and behavior, the Lagrangian must be determined. The Lagrangian is constrained by the Standard Model. The Standard Model has $\mathrm{SU}(3) \times \mathrm{SU}(2)_L \times \mathrm{U}(1)_Y$ symmetry, which means the Lagrangian must behave identically under the different transformations of these symmetry. Gauge bosons are produced by these symmetries. The $\mathrm{SU}(3)$ symmetry produces gluons which mediate the strong force between quarks [13]. The $\mathrm{SU}(2)_L \times \mathrm{U}(1)_Y$ symmetry produces two gauge fields. The first interacts with left handed fermions, and the second interacts with all fermions through their hypercharge, Y. These forces are ultimately mixed into what is known as the electroweak force due to electroweak symmetry breaking [2]. Electroweak symmetry breaking is the required solution of the problem that vector gauge bosons of the electroweak force had mass. It was not possible to grant these bosons mass while maintaining the symmetries of the Standard Model without the Higgs boson [14–16]. The granting of mass happens for two reasons: the Higgs field has a non-zero vacuum expectation value, and the Higgs field couples directly to vector boson and massive fermion fields.

2.1 Electroweak Symmetry Breaking

First, consider the SU(2) × U(1) symmetry where the Higgs interacts with the electroweak bosons. The L and Y of the $SU(2)_L \times U(1)_Y$ symmetry can be forgotten for the moment, since they describe fermion interactions with the electroweak force. To preserve SU(2) symmetry, the Higgs boson is described as two complex scalar fields in a weak isospin doublet with a quartic potential. The SU(2) symmetry means rotations between the doublet states must be equivalent in the Lagrangian. The Lagrangian for a free Higgs is then

$$\mathcal{L} = (\delta_\mu \phi)^\dagger (\delta^\mu \phi) - (\mu^2(\phi^\dagger \phi) + \lambda(\phi^\dagger \phi)^2) \tag{2.2}$$

Through the virial theorem [17], the potential has a minimum value when

$$\phi^\dagger \phi = \frac{-\mu^2}{2\lambda} = \frac{v^2}{2} \tag{2.3}$$

This potential of the Higgs field breaks the SU(2) × U(1) symmetry of the Standard Model Lagrangian. Through this non-zero vacuum expectation value, the Higgs then has a constant influence in other parts of the Standard Model Lagrangian. In this way, it gives mass to electroweak vector bosons, to itself, and to massive fermions.

The first two sets of masses manifest when we force the SU(2) × U(1) symmetry back onto the Lagrangian in Equation (2.2). The derivatives must be replaced.

$$\delta_\mu \to D_\mu = \delta_\mu + i\frac{g_W}{2}\boldsymbol{\sigma} \cdot \mathbf{W}_\mu + ig'\frac{Y}{2}B_\mu \tag{2.4}$$

To simplify the expansion of Equation (2.2), a particular gauge, or particular doublet state, is chosen. The unitary gauge gives the massless neutral boson known as a photon [18]. In the unitary gauge, ϕ is written as

$$\phi(x) = \frac{1}{\sqrt{2}}\begin{pmatrix} 0 \\ v + h(x) \end{pmatrix} \tag{2.5}$$

This leads to the following expansion for the kinetic term of the Lagrangian.

$$(D_\mu \phi)^\dagger (D^\mu \phi) = \frac{1}{2}(\delta_\mu h)(\delta^\mu h) + \frac{1}{8}g_W^2 (W_\mu^{(1)} + iW_\mu^{(2)})(W^{(1)\mu} - iW^{(2)\mu})(v+h)^2$$
$$+ \frac{1}{8}(g_W W_\mu^{(3)} - g'B_\mu)(g_W W^{((3)\mu} - g'B^\mu)(v+h)^2 \tag{2.6}$$

Terms that are quadratic in terms of the gauge boson fields reveal the mass of the fields. Taking $h(x) \to 0$, the terms for $W^{(1)}$ and $W^{(2)}$ are just

$$\frac{1}{4}g_W^2 v^2 W_\mu^{(1)} W^{(1)\mu} \quad \text{and} \quad \frac{1}{4}g_W^2 v^2 W_\mu^{(2)} W^{(2)\mu},$$

giving the mass.

$$m_W = \frac{1}{2}g_W v \tag{2.7}$$

The quadratic terms for $W^{(3)}$ and B mix to give a non-diagonal mass matrix $\mathbf{M}$.

$$\frac{v^2}{8}\begin{pmatrix} W_\mu^{(3)} & B_\mu \end{pmatrix} \mathbf{M} \begin{pmatrix} W^{(3)\mu} \\ B^\mu \end{pmatrix} = \frac{v^2}{8}\begin{pmatrix} W_\mu^{(3)} & B_\mu \end{pmatrix} \begin{pmatrix} g_W^2 & -g_W g' \\ -g_W g' & g'^2 \end{pmatrix} \begin{pmatrix} W^{(3)\mu} \\ B^\mu \end{pmatrix} \tag{2.8}$$

The non-diagonal matrix allow $W^{(3)}$ and B to mix. Physical states must be represented by a diagonal Hamiltonian. Diagonalizing the term above gives masses of the physical states.

$$\frac{1}{8}v^2 \begin{pmatrix} A_\mu & Z_\mu \end{pmatrix} \begin{pmatrix} 0 & 0 \\ 0 & g_W^2 + g'^2 \end{pmatrix} \begin{pmatrix} A^\mu \\ Z^\mu \end{pmatrix} = \frac{1}{2}\begin{pmatrix} A_\mu & Z_\mu \end{pmatrix} \begin{pmatrix} m_A^2 & 0 \\ 0 & m_Z^2 \end{pmatrix} \begin{pmatrix} A^\mu \\ Z^\mu \end{pmatrix} \tag{2.9}$$

This gives us the masses of the neutral gauge bosons.

$$m_A = 0 \quad \text{and} \quad m_Z = \frac{1}{2}v\sqrt{g_W^2 + g'^2} \tag{2.10}$$

From the simple act of requiring SU(2) × U(1) symmetry on the Lagrangian of a scalar doublet with non-zero vacuum expectation value, the masses of all the

electroweak gauge bosons have been produced.

A similar procedure produces the masses of fermions due to their coupling to the electroweak force. Though there are some differences since left-handed and right-handed chiral fermions are treated differently in the Standard Model Under the $SU(2)_L$ symmetry, left-handed fermions are a doublet L, and right-handed fermions are a singlet R. The combination $\bar{L}\phi$ is invariant under $SU(2)_L$ transformations, so the following symmetric term coupling fermions and the Higgs field can be added to the Lagrangian.

$$\mathcal{L} = -g_f(\bar{L}\phi R + \bar{R}\phi^\dagger L) \tag{2.11}$$

Using the unitary gauge in Equation (2.5), this expands to the following.

$$\mathcal{L}_f = -\frac{g_f}{\sqrt{2}}v\left(\bar{f}_L f_R + \bar{f}_R f_L\right) - \frac{g_f}{\sqrt{2}}h\left(\bar{f}_L f_R + \bar{f}_R f_L\right) \tag{2.12}$$

In Equation (2.12), f refers to the lower field of the fermion's $SU(2)_L$ doublet. The Lagrangian also includes terms for the upper field since the conjugate of ϕ has the same symmetries as ϕ. This behaves like a particle with a fixed mass from the Higgs vacuum expectation value interacting with the Higgs boson. In particular using

$$m_f = \frac{g_f}{\sqrt{2}}v \tag{2.13}$$

gives

$$\mathcal{L} = -m_f\bar{\psi}_f\psi_f - \frac{m_f}{v}\bar{\psi}_f\psi_f h \tag{2.14}$$

2.2   Associated Production

The next thing to consider is the couplings also produced by this process. The couplings will allow us to determine more precisely the parameters above by measuring cross sections.

The physical states of W^+ and W^- bosons can be written as the raising and lowering operators for isospin.

$$W^\pm = \frac{1}{\sqrt{2}} \left(W^{(1)} \mp i W^{(2)}\right) \tag{2.15}$$

The second term of Equation (2.6) can be further expanded.

$$\frac{1}{4} g_W^2 W_\mu^- W^{+\mu} (v+h)^2 = \frac{1}{4} g_W^2 v^2 W_\mu^- W^{+\mu} + \frac{1}{2} g_W^2 v W_\mu^- W^{+\mu} h + \frac{1}{4} g_W^2 W_\mu^- W^{+\mu} h^2 \tag{2.16}$$

The second term on the right hand side of Equation (2.16) gives us the coupling strength of a vertex with a Higgs and two W bosons.

$$g_{HWW} = \frac{1}{2} g_W^2 v = g_W m_W \tag{2.17}$$

The coupling to the Z boson can also be found from Equation (2.9) by substituting $(v+h)^2$ back in for v^2 and extracting the terms proportional to $h Z_\mu Z^\mu$.

$$g_{HZZ} = \frac{1}{2} \left(g_W^2 + g'^2\right) v = \sqrt{g_W^2 + g'^2} m_Z \tag{2.18}$$

When arranged in a way that the W or Z boson radiates the Higgs, as opposed to a Higgs decaying into a pair of W or Z bosons, the process is called associated production or *Higgstrahlung*. The vertex showing associated production is pictured in Figure 2-1.

2.2.1 Coupling Between Vector Bosons and Fermions

The W and Z bosons are themselves intermediate states, never existing in a directly observable manner. They must be produced through interactions with stable fermions. Since the LHC is a hadron collider, the vector bosons' couplings with quarks need to be understood to accurately predict production mechanisms.

Quarks couple to each other through the strong force, resulting from a SU(3)
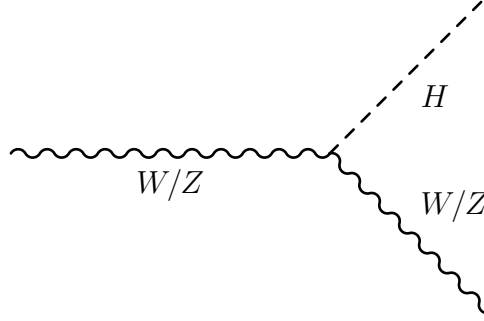
Figure 2-1: Above is the Feynman diagram for associated production. The W or Z boson radiates a Higgs boson. Both bosons later decay into particles detected by CMS.

symmetry. There are three generations of quarks each consisting of a pair of quark types. Their mass eigenstates are denoted as down-type or up-type. A feature of quarks is that their mass eigenstates do not match their weak eigenstates. There is a mixing among the down-type quarks that is parameterized by the Cabibbo-Kobayashi-Maskawa (CKM) matrix.

$$\begin{pmatrix} d' \\ s' \\ b' \end{pmatrix} = \begin{pmatrix} V_{ud} & V_{us} & V_{ub} \\ V_{cd} & V_{cs} & V_{cb} \\ V_{td} & V_{ts} & V_{tb} \end{pmatrix} \begin{pmatrix} d \\ s \\ b \end{pmatrix} \tag{2.19}$$

The mass eigenstates are denoted as d, s, and b, while d', s', and b' are the weak eigenstates. This mixing allows quarks to change generations through interaction with $W^{\pm}$ bosons, which raise or lower the weak isospin. The following is the charge current vertex interaction with the W propagator, which affects the overall cross section based on the propagator four-momentum q.

$$-i\frac{g_W}{\sqrt{2}} \begin{pmatrix} \bar{u} & \bar{c} & \bar{t} \end{pmatrix} \gamma^{\mu} \frac{1}{2}(1-\gamma^5) \begin{pmatrix} V_{ud} & V_{us} & V_{ub} \\ V_{cd} & V_{cs} & V_{cb} \\ V_{td} & V_{ts} & V_{tb} \end{pmatrix} \begin{pmatrix} d \\ s \\ b \end{pmatrix} \times \frac{1}{q^2 - m_W^2}$$

The vertices for this interaction is shown in Figure 2-2 arranged in a way to show
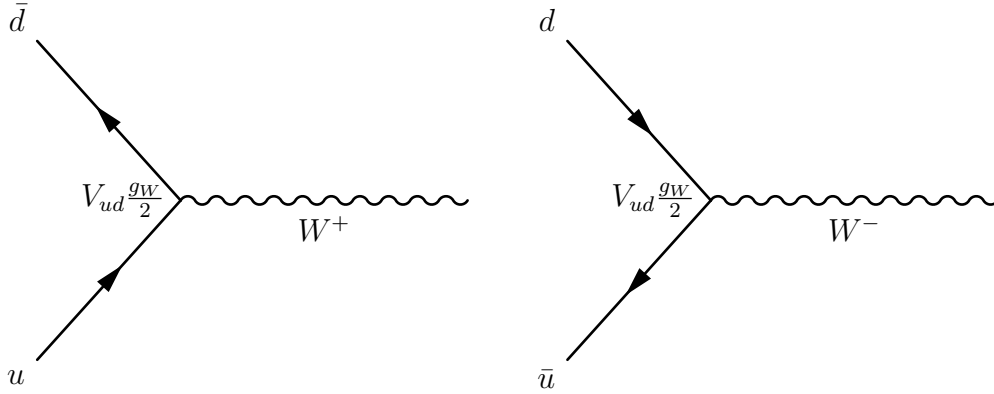
Figure 2-2: Above are diagrams for generating W^+ and W^- bosons. The u and d quarks in the diagram can be replaced with any up-type or down-type quark, respectively. The CKM matrix element would in the vertex element would be changed accordingly.

the processes of generating a W^+ or W^- boson from annihilating quarks. The γ matrices in the interaction are present because the SU(2) component of the Standard Model only interacts with left-handed fermions and right-handed anti-fermions. For this reason, the SU(2) component is more accurately labeled SU(2)$_L$. From Equation (2.15), the $W^\pm$ bosons are completely made up of the $W^{(1)}$ and $W^{(2)}$ components of the SU(2)$_L$, so they also only interact with left-handed fermions and right-handed anti-fermions.

For this analysis, it is important to note that the CKM matrix is unitary and $V_{tb} > 0.975$ [19] is close to unity. The couplings between the third generation of quarks and lighter generations, in particular V_{cb} and V_{ub} are small. This means the matrix element weak decays of the b hadrons is small. Since t quarks are so massive, b quarks or hadrons cannot decay into t quarks. Going to c or u is the only decay channel available to the lightest b hadrons, so their lifetimes are relatively long.

Next, the derivation of Z boson couplings to fermions is considered. The photon and the Z boson are both linear combinations of the SU(2)$_L$ and U(1)$_Y$ components of the Standard Model. From the mixing in Equation (2.8), the photon and Z fields

31

can be expressed as the following.

$$A_\mu = B_\mu \cos\theta_W + W_\mu^{(3)} \sin\theta_W \qquad (2.20)$$

$$Z_\mu = -B_\mu \sin\theta_W + W_\mu^{(3)} \cos\theta_W \qquad (2.21)$$

θ_W is known as the weak mixing angle. The relative strengths of the B and $W^{(3)}$ couplings are determined directly through lepton electromagnetic characteristics, keeping in mind that $W^{(3)}$ only interacts with left handed particles. The strength of the Z boson couplings can be extracted by exploiting the symmetry of photon interactions. The photon interacts the same with left and right handed charged fermions, and not at all with neutral fermions. This is shown directly with experiments with leptons. The charged leptons, electrons, muons, and taus, interact with photons, while the respective neutrinos do not. The following are therefore the electromagnetic interaction strengths of left- and right-handed electrons and neutrinos.

$$e_L: \qquad Qe = \frac{1}{2}g'Y_{e_L}\cos\theta_W - \frac{1}{2}g_W\sin\theta_W \qquad (2.22)$$

$$\nu_L: \qquad 0 = \frac{1}{2}g'Y_{\nu_L}\cos\theta_W - \frac{1}{2}g_W\sin\theta_W \qquad (2.23)$$

$$e_R: \qquad Qe = \frac{1}{2}g'Y_{e_R}\cos\theta_W \qquad (2.24)$$

$$\nu_R: \qquad 0 = \frac{1}{2}g'Y_{\nu_R}\cos\theta_W \qquad (2.25)$$

Y_{e_L} and Y_{ν_L} must be equal to maintain $\mathrm{SU}(2)_L$ symmetry. To satisfy these constraints, the follow definition of Y is needed.

$$Y = 2\left(Q - I_W^{(3)}\right) \qquad (2.26)$$

The following relationship also arises from these experimental constraints.

$$e = g_W \sin\theta_W = g' \cos\theta_W \qquad (2.27)$$
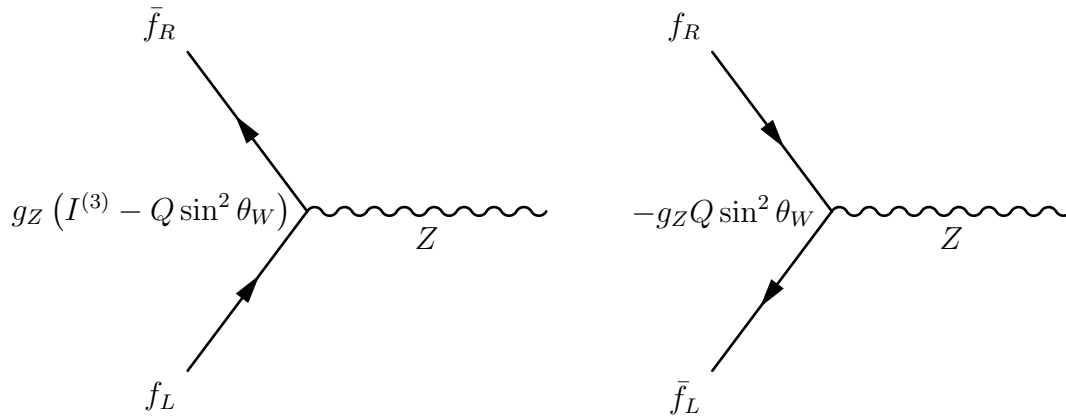
Figure 2-3: Above are diagrams for generating Z bosons. Left- and right-handed fermions are both coupled to, but with different coupling strengths.

Returning to the Z boson, from Equation (2.21), and defining

$$g_Z = \frac{e}{\sin\theta_W \cos\theta_W}, \tag{2.28}$$

we have the following couplings to left- and right-handed fermions.

$$-\frac{1}{2}g'\sin\theta_W (Y_{f_L}\bar{u}_L\gamma^\mu u_L + Y_{f_R}\bar{u}_R\gamma^\mu u_R) + I_W^{(3)}g_W \cos\theta_W (\bar{u}_L\gamma^\mu u_L) =$$
$$g_Z \left(\left(I^{(3)} - Q\sin^2\theta_W \right) \bar{u}_L\gamma^\mu u_L - Q\sin^2\theta_W \bar{u}_R\gamma^\mu u_R \right) \tag{2.29}$$

Now the coupling of the Z to left- and right-handed quarks can be calculated from Table 2.1, remembering that $I_W^{(3)}$ for right-handed fermions is 0. Diagrams showing the interaction strengths of fermion-Z vertices are shown in Figure 2-3.

Thus vector bosons couple to quarks, the constituents of hadrons, which means they can be produced at the LHC. As mentioned earlier in this section, quarks interact through an SU(3) symmetry that results in the strong force. The three states that this symmetry supports are known as color states, and they are labeled red, green, and blue, or r, g, and b. There are also anti-states for each color state, labeled $\bar{r}$, $\bar{g}$, and $\bar{b}$. The resulting gauge bosons are known as gluons, and they carry the following color states.

$$r\bar{g}, g\bar{r}, r\bar{b}, b\bar{r}, g\bar{b}, b\bar{g}, \frac{1}{\sqrt{2}}(r\bar{r} - g\bar{g}) \text{ and} \frac{1}{\sqrt{6}}(r\bar{r} + g\bar{g} - 2b\bar{b})$$

Since gluons carry color charge, they interact with other gluons. As the distance between two color-charged particles grows, the energy density of the self-interacting gluon field remains constant. It soon becomes energetically favorable for new a particle/anti-particle pair to pop into existence if it simultaneously reduces the distance that the strong force is interacting. As a result, all observable hadronic states are color singlets. The most common hadronic states are mesons, made of a quark/anti-quark pair with the color singlet state

$$\psi(q\bar{q}) = \frac{1}{\sqrt{3}}(r\bar{r} + g\bar{g} + b\bar{b}), \tag{2.30}$$

and baryons, made of three quarks with the following color singlet state.

$$\psi(qqq) = \frac{1}{\sqrt{6}}(rgb - rbg + gbrgrb + brg - bgr) \tag{2.31}$$

Baryons can also be composed of three anti-quarks, which has a state corresponding to Equation (2.31), but with anti-color. The resulting sprays of hadronic particles generated by the vacuum to screen the color charge of high energy quarks are called jets.

For this measurement, protons are collided at the LHC. The proton consists of two u quarks, and one d quark. Since the three quarks inside the proton interact strongly, there are also many virtual gluons and quark/anti-quark pairs present at all times. The quantity and energies of all these partons are not able to be calculated since QCD is non-perturbative. They can be measured in deep inelastic scattering experiments though. In these, electrons are scattered off of protons, and parton distribution functions (PDFs) can be measured. The PDFs for protons are shown in Figure 2-4.

Combining the known proton energy, PDFs, the CKM matrix, and the theory of the electroweak force, we can predict the cross section of generating W and Z bosons at the LHC. These initial vector bosons will be off-shell, which means they will have a mass much different than the resonance peak. Then they will radiate a Higgs in order to most commonly produce an on-shell vector boson and on-shell Higgs boson. The
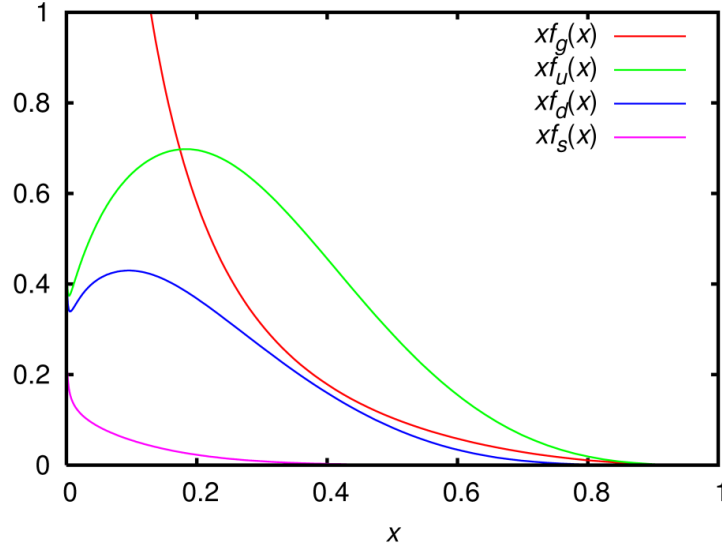
Figure 2-4: The Parton Distribution Function for protons is shown above. Most of the proton's momentum is carried by u and d quarks, but virtual s quarks as well as gluons can also interact with particles passing through the proton.

cross section of generating off-shell particles are suppressed according to the required center-of-mass energy, E, and the resonance mass M. The suppression is in the form of the relativistic Breit-Wigner formula.

$$f(E) = \frac{k}{(E^2 - M^2)^2 + M^2\Gamma^2} \tag{2.32}$$

This associated production is one of three production mechanisms of the Higgs Boson. The other two are gluon fusion, where gluons form a top loop, and vector boson fusion, both shown in Figure 2-5. In these other two production mechanisms, only the Higgs is in the final state. These events can only offer additional identification through initial state radiation. In contrast, associated production also results in leptons from the vector boson decay, which allow for tighter selection criteria for event identification.

2.2.2 Decay Channels of Vector Bosons

Due to the couplings described in Section 2.2.1, the vector bosons decay predominantly into quarks. However, in the hadronic environment produced at the LHC these are not the best indicators of a vector boson intermediate state. This mea-
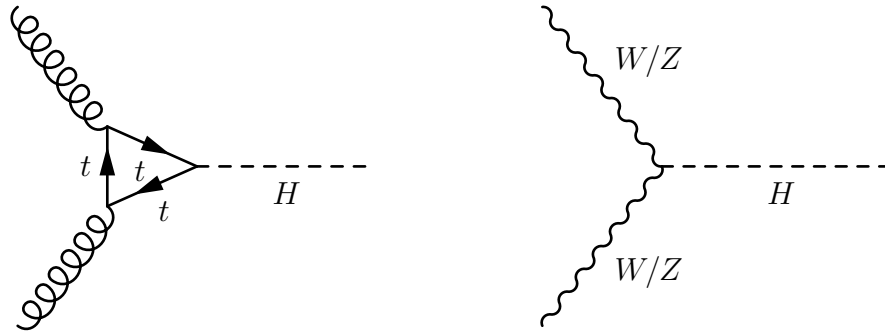
Figure 2-5: Above are the Feynman diagrams for other production mechanisms of the Higgs boson. Gluon fusion is shown on the left, and vector boson fusion is shown on the right.

surement uses leptonic decays in the final state since they are easier to identify and separate from background processes.

As shown earlier in Table 2.1, there are three generations of leptons, just as there are three generations of quarks. Each generation consists of a charged lepton, and a neutral lepton, also referred to as a neutrino. The left-handed charged lepton and neutrino of each generation form an electroweak $SU(2)_L$ doublet. In order of increasing mass, the three generations are called electron, muon, and tau. Heavier charged leptons decay into lighter leptons via the weak force. Two neutrinos result from this decay, as shown in Figure 2-6, making the characteristics of the parent lepton's parent difficult to reconstruct. The tau lepton has a short enough lifetime to consistently decay before reaching the CMS detector. The accompanying neutrino cannot be measured by the detector. The tau lepton is also massive enough to also decay into quarks, making its measurement even more complicated. Muons have an average lifetime long enough to penetrate the entire detector, and electrons are stable particles. As a result, only final states with muons and electrons are considered in this analysis. The Feynman diagrams for the decay channels of interest are shown in Figure 2-7.
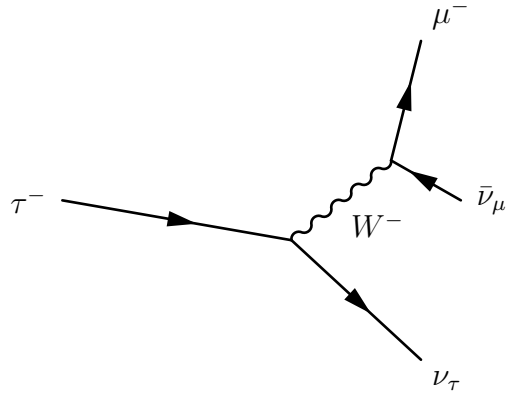
Figure 2-6: Heavier leptons can decay to lighter leptons while emitting two neutrinos. Above is an example of a decay of $\tau \to \nu_\tau \mu \bar{\nu}_\mu$.
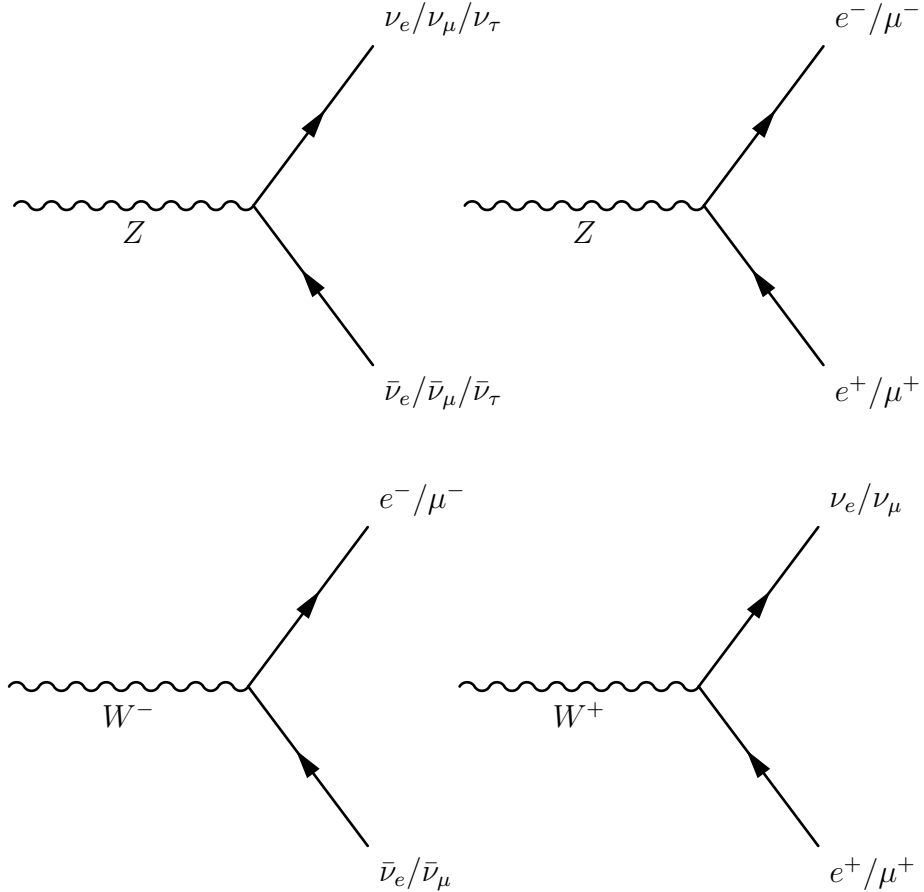
Figure 2-7: Above are the three different vector boson decays we are interested in. τ decays do also contribute to the charged lepton final states as seen by a detector, but the energy carried away by neutrinos significantly reduces those decay modes' contribution to the accepted states.
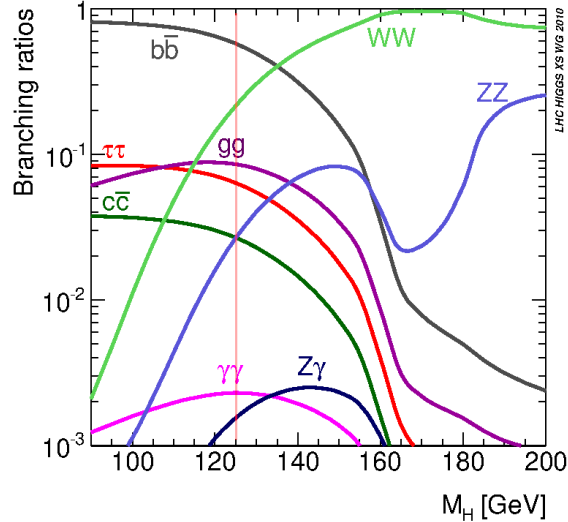
Figure 2-8: The predicted ratios of the Higgs are shown above [20]. A vertical line is added at the mass point where the Higgs boson has been observed. At this point, $b\bar{b}$ is the most common decay mode for the Higgs.

2.3 Decay Channels of the Higgs

What we are ultimately interested in measuring is the contribution of the Higgs intermediate state to the final state of $b\bar{b}$. As shown earlier in Equation (2.12), the Feynman rule for the interaction vertex between the Higgs and fermions is proportional to the fermion's mass. Of the quarks, the b quark is the second most massive. The most massive t quark is too massive to be the final decay product of an on-shell Higgs. The required off-shell Higgs would need a center of mass energy, E, at least $350\,\text{GeV}$, and the cross section for this production drops off as the tail of a relativistic Breit-Wigner function in Equation (2.32) with a mass M of $125\,\text{GeV}$. The Higgs can also decay to two vector bosons. In this case, instead of requiring the Higgs to be off-shell, one of the unstable vector bosons can be less massive than its resonance mass. However, this still results cross section suppression in the form of Equation (2.32). The predicted branching ratios, before the Higgs discovery, and their dependence on the Higgs mass is shown in Figure 2-8, showing $b\bar{b}$ to be the most common decay mode of the discovered Higgs.

As the decay mode with the highest coupling requiring no off-shell particles, the
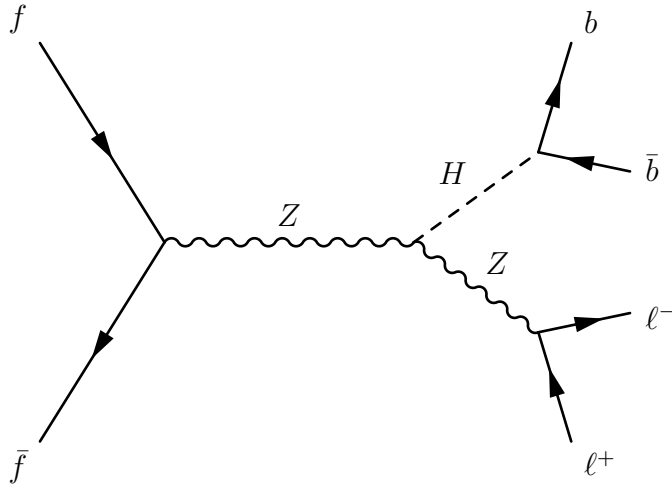
Figure 2-9: Above is the full Feynman diagram for $ZH \to \ell^+\ell^- b\bar{b}$.

predicted branching ratio of $H \to b\bar{b}$ is 57.8%. Therefore, measuring $H \to b\bar{b}$ is the most direct measurement to confirm this theory of quark masses. The diagram for this decay can be combined with the Feynman diagrams in Figure 2-1, Figure 2-2 or 2-3, and one of the decays in Figure 2-7 in order to generate the full Feynman diagrams for the processes being measured in this analysis. One such full diagram is shown in Figure 2-9.

2.4   The *VHb̄b* Measurement

The principles outlined in this chapter, along with higher order calculations, should be able to predict the frequency of $H \to b\bar{b}$ in associated production, as well as the energy dependence of these events. If there are discrepancies between the predictions and what is measured, that means there must be additional interactions beyond the Standard Model. An efficient selection of the $b\bar{b}$ decay mode will yield a large number of Higgs events, since it is the most common decay mode. Looking for associated production not only gives a clear detector signature to select on, but it also allows for accurate measurements of the vector boson momentum spectra. Since these vector bosons are directly coupling to the Higgs through *Higgstrahlung*, their momentum will be directly affected by any additional interactions affecting the Higgs.

Chapter 3

The CMS Detector

The Compact Muon Solenoid (CMS) detector, located at the LHC, consists of multiple sub-detectors. The analysis in this work is quite complex, and depends on all parts of the detector. Therefore, a full description of CMS is presented in this chapter.

First, a brief description of the LHC is given in Section 3.1. Then design requirements and considerations are outlined for the CMS detector in Section 3.2. Specific design decisions and descriptions of subdetectors are given in Section 3.3. Section 3.4 describes event reconstruction algorithms, Section 3.5 describes the triggers used to collect data, and Section 3.6 outlines simulation techniques used for CMS. Finally, Section 3.7 describes how data is stored and accessed by members of the collaboration. More can be learned about the design and motivations for the detector in the TDR [21]. Information presented on the physical CMS design parameters are taken directly from that document unless otherwise noted.

3.1   The Large Hadron Collider

The CMS detector only observes events. Before describing the devices that are used to observe and record events, the method of generating interesting events must be described. The CMS detector is located at the Large Hadron Collider (LHC). Described in detail in multiple publications [22], a brief description is given here.

The LHC, with a circumference of 26.7 km, is large enough to be considered

located in multiple towns and countries, but it will suffice to say it is near Geneva, Switzerland at the European Organization for Nuclear Research (CERN), the main campus of which is addressed in Meyrin, Switzerland. This campus itself also spans the border between Switzerland and France. This large circumference is needed since charged particles traveling in a circular path with radius r emit synchrotron radiation at the following rate.

$$P = \frac{q^2 p^4}{6\pi\epsilon_0 m^4 c^5 r^2} \tag{3.1}$$

The amount of power lost by the particles decreases quadratically with the size of the collider. In addition, the energy lost decreases with the mass of the accelerated particles to the fourth power. The LHC was built in the same tunnels that were used for LEP, which was a collider for electrons and positrons that took much of its data at $\sqrt{s} = 91\,\text{GeV}$ in order to study the Z boson resonance. The resulting LHC is designed to collide protons at energies of $\sqrt{s} = 14\,\text{TeV}$, with the data for this analysis taken at $\sqrt{s} = 13\,\text{TeV}$. This is more than enough energy to generate the massive off-shell vector bosons that are needed for *Higgstrahlung*, as well as many accompanying jets, via the mechanisms described previously in Chapter 2.

The luminosity of the LHC is given by the following formula.

$$\mathcal{L} = \frac{N_B^2 f_{\text{rev}} k_B}{4\pi B^* \epsilon_{xy}} \times F \tag{3.2}$$

N_B is the number of protons per bunch, f_{rev} is the frequency of beam revolutions, k_B is the number of bunches per beam, B^* and ϵ_{xy} describe the goodness of the beam, and F is a geometric collision factor.

$$F = \frac{1}{\sqrt{1 + \frac{(\sigma_s \tan\phi)^2}{\epsilon_{xy}\beta^*}}} \tag{3.3}$$

σ_s is the length of each bunch, ϕ is the crossing angle, and β^* is the value of the amplitude function at the focal point. In order to generate as much collision data as possible, the LHC operates at a high frequency of collisions, and generates many

simultaneous collisions. For Run 2, there is a proton bunch crossing every 25 ns. The CMS detector must be able to read out and process data on that timescale. Each proton bunch includes over 100 billion protons [23].

3.2   Detector Requirements

One configuration of possible final state particles was shown previously in Figure 2-9. There, two oppositely charged leptons and two b quarks are the end decay products. The b quarks also hadronize to form color singlets well before reaching the detector, but the resulting jets can actually be distinguished well from the jets resulting from the fragmenting protons.

Hadrons containing b quarks decay through the weak force since they require a flavor change. As mentioned before, the CKM matrix in Equation 2.19 quantifies the mixing between the different quark flavors, and b hadrons have a relatively long lifetime due to small values of V_{cb} and V_{ub}. The delayed decay results in a jet with a secondary vertex where many of its particles are generated from the vacuum at a distance from the initial collision point.

Alternate signatures of interest can be seen by substituting other vector boson final states from Figure 2-7. In these, there may be one or zero charged leptons, with one or two neutral leptons, respectively. Neutral particles are difficult to detect, with neutral leptons being capable of passing through the entire Earth without being part of a detectable interaction. The CMS detector therefore ignores the neutrinos, but their presence can still be inferred. Even with the variation in momentum along the beam direction, all partons in each proton have approximately zero momentum in the transverse direction. Therefore, the sum of the transverse momenta of all final state particles must also be zero. Many events in CMS have an overall imbalance in the transverse plane. This imbalance is labeled Missing Transverse Energy, E_T^{miss}, or MET. Large MET in an event is often a sign of high energy neutrinos that the detector cannot detect.

We need to identify all of these interesting particles, as well as be able to recon-

struct missing transverse momentum. In addition, the additional hadronic activity in the event, called pileup, must be mitigated. The energy of the decay products have energies on the scale of the masses of the parent particles. The detector must be capable of measuring jets and leptons with energies on the order of 10s or 100s of GeV. Better energy resolution for each of these decay products allows better separation of our signal process from background processes that generate very similar final states.

3.3   Detector Design

The CMS detector as a whole has cylindrical symmetry around the proton beams. It is 21 meters long and 15 meters in diameter. There are gaps at either end to allow the beams to enter and leave, but otherwise the design tries to cover the full solid angle around the collision point. The azimuthal angle of a particle relative to the beam axis is described by pseudorapidity, η.

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

The barrel portion of CMS detects particles up to $|\eta| < 1.5$, while the forward caps of the detector can reach $|\eta| < 5.0$. The muon and silicon trackers reach up to $|\eta| < 2.5$, with additional space covered by calorimetry.

Different technologies are better for measuring the energy or other kinematics variables of different particles. As a result, the CMS detector is made up of different sub-detector systems, arranged in cylindrical layers. Each layer consists of a "barrel" portion and two end caps on either side.

The innermost sub-detector is designed to extrapolate the tracks of charged particles back to their point of origin. This is called the Silicon Tracker. One key design feature of the silicon detector is that it is non-destructive. Particles it detects pass through to the rest of the detector for additional measurement and identification. The next sub-detector encountered by most particles is the Electromagnetic Calorimeter, which is designed to measure the energies of photons and electrons. The
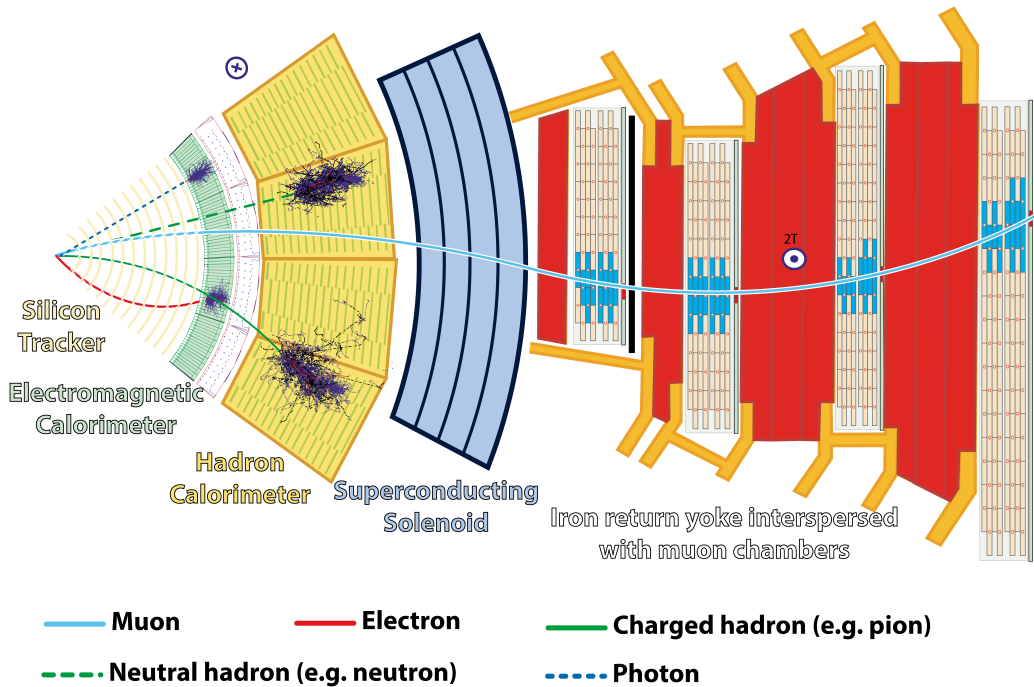
Figure 3-1: A slice of the CMS detector is shown above [24]. The four detector layers are labeled and show the penetration depths of various particles stable enough to travel a measurable distance.

next sub-detector, the Hadronic Calorimeter, measures the energies of both charged and neutral hadrons. The two calorimeters are destructive because they absorb the particles that interact with them in order to measure their full energy. Outside of these three sub-detectors is a superconducting solenoid, which generates a magnetic field for the entire detector. On the very outside of the detector are gas chambers designed to detect muons interspersed with the iron return yoke for the solenoid. A slice of the CMS detector showing the relative positions of each layer is shown in Figure 3-1.

The magnet is described first since the magnetic field it produces is a key part of most of the rest of the detector. After that, the sub-detectors are summarized in the order of closest to farthest from the beamline, since this is the order that particles would interact with the layers. Each sub-detector section also describes the measured performance during Run 2. Note that this measurement is an iterative process that depends on the event reconstruction described in Section 3.4, which in turn depends on the performance of the entire detector. The performance numbers are presented

with each sub-detector design though so that it is immediately clear how effective each design has been.

3.3.1 Solenoid Magnet

A key feature of CMS is the solenoid which generates a strong magnetic field throughout the detector. The presence of a magnetic field is paramount for accurate measurements of charged particles passing through the silicon detector and the muon chambers.

The magnetic field generated is designed to cause the path of a muon with $1\,\mathrm{TeV}$ of energy to bend enough to have a momentum resolution of 10%. Inside the solenoid, the magnetic field operates at $3.8\,\mathrm{T}$, with the solenoid design being capable of achieving $4\,\mathrm{T}$. The return field is large enough to cause muon tracks to curve throughout the muon chambers outside the magnet.

A super-conducting solenoid enables the creation of a magnetic field with the required strength. A current of $19.5\,\mathrm{kA}$ is sent through 2168 turns over $12.9\,\mathrm{m}$. The magnetic field stores $2.7\,\mathrm{GJ}$ of energy. In order to hold this, the structural components holding the magnet and the detector in place are strong enough to withstand $64\,\mathrm{atm}$ of hoop pressure.

3.3.2 Silicon Tracker

The layer closest to the beamline is designed to obtain a precise track pointing to the origin of particles passing into the detector. It is made up of layers of many small pixels to do this. As distance from the interaction point increases, the pixel size also increases since the absolution spacial resolution does not need to be as fine. The innermost three layers, with the closest layer being a distance of $r = 4\,\mathrm{cm}$ from the interaction point, are made of hybrid pixel detectors. Each pixel has dimensions of $100 \times 150\,\mathrm{\mu m}$ in order to achieve fine resolution of where the detected particles originated. The TDR also claims an occupancy of 10^{-4} per pixel per LHC bunch crossing, which improves the pixel's longevity and reduces problems from detector

deadtime.

Outside of the pixel detector layers are the strip detectors. Silicon strip detectors that use coincident signals to more precisely locate the locations of charged tracks. These are placed in the region that is $20 < r < 55\,\mathrm{cm}$ from the beamline. Strip dimensions give a cell size of approximately $10\,\mathrm{cm} \times 80\,\mu\mathrm{m}$. 2–3% of cells are activated during a typical bunch crossing. The outermost layers are made of larger strips with cell sizes of $25\,\mathrm{cm} \times 180\,\mu\mathrm{m}$. About 1% of these strips are triggered each bunch crossing.

The active material of the silicon detector is semi-conducting silicon. When charged particles pass through, electron-hole pairs are generated and drift apart due to a bias voltage. The voltage change when these pairs reach their respective electrodes indicates a charged particle passed through. Because of this, the silicon detector cannot detect neutral particles, but it gives a precise point of origin for charged particles. The points of origin allow for the determination of locations of primary and secondary collision vertices, which plays an important role in the identification of pileup.

In the beginning of 2017, the pixel detector was upgraded to handle the higher radiation environment of Run 2 [25]. A layer was added to both the barrel and endcap sections of the pixel detector. Firmware was also upgraded to keep the pixel detector operating at a frequency higher than the Run 2 collision frequency. With this upgrade, the detector operated with a 97% hit efficiency for all layers at the highest instantaneous luminosity. Layers beyond the first performed with greater than 99% hit efficiency [26].

3.3.3   Electromagnetic Calorimeter

The next layer of the detector is called the Electromagnetic Calorimeter or ECAL. This layer is designed to fully capture and accurately measure the energy of photons and electrons. The ECAL is made of crystals of the scintillating material Lead Tungstate ($PbWO_4$). Each crystal is placed in the detector so that its smallest face is facing the collision point. These small faces have dimensions of $22 \times 22\,\mathrm{mm}$. The length of each crystal is $230\,\mathrm{mm}$, and the far face is slightly larger at $26 \times 26\,\mathrm{mm}$.

PbWO$_4$ has a radiation length of $\chi_0 = 8.9\,\text{mm}$ and a Moliere radius of $21\,\text{mm}$. This means each crystal is 25.8 radiation lengths, containing the full shower within the ECAL, and each shower is also localized to within one crystal from the initial ionization. There is are gaps in active detecting volume of the ECAL, which are needed to accommodate various electronics and structural components. The gaps are located in symmetric locations on either side of the ECAL from $|\eta| = 1.4442$ to $|\eta| = 1.5660$.

The scintillating properties of PbWO$_4$ are also desirable for observing LHC collisions. The photodiodes at the far end of the crystals ultimately detect 4.5 photons for every MeV of energy deposited in the ECAL. This is a low number for other experiments, but the only photons and electrons of interest in this measurement deposit at least tens of GeV of energy. This gives the ECAL energy resolutions in the range of $5 - 10\%$. More importantly, the scintillation is very fast. About 80% of the light from an interaction is emitted within the $25\,\text{ns}$ between bunch crossings, making it easy to associate the readouts with the appropriate bunch crossing.

When exposed to the high radiation environment of the LHC, the ECAL crystals are damaged by radiation. Damage to the crystal structure causes it to become more opaque to the scintillated light. Much of this damage happens within the first 30 minutes of operation. Some recovery occurs as the crystal structure falls back into the ground state, but over time, the performance of the crystals degrade. That degradation happens at different rates in different areas of the detector, but, aside from the initial darkening, is slow enough to be able to correct for it during the run. Lasers are used to calibrate the ECAL online during the gaps between beams [27]. Resolution is measured by looking at $Z \rightarrow ee$ events. For Run 2, the barrel region of the ECAL performed with 1.6% resolution, and the other regions had a 5% resolution [28].

3.3.4 Hadronic Calorimeter

The ECAL absorbs electromagnetic particles and measures their energies which are dispersed in electromagnetic showers. Hadrons deposit energy in hadronic showers, which require a different mechanism to contain and measure. The Hadronic Calorime-

ter (HCAL) does this. Like the ECAL it contains particles and measures their energy destructively. However, it does this for hadrons, such as protons, neutrons, and stable mesons. Since hadrons are much more massive than electrons, the ionizing collisions in a typical scintillator does not slow them down enough to contain them. Instead, they must interact via nuclear collisions to be attenuated. CMS uses brass for its HCAL due to its relatively short interaction length, the fact that it is non-magnetic, and its affordability.

The barrel of the HCAL is jacketed in stainless steel for structural support. This layer is 61 mm thick on the layer immediately next to the ECAL and 75 mm thick on the outer edge. The inside of the HCAL consists of brass absorber plates interspersed with plastic scintillator tiles. The layers closer to the beamline alternate 50.5 mm brass plates with 3.7 mm scintillator plates. Farther away, the brass plates are instead 56.5 mm thick. Wavelength shifting fibers are run through the scintillator tiles to allow photons to travel to the outside of the HCAL where they are detected by photodiodes.

Like the ECAL, the HCAL performance also degrades as it is exposed to radiation. The calibration for HCAL is performed using an embedded radioactive source, lasers and LEDs, and an *in situ* calibration using assumed symmetry in ϕ. With these methods, a response within 3.4% was maintained in the HCAL barrel and within 2.6% in the HCAL endcap up to $|\eta| < 2$ [29].

3.3.5 Muon Chambers

Muons are the most penetrative particles that CMS detects. Through the calorimeters, muons act as minimum-ionizing particles [30]. They are heavier than electrons, so they are not stopped in the ECAL. They do not interact via the strong nuclear force, so the high density of the HCAL also does not cause significant interactions. Instead of stopping and measuring muons in calorimeters, CMS tracks their trajectory with both the silicon tracker on the inside of the detector and the muon chambers that make up the outer layer of the detector.

This is the only sub-detector system outside of the solenoid, but the returning magnetic field is still present outside of the return yoke [31, 32], allowing the momen-

tum of the muons to be extracted from the curvature of their trajectory. Layers of muon chambers act much like the silicon tracking detectors, but at a larger and more distant scale. The muon chambers in the barrel region of $|\eta| < 1.2$ consist of drift tube chambers. In the endcaps, cathode strip chambers are used. The difference is to account for higher neutron backgrounds in the endcap, as well as a greater magnetic field. In both regions, resistive plate chambers are spaced between the layers of the other muon chambers.

Each muon chamber has a detection efficiency greater than 95%. The overall efficiency of the muon trigger, which relies heavily on the muon system and is described in more detail in Section 3.5, increases as a function of muon p_T and plateaus around 90%. The timing of the muon system leads to 1% of muons to be assigned as originating from the wrong bunch crossing [33].

3.4 Event Reconstruction

Each sub-detector reconstructs the particles that passes through it. The independent reconstructions are then linked across the different detector components to identify particles. This overview follows the overview in reference [34].

3.4.1 Charged Particle Tracks

Both the Silicon Tracker and the Muon Chambers are designed for charged particles to leave tracks. In both sub-detectors, the basic steps for track reconstruction are the same. First, a track must be seeded. Usually, this is done by finding hits in consecutive layers that are consistent with a particle coming from the beamline. Particles loose energy as they pass through matter, and they can also be redirected through multiple scattering, so the extrapolation is non-trivial. A Kalman Filter is used to find hits in the other layers of the appropriate sub-detector that are consistent with the initial seed. Once more hits have been found, a fit is performed for the precise trajectory of the track.

Tracks are kept or discarded based on the number of layers that are missing

hits and on the momentum of a charged track in the magnetic field of the solenoid. Making these parameters looser results in better recovery of tracks, but the high activity within the detector results in a combinatorial background. This background increases exponentially when the momentum cut is reduced, for example. To help reduce this background, tracks with missing hits use an iterative fit. Different seeds are found for each track to make sure that the resulting collections of hits remain the same.

Additional complications arise for each the inner tracker and muon detectors. The Silicon Tracker is the only tracking detector that deals with electrons. Because of their small mass, electrons are likely to radiate energy while traveling through the magnetic field of CMS. This leads to complications in the calorimeters described in Section 3.4.2. It also means that the radius of curvature of an electron track can decrease appreciably within the inner tracker. This can lead to the Kalman Filter approach missing tracks entirely, depending on the number of hits required. Tracks with a large χ^2 and a certain number of hits are fit again using a Gaussian-Sum Filter (GSF). The GSF allows for fitting tracks that have significant energy loss, recovering electron and positron tracks.

A change of trajectory may also happen in the muon chambers, but this is due to multiple scattering in the return yoke. No specialized tracking algorithm is used to account for it. The muon tracking performs best when the tracks in the muon chambers are successfully linked to a track in the inner silicon detector. The most common backgrounds in the muon chambers is caused when hadrons manage to punch through the HCAL. This is often mitigated by considering the amount of energy deposited along the particle track in the other sub-detector systems.

3.4.2 Calorimeters

Clusters are identified in calorimeters, also using a seeding algorithm. First clusters with a large energy deposit are identified, and then nearby crystals are checked against noise thresholds. The energy deposition is assumed to have a Gaussian profile, and a fit is performed to disentangle overlapping energy depositions.

The dimensions of the ECAL crystals are comparable to the Moliere radius, keeping clusters localized. Although a complication arises due to curvature of the electromagnetic shower caused by the magnetic field, leading to the need for using GSF to find tracks. Superclusters that are linked to electrons have a larger allowed range over ϕ to account for this.

The calibration of HCAL is complicated by the fact that particles reaching it have to first pass through the ECAL. Initial calibration was done with a $50\,\text{GeV}$ pion test beam, but the actual response is non-linear in energy as well as different for charged and neutral particles. Reasons for this difficulty include particles losing energy in the region between the ECAL and the HCAL, in addition to the energy lost in the ECAL. Therefore, there are calibration coefficients that are used depending on if the energy deposits are all in the HCAL, or in preceding sub-detectors as well.

3.4.3 Linking and Particle Identification

An important step in making sense of the various sub-detector readouts is linking tracks to calorimeter clusters. The general procedure is to extrapolate tracks from the inner tracker out to each calorimeter. A shower that originates within one radiation or interaction length in the calorimeter along that track is linked with the track.

The bremsstrahlung from GSF electrons is linked to the track by looking along track tangents. A dedicated conversion finder is used to identify pair production within the inner tracking detector caused by either bremsstrahlung or prompt photons in order to not mistakenly link a charged particle track with what should otherwise be measured as a photon. This step, in addition to some ECAL clusters that do not have a track make it possible to identify isolated photons. On the other hand, it is still difficult to determine whether an electron track is well isolated or not. The large number of variables that go into identifying an electron leads to the training of a Boosted Decision Tree (BDT) to identify an electron. Separate BDTs are needed for the barrel and endcap regions of the detector.

For accurate HCAL readings, the linking algorithm also ECAL clusters and HCAL clusters along a path. These may not always be along a charged particle track.

Multiple calorimeter links of this nature may be found, but only a single link is kept based on a distance assigned to each link. HCAL hits without a track are identified as neutral hadrons. HCAL hits with a linked track are likely charged hadrons. Though the ECAL clusters must be linked in order to determine the energy coefficient to calibrate the HCAL, ECAL hits without tracks are still identified as photons because photons carry some of the energy of jets.

Of particular interest to this analysis is also the secondary vertex step of linking. Charged particle tracks that do not go back to the interaction vertex are linked together if they share a common secondary vertex. These tracks must have a mass greater than $0.2\,\mathrm{GeV}$ to be kept. There must also be a track from the secondary vertex to the primary vertex, which would belong to a long-lived hadron. As mentioned in Section 3.2, this is the signature of a b jet. It is possible, however, that the secondary vertex is generated by a nuclear scattering, pair production, or other long-lived particles like K_S or Λ within the silicon detector, so additional analysis is needed for each secondary vertex.

The final link is made between tracks in the muon chambers and tracks in the inner tracker. Muons are identified as tracker muons if they only leave tracks in the inner tracker. This can often happen with low energy muons. They are called standalone muons when only the track in the muon chamber is identified. When tracks are successfully linked in both sub-detectors, the resulting reconstructed particle is called a global muon. When a global muon is not well-isolated from other energy deposits, it must have tighter requirements on how it behaves in the muon chambers. This is to prevent energy from a jet from being attributed to a muon or vice versa. This is important for b jets because the decay that happens at a displaced vertex is a decay through the weak nuclear force, which can result in leptons being present inside of a jet.

3.5   Trigger

Bunch crossings happen every 25 ns, with each bunch crossing producing on average 20 collisions. The amount of data that the detector generates for each bunch crossing is too large to store all of it at this rate. Luckily, most collisions result in processes that are well understood, so the corresponding additional data is not needed. A trigger system is used to identify interesting events and reduce the frequency of event writing to 1 kHz. This is done using two stages. The Level-1 (L1) trigger passes events with a frequency of 100 kHz, and the High Level Trigger (HLT) picks from the remaining events with a frequency of 1 kHz [35].

The L1 trigger is implemented in hardware. It was upgraded for Run 2 of the LHC to run on FPGAs on an Advanced Mezzanine Card. There are two main components of the L1 trigger. One considers calorimeter deposits, and the other examines the muon chambers. The overall L1 trigger fires when there are high energy, resolved calorimeter hits or if a possible muon is reconstructed. Due to the flexibility of FPGAs, the exact conditions of the firing are configurable [36].

After the L1 is fired, the data is sent to the High-Level Trigger (HLT), which is a computing farm that makes a final decision on whether or not to save the data using a rough event reconstruction. The use of 30,000 cores in the HLT allows for buffering data so the HLT has plenty of time to make this decision [37].

For this analysis, only a few of the possible HLT paths are of interest. The exact trigger names are given in Chapter 4, but for the most part, they only depend on three different identifiable objects. Figures 2-7 and 2-9 show the different final states of interest. b jets are difficult to identify quickly because we must rely on the Silicon Tracker's reconstruction of the secondary vertex, but the decay mode of the vector boson can be used for the trigger. More boosted vector bosons leave a signature with a higher trigger efficiency. They also will cause the b jets to have a higher p_T, leading to easier identification and measurement. In that case, only events with one of the following are worth saving and examining for $VH \rightarrow b\bar{b}$.

- ECAL deposits consistent with a high p_T electron

- muon chamber hits consistent with a high p_T muon

- an overall energy imbalance consistent with MET from neutrinos

For the specific decay channel in Figure 2-9, the HLT also includes paths where two electrons or two muons are identified.

3.6   Simulation

After the detector is well understood, predictions on how it responds in the LHC environment can be made. The number of ways the detector could possibly respond are nearly infinite. Therefore, simulation is performed using Monte Carlo methods, and the resulting analysis is statistical in nature. The data format for simulation results is similar to the data format for data collected from the detector. Unobservable information about intermediate steps in the simulation is also stored, but otherwise the data is the predicted output of a collection of events.

The simulation itself consists of several steps each outlined in a separate section of this chapter. First the background processes that will appear in our analysis must be known. Identifying all of these processes is necessary to quantify and characterize the signal events that are also mixed in to our selection. Then each of these processes must be simulated to determine the final state particles that the detector will observe. Each process looks slightly different in our signal selection. Events outside of the selection must also be simulated so that they can be studied for accuracy in separate phase spaces that do not include the Higgs events. After the final state particles are predicted, the detector response to those particles passing through must be simulated. This allows researchers to compare the physical readouts they can observe to predicted detector signals. Finally, using the phase spaces outside of the desired signal process, minor corrections to the simulation can be made. Simulated energies from the detector model might not be the exact same as what the physical detector produces, for example, and they must be made to match to make the signal process cleanly appear in the analysis.

The physical processes that occur at the LHC all contain QCD-driven phenomena. As a result, the part of the simulation that predicts the particles present in the detector has two distinct parts. QCD is perturbative at small distances, and other forces are perturbative at all distances. The collisions themselves are in this regime, so the initial- and final-state particles over a distance of femtometers can be simulated using typical calculations using perturbative rules described by Feynman diagrams. The decay of unstable particles can also be simulated this way. Once particles interacting through QCD exceed this distance, well before reaching the detector, hadronization, or parton showers, must be simulated differently. The following two sub-sections describing these techniques. The exact generators and configurations used to simulate each process for this analysis are detailed in Appendix B. After final state particles are generated, their propagation through the detector is simulated in a third step.

3.6.1   Short-Scale Simulation

Events are generated by selecting results and assigning weights in a way proportional to the phase space and the matrix element squared of the event. The phase space integral has the following form [38].

$$\int d\Pi_n = \left(\prod_f \int \frac{d^3 p_f}{(2\pi)^3} \frac{1}{2E_f} \right) (2\pi)^4 \delta^{(4)}(P - \sum p_f) \tag{3.5}$$

P is the total initial 4-momentum and f runs over all final state particles. This phase space integral is Lorentz invariant. Once an event has been selected, the available phase space can then be used to assign directions to the final particles.

The proportional matrix elements are described in Chapter 2, but most of the diagrams described were Leading Order (LO). Generators used in this analysis can also simulate Next to Leading Order (NLO) processes thanks to the FKS method of subtracting particles to avoid double counting them during the showering calculation [39]. However, the option to use NLO simulation is not always used. The results of NLO calculations more accurately predict physical processes, but they also take more computational resources, resulting in larger measurement uncertainties due to

statistical limitations.

The generators used for short-scale simulation in this analysis are POWHEG [40] and MadGraph5 [41].

3.6.2 Parton Showers

The final state particles from the short-scale simulation include a number of free quarks and gluons. As mentioned in Section 2.2.1, it is energetically favorable for color-charged particles to create additional quark/anti-quark pairs to screen the color charge. This is known as hadronization or parton showering and is simulated separately from the calculation of tree-level processes. Hadronization happens well before particles reach the CMS detector, so the results are needed to predict the detector response. Accurate simulation of this process is important for all collisions at the LHC, which produces much QCD background. It is also important to accurately simulate the constituents of individual jets because this analysis includes detailed inspection of each jet in order to identify b-jets and to estimate the amount of energy carried away by neutrinos.

To be able to analyze the simulation in the same way data is processed, Monte Carlo simulation is used to predict precise final states of the jets. CMS uses the Lund model [42] as implemented in `PYTHIA8` [43].

3.6.3 Detector Simulation

After determining all of the final state particles that will reach the detector, the interaction between these particles and the detector components must be simulated. Multiple simulations of pp collisions are combined to simulate pileup, and then the particle propagation through various materials is done with `GEANT4` [44]. The full CMS detector geometry is maintained within CMSSW [45] using a framework written in the Unified Modeling Language [46]. To be able to process the simulated data in the exact same fashion as the measured data, the readout of the electronics is also simulated.

3.7 Accessing Data

The final important piece of the CMS detector is its offline computing resources. The data that is gathered by the detector must be processed and stored. This is done by using computing resources spread around the world. They are grouped into Tier-1 and Tier-2 sites, based largely on the geographic space that they are meant to provide computing services for. Tier-1 sites typically provide 30,000 CPU cores, while the more numerous Tier-2 sites provide another 60,000 [47]. Together they also provide around 100 Petabytes of space [48]. Keeping these services running smoothly is important for the CMS collaboration to function, and two projects that aid in this are outlined in Appendix A. Since the data is event-based, they are stored in n-tuple format [49] supported by ROOT [50]. These n-tuples are created and read by the CMS Software (CMSSW) [51].

Chapter 4

Event Selection

This chapter gives the specific selection requirements on each physics object that allows us to count particle candidates and to reject or otherwise classify events, based on the description of physics processes described thus far. First, objects are defined in terms of variables and particle candidates provided by the detector reconstruction algorithms. Using these object definitions, selection requirements are then used to reject events from the analysis entirely. Of the events that remain, different selection requirements are applied to classify events into different decay channels of the vector boson. Finally, there are also selection requirements that allow events to be treated separately when the Higgs decay products can be resolved as separate jets and where they are merged into a single massive jet.

4.1   Object Definitions

Detector responses are linked to possible physical particles. Most of the particle ID techniques described so far can give false positives for individual particle candidates or provide composite physics objects that are in reality composed of background particles. What follows are tighter selections used in order to reduce these backgrounds. Once objects are more strictly defined, they can be used for more reliable event classification.

Each type of object generally has a method of loose pre-selection and additional

tighter requirements for a selection. The distinction is particularly useful for categorizing events. Each category is designed to be enriched with a particular physics process. Each physics process would result in certain final states with specific multiplicities for some particles. If an object passes the loose selection, it is defined well enough to count for the event selection that would not include the corresponding particle. Additional selection requirements are added for objects to classify as a particular particle candidate in order to reduce false positives of events that are included in a given category that instead requires the particle to be present. For example, if an event has a single muon that passes the loose selection, but not the tight, the event would be rejected for the 0-lepton category, requiring no charged leptons, but would also be rejected from the 1-lepton category in case the muon is a false positive.

4.1.1 Variable Definitions

Many of the object definitions use variables that are derived from reconstructed quantities. They can be understood in terms of the reconstruction described in Section 3.4. Lepton isolation is quantified using the following formula.

$$I = \frac{1}{p_T^\ell} \left(\sum p_T^{\text{charged}} + \max \left[0, \sum p_T^{\text{neutral}} + \sum p_T^\gamma - p_T^{\text{PU}} \right] \right) \qquad (4.1)$$

The sums are over charged hadrons originating from the primary vertex and all neutral hadrons and photons within a distance of $\Delta R < 0.4$ from the lepton if it is a muon, or $\Delta R < 0.3$ from an electron, where ΔR is a distance on the (η, ϕ) plane.

$$\Delta R = \sqrt{\Delta \eta^2 + \Delta \phi^2} \qquad (4.2)$$

The term p_T^{PU} is defined as the following for muons.

$$p_T^{\text{PU}} = 0.5 \times \sum p_T^{\text{PU,charged}} \qquad (4.3)$$

Electrons use a different definition.

$$p_T^{\text{PU}} = \rho \times A_{\text{eff}} \tag{4.4}$$

A_{eff} is the area of the isolation cone, and ρ is the median of the p_T density of neutral particles in that area.

Particles can also be defined as coming from the primary vertex of an event or from pileup. Vertices are defined through deterministic annealing [52], using the closest approach of tracks to the beamline [53]. The primary vertex is the vertex with the greatest sum of E_T of the charged particles originating from it. After identification of the primary vertex, charged particles are classified as originating from the primary vertex or as pileup using their extrapolated track's distance in the transverse plane, d_{xy} and distance along the beamline, d_z.

4.1.2 Isolated Muons

An isolated muon gives one of the cleanest signatures in CMS, with only perhaps the exception of an isolated photon that does not undergo pair production in the pixel tracker. Muons can also show up in jets from weakly decaying hadrons, in which case they are not isolated. Since weakly decaying b jets are central to this analysis, events with non-isolated leptons are not rejected, but the distinction is important. Loosely selected muons must meet the following requirements so that they are more likely from the decay of W and Z bosons.

- The muon must have a relatively high transverse momentum of $p_T > 5\,\text{GeV}$.

- The muon should pass through the inner tracker within $|\eta| < 2.4$.

- The muon originates from the primary vertex, satisfying both $d_{xy} < 0.5\,\text{cm}$ and $d_z < 1.0\,\text{cm}$.

- The muon must pass a loose isolation requirement of $I < 0.4$.

- The muon must be a PF muon.

- The muon is either a global muon or a tracker muon.

Tightly identified muons have some additional criteria they must pass.

- They must have a higher transverse momentum at $p_T > 25\,\text{GeV}$. In events with two muons, such as caused by $Z \to \mu\mu$, only one muon needs to satisfy this p_T requirement. A $p_T > 15\,\text{GeV}$ is required of the second muon.

- The muon must be a global muon, leaving tracks in both the central tracker and the muon chambers.

- There must be more than five hits in the inner tracker with at least one hit in the pixel detector.

- The fit for the global muon track must be good with $\chi^2/ndof < 10$.

- The muon must be well isolated with $I < 0.06$

These definitions are an accepted standard in the CMS collaboration as loose and tight working points, respectively. This allows the analysis to use efficiency measurements created for wider use.

4.1.3 Isolated Electrons

The kinematic variables associated with an electron are extracted from the GSF fit. Loosely selected electrons must meet the following requirements.

- They must have a transverse momentum satisfying $p_T > 7\,\text{GeV}$.

- They should be centered in the detector with $|\eta| < 2.4$.

- The distance from the primary vertex is limited, requiring $d_{xy} < 0.05\,\text{cm}$ and $d_Z < 0.2\,\text{cm}$.

- They pass a loose isolation requirement of $I < 0.4$.

To optimize the electron selection, electrons are identified with the aid of an MVA [54]. Fully selected electrons pass the tight working point used by the CMS collaboration. In order to also match the samples of simulated electrons used in the training sample, the selected electrons must also pass the following criteria.

- The electron must have higher energy with $p_T > 15\,\text{GeV}$.

- The deposit of HCAL energy must be less than 9% of the ECAL energy deposit along the electron track.

- The track sum p_T component of the isolation must be less than 18% of the electron p_T.

- There is a gap in the ECAL geometry, so the electron must be outside the gap $1.4442 < |\eta| < 1.5660$. The gap is between the barrel and endcap regions of the detector. Due to detector geometry and different conditions in each region, the remaining selection criteria differs based on the electron's location.

- For electrons with $|\eta| < 1.4442$ in the detector barrel:

    - The shower shape must satisfy $\sigma_{i\eta i\eta} < 0.012$

    - Isolation in the ECAL cluster must be less than 0.4, and isolation in the HCAL must be less than 0.25.

    - The difference between super cluster and track location of the electron must be small with $\Delta\eta < 0.0095$ and $\Delta\phi < 0.065$.

- For electrons with $|\eta| > 1.5660$ in the detector endcap:

    - The shower shape must satisfy $\sigma_{i\eta i\eta} < 0.033$

    - Isolation in the ECAL cluster must be less than 0.45, and isolation in the HCAL must be less than 0.28.

4.1.4 Jets

Strong interactions cause jets of particles when quarks or gluons hadronize. Conservation of energy and momentum means that the sum of jet constituents give the kinematics of the initial parton that produced them. Jets are constructed by clustering all particle-flow candidates with the anti-k_T algorithm [55] using the jet clustering parameter $R = 0.4$. Due to factors like pileup, imperfect detector response, and differences in various jet-clustering configurations the energy of the reconstructed jets are corrected [56].

Loose jet identification, based on the constituents, is applied to remove jets constructed from detector noise. Jets that get a significant fraction of their energy from pileup are also removed. Pre-selected electrons and muons are also often reconstructed as jets. Any jet within $\Delta R < 0.4$ from a pre-selected lepton as described in Sections 4.1.2 and 4.1.3 is removed.

To be considered for the study of decay products of the Higgs boson, jets must be within the inner tracker of the detector with $|\eta| < 2.5$. This allows pileup to be removed and for accurate vertexing of the constituents. The jets must satisfy $p_T > 25\,\mathrm{GeV}$ for the zero and one lepton signatures. The two lepton signature from $Z(\ell\ell)H$ is cleaner, and looser jet selection criteria of $p_T > 20\,\mathrm{GeV}$ are applied to the jets.

4.1.5 Identification of b Jets and Energy Regression

Jets containing b hadrons have a distinct signature. This includes secondary vertices displaced from the beamline, as well as non-isolated leptons from weak decays. When looking at jets in the inner tracker, all of these features are considered in a deep neural network (DNN) called Deep Combined Secondary Vertex (DeepCSV) designed to identify b jets [57]. The output of DeepCSV has three working points that are defined based on the amount of false positives that are expected in a collection of jets meeting the requirement. The specific values are different for each year of operation. The detector and collision conditions change, and a separate model is trained each

Table 4.1: The minimal value on the neural network output for each DeepCSV working point are defined for each year of Run 2 of the LHC. The working points are defined by their mistag rates in simulation.

Working Point	Mistag Rate	2016	2017	2018
Loose	10%	0.2219	0.1522	0.1241
Medium	1%	0.6324	0.4941	0.4184
Tight	0.1%	0.8958	0.8001	0.7527

year to account for that. The working points for each year are given in Table 4.1.

The non-isolated leptons within b jets are caused by flavor-changing weak decay of b hadrons. This decay mode also results in neutrinos which carry away a portion of the jet energy. In order to more accurately reconstruct the di-jet mass of a candidate Higgs decay, a Deep Neural Network (DNN) is trained in Tensorflow [58], which is designed to improve the energy measurement and resolution of all b jets for CMS [59]. The target is the transverse momentum of the generator-level particles that make up the jet with simulated neutrinos included in the calculation. The Huber loss function is used to avoid training too aggressively for outliers. The following variables are used as inputs to the regression:

- the jet's p_T, η, mass and transverse mass

- the event's median energy density, commonly denoted as ρ

- information about the hardest lepton clustered into the jet, including momentum perpendicular to the jet, distance ΔR from the center of the jet, and the lepton's flavor

- the p_T, mass, and number of tracks from any secondary vertex linked to the jet, as well as the secondary vertex's distance from the collision point and associated uncertainties

- the fractions energy in the jet due to charged and neutral hadrons and electromagnetic constituents

- the highest p_T of charged hadron constituents

- the energy fraction contained in five concentric rings around the jet center binned by $\Delta R \in [0, 0.05, 0.1, 0.2, 0.3, 0.4]$

- number of PF candidates in a jet

- energy sharing computed by

$$\frac{\sqrt{\sum_i p_{T,i}^2}}{\sum_i p_{T,i}}$$

where i runs over the jet constituents

This list results in 41 input variables for the DNN. All selections depending on the p_T values of b-tagged jets are evaluated after applying the b-jet energy regression.

4.1.6 Fat Jets

When the Higgs boson is highly boosted, its decay products are closer together in the frame of the detector. At some point, the jets from the b quarks would not be clustered separately. These highly boosted events are interesting for the differential cross section measurements, and they are more enriched with the signal associated production process than events with lower p_T intermediate particles. To not loose the events, single jets are analyzed for evidence of containing two b hadrons. In order to handle the transition from resolved jets to boosted single jets, larger jets, labeled fat jets, are used. A second collection of jets is made from the same set of PF candidates that are clustered to create the jets described previously. This collection also uses the anti-k_T algorithm, but with a larger jet clustering parameter of $R = 0.8$ to capture the majority of boosted events. The requirement on the fat jet which ensures that Higgs decay products are contained in the jet when they are at a maximum opening angle is $p_T > 250\,\mathrm{GeV}$.

Being significantly larger in area, the fat jets also contain much additional radiation from the underlying event. As a result, the mass of the constituents collected into the jet is also significantly larger than the original mass of the primary parent

66

particle whose daughters make up most of the jet constituents. These jets make use of the PUPPI algorithm [60] for more accurate removal of pileup contributions to the calculation of kinematics. A number of other grooming algorithms, which are also designed to remove particles from pileup from the jets, were also considered within CMS for additional refinement of the jet mass calculation [61]. The soft drop algorithm [62] was chosen as the standard in the experiment and is used in this analysis. Since the resulting groomed mass of the jet is close to its original parent particle, the soft drop algorithm has the additional benefit of forcing pileup jets to low mass, Therefore, a second requirement on fat jets considered for the analysis is $m_{\mathrm{SD}} > 50\,\mathrm{GeV}$.

4.1.7 Missing Transverse Momentum

Missing transverse energy, which is actually the missing transverse momentum, also labeled E_T^{miss} or MET, is a vector that takes advantage of the fact that momentum transverse to the beamline is conserved. Particle flow MET is calculated by taking the negative vector sum of the transverse momentum of all particle flow candidates in the event. The resulting vector is then adjusted by taking into account the difference between the uncorrected and corrected jet energies [63]. The resulting magnitude and direction is a proxy for the transverse momentum of any neutrinos in the event. Since particle flow MET is the type of MET that is used most often in this analysis, it is referred to throughout as just MET. Large MET values can be generated by instrumental and beam effects as well. Therefore, there are additional event filters applied to events with large MET that removes events where these known instrumental and beam effects have been identified.

An additional technique to select interesting events with large MET is to compare the particle flow MET to other calculations of MET. This analysis also uses track MET, or trkMET. Track MET is calculated using just the reverse vector sum of charged particles that are detected by the tracker. In collisions without leptons, charged and neutral particles energy distributions are similar in the detector, so differences in track and particle flow MET directions can indicate instrumental failures in purely hadronic environments.

4.1.8 Soft Hadronic Activity

In signal VH events, hadronic activity outside of the $b\bar{b}$ decay of the Higgs is expected to be low. This hadronic activity is defined by considering the additional charged PF tracks coming from the primary vertex. An exclusion region is defined in an ellipse in (η, ϕ) space containing the two selected b jets with a major axis length of $\Delta R(b\bar{b}) + 1$ and a minor axis length of 1. All charged tracks outside of this ellipse that also do not correspond with the selected leptons and that satisfy $p_T > 300\,\text{MeV}$ and $d_Z < 0.2\,\text{cm}$ are clustered using the anti-k_T algorithm [55] with $R = 0.4$. The resulting collection of soft jets is used to define four variables:

- H_T^{soft} – The scalar sum of soft jets' p_T for jets with $p_T > 1\,\text{GeV}$

- N_2^{soft} – The number of soft jets with $p_T > 2\,\text{GeV}$

- N_5^{soft} – The number of soft jets with $p_T > 5\,\text{GeV}$

- N_{10}^{soft} – The number of soft jets with $p_T > 10\,\text{GeV}$

These variables are used in the training of the BDT that discriminates signal and background events.

4.1.9 Kinematic Fit

In the two-lepton region, the Z boson's decay products are observed directly, and the Z boson momentum is precisely reconstructed. A kinematic fit is performed to constrain the momenta of the two leptons by requiring their combined mass to match the Z boson mass. After this constraint is applied, the transverse momentum of the di-jet system, along with identified initial state radiation jets, is balanced with the transverse momentum of the di-lepton system. The fit adjusts the momentum of each identified component of the event in a way that minimizes the chi-squared value meeting measured and theoretical physics constraints.

For electrons and muons, the corrections defined in centralized efforts by the CMS collaboration include uncertainties. The energy uncertainty used for the b-jets and
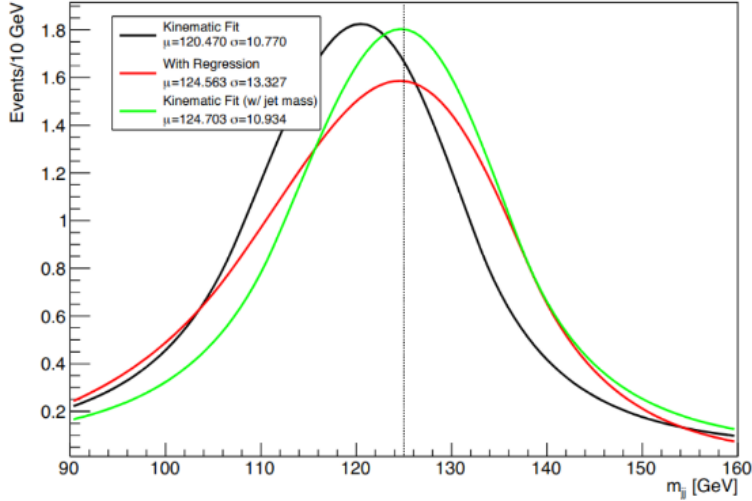
Figure 4-1: The Higgs di-jet mass in the 2-lepton signal samples is shown above. Peaks from the raw jet, the regressed jet energy, and the kinematic fit are compared.

recoil jets is the same as that used by the $HH \to b\bar{b}b\bar{b}$ analysis from CMS [64]. The mass of the Z boson is given a Gaussian uncertainty of $5\,\text{GeV}$, and it is assumed that the MET in the event is less than $20\,\text{GeV}$. The kinematic fit minimizes the chi-squared value of these constraints, resulting in new energies for all particles in the fit. The full implementation is in the `PhysicsTools/KinFitter` package of `CMSSW_10_2_0_pre3` [65]. The fitted b-jet energies are then used for all values in the remainder of the two-lepton channel analysis. The improvement of the di-jet mass in the signal sample as a result of the kinematic fit are shown in Figure 4-1.

4.2   Simplified Template Cross Section Bins

The measurement performed in this analysis is a binned differential cross section of Higgs production using the Simplified Template Cross Section (STXS) approach [66]. The allows the result to be compared more thoroughly to Standard Model predictions while reducing the model dependence of the measurements. For this measurement, the vector boson produced as well as the p_T of the vector boson separates data points into different bins. The clean signal of the $Z \to \ell\ell$ decay channel allows for more bins to be measured for ZH processes. In addition, the middle p_T bin for Z boson

production is split by multiplicity of additional jets. There are eight bins overall:

- $WH, 150\,\text{GeV} < p_T(V) \leq 250\,\text{GeV}$

- $WH, 250\,\text{GeV} < p_T(V) \leq 400\,\text{GeV}$

- $WH, 400\,\text{GeV} < p_T(V)$

- $ZH, 75\,\text{GeV} < p_T(V) \leq 150\,\text{GeV}$

- $ZH, 150\,\text{GeV} < p_T(V) \leq 250\,\text{GeV}, n_{\text{jet}} = 0$

- $ZH, 150\,\text{GeV} < p_T(V) \leq 250\,\text{GeV}, n_{\text{jet}} \geq 1$

- $ZH, 250\,\text{GeV} < p_T(V) \leq 400\,\text{GeV}$

- $ZH, 400\,\text{GeV} < p_T(V)$

All of the selections in the following sections are also divided into the appropriate set of STXS bins for the generation of datacards and fits. Since the fat jets are most helpful in events where intermediate particles are highly boosted, they are only considered in selections for the bins where $p_T(V) > 250\,\text{GeV}$.

4.3 Treatment of Background Yields

In order to effectively measure Higgs production, we need to be able to accurately estimate other events that end up in our selection. To start, the most significant contaminant physics processes with a final state identical to the $VHb\bar{b}$ process are identified. Then, selections that are enriched with these processes, called control regions, are created, which allows more direct comparison between simulation of these processes and collected data. When fitting for the signal strength, the background process yields are simultaneously fit from these control regions.
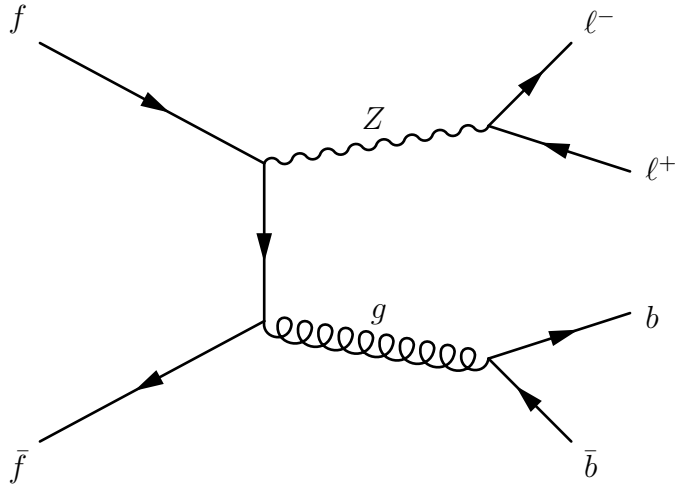
Figure 4-2: Above is the Feynman diagram matching the two lepton final state coming from Drell-Yan and jets.

4.3.1 Identification of Background Processes

The final state for the two lepton decay in Figure 2-9 contains two oppositely-charged, same flavor leptons and two b-tagged jets. It can also be achieved by a Drell-Yan process radiating b jets or a $t\bar{t}$ event where both W bosons from the top decays decay leptonically. Feynman diagrams in Figure 4-2 and Figure 4-3 show how the two respective processes can result in the same final state as the signal process. The Drell-Yan process can also radiate jets initiated by lighter flavor quarks that are mistakenly identified as b-jets, and those make up a significant portion of the backgrounds as well. Less significant, but still important backgrounds include processes like di-boson production, QCD jets, and single top processes.

The backgrounds for the one- and zero-lepton signal decay channels are caused by similar processes. For the one-lepton decays of WH, the Drell-Yan background in Figure 4-2 is replaced with a flavor changing current of W + jets. The $t\bar{t}$ background would instead be caused by either a hadronic decay of one of the W bosons in Figure 4-3 or by one of the pictured leptons traveling out of the detector without being observed. For the zero-lepton channel, the Drell Yan process is instead replaced with $Z \rightarrow \nu\bar{\nu}$. The $t\bar{t}$ process still needs high MET in order to appear to contain a hard neutrino presence, so it is most often caused when a single W decays leptonically with the lepton falling outside of the detector acceptance. For both of these channels,

71
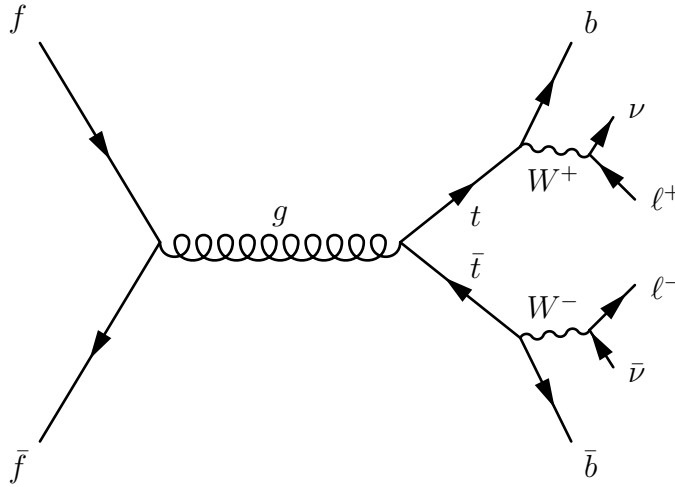
Figure 4-3: Above is the Feynman diagram matching the two lepton final state coming from fully leptonic $t\bar{t}$ decay. In events with little energy carried away by the neutrinos, this can appear to be the same as the two-lepton signal process.

di-boson, QCD, and single top backgrounds can also contribute, but not significantly enough to require a tight constraint on their yields.

4.3.2 Goals for Control Regions

To accurately estimate the contribution of each of these backgrounds, control regions are used. These are selections that are in similar phase spaces as the signal selection, but are instead enriched with background events. By comparing the prediction from Monte Carlo to data, scaling corrections to the simulation are made for the relevant phase space and extrapolated to the signal region.

For each channel, three control regions are defined. There are two control regions for the appropriate vector boson radiating jets. The region is enriched with light flavor jets, labeled $udsg$ for up, down, and strange quarks, and gluons, by requiring a di-jet system that fails the b-tagging working points. The heavy flavor $V +$ jets control regions are differentiated from the signal selection primarily by requiring the di-jet mass to be different than Higgs boson mass. The $t\bar{t}$ process tends to radiate more jets than the signal process, but the lack of a di-lepton resonance is taken advantage of in the two-lepton regions. Specific selection requirements for each signal region and the control regions are given in the next section.

72

4.4 Resolved Analysis Selection

First, the selection for events with two b-tagged jets, also known as resolved events, will be given. The next section will explain the selections relying on a single fat jet. With objects defined, the selections differ mostly in counting the number of charged leptons present in the event. However, other adjustments are made per channel to optimize the presence of signal events. Therefore, the first channels described will have the most thorough selection description, with later channel sections noting many similarities and differences. For each channel, multiple control region selections are also used in order to more accurately estimate the contribution of each physics process to the events in each phase space, and these will be described after each channel's signal region. The most significant contributions to background contamination of the signal region are from $Z/W + $ jets (depending on the number of leptons in the selection) and $t\bar{t}$ processes. The vector boson plus jets regions are separated into heavy-flavor and light-flavor jets to control the relative contributions independently since they are not well-known. For each channel, there are therefore three control regions in addition to the signal region.

A summary of requirements for each region in each channel is given in Table 4.2. A few channel- or region-specific requirements are left out, but are described in the appropriate sections. The common variables that are used in the event categorization are the following:

- $p_T(V)$ – The transverse momentum of the reconstructed vector boson

- $p_{T,j}$ – The minimum transverse momentum of both jets, which may be an asymmetric requirement

- max b – The working point that the jet with the higher b-tag value must satisfy

- min b – The working point that the jet with the lower b-tag value must satisfy

- $p_{T,jj}$ – The summed transverse momentum of the di-jet system

- m_{jj} – The mass of the di-jet system

Table 4.2: Below is a summary of common requirements for all regions in the resolved channels. See the text for each channel for an explanation of variables. All energy equivalent values are in GeV. A "!" in front of a working point label means the jet is required to fail the relevant b-tagging criteria.

0-lepton channel								
Region	$p_T(V)$	$p_{T,j}$	max b	min b	$p_{T,jj}$	m_{jj}	N_{aj}	$\Delta\phi(jj,V)$
Signal	170	60, 35	med.	loose	120	$>90, <150$	≤ 1	> 2.0
$Z + b$	170	60, 35	med.	loose	120	<90 or >150	≤ 1	> 2.0
$Z + udsg$	170	60, 35	!med.	loose	120	$>50, <500$	≤ 1	> 2.0
$t\bar{t}$	170	60, 35	med.	loose	120	$>50, <500$	≥ 2	$-$

1-lepton channel								
Region	$p_T(V)$	$p_{T,j}$	max b	min b	$p_{T,jj}$	m_{jj}	N_{aj}	$\Delta\phi(jj,V)$
Signal	150	25	med.	loose	100	$>90, <150$	≤ 1	> 2.5
$W + b$	150	25	med.	loose	100	<90 or >150	≤ 1	> 2.5
$W + udsg$	150	25	!med.	loose	100	$>50, <250$	$-$	> 2.5
$t\bar{t}$	150	25	med.	loose	100	$>50, <250$	≥ 2	$-$

2-lepton channel								
Region	$p_T(V)$	$p_{T,j}$	max b	min b	$p_{T,jj}$	m_{jj}	N_{aj}	$\Delta\phi(jj,V)$
Signal	50	20	med.	loose	50	$>90, <150$	$-$	> 2.5
$Z + b$	50	20	med.	loose	50	<90 or >150	$-$	> 2.5
$Z + udsg$	50	20	!loose	!loose	50	$>50, <250$	$-$	> 2.5
$t\bar{t}$	50	20	tight	loose	50	$>50, <250$	$-$	$-$
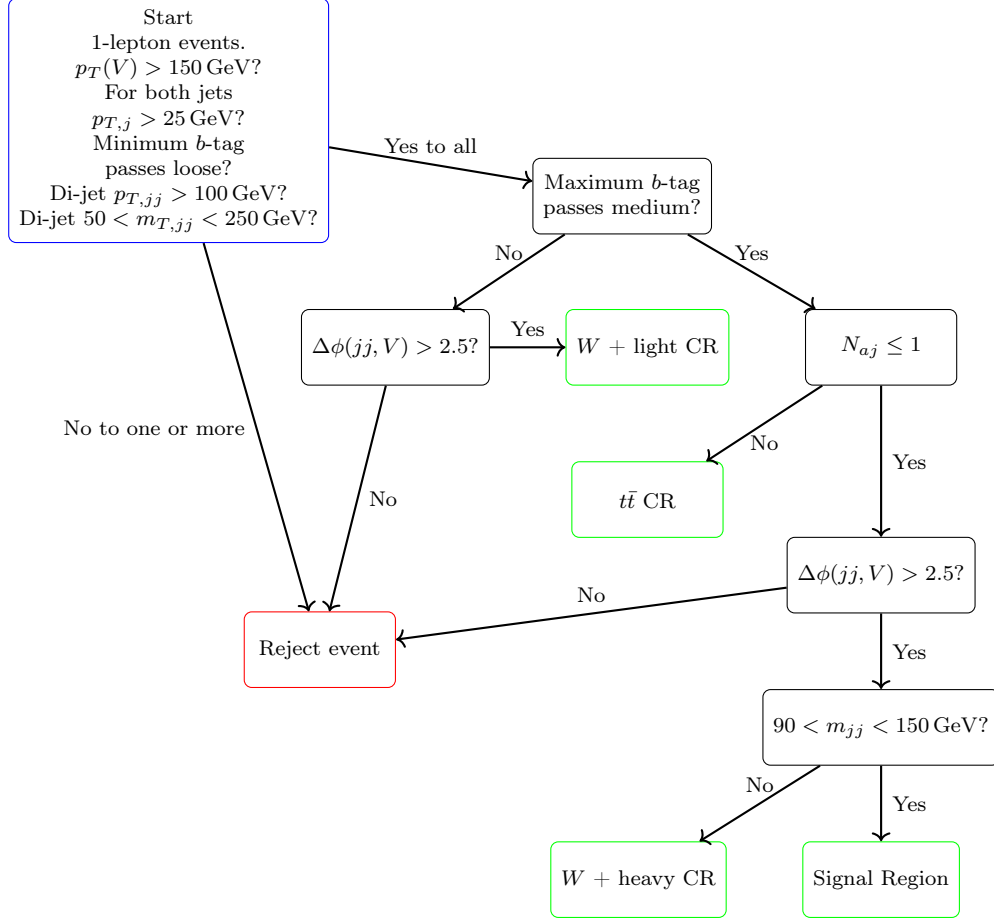
Figure 4-4: Above a flowchart shows the process of using selection requirements to categorize events into the signal region or different control regions. The example is of the 1-lepton resolved channel, as detailed in Table 4.2. The other channels follow similar patterns.

- N_{aj} – The number of additional jets outside of the selected di-jet system

- $\Delta\phi(jj, V)$ – The distance in angle ϕ between the di-jet system and the recoiling reconstructed vector boson

A flow chart to help visualize the process of categorizing events into different regions is shown in Figure 4-4.

4.4.1   0 Leptons

In the 0-lepton channel, the transverse momentum carried away by the neutrinos in the Z boson decay results in a large amount of MET. CMS has triggers that identify

Table 4.3: Below are the trigger paths used for the 0 lepton selections for all three years of Run 2 of the LHC.

Year	Trigger path(s)
2016	HLT_PFMET110_PFMHT110_IDTight
	HLT_PFMET120_PFMHT120_IDTight
	HLT_PFMET170_NoiseCleaned
	HLT_PFMET170_BeamHaloCleaned
	HLT_PFMET170_HBHECleaned
2017	HLT_PFMET120_PFMHT120_IDTight
	HLT_PFMET120_PFMHT120_IDTight_PFHT60
2018	HLT_PFMET120_PFMHT120_IDTight

events with large values of MET. Slightly different triggers are used for each year of Run 2 of the LHC, and they are listed in Table 4.3. The MET for the event must be larger than $170\,\mathrm{GeV}$. The trigger efficiency does not quite plateau at that point, as can be seen in Figure 4-5. However, the extracted efficiency scale factor for MC matches for both $t\bar{t}$ and $W + $ jets simulation at and above this value. During the 2018 run, a number of HCAL endcap modules were taken offline due to power supply problems. These modules were all located in the region $-1.57 < \phi < -0.87$, so an excess number of high MET events with ϕ_{MET} in that region were recorded in later 2018 runs when jets would have been registered in the deactivated detector elements. The resulting peak can be seen in Figure 4-6. To handle this, all events with $-1.86 < \phi_{\mathrm{MET}} < -0.7$ that occurred during and after run 319077, when the faulty detector elements were shut off.

In addition to the neutrino decay of the Z boson, the $b\bar{b}$ decay of the Higgs also needs to be selected and backgrounds need to be removed. Many of the following requirements are similar for all of the channels. For the Higgs decay, two jets are selected. One jet is the one with the highest b-tag score out of jets with $p_T > 60\,\mathrm{GeV}$. The other jet is the highest b-tag score out of jets remaining with $p_T > 35\,\mathrm{GeV}$. This distinction is important, because these two jets may not have the highest b-tag score of the event if the two highest b-tag scored jets both have $p_T < 60\,\mathrm{GeV}$, for example. However, we are not interested in such relatively low p_T systems since the di-jet system should be reasonably balanced against the MET requirement. To additionally enforce
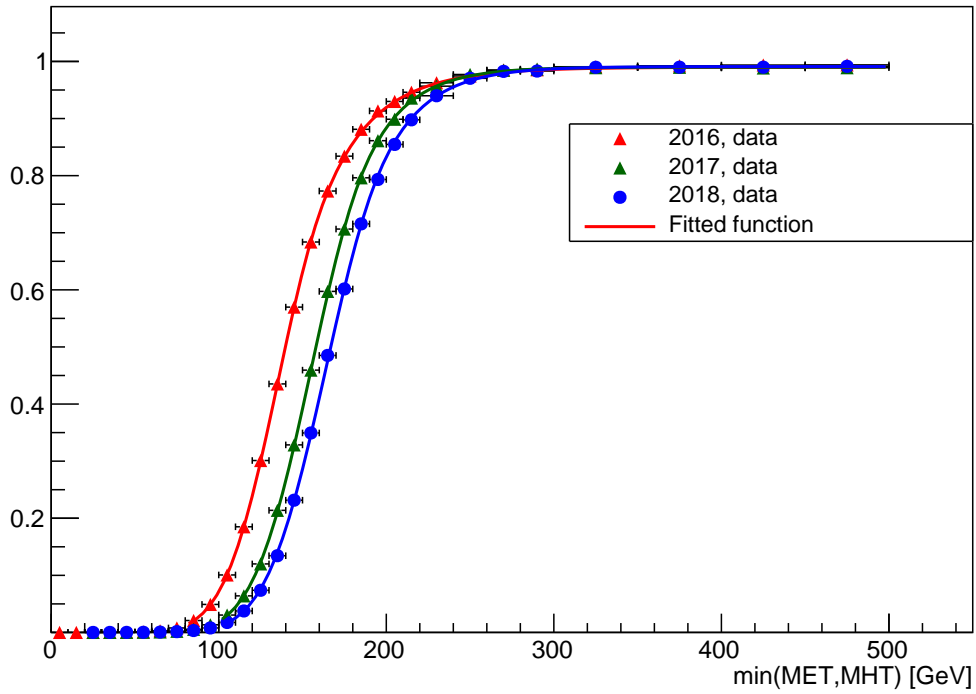
Figure 4-5: The efficiency for the MET triggers are shown for all three years in data.
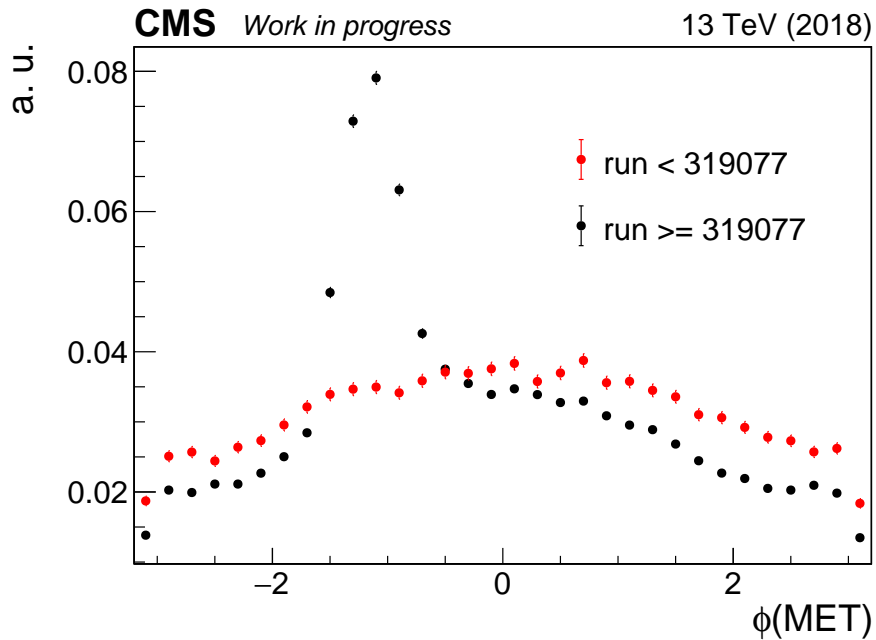
Figure 4-6: Above compares the MET ϕ distributions before and after shutting off problematic HCAL modules during for run 319077. The excess of events in the region $-1.86 < \phi_{\mathrm{MET}} < -0.7$ is caused by mismeasuring the momenta of forward jets in that region.

this balance, the di-jet system must satisfy $p_T > 120\,\text{GeV}$. Also the di-jet system is selected to be back-to-back with the MET by requiring $\Delta\phi(\text{MET}, jj) > 2.0$. The di-jet mass is required to be less than $500\,\text{GeV}$ to remove high energy combinatoric backgrounds. To increase purity of $Z\nu\nu$ processes, events are not considered if there are any isolated leptons with $|\eta| < 2.5$ and $p_T > 15\,\text{GeV}$. Finally, to reduce QCD background contributions, all jets in the event with $p_T > 30\,\text{GeV}$ must be a minimum distance from the event MET satisfying $\Delta\phi(\text{MET}, j) > 0.5$.

In the signal region, it is further required that the selected b-jets to be of high quality, the di-jet system has a mass close to the Higgs, and low additional hadronic activity. At least one of the b-jets must pass the medium working point, and the other b-jet must pass the loose working point. The value of the di-jet mass must satisfy $90\,\text{GeV} < m_{jj} < 150\,\text{GeV}$. At most one jet with $p_T > 30\,\text{GeV}$ is allowed in addition to the selected b-jets. Since the event has no high energy leptons, the particle flow MET and track MET must have agreement with $\Delta\phi(\text{MET}, \text{trackMET}) < 0.5$.

As mentioned earlier, three processes make up the majority of the backgrounds in the signal selection, and control regions must be defined for them. Two of them when a $Z \rightarrow \nu\bar{\nu}$ occurs recoiling off of jets. In one process, the recoiling jets are b-jets, and in the other, the jets are light jets. Since the fraction of actual $Z \rightarrow \nu\bar{\nu}$ events containing b-jets is not well known, they are scaled separately. The third set of background events that needs to be separately measured in this phase space are the $t\bar{t}$ semi-leptonic decays where the lepton falls outside of the detector acceptance. This type of background also results in a final state with high MET and b-jets.

Of the three processes, the Z + heavy flavor jets selection is the most similar to the signal selection. The only difference is the di-jet mass. Events with a mass between $50\,\text{GeV} < m_{jj} < 500\,\text{GeV}$ but not between $90\,\text{GeV} < m_{jj} < 150\,\text{GeV}$, which is the signal region window, are selected to quantify the Z + heavy flavor. The Z + light flavor selection includes the entire mass range, without a veto for the Higgs mass. It is instead enriched with light jets by requiring that the selected b-jet with a higher b-tag score fails the medium working point. The selection for $t\bar{t}$ events is different from the signal selection mostly because more hadronic activity is expected.

Table 4.4: Below are the trigger paths used for the 1 lepton selections for all three years of Run 2 of the LHC. Different triggers are used for the electron and muon channel, and are labeled separately.

Year	Lepton	Trigger path(s)
2016	e	HLT_Ele27_WPTight_Gsf
2016	μ	HLT_IsoMu24
		HLT_IsoTkMu24
2017	e	HLT_Ele32_WPTight_Gsf_L1DoubleEG
2017	μ	HLT_IsoMu27
2018	e	HLT_Ele32_WPTight_Gsf
2018	μ	HLT_IsoMu24

It uses the same full di-jet mass window as the Z + light flavor region, but requires at least two additional jets with $p_T > 30\,\text{GeV}$ instead of zero or one. Also, the $\Delta\phi(\text{MET}, \text{trackMET})$ requirement is dropped.

4.4.2   1 Lepton

In the 1-lepton channel, a single fully-selected isolated lepton as defined in Section 4.1.2 or Section 4.1.3 is required. CMS has triggers for isolated leptons, and the ones used in the analysis are listed in Table 4.4. That lepton must point in a similar direction of the MET, satisfying $\Delta\phi(\ell, \text{MET}) < 2.0$. The p_T of the reconstructed W boson, consisting of the vector sum of MET and the lepton, must satisfy $p_T > 150\,\text{GeV}$. If there are any additional leptons, the event is not used. The presence of an isolated lepton provides a much cleaner signal than in the zero lepton channel. Therefore, the kinematic requirements on the selected b-jets can be looser. The b-jets only need to satisfy $p_T > 25\,\text{GeV}$, and the di-jet system only needs $p_T > 100\,\text{GeV}$. Any events with $m_{jj} \geq 250\,\text{GeV}$ are not considered for any regions. The b-tagging requirement is the same as the one in the 0-lepton channel. There is a slightly tighter requirement on the di-jet direction of $\Delta\phi(\text{MET}, jj) > 2.5$.

In the signal region, the additional selection criteria are again similar to the 0-lepton channel. The di-jet mass window is the same, as is the requirement of at most one additional jet. The only other difference from the 0-lepton selection aside from the adjusted criteria listed above is that there is no dependence on the track MET

Table 4.5: Below are the trigger paths used for the 2 lepton selections for all three years of Run 2 of the LHC. Different triggers are used for the electron and muon channel, and are labeled separately.

Year	Leptons	Trigger path(s)
2016	e^+e^-	HLT_Ele23_Ele12_CaloIdL_TrackIdL_IsoVL_DZ
2016	$\mu^+\mu^-$	HLT_Mu17_TrkIsoVVL_Mu8_TrkIsoVVL
		HLT_Mu17_TrkIsoVVL_TkMu8_TrkIsoVVL
		HLT_Mu17_TrkIsoVVL_Mu8_TrkIsoVVL_DZ
		HLT_Mu17_TrkIsoVVL_TkMu8_TrkIsoVVL_DZ
2017	e^+e^-	HLT_Ele23_Ele12_CaloIdL_TrackIdL_IsoVL_DZ
2017	$\mu^+\mu^-$	HLT_Mu17_TrkIsoVVL_Mu8_TrkIsoVVL_DZ_Mass3p8
		HLT_Mu17_TrkIsoVVL_TkMu8_TrkIsoVVL_DZ_Mass8
2018	e^+e^-	HLT_Ele23_Ele12_CaloIdL_TrackIdL_IsoVL_DZ
2018	$\mu^+\mu^-$	HLT_Mu17_TrkIsoVVL_Mu8_TrkIsoVVL_DZ_Mass3p8

direction. The change to create the W + heavy flavor jets control region is the exact same as the Z + heavy flavor region in the 0-lepton channel. The mass window for the Higgs is vetoed. The requirement for the W + light flavor control region is also the same in terms of b-tagging, but the additional jet requirement is also removed. The $t\bar{t}$ control region is also the same in that the only changes from the signal region are a relaxed mass window, the requirement of at least two additional jets, and no requirement on the di-jet direction relative to MET.

4.4.3   2 Leptons

For the 2-lepton channel, two oppositely charged leptons with the same flavor are required. The triggers used that are designed to match this requirement are given in Table 4.5. They must have a mass satisfying $75 < m_{\ell\ell} < 105\,\mathrm{GeV}$. This process is clean enough to relax the kinematic requirements on the selected b-jets even further than the relaxed selection criteria of the 1-lepton channel. The selected b-jets only need to have a $p_T > 20\,\mathrm{GeV}$, and the di-jet system only needs $p_T > 50\,\mathrm{GeV}$. There are no requirements on the number of additional jets, outside of categorization for STXS bins. The di-jet system still needs to satisfy the tighter criterion of $\Delta\phi(jj, V) > 2.5$.

The signal region uses the same b-tag and mass value requirement as the other two channels. For the Z + heavy flavor control region, the di-lepton mass is narrowed

to $85 < m_{\ell\ell} < 97\,\mathrm{GeV}$ in order to remove $t\bar{t}$ events, and the usual Higgs mass veto is applied. The MET is also required to be low with MET $< 60\,\mathrm{GeV}$ for the $Z +$ heavy region, but no other. For the $Z +$ light flavor region, purity is achieved by requiring that both selected jets fail the loose working point for b-tagging. The $t\bar{t}$ region is then selected by requiring one selected jet to pass the tight working point. The di-lepton mass value also must be either $10 < m_{\ell\ell} < 75\,\mathrm{GeV}$ or $m_{\ell\ell} > 120\,\mathrm{GeV}$.

4.5   Boosted Analysis Selection

When the Higgs has an adequately high p_T, the jet clustering algorithms will often find both daughter particles as being part of a single jet. The boosted analysis only targets the STXS bins with $p_T(V) > 250\,\mathrm{GeV}$. The selection differs primarily in the fact that a single fat jet which passes a double b-tag test [57] is used to reconstruct the potential Higgs instead of two b-tagged jets. The double b-tagger used in this analysis is DeepAK8, a DNN as opposed to a BDT tagger.

Additional b jets outside of the fat jet are also counted to define selections. These come from the regular jet collections of Section 4.1.4. In order to count as an additional b-jet, the jet must pass the DeepCSV medium working point, have a $p_T > 25\,\mathrm{GeV}$, and be outside of the selected fat jet with $\Delta R(j, fj) > 0.8$.

4.5.1   0 Leptons

The boosted selection regions are designed to study the same processes as the resolved selection regions. The 0-lepton channel is enriched with signal events where the Z boson decays to neutrinos, so the 0-lepton channel has high MET. A requirement of MET $> 250\,\mathrm{GeV}$ is applied, which is consistent with the p_T requirement of the fat jet. As for the resolved analysis, the presence of any leptons leads to the event not being considered for the 0-lepton channel. To remove QCD background for all regions, the same selection requirement from the resolved analysis of $\Delta\phi(\mathrm{MET}, j) > 0.5$ for all jets with $p_T > 30\,\mathrm{GeV}$ is used.

In the signal region, jets must have a score of 0.8 or higher in the bbVsLight

output node of the DeepAK8 tagger. They must also have a soft drop mass in the range of $90 < m_{\mathrm{SD}} < 150 \,\mathrm{GeV}$. No additional jets outside of the fat jet are allowed in the event. The control regions are the same as for the resolved analysis: $Z +$ heavy flavor, $Z +$ light flavor, and $t\bar{t}$. For the $Z +$ heavy flavor control region, the requirement on the mass is changed to instead veto the Higgs mass window. For the $Z +$ light flavor, there is no mass requirement outside of the $m_{\mathrm{SD}} > 50 \,\mathrm{GeV}$ required for all fat jets. Orthogonality is enforced by requiring the bbVsLight score to be less than 0.8. For $t\bar{t}$, the lack of a mass requirement is also present, but there must be at least one b jet outside of the fat jet.

4.5.2   1 Lepton

For the single lepton channel, exactly one selected lepton must be present. It must also point in the same direction as the MET with $\Delta\phi(MET, \ell) < 2.0$. Otherwise, the selection criteria for the different 1-lepton regions are the exact same as for the boosted 0-lepton regions.

4.5.3   2 Leptons

For the two lepton channel, two oppositely charged, same flavor leptons must be present, as described in Section 4.4.3. These leptons must also have a mass near the Z boson mass for the signal region and the two $Z +$ jets regions of $87 < m_{\ell\ell} < 95 \,\mathrm{GeV}$. The selection for the signal and control regions are otherwise similar to the selections for the 0- and 1-lepton channels in the boosted analysis. The only difference is that instead of requiring a b jet outside of the fat jet for the $t\bar{t}$ control region, a mass veto of the di-lepton mass is applied, just as was done for the resolved analysis.

4.6   Overlap in Resolved and Boosted Selections

An important note is that each signal and control region described in this chapter must be orthogonal to all other regions. To prevent any statistical bias in the analysis,

both simulated and measured events must not be double counted. It is easy to make orthogonal selections within resolved or boosted categories by counting the number of leptons, comparing their flavor, and reversing other selections to go from signal to control regions. However, when making selections for the resolved and boosted channels, enforcing orthogonality is not as straightforward. The same PF candidates are reused to define two different kinds of jet collections, making it harder to reverse a single selection requirement to define a different category. The selections described so far result in some events simultaneously passing both boosted and resolved selections. To maximize the expected sensitivity of this analysis, most events that are in both resolved and boosted categories are assigned to the resolved category and omitted from the boosted. The only exception to this rule is when the event is assigned to one of the resolved control region and a boosted signal region. In that case, the event is assigned to its boosted category. This may be unexpected since the boosted regime tends to have a larger purity of signal than the resolved. However, this is likely due to the difference in classification strategy between the boosted and resolved regions. As will be described later, the resolved classifier makes use of more variables within the event than the boosted classifier does.

Chapter 5

Corrections and Uncertainties

Despite efforts to simulate LHC collisions accurately, a number of differences in the distributions predicted by MC and present in data arise. There are many reasons for this. The theory calculations used for simulation are not accurate since only leading order or next-to-leading order perturbations are used by the generators. After the generator step, the detector must be simulated, which is difficult due to the many composite parts, both active detector material and inactive electronics, that must be modeled in GEANT4. This is made more difficult since the detector degrades in its high radiation environment. The relation of generating particles and detector geometry also depends on the beam conditions, which are also not predicted exactly. The most accurate measurements of the beam profile are destructive to the beam, so measurements cannot be made during the run. Predictions from simulations are also made harder because there are so many events recorded at the LHC, so tails will be generated in obscure areas of phase space. These tails are difficult to model correctly, yet often contain the most interesting events. Corrections are made to the simulation by re-weighting based on detector conditions and efficiencies of particle identification, and by scaling the predicted energies based on particle type.

5.1 Efficiency Scale Factors

The identification of particles based on working points happen at different rates between simulation and collected data. These efficiencies are measured separately in data and Monte Carlo. The difference depends on a scale factor, which is measured as a function of particle energy and location within the detector. Inclusive selections are used for the measurement, which is then transferred to the events in the analysis. Since this analysis relies on counting leptons for its categorization, selection efficiency for both muons and electrons must be measured. A scale factor for the MET trigger efficiency is also derived for the 0-lepton categories.

5.1.1 Muons

To remove fake muons from events, muons are selected in three ways. Muon identification cuts are applied, isolation cuts are applied, and certain triggers are required. These three things each behave differently in MC and data. For each of these, a separate efficiency is derived in MC and data, and then a total scale factor is derived via the following formula.

$$\epsilon^\mu = \epsilon^\mu_{\text{ID}} \times \epsilon^\mu_{\text{ISO|ID}} \times \epsilon^\mu_{\text{Trig|ISO}} \tag{5.1}$$

The scale factor for muon ID, which is tight for positively counting muons, and loose for counting muons to veto, is ϵ^μ_{ID}. Passing the isolation cut, given the ID is scaled by $\epsilon^\mu_{\text{ISO|ID}}$, and the scale factor for the trigger, after passing the isolation cut is $\epsilon^\mu_{\text{Trig|ISO}}$.

These are each separately measured via the tag and probe method. A single tag muon is selected in events with a single muon trigger. For events with a second, oppositely-charged muon that reconstructs the Z boson resonance with the first, the second muon, called a probe, is checked for identification efficiency. Since the probe muon does not rely on a categorization for the selection, only on the successful reconstruction of a Z boson, the collection of probe muons make up an unbiased sample for the efficiency measurement. A Breit-Wigner convoluted with a Gaussian and a falling
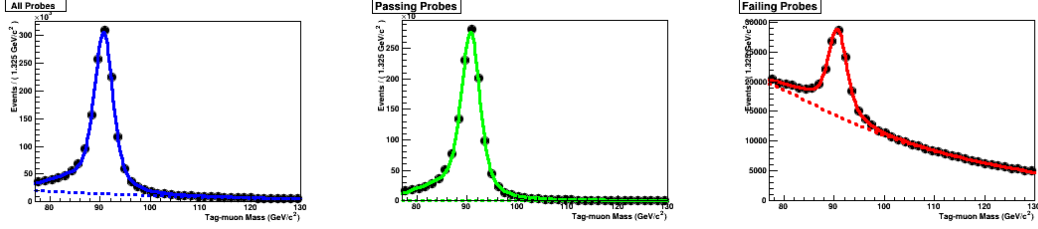
Figure 5-1: Example tag and probe fits are shown above. The left plot shows the dilepton mass fit for all tagged events. The tagging procedure yields little background, which allows for a precise measurement. The middle plot shows the events where the probe passed muon identification, and the right plot show the events when the probe failed.

combinatoric background is fit to the peak to estimate the contribution of the Z boson resonance in both the passing and failing probes. A scale factor is then applied to MC to match the data efficiency. Figure 5-1 shows one of these measurements of efficiency in Data. Each of the efficiency measurements and scale factors are binned in muon p_T and η. The p_T bin boundaries are $[20, 25, 30, 40, 50, 60, \infty)$ in GeV. Bins in $|\eta|$ are delimited at $[0, 0.9, 1.2, 2.1, 2.4]$.

The uncertainties from the scale factors are also combined in the final analysis. The identification efficiencies in data are slightly lower than in Monte Carlo. Loosely identified muons require an average scale factor of 0.998 ± 0.002 applied to simulated events, and tightly identified muons require an average scale factor of 0.98 ± 0.005 [67]. In general, the scale factors are half a percent or one percent lower, respectively, at values of $|\eta| > 2.0$, and flat across p_T bins.

5.1.2 Electrons

The electron scale factors are measured in a similar manner as the muon scale factors, using the tag and probe method. The relative difficulty in reconstructing electrons, which are contained in the ECAL, compared to muons leaving tracks through the muon chambers means that a reconstruction scale factor, ϵ^e_{RECO}, is also factored into the full scale factor.

$$\epsilon^e = \epsilon^e_{\text{RECO}} \times \epsilon^e_{\text{ISO+ID|RECO}} \times \epsilon^e_{\text{Trigger|ISO+ID+RECO}} \tag{5.2}$$

The rest of the efficiency scale factors follow the same definition patterns given for the muons. The binning for the electron scale factors is also the same as the binning for the muon scale factors. The scale factor is around 1.0 ± 0.02 per electron for most of the bins.

5.1.3 MET Trigger Scale Factors

The MET trigger efficiencies for each year are measured in events with a single electron and large MET. This allows tagging events with the single electron triggers, and selecting mostly events from W+jets and $t\bar{t}$ events. The events are additionally required to be similar to the events of interest in the analysis. There must be two jets with $p_T > 20\,\text{GeV}$ and $|\eta| < 2.5$, and the events must pass the MET filters. In addition, it is required that the $\Delta\phi$ between the electron and MET is less than 2.5. The nominal values for the scale factors are derived by measuring the MET trigger efficiency in data and scaling the trigger efficiency in the $t\bar{t}$ Monte Carlo sample to match. Uncertainties are derived by repeating the measurement using a W+jets sample. The scale factors for the nominal measurement and the W+jets sample are shown in Figure 5-2. The MET trigger uncertainty is derived by finding the maximum deviation between the two lines above the $170\,\text{GeV}$ requirement in the 0-lepton channel, and applying that difference as a shape uncertainty.

5.1.4 Scale Factors for b-Tagging Identification

The identification efficiency of b-tagged jets can be measured in data by selecting fully leptonic $t\bar{t}$ decays with two jets. This sample is highly enriched in b-jets [57]. The misidentification probability in data is obtained by selecting inclusive multi-jet events, and plotting a tagging distribution that is trained to identify light-flavored jets, called the "negative tagger" [68], alongside the b-tagging discriminator. Comparing these distributions between data and simulation give a scale factor that can be applied to simulation. Because of the large number of events used for these studies, efficiencies are applied as a function of p_T and η of each b-tagged jet in the simulated event. The
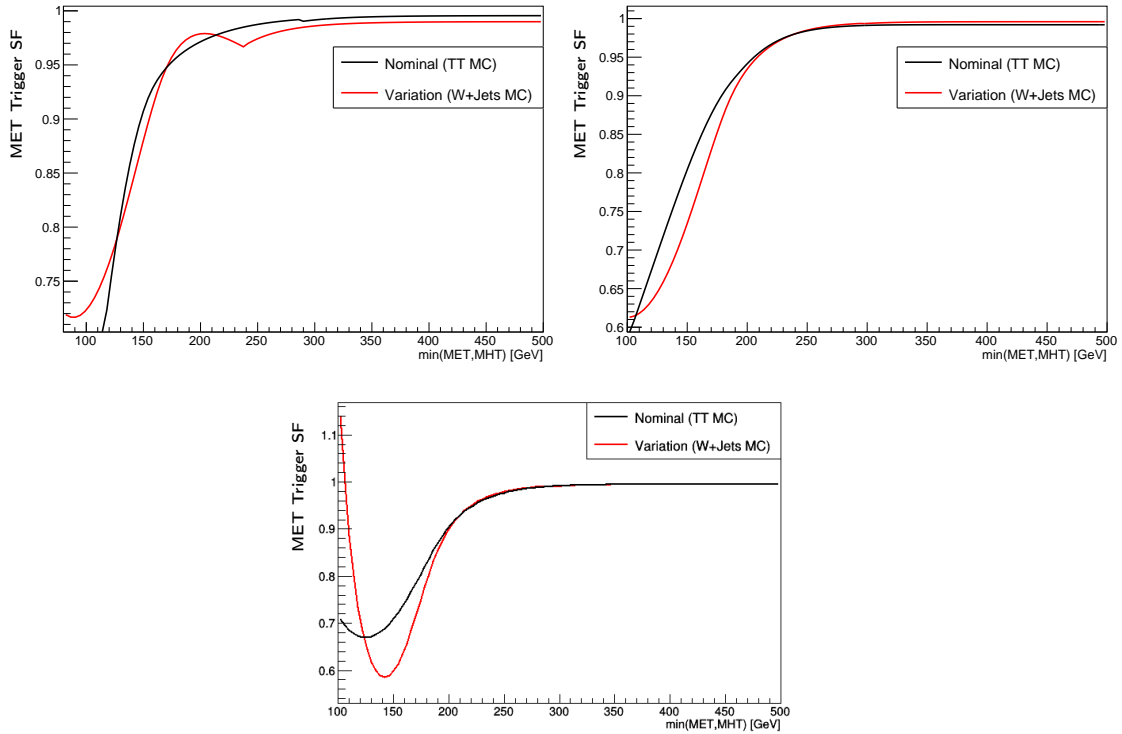
Figure 5-2: The MET trigger scale factors for two varying simulations are given. The $t\bar{t}$ sample is used for the nominal scale factor, and the W+jets is used to estimate systematic uncertainties. The top left plot shows the measurements for 2016, the top right plot shows 2017, and the bottom plot shows the MET scale factors for 2018. Each scale factor is a function of the minimum MET or missing hadronic energy. Remember the MET cut for the 0-lepton channel is at $170\,\mathrm{GeV}$, so the different behavior in 2018 below $150\,\mathrm{GeV}$ does not affect the analysis. The kink in the 2016 W+jets line is caused by the merging of two fits, but since this is a variation to assess uncertainties, it is not necessary to fix.

double-b tagger used for fat jets is evaluated in data by selecting events with gluons decaying into $b\bar{b}$ pairs that cluster into the fat jet [69].

5.2 Energy Corrections of b-Jets

This analysis does depend in particular on b-jet energy predictions, which do not have recommended corrections from the CMS collaboration, so a bespoke measurement is done for this analysis. Even when distributions of individual variables agree between MC and data, correlations are often different. These correlations are also important in the evaluation of a DNN. The DNN used to estimate the energy of b-jets therefore has differing performance in MC and data. In particular, it is better at estimating the true energy of a b-jet in MC. The energies evaluated in MC must be smeared in order to accurately simulate the resolution of jets in data after they have been modified by the DNN regression.

One way to measure jet energy resolution is to consider an event where a jet is recoiling off of a Z boson that decays into leptons. In principle, the Z boson's transverse momentum is balanced with the jet's transverse momentum. Measurements of lepton energies in the CMS detector is relatively precise, so the ratio of the reconstructed jet's p_T to the Z boson's p_T allows measurement of the jet energy resolution. Ideally, this measurement would be done with an collision resulting in one Z boson decay, and one jet. However, this is an infrequent occurrence. Instead, events with two jets are selected, with one jet having relatively low p_T. A fit is performed to estimate resolution characteristics where the second jet would have $p_T = 0\,\mathrm{GeV}$.

These events are selected using the following requirements:

- Exactly two muons or two electrons must pass the selection criteria for the di-leptons channels described in Section 4.4.3.

- The two selected leptons must be oppositely charged.

- The di-lepton kinematics must satisfy
  $p_{T,\ell\ell} > 100\,\mathrm{GeV}$ and $71\,\mathrm{GeV} < m_{\ell\ell} < 111\,\mathrm{GeV}$.
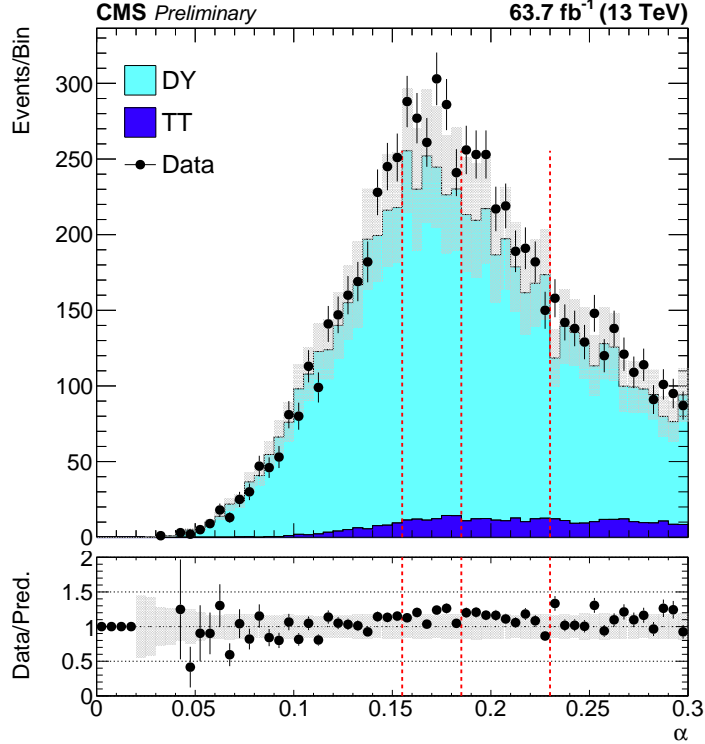
Figure 5-3: The distribution for α is shown above, along with the binning boundaries.

- Exactly two jets must pass the pre-selection described in Section 4.4.3.

- The leading jet must also satisfy $\Delta\phi(j, \ell\ell) > 2.8$

- The ratio between the sub-leading jet p_T and the di-lepton p_T must be less than 0.3.

- The leading jet must pass the tight working point for the b-tagger, as defined for each year in Table 4.1.

The selected events are divided into four bins of $\alpha = p_{T,j2}/p_{T,\ell\ell}$ with bounds $(0, 0.155, 0.185, 0.23, 0.3)$. The distribution of α is shown in Figure 5-3, along with the bin boundaries. These bins were selected to give approximately the same number of data events in each bin. The jet response $(p_{T,j1}/p_{T,\ell\ell+j2})$ is plotted in each bin, with uncertainties from renormalization and refactorization scale weights and parton shower weights. These histograms of jet response are shown in Fig. 5-4. From each plot, the mean (μ) and the standard deviation (σ) are extracted. The relative resolution (σ/μ) is fit as a function of α.

91
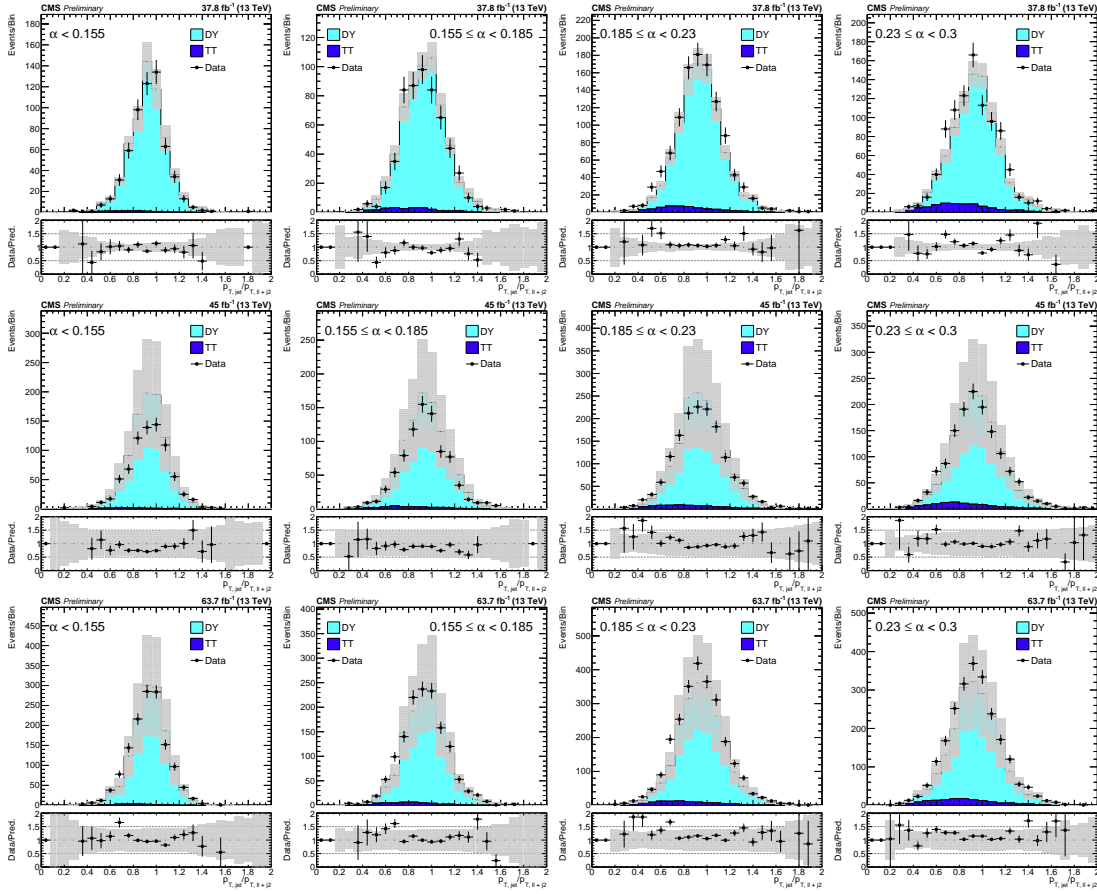
Figure 5-4: The histograms of response for each event are shown above. The top row shows 2016, the middle shows 2017, and the bottom row shows 2018 histograms.
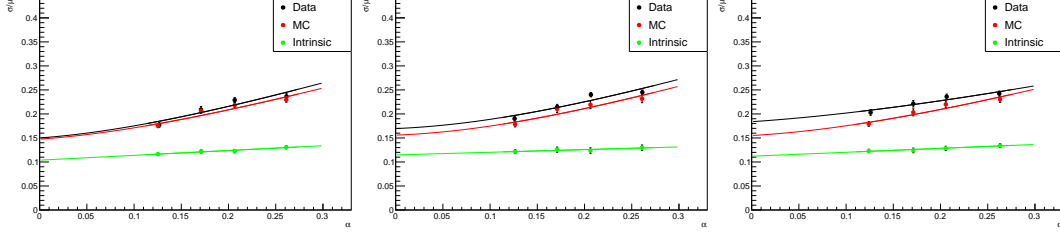
Figure 5-5: The fits to Data, MC, and intrinsic resolutions are shown. From left to right are the fits for 2016, 2017, and 2018.

Table 5.1: The extracted corrections needed for each year of data as a percent of the jet's p_T.

Year	Scaling	Resolution Difference
2016	0.998 ± 0.019	0.017 ± 0.060
2017	1.020 ± 0.023	0.088 ± 0.071
2018	0.985 ± 0.019	0.080 ± 0.073

$$f(\alpha) = (m \times \alpha) \oplus b \times (1 + c_k \times \alpha) \tag{5.3}$$

The c_k term is fixed by a linear fit to the MC's intrinsic jet resolution ($p_{T,reco}/p_{T,gen}$) over α as $c_k = m_0/q_0$. The fit results are shown in Fig. 5-5. The resolution is corrected by scaling the difference between $p_{T,reco}$ and $p_{T,gen}$ by b_{data}/b_{MC}. This causes the post-smearing fits to agree at $\alpha = 0$. Uncertainties are extracted from the fit uncertainties of b for data and MC. The resulting smearing values are in Table 5.1. The difference in smearing values for 2016 can be explained by a few characteristics of the simulation and data. First, the 2016 data has been more thoroughly studied and can generally be expected to have more robust calibration. The detector behavior also degrades over time as it accumulates radiation damage. Though this is accounted for in calibrating, it does make more precise simulations of detector response less accurate. Finally, the simulation for 2016 uses a different PYTHIA tune.

93

5.3 Theoretical Corrections

There are known inaccuracies in the simulations used in this analysis that can be understood at the theoretical level. However, less accurate calculations are run because they are simpler and faster to compute the results for. A trade-off must be made between the accuracy of the simulation and being able to generate enough simulated events to fill the phase space relevant for all analyses CMS collaborators are undertaking.

5.3.1 LO to NLO Reweighting

For each V+jets process, $Z \to \nu\nu$, $W \to \ell\nu$, and $Z \to \ell\ell$, there are three distinct processes that are simulated separately.

- b-quarks are generated in the matrix element

- b-quarks are not in the matrix element, but are produced in the parton shower

- No b-quarks are present in the event

For the first two, VBJets and VJetsBGenFilter datasets are used respectively, as listed in Appendix B, but they are only generated for $p_T(V) > 100\,\text{GeV}$. This small phase space and low relative cross section for generating b quarks means that it is feasible to simulate these processes to Next to Leading Order (NLO) diagrams. However, expected events with no b-quarks and events with $p_T(V) < 100\,\text{GeV}$ are more numerous, so more simulated events must be generated. To speed up calculations, Leading Order (LO) samples are used for these processes.

One common way to make LO samples more accurate is to generate NLO in QCD samples and reweight the LO sample to match generator-level kinematics. This can be done with an inclusive selection so that the statistical limitations of the available NLO samples are not significant. While the samples in this analysis were generated using just MadGraph5 [41], the following inclusive samples using aMC@NLO [70] are used to reweight the LO simulation as a function of $p_T(V)$.

94

- Z1JetsToNuNu_M-50_LHEZpT_50-150_TuneCP5_13TeV-amcnloFXFX-pythia8

- Z1JetsToNuNu_M-50_LHEZpT_150-250_TuneCP5_13TeV-amcnloFXFX-pythia8

- Z1JetsToNuNu_M-50_LHEZpT_250-400_TuneCP5_13TeV-amcnloFXFX-pythia8

- Z2JetsToNuNu_M-50_LHEZpT_50-150_TuneCP5_13TeV-amcnloFXFX-pythia8

- Z2JetsToNuNu_M-50_LHEZpT_150-250_TuneCP5_13TeV-amcnloFXFX-pythia8

- Z2JetsToNuNu_M-50_LHEZpT_250-400_TuneCP5_13TeV-amcnloFXFX-pythia8

- WJetsToLNu_0J_TuneCP5_13TeV-amcatnloFXFX-pythia8

- WJetsToLNu_1J_TuneCP5_13TeV-amcatnloFXFX-pythia8

- WJetsToLNu_2J_TuneCP5_13TeV-amcatnloFXFX-pythia8

- DYJetsToLL_M-50_TuneCP5_13TeV-amcatnloFXFX-pythia8

The change to the $p_T(V)$ spectra and the resulting improved agreement between data and MC are shown in Figure 5-6.

Note that these NLO corrections are only applied to the HT-binned V + jets samples. These samples are also combined with b-enriched samples that are already simulated in NLO. It is common to remove events with simulated b hadrons from the HT-binned samples in this case and use only the b-enriched samples for those events. However this is not possible in this analysis because there are not enough simulated events with high vector boson p_T in the b-enriched samples. Therefore, both the NLO samples that are used to reweight the LO samples also need to simulate events containing b hadrons.

5.3.2 Electroweak and QCD Corrections to Background

Though the LO to NLO reweighting does improve the agreement between MC and data, there is also an additional small correction. Higher order electroweak and QCD corrections predict a slightly softer $p_T(V)$ spectrum [71]. The reweighting function is shown in Figure 5-7.
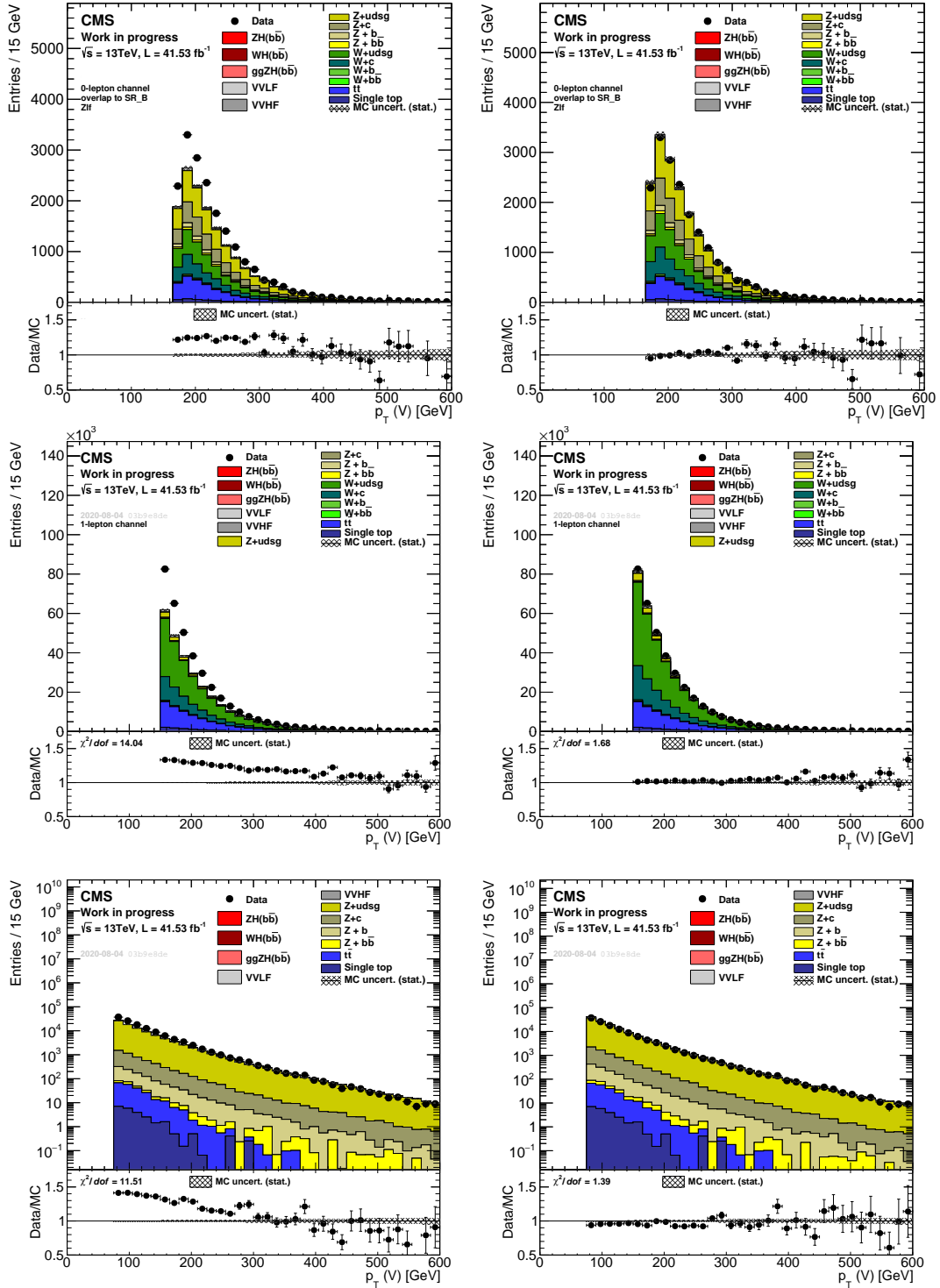
Figure 5-6: The shapes for $p_T(V)$ in the V+light control regions are shown for 2017. The left plots are before LO to NLO reweighting, and the right plots are after the correction factor is applied. The top row of plots show the 0-lepton control region, the middle row shows the 1-lepton selection, and the bottom row shows the 2-lepton control region.
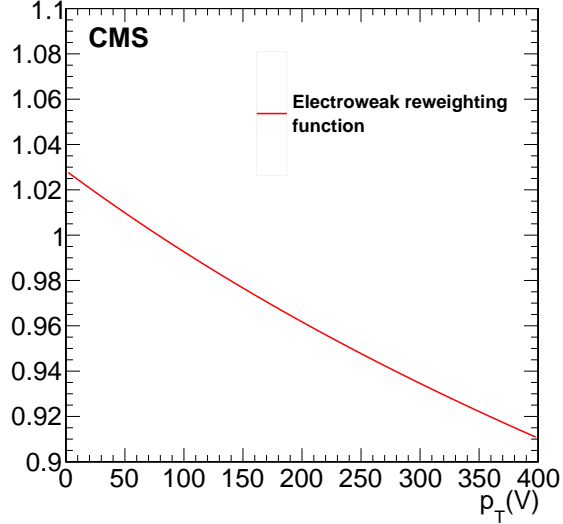
Figure 5-7: Electroweak correction as a function of boson $p_T(V)$ for the V+jets samples.

5.3.3 Corrections to Signal Process

The samples produced for signal are NLO. For an accurate measurement of the *VH* cross section, the simulation is scaled to Next to Next to Leading Order (NNLO) QCD effects, where two gluons are added to the Feynman diagrams. An example NNLO VH process is shown in Figure 5-8. Electroweak corrections can still be factored out, and total *VH* production cross sections are given as the following [72].

$$\sigma^{WH} = \sigma^{WH,DY}_{\text{NNLOQCD}}(1 + \delta_{\text{EWK}}) + \sigma_{t-\text{loop}} + \sigma_\gamma \tag{5.4}$$

$$\sigma^{ZH} = \sigma^{ZH,DY}_{\text{NNLOQCD}}(1 + \delta_{\text{EWK}}) + \sigma_{t-\text{loop}} + \sigma_\gamma + \sigma^{ggZH} \tag{5.5}$$

Each σ on the RHS of the equations refers to a different Higgs production mechanism. The full correction shape as a function of $p_T(V)$ is shown in Figure 5-9.

5.4 Application of Uncertainties

In addition to the corrections and their associated uncertainties listed so far, there are additional experimental and theoretical uncertainties. Some of these uncertainties only affect the normalization of samples. Other uncertainties affect the shape of
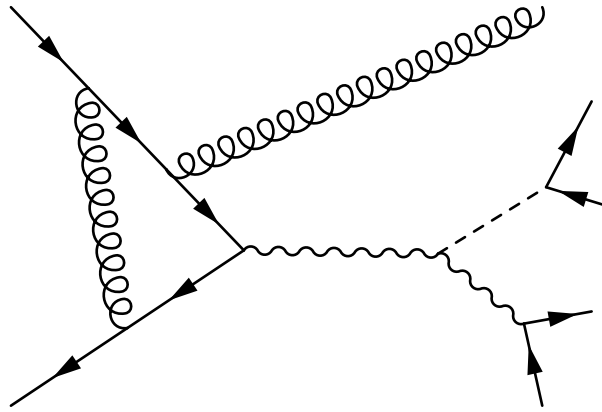
Figure 5-8: Above is an example diagram used to calculate NNLO QCD corrections to the VH production cross section.
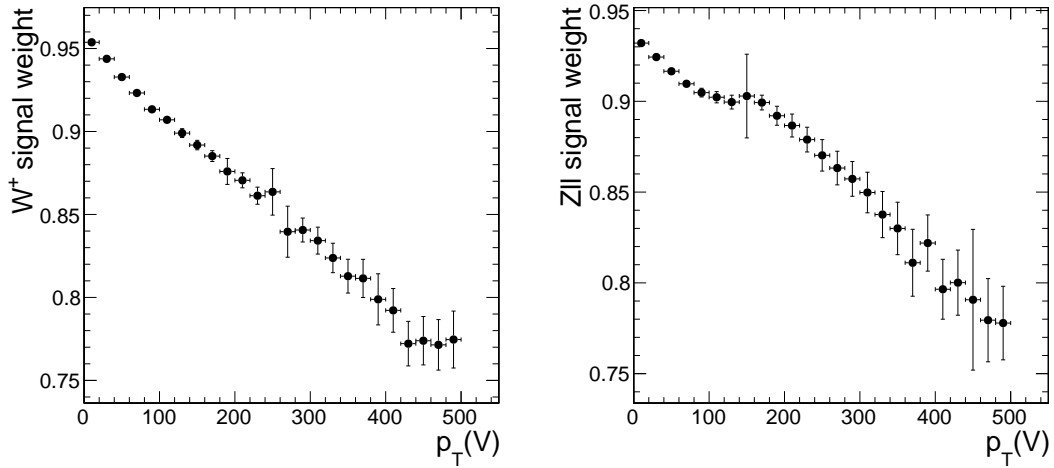
Figure 5-9: The full set of corrections to two signal samples are shown above. On the left is production of the Higgs with a W^+ boson. On the right is Higgs production with a Z decaying to two leptons.

variables that are used in the DNN discriminator described in the following chapter. The final set of uncertainties is a set of migration uncertainties to account for variation near the STXS bin boundaries.

5.4.1 Normalization Uncertainties

The following uncertainties only affect the normalization of the simulated samples.

- The luminosity measurement has an uncertainty of 2.5% for 2016 and 2018, and 2.3% for 2017 [73–75]. Since the method of luminosity measurement was the same for all three years, the uncertainties are partially correlated across the three years.

- The theoretical uncertainty of the branching ratio of the Higgs Boson to bottom quarks is 0.5% [72].

- QCD scale uncertainties for the signal production cross section are implemented as acceptance uncertainties.

- The uncertainties to the proton's PDF and α_s for the signal processes are 1.6% for *ZH* production, and 1.9% for *WH* production.

- At the high $p_T(V)$ regions of this analysis, the theoretical uncertainties grow. For the NLO electroweak corrections, the uncertainties are 2%. For the NNLO QCD correction, the uncertainty is 5%.

- For smaller background processes without a dedicated control region, primarily single-top and di-boson processes, a 15% normalization uncertainty is applied. This number comes from measured cross sections for single-top [76] and di-boson [77].

- The lepton identification efficiency uncertainties are applied as a flat uncertainty to each channel, as appropriate.

- The MET trigger efficiencies are also applied as a normalization uncertainty to the 0-lepton regions, as mentioned before.

5.4.2 Shape Uncertainties

In addition to normalization, there is some uncertainty in the shapes of the various distributions that are fit. This is typically caused by uncertainties in variables that are used in the event selection, but can also be a result of normalization uncertainties that only affect a subset of events in a selection. The shape variations are generated by varying different parameters up and down and re-applying selections and weights to determine the final shape of each distribution. Uncertainties of simulated processes that cause shape differences are the following.

- The energy resolution, and therefore the exact final energy, of simulated b jets are varied by increasing or decreasing the distance between the generator-level p_T and the reconstructed jet p_T.

- The energy scale is varied by varying all jet energies up and down by one standard deviation [56].

- The b-tagging efficiency uncertainties are calculated for multiple p_T and η bins. The uncertainties for these different bins are uncorrelated [78], so they each affect a different subset of events when applied.

- The b-tagging efficiencies for fat jets are applied in bins of p_T [79]. The uncertainties are decorrelated in these p_T bins.

- The uncertainties from the limited number of Monte Carlo events is handled using an approximation of the Barlow-Beeston method [80].

5.4.3 Migration Uncertainties

An additional uncertainty is a applied that accounts for simulated events crossing the STXS bin boundaries due to uncertainties in the vector boson p_T. The effect of each nuisance parameter is anti-correlated between the bins on either side of the boundary. Most bins have two uncertainties allowing events to cross into lower or higher $p_T(V)$ bins, with the exception of the highest STXS bins, which have no upper bounds.

Due to limited availability of studies on this effect, these systematic uncertainties are given large values before the global fit is performed, ranging from 20% to 50% for each individual boundary. These are worth noting since they have the largest affect of all systematic uncertainties on the pre-fit simulation uncertainties, though they are significantly constrained by the fit process.

Chapter 6

Analysis Results

With the selection and correction procedure outlined in the previous chapters, data gathered at CMS is gathered and compared to predictions of the Standard Model. Before looking at the signal events, various distributions are plotted in each of the control regions to ensure that the background events are simulated accurately. After some example distributions, discriminating variables to be used in the signal selections are defined. Right before getting the final measurement, a cross-check analysis is done using the same techniques on the more common VZ process.

6.1   Run 2 Data Collection

The CMS detector collected proton-proton collision data at $\sqrt{s} = 13\,\mathrm{TeV}$ over three years during Run 2 of the LHC. In 2016, CMS collected $35.9\,\mathrm{fb}^{-1}$ of data [73]. In 2017, data corresponding to $41.5\,\mathrm{fb}^{-1}$ were collected [74]. In 2018, CMS collected $60\,\mathrm{fb}^{-1}$ of collision data [75]. This data, totaling about $137\,\mathrm{fb}^{-1}$ was used for the analysis.

As noted in defining the MET selection of 2018, the CMS detector performance does change over time as the systems are exposed to radiation from the LHC environment. Therefore, each of the three years is treated separately with many systematic uncertainties related to the detector uncorrelated between the years. Even though these experimental values float separately, the final result is that of a combination of all three years.

6.2 Control Region Plots

A number of variables are compared in all of the control regions to ensure that the data and simulation are in reasonable agreement. A selection of these plots are shown in Figures 6-1 through 6-6. These are the distributions before fitting the control and signal regions to data, so agreements improve after the fit. The number of variables, selections, and separation of the data collection years means that this is just a fraction of the total number of plots produced. However, this subset shown demonstrate several discrepancies that needed to be accounted for in our systematics, such as the $p_T(V)$ trend which can be seen in Figure 6-1, the p_T of individual jets in Figures 6-2 and 6-4, and normalization of certain background samples, like the $Z +$ jets samples in Figure 6-6. Figure 6-3 shows how difficult it is to model the b-tagging variables. Systematic uncertainties are applied to all of these variables to account for the modeling difficulties, but the plots in Figures 6-1 through 6-6 only show the statistical uncertainties in the simulation.

6.3 Multivariate Discriminator

In each STXS bin, a multivariate discriminator is evaluated for the signal region which separates the signal events from background events. A Deep Neural Network (DNN) is trained for the resolved selection, and a Boosted Decision Tree (BDT) is trained for the boosted selection. The DNN was also used to train a multi-classifier for use in the resolved $Z +$ heavy jets control region. Half of the simulated events are used in these trainings. To prevent biases, as well as to avoid over-training, the other half of the simulated events are used to compare classifier distributions with data.

6.3.1 Resolved DNN

The DNN classifier for distinguishing background and signal events is prepared using Keras with a Tensorflow back-end using an Adam optimizer. It has five hidden layers. The number of nodes in each layer, from input to output, is 512, 256, 128, 64, 64, and
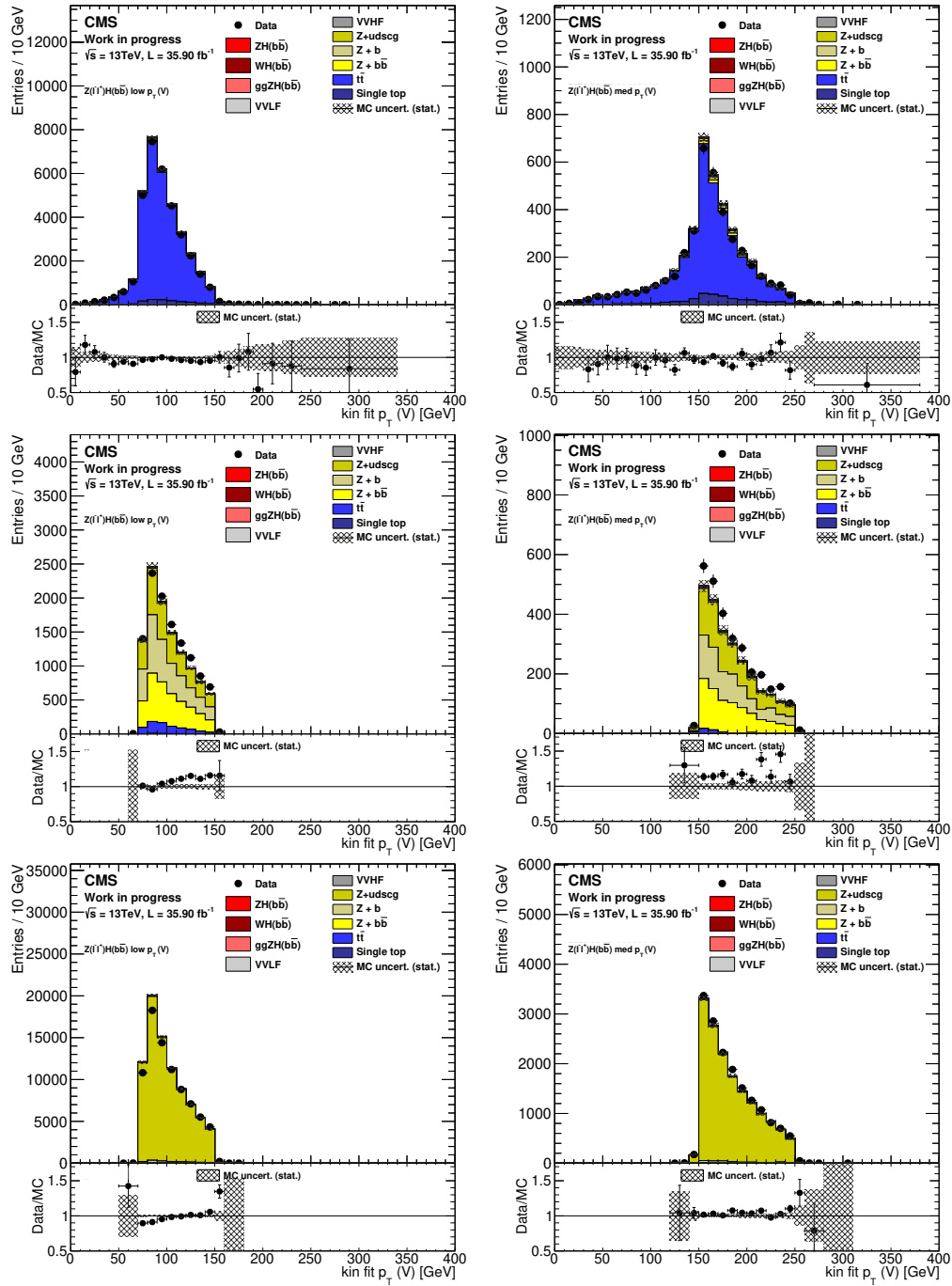
Figure 6-1: The vector boson p_T is plotted for multiple control regions and bins in an inclusive selection of both the $Z \to \mu\mu$ and $Z \to ee$ two lepton channels with 2016 data. The top row shows the $t\bar{t}$ control region, the middle row shows the $Z+$ heavy jets, and the bottom row shows the $Z+$ light jets control region. The left column shows the low $p_T(V)$ bin of 75 to 150 GeV, and the right column shows the medium $p_T(V)$ bin of 150 to 250 GeV.
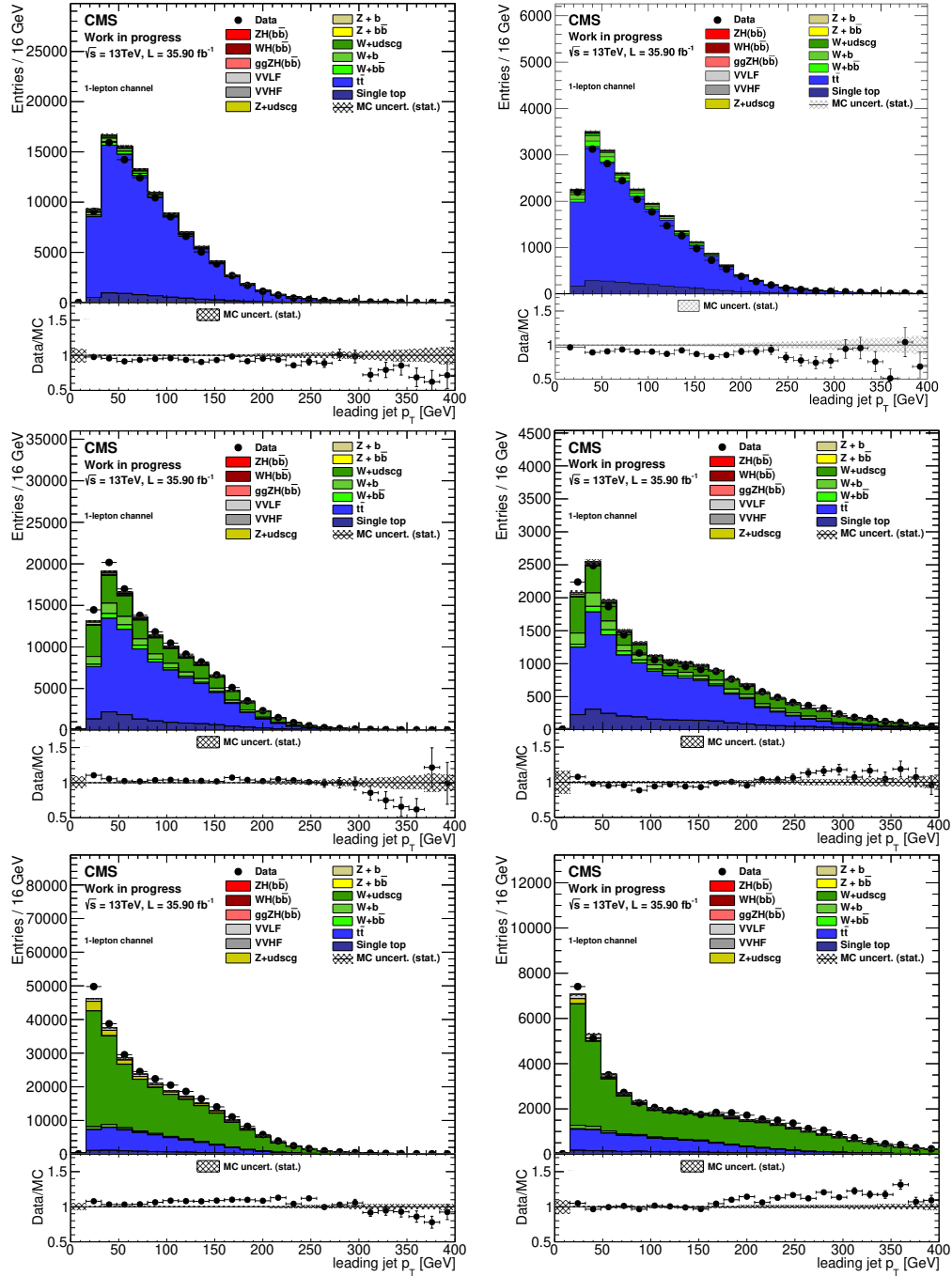
Figure 6-2: The p_T of the jet with a higher DeepCSV value is plotted for multiple control regions and bins in an inclusive selection of both the $W \to \mu\nu$ and $W \to e\nu$ one lepton channels with 2016 data. The top row shows the $t\bar{t}$ control region, the middle row shows the $W+$ heavy jets, and the bottom row shows the $W+$ light jets control region. The left column shows the medium $p_T(V)$ bin of 150 to 250 GeV, and the right column shows the high $p_T(V)$ bin of 250 GeV and above.
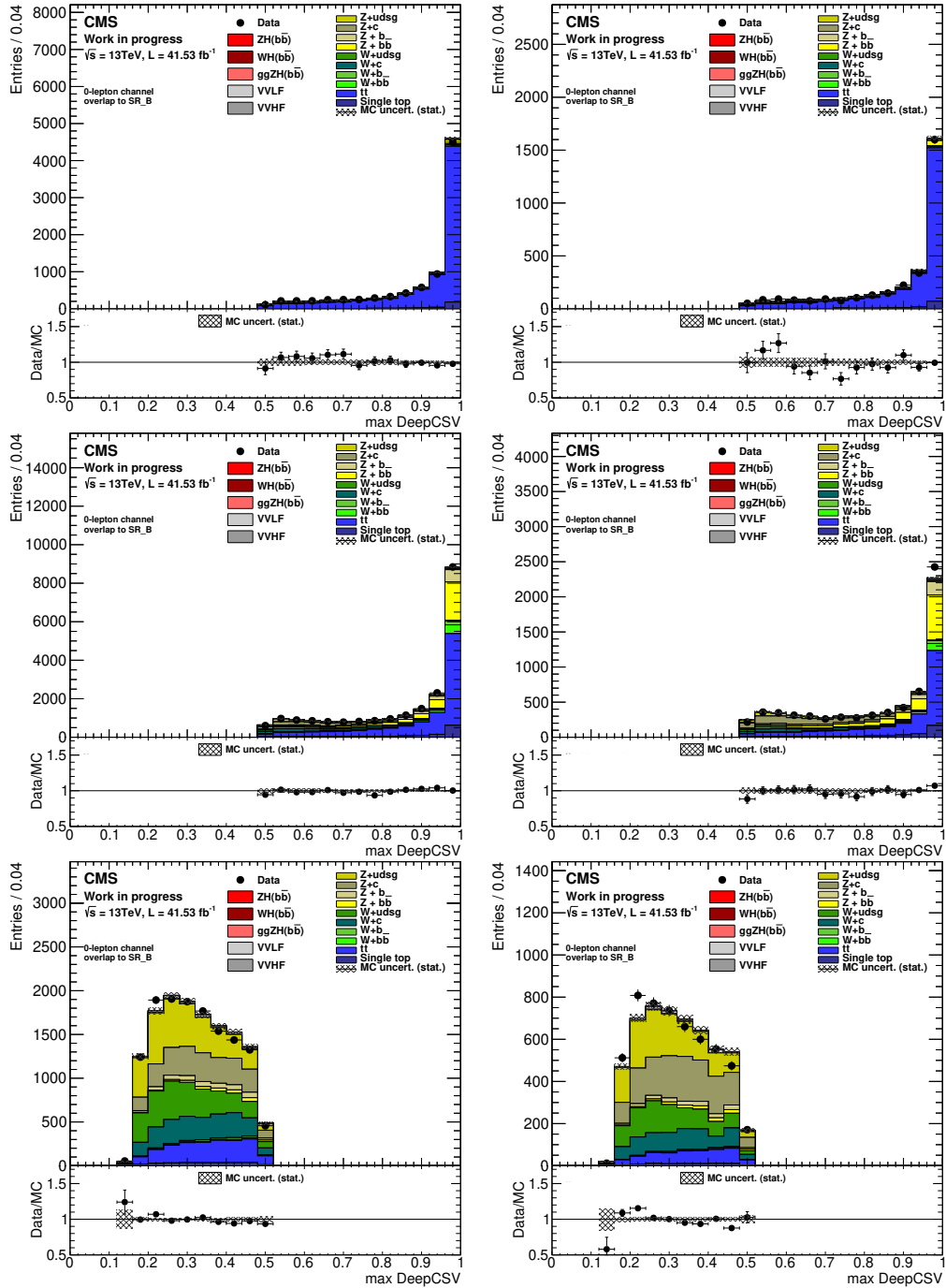
Figure 6-3: The DeepCSV value of the jet with a higher DeepCSV value is plotted for multiple control regions and bins in the $Z \to \nu\bar{\nu}$ zero lepton channel with 2017 data. The top row shows the $t\bar{t}$ control region, the middle row shows the $Z+$ heavy jets, and the bottom row shows the $Z+$ light jets control region. The left column shows the medium $p_T(V)$ bin of 150 to 250 GeV, and the right column shows the high $p_T(V)$ bin of 250 GeV and above.
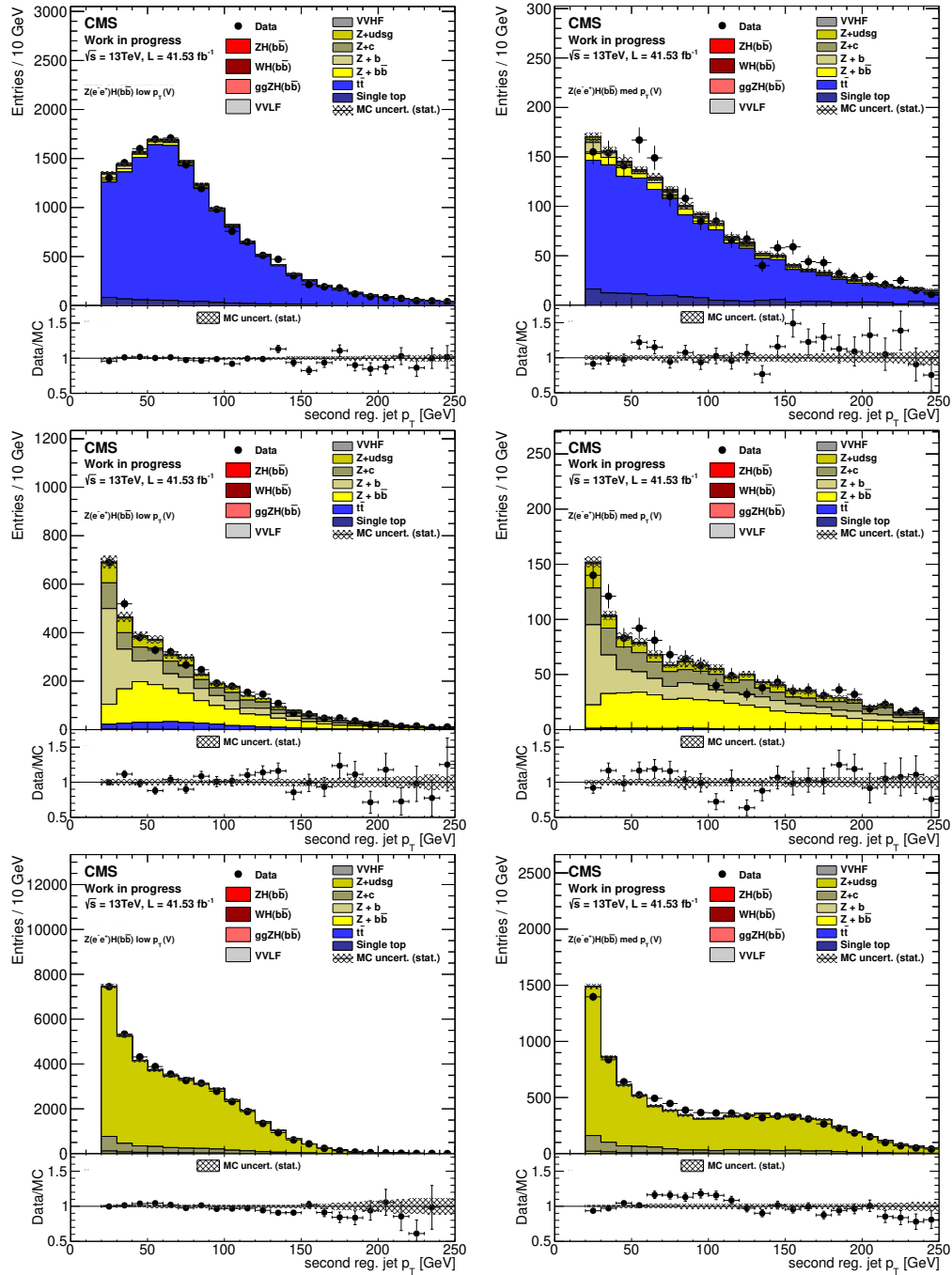
Figure 6-4: The p_T of the jet with a lower DeepCSV score is plotted for multiple control regions and bins in the $Z \to ee$ two lepton channel with 2017 data. The top row shows the $t\bar{t}$ control region, the middle row shows the $Z+$ heavy jets, and the bottom row shows the $Z+$ light jets control region. The left column shows the low $p_T(V)$ bin of 75 to 150 GeV, and the right column shows the medium $p_T(V)$ bin of 150 to 250 GeV.
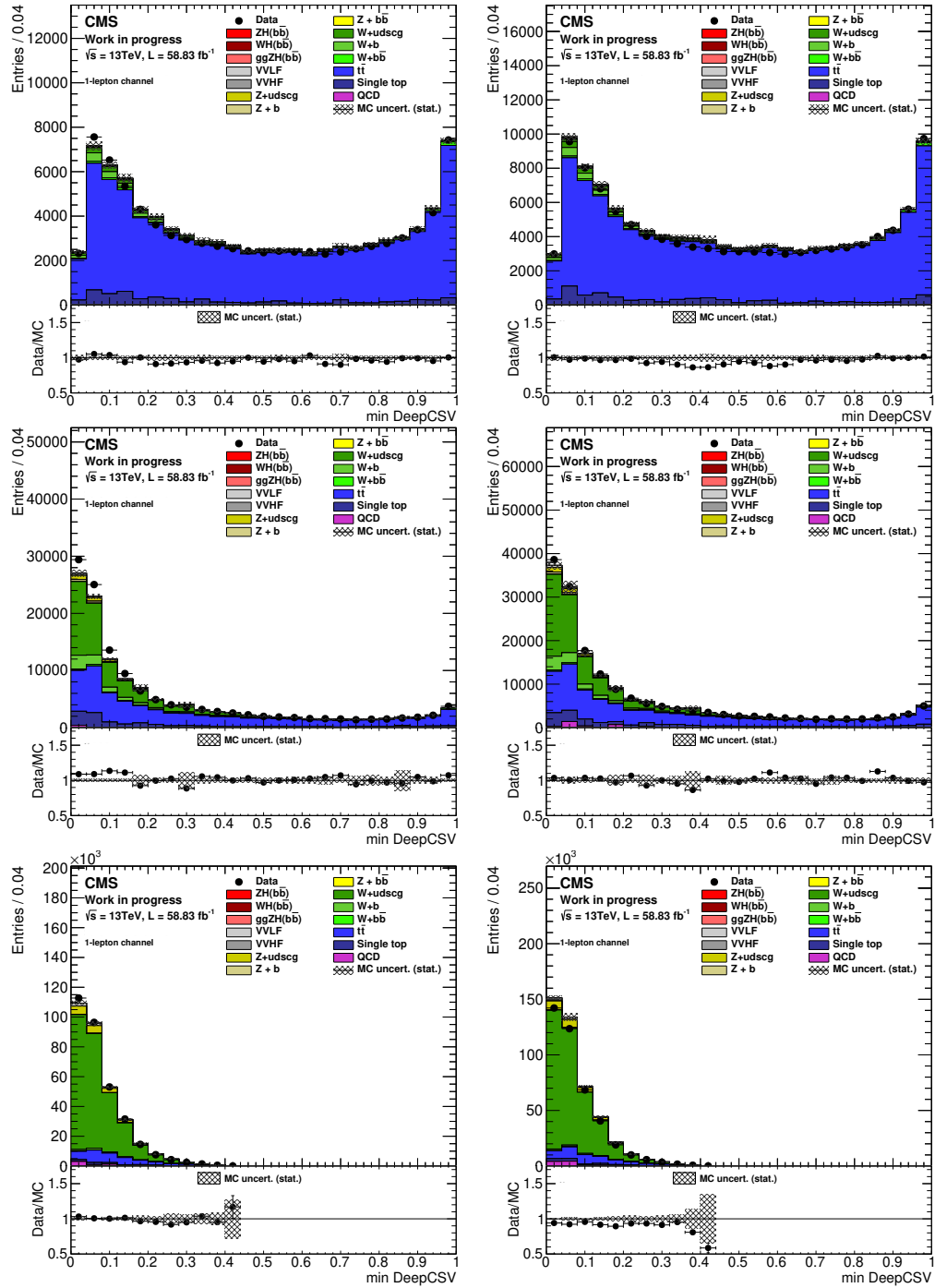
Figure 6-5: The DeepCSV value of the jet with a higher DeepCSV value is plotted for multiple control regions in an inclusive selection that contains all STXS bins with $p_T(V)$ greater than $150\,\mathrm{GeV}$ in the one lepton channels with 2018 data. The top row shows the $t\bar{t}$ control region, the middle row shows the $W+$ heavy jets, and the bottom row shows the $W+$ light jets control region. The left column shows the $W \to e\nu$ channel, and the right column shows the $W \to \mu\nu$ channel.
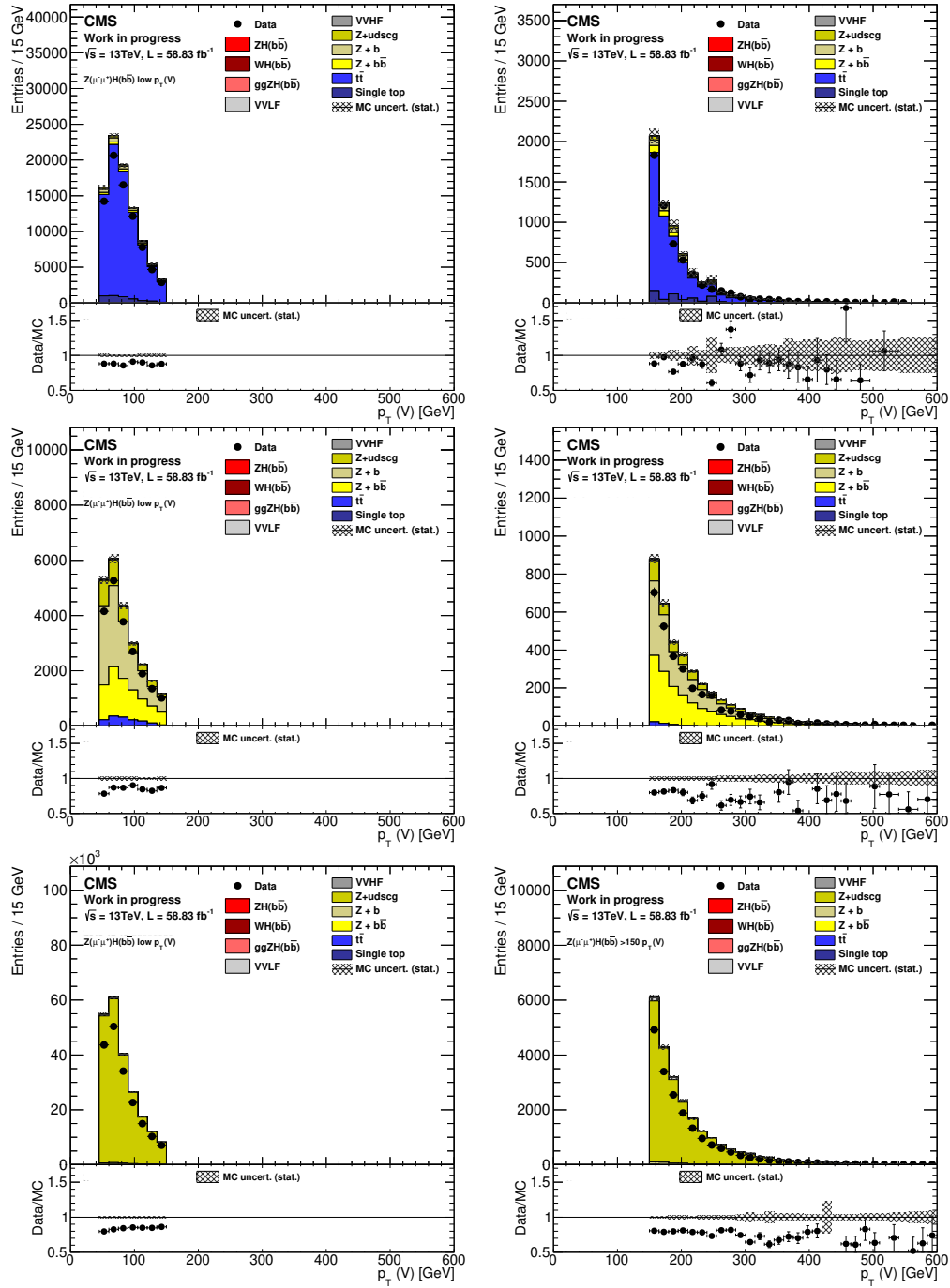
Figure 6-6: The vector boson p_T is plotted for multiple control regions and bins in the $Z \to \mu\mu$ two lepton channel with 2018 data. The top row shows the $t\bar{t}$ control region, the middle row shows the $Z+$ heavy jets, and the bottom row shows the $Z+$ light jets control region. The left column shows the low $p_T(V)$ bin of 75 to 150 GeV, and the right column shows an inclusive selection of the medium and high $p_T(V)$ bins of greater than 150 GeV.

64. The final layer is a softmax layer with the target of predicting the probability of each event belonging to a particular class. The cross entropy loss function is used for the training.

A separate training using the same architecture is performed on the $V +$ heavy jets control regions. Each class in the output is designed to control the normalization and input variable shapes of different background processes. There are five classes that attempt to separate the following processes:

- $V +$ light-flavored ($udsg$) jets

- $V + c$ jets

- $V + b$ jets

- Single-top events

- $t\bar{t}$ events

Each channel of 0-, 1-, and 2-leptons is trained separately, and has slightly different input variables. The list of input variables is given in Table 6.1. All variables that are affected by the kinematic fit in the 2-lepton region use the values calculated by the fit.

6.3.2 Boosted BDT

The BDT used to classify signal and background events in the boosted region was trained using ROOT. The model uses 100 trees with 20 cuts and a minimum node size of 0.05. The QCD multi-jet backgrounds were not used in the training since the sample's large weights of individual events affected the training.

The list of input variables for the BDT is the following:

- Soft-drop mass of the reconstructed fat jet

- Transverse momentum of the fat jet

- Transverse momentum of the reconstructed vector boson

111

Table 6.1: The list of input variables used for each DNN training is shown.

Variable	Explanation	0-lepton	1-lepton	2-lepton
m_{jj}	Di-jet mass	✓	✓	✓
$p_{T,jj}$	Di-jet transverse momentum	✓	✓	✓
MET	Missing transverse energy	✓	✓	✓
$m_{T,V}$	Vector boson transverse mass		✓	
$p_T(V)$	Vector boson p_T		✓	✓
$p_{T,jj}/p_T(V)$	Redundant ratio		✓	✓
$\Delta\phi(V,jj)$	Azimuthal angle between vector boson and di-jet	✓	✓	✓
b-tag$_{\mathrm{max}}$ WP	1, 2, or 3 if higher b-tag discriminate meets the tight, medium, or loose working point respectively	✓	✓	✓
b-tag$_{\mathrm{min}}$ WP	1, 2, or 3 if lower b-tag discriminate meets the tight, medium, or loose working point respectively	✓	✓	✓
$\Delta\eta(jj)$	η difference between jets	✓	✓	✓
$\Delta\phi(jj)$	Azimuthal angle between jets	✓	✓	
$p_{T,\mathrm{lead}}$	Leading jet p_T	✓	✓	✓
$p_{T,\mathrm{trail}}$	Trailing jet p_T	✓	✓	✓
SA5	Number of soft jets, $p_T > 5\,\mathrm{GeV}$	✓	✓	✓
N_{aj}	Number of additional jets	✓	✓	
b-tag$_{\mathrm{add}}$	Maximum b-tag of additional jets	✓		
$p_{T,\mathrm{add}}$	Maximum p_T of additional jets	✓		
$\Delta\phi(\mathrm{add},\mathrm{MET})$	Azimuthal angle between additional jet and MET	✓		
$\Delta\phi(\ell,\mathrm{MET})$	Azimuthal angle between lepton and MET		✓	
m_t	Reconstruction top mass		✓	
m_V	Vector boson mass			✓
$\Delta R(V,jj)$	Separation between vector boson and di-jet			✓
ΔR_{jj}	Separation between jets			✓

- Number of soft-track jets with $p_T > 5\,\mathrm{GeV}$

- Double b-tagger output node for boosted jets

All of these same variables were used in the 0-, 1-, and 2-lepton regions, even though the regions were trained separately. Because of the decreased complexity of the boosted BDT compared to the DNN, there is no need to attempt to simultaneously control the input variables from the backgrounds in the V + heavy flavor control regions.

6.4 Combination Fit

A simultaneous fit is run over all channels, control regions, and the signal selection region in order to determine the most likely values for all parameters with systematic uncertainties, called nuisance parameters, as well as the most likely scale factors for all the MC backgrounds and signal. The fit is done by using the `combine` tool [81] as part of `CMSSW_10_2_13` [65]. Included in this fit is the strength of the signal in the STXS bins. This is how the final result of this analysis is measured.

There are a total of 243 distributions that make up the fit. Table 6.2 shows how the different channels and control regions contribute to this number. As mentioned before, there are a number of uncertainties that also produce variations in the fit histograms. A separate fit is run with each corresponding nuisance parameter frozen to their most likely value, which gives the purely statistical uncertainty on the measurement result. The systematic uncertainty is then extracted.

Running this fit without data, but instead toy distributions gathered directly from the initial simulated distributions gives the expected sensitivity of the analysis. Analyzers can see the statistical and systematic uncertainties before measuring the final result in this way. It is under this limitation that the analysis strategy is developed and optimized for sensitivity. For Higgs analyses, there is an additional measurement which is done before the fit is run with measured data to give the unblinded result. In addition to the ability to fit distributions , `combine` includes a variety of

Table 6.2: Below, a table indicates the distributions from each control region (CR) and signal region (SR) selection. Each of the three control regions contributes the same number of distributions to the fit. The channels containing charged leptons can either be electron or muon flavored. These two considerations determine the multipliers shown in the table.

$p_T(V)$ [GeV]	0-leptons				1-lepton				2-leptons			
	Resolved		Boosted		Resolved		Boosted		Resolved		Boosted	
	CR	SR	CR	SR	CR	SR	CR	SR	CR	SR	CR	SR
$75-150$									X	X		
$150-250$	X				X	X			X			
with jet		X								X		
without		X								X		
$250-\infty$	X		X		X		X		X		X	
$250-400$		X		X		X		X		X		X
$400-\infty$		X		X		X		X		X		X
number X	2	4	1	2	2	3	1	2	3	5	1	2
multiplier	3	1	3	1	6	2	6	2	6	2	6	2
regions/year	6	4	3	2	12	6	6	4	18	10	6	4

diagnostic tools to ensure the assumed model is adequate. Before unblinding results, multiple toy distributions are generated using the best-fit nuisance parameter values and comparing a test statistic, which is of a generalized χ^2 test [82], of these toys to that of the observed data. The distributions of the values from the set of toys allows the evaluation of the p-value of the test statistic observed in data. The `combine` tool also evaluates the impacts of each nuisance parameter on the likelihood by varying them individually. If nuisance parameters with large impacts are pulled far from their initial value, that means that an important aspect of the measurement was not well understood. Both of these tests can be run without revealing the measured values of cross sections. This prevents biases in the analysis strategy which is finalized in response to these tests.

The final results will therefore show comparison of data and best-fit toys, called Goodness of Fit, as well as the final scan results. Post-fit plots will also reveal the way that the fit moves the background and signal distributions within uncertainties in order for the background to describe the data. Impacts plots are shown in Appendix D due to the high multiplicity of nuisance parameters. Before running these checks and

measurements for the *VH* measurement, they were first tested on *VZ*.

6.4.1 Test of Methodology Using a *VZ* Selection

Before unblinding the *VH* signal region and results, the techniques used in the analysis are verified by performing a similar analysis. The process $VZb\bar{b}$, where a vector boson radiates a Z boson that decays into b quarks, looks very similar to the $VHb\bar{b}$ process. The di-jet mass is just in a different location. The *VZ* cross check analysis uses a strategy that is similar to what has been presented so far. The primary differences are that a di-boson sample with generator-level b hadrons produced is used as the signal sample, and the signal mass window is moved from $90\,\mathrm{GeV} < m_{jj} < 150\,\mathrm{GeV}$ to $60\,\mathrm{GeV} < m_{jj} < 120\,\mathrm{GeV}$. However, other differences should be mentioned that may affect the final result. Additional measured electroweak corrections are applied to the di-boson samples for each STXS bin. The values of these corrections vary from as low as 4% for the low $p_T(V)$ bins to as high as 23% for the high $p_T(V)$, but all will increase the measured signal strength compared to *VH*. Other differences, the effect of which are more difficult to quantify, include the dedicated *VZ* DNN training and the fact that the Z boson will decay into light flavored jets at a higher proportional rate than the Higgs boson. The DNN training for *VH* has to account for the similar *VZ* background, while the *VZ* is not influenced as much by the rare *VH* process. Also, the light flavor *VZ* process will have more influence over the control region fits than the *VH* process has.

Data from all three years of Run 2 are used so that any differences that may affect the *VH* analysis can also be investigated using *VZ*. The Goodness of Fit plots are shown in Figure 6-7. The high p-value means that the simulated processes were able to match the data very well while staying within the fit constraints of the systematic uncertainties. This suggests that the systematic uncertainties applied may be much larger than needed for this cross check analysis.

Due to the splitting of STXS bins, the multiple channels for each bin, and the combination of three different years, 243 distributions are fit in the *VZ* cross check analysis. For the sake of brevity, the effect of the fit will only be shown on eight of
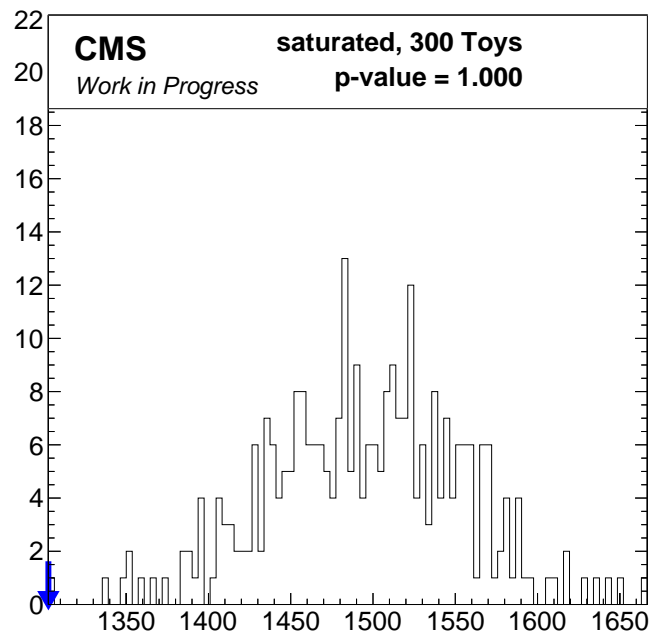
Figure 6-7: The Goodness of Fit test results are shown for *VZ*. The fit is performed using STXS bins, and the test statistic is generated by toys for all control regions and signal regions. The high *p*-value indicates that the model may be over-fitting to data due to large systematic uncertainties. This is not the case for the *VH* analysis.
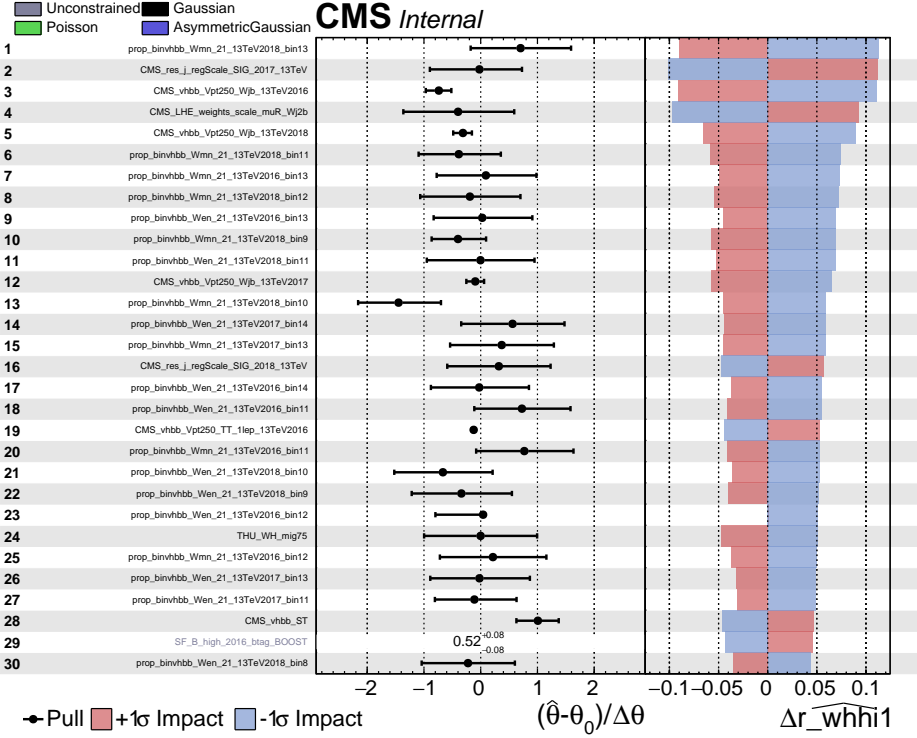
Figure 6-8: The most significant impacts in on WH $250 < p_T(V) < 400\,\text{GeV}$. Note the decrease in migration uncertainties, such as `CMS_vhbb_Vpt250_TT_1lep_13TeV2016`.

these distributions for the *VZ* cross-check analysis. These are split into two figures, each showing distributions closely related to each other. On the left of each figure are simulation values and uncertainties before the nuisance parameters are fit to match the simulation to data. On the right are the simulated values after fitting the nuisance parameters. In the lower parts of each plot, ratios between observed data and expected yields improve as a result of the fit. The uncertainties of the simulations also decrease as a result of the fit. Some nuisance parameters end up with smaller individual uncertainties from the fit, including the initially large migration uncertainties. Figure 6-8 shows this decrease in bin migration uncertainties, especially in $t\bar{t}$ in line 19. The overall uncertainty also decreases slightly since each named nuisance parameter is at first assumed to be uncorrelated to the other nuisance parameters. The fit reveals many of the nuisance parameters to be correlated, decreasing the overall possible variation.

Figure 6-9 shows all of the distributions fit for the low p_T bin in the $Z \to \mu\mu$ chan-

nel in 2017. Several additional features aside from the pre-fit and post-fit differences are worth noting. There is only one bin in the $t\bar{t}$ control region, so it is only used to constrain the normalization of the $t\bar{t}$ background. In the Z + heavy jets control region, the multi-classifier DNN with the same inputs as the signal region DNN is shown. The signal distributions are binned so that the output of the DNN is flat in the VH signal. The VZ signal distribution is close to flat when following the same binning scheme. The important feature in that distribution is that the backgrounds are all falling as the DNN value increases. In all of the regions, we see improved agreement between the simulation and data after the fit, and the combination fit also decreases the uncertainty on each bin. The relative contributions of the main background processes can also be seen at a glance at the signal region plots. These distributions can be compared to those in Figure 6-10, which shows distributions from the boosted selection of the $W \rightarrow e\nu$ channel in 2018. In this case, the W + heavy jets control region is not shown because the related process contributes less in the signal region. Instead, the $t\bar{t}$ background is dominant in this boosted topology. The control regions use the double b-tag score of the fat jet instead of $p_T(V)$ for the distribution. The signal region is also split into two higher $p_T(V)$ bins, while the control regions stay combined to constrain both STXS bins. As in the low p_T bin, though, both agreement between simulation and data and estimated uncertainties are improved as a result of the fit.

From these post-fit distributions, the strength and uncertainty of the signal samples are extracted. A value of 1.0 corresponds to the expectation based on simulation of the Standard Model. By correlating all of the signal samples across STXS bins, an inclusive signal strength is obtained. The likelihood scan of this strength is shown in Figure 6-11. The maximum likelihood with respect to the Standard Model is $1.182^{+0.118}_{-0.112}(\text{stat})^{+0.098}_{-0.097}(\text{sys})$. The uncertainty of this measurement means the result has a p-value of 22%, assuming Standard Model couplings, or within 1.2 standard deviations of the Standard Model. Figure 6-12 shows the measured kinematic distributions of the various VZ processes. Assuming the Standard Model, the measurement has a p-value of 23%. The WZ STXS bins are consistently slightly above the Standard
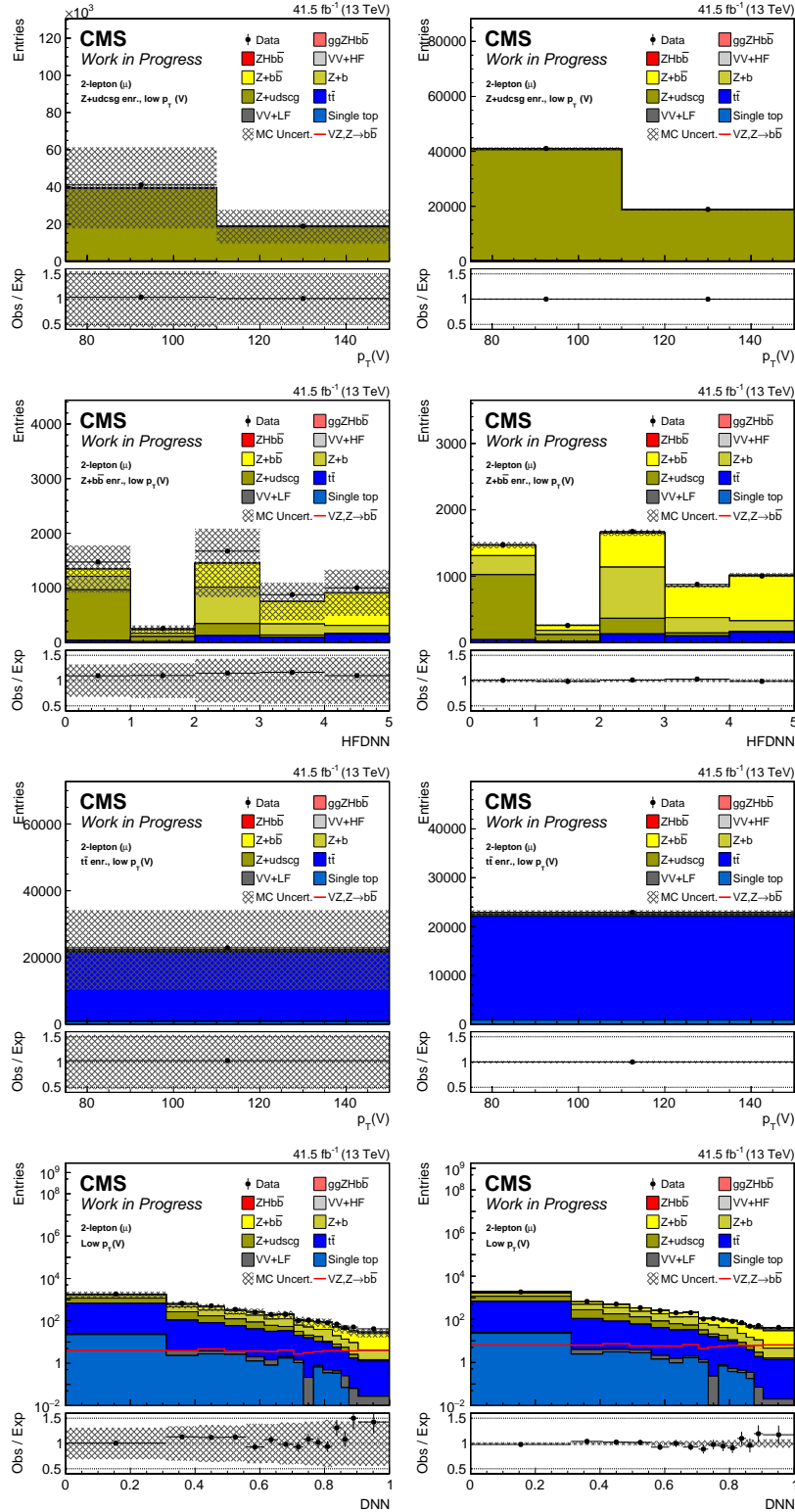
Figure 6-9: Above are $Z \to \mu\mu$ pre-fit (left) and post-fit distributions (right) for the low p_T STXS bin in 2017 in the VZ cross-check analysis. The top row shows the Z + light jets control region, the second row shows the Z + heavy jets control region, and the third row shows the $t\bar{t}$ control region. The bottom row shows the signal region.
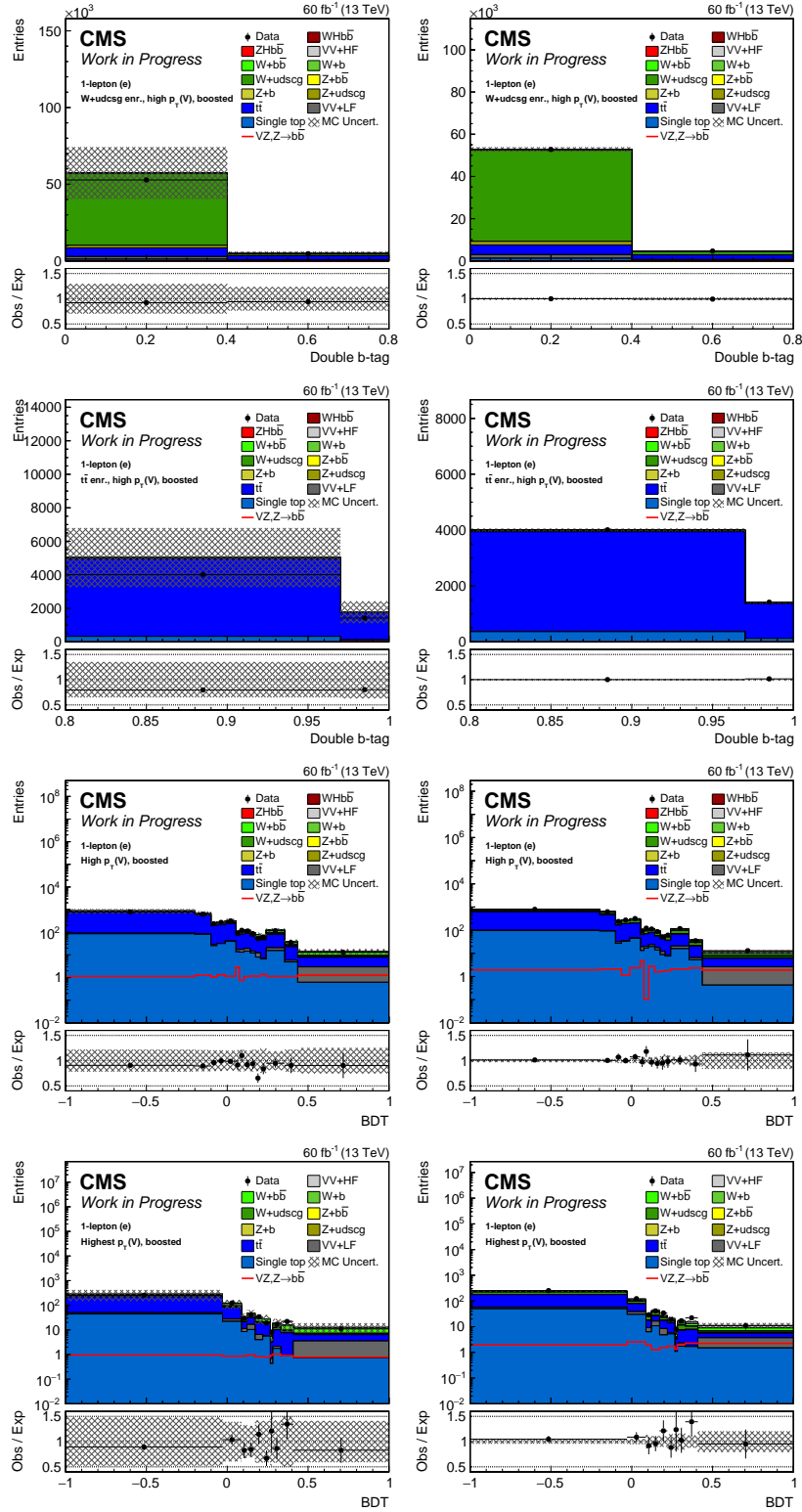
Figure 6-10: Above are $W \to e\nu$ pre-fit (left) and post-fit (right) distributions in the boosted selections in 2018 for the VZ analysis. The top row shows the W + light jets control region, and the second row shows the $t\bar{t}$ control region. The third row shows the signal region for the $250\,\text{GeV} < p_T(V) < 400\,\text{GeV}$ bin, and the bottom row shows the signal region for the bin of $p_T(V) > 400\,\text{GeV}$.
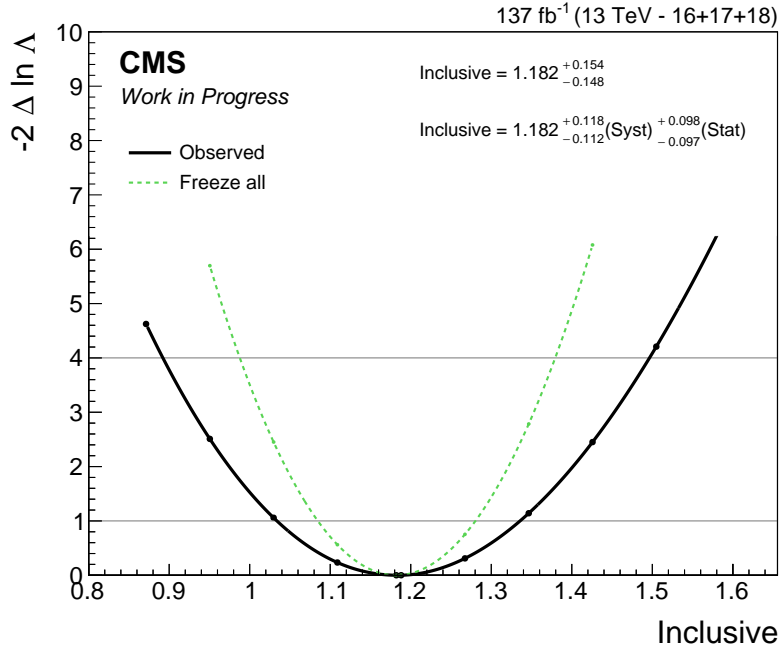
Figure 6-11: The likelihood scan of the inclusive signal strength of the VZ cross check analysis is shown above. The dashed line is generated by freezing all nuisance parameters to their most likely value, so the likelihood variation for that curve is from statistical uncertainties only. Using quadratic subtraction between the two curves allows the systematic uncertainties to be evaluated.

Model prediction. When the VZ process is split into inclusive WZ and ZZ parameters though, the WZ process is seen to be 2.1 standard deviations away from the Standard Model. Overall the VZ cross-check is in agreement with the Standard Model.

6.4.2 VH Combination Fit Results

To extract the VH results, the same procedure is followed as was done in the VZ cross-check analysis. The first step to show that the background processes for VH in the detector are well-modeled by simulation is the Goodness of Fit test, which is shown in Figure 6-14. The observed value for the saturated test statistic in data is in the bulk of the distribution created by throwing post-fit toys, with an associated p-value of 70%. This suggests both good modeling, since the p-value is not low, and reasonable values for post-fit uncertainties, since the p-value is not too high.

Again, only a sub-set of the 243 pre-fit and post-fit distributions are shown. Like,
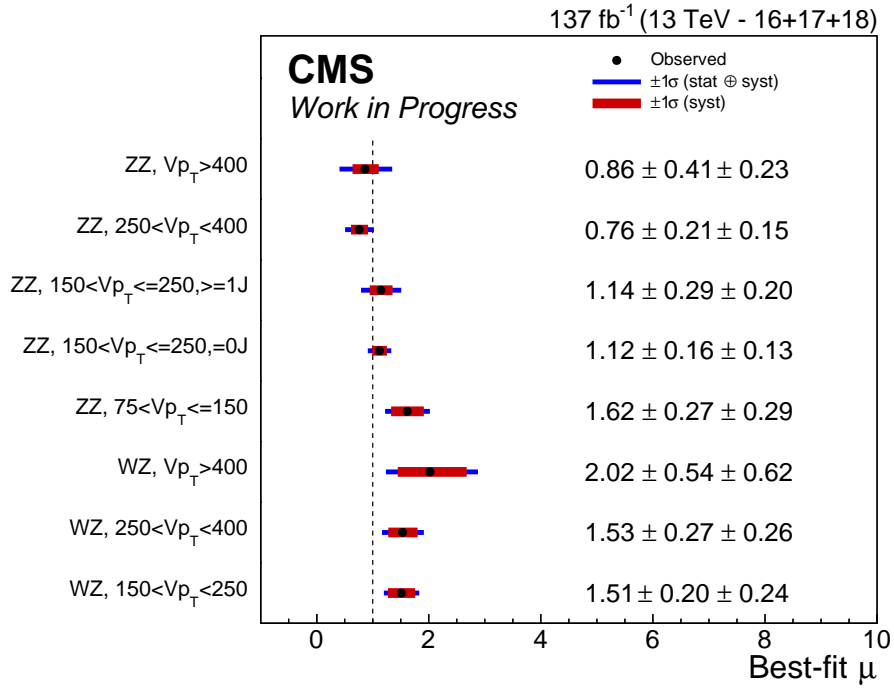
Figure 6-12: The measured most likely values of all STXS bins in the *VZ* cross check analysis.
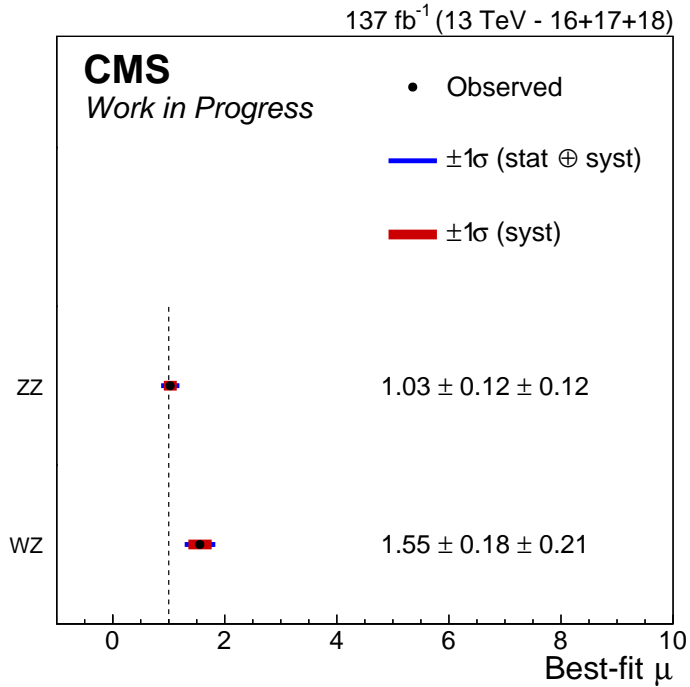
Figure 6-13: The measured most likely values of *WZ* and *ZZ* in the *VZ* analysis.
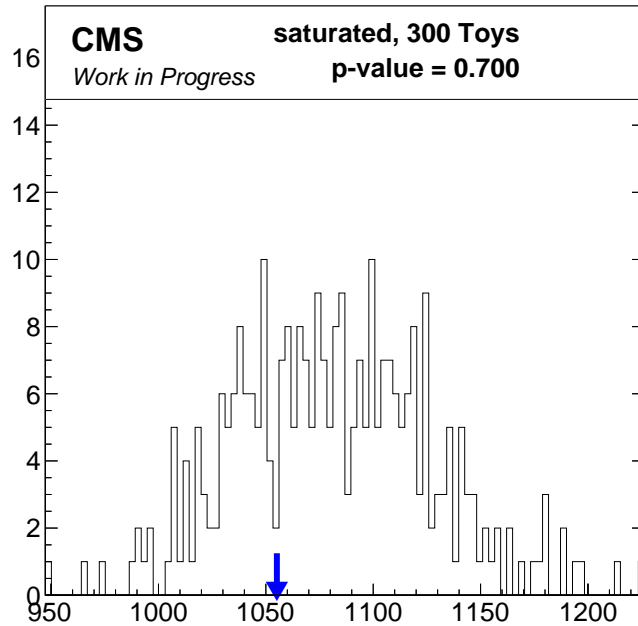
Figure 6-14: The Goodness of Fit test results are shown for VH. The fit is performed using STXS bins, and the test statistic is generated by toys for all control regions and signal regions.

the VZ analysis, there are too many distributions to show all of them in this work, so what follows are only a few illustrative examples. Figure 6-15 shows the zero-lepton channel in the high $p_T(V)$ regime. Unlike the two lepton selections, the signal distributions display contamination from W + jets background processes. There is no zero-lepton control region for W + jets. Instead, these backgrounds are controlled only by one-lepton control regions, so the five channels are correlated in the fit. Figure 6-16 shows the $W \to \mu\nu$ distributions for the medium $p_T(V)$ STXS bin. Like in the VZ cross-check analysis, the single lepton signal selection is largely made up of $t\bar{t}$ background. Finally, Figure 6-17 shows the $150 < p_T(V) < 250\,\text{GeV}$ STXS bin for $Z \to ee$ in 2018. In this $p_T(V)$ bin, the ZH the signal region is split between events with no extra jets, and events with extra jets, which can be seen in the distributions. The other thing to note is that the post-fit signal values are pulled to low values, and disappear in the post-fit plot. A deficit of signal-like events is observed in data for events with jets. Even though there is an excess of signal-like events in events
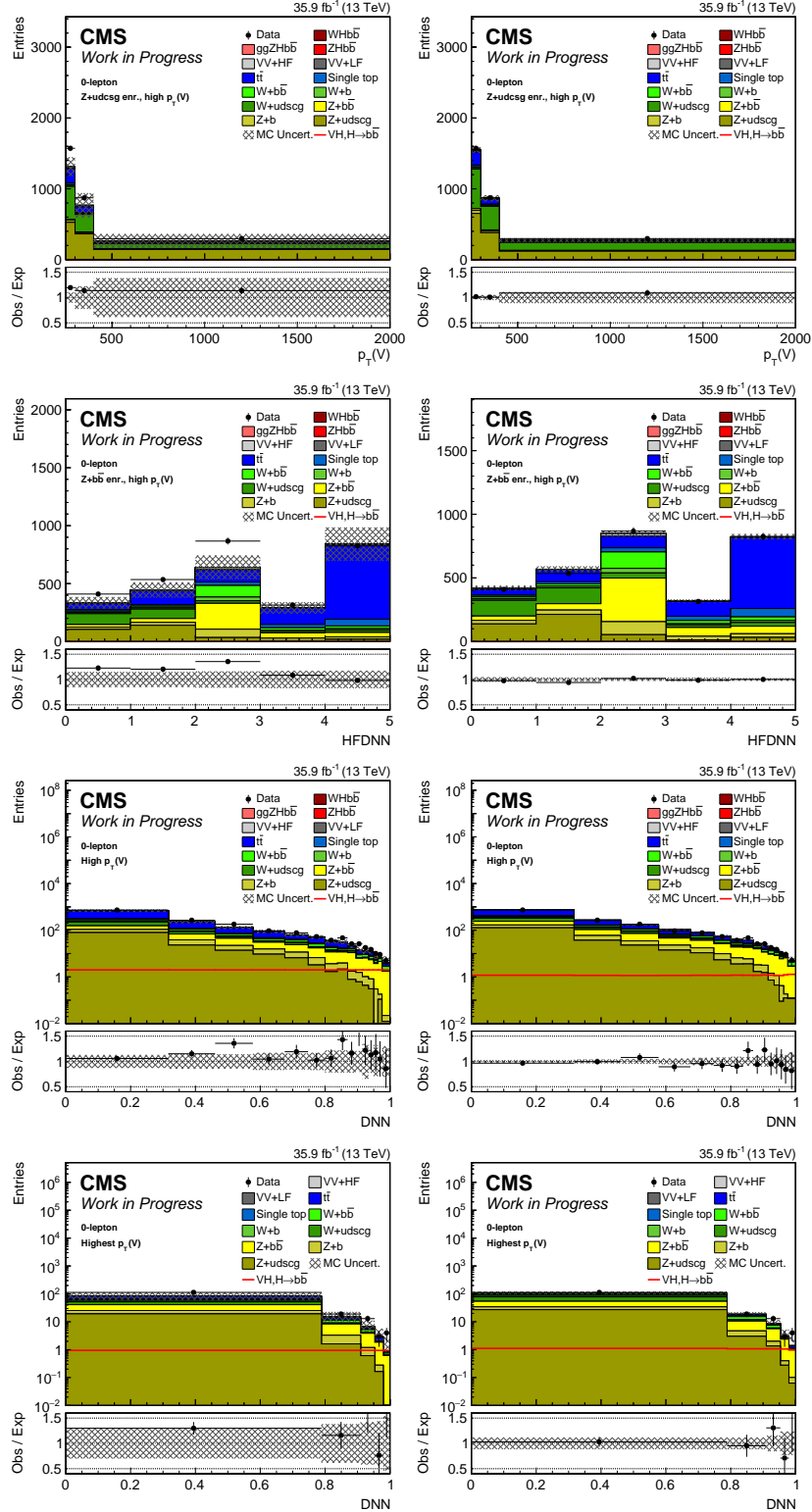
Figure 6-15: Above are $Z \to \nu\nu$ pre-fit (left) and post-fit distributions (right) for the high p_T STXS bins in 2016 in the VH measurement using the resolved selection. The top row shows the Z + light jets control region, and the second row shows the Z + heavy jets control region. The third row shows the signal region for the $250\,\mathrm{GeV} < p_T(V) < 400\,\mathrm{GeV}$ bin, and the bottom row shows the signal region for the bin of $p_T(V) > 400\,\mathrm{GeV}$.

124
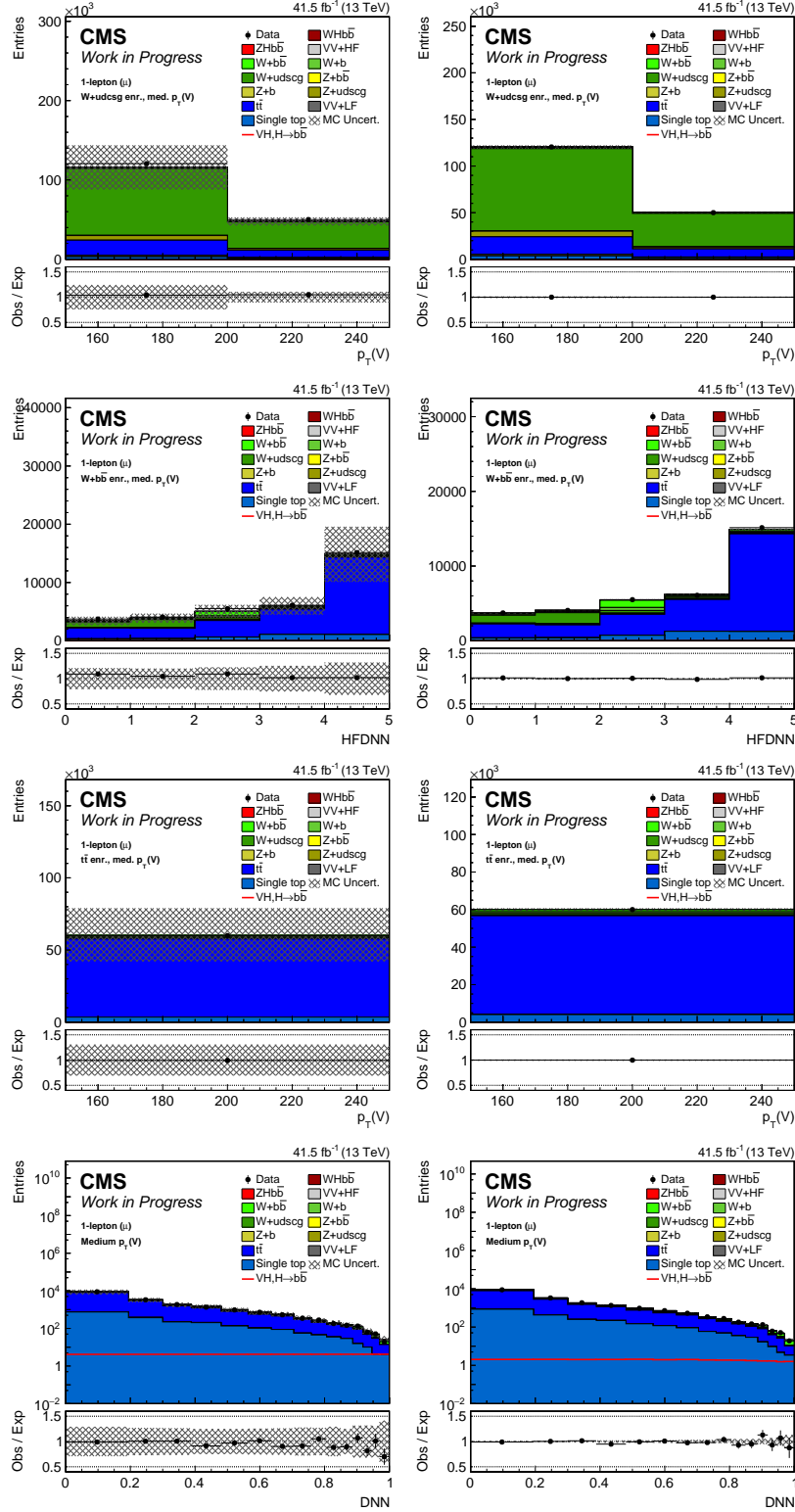
Figure 6-16: Above are $W \to \mu\nu$ pre-fit (left) and post-fit distributions (right) for the medium p_T STXS bins in 2017 in the VH measurement. The top row shows the W + light jets control region, and the second row shows the W + heavy jets control region, and the third row shows the $t\bar{t}$ control region. The bottom row shows the signal region.

125

without jets, the signal strength is affected by the fit in other channels ($\mu\mu$ and $\nu\nu$) and other years.

The signal strength extracted from the fit can be separated a number of ways. The inclusive signal strength is shown in Figure 6-18 to be $0.568^{+0.154}_{-0.147}(\text{stat})^{+0.134}_{-0.133}(\text{sys})$. The uncertainties put this measurement within 2.1 standard deviations of the Standard Model. Figure 6-19 shows that this low value is caused primarily by a low cross section measurement of ZH, which is 2.9 standard deviations below the Standard Model. This single measured parameter of interest is the largest discrepancy measured from the Standard Model. Additional investigation splits the production mechanism of the vector boson. The results of this study are shown in Figure 6-20, where the $ggZH$ process is shown to produce this deviation with the Standard Model. However, the uncertainty for this process alone is so large that the measured result is only 1.5 standard deviations below the Standard Model. Figure 6-21 shows the full STXS measurement results. Both ZH and WH show increasing trends in differential cross section as $p_T(V)$ increases. However, substituting the inclusive signal strength in for the Standard Model, the STXS scans have an agreement with a p-value of 9.3%. Standard Model kinematics cannot be entirely ruled out by this measure.

The total coupling between the Higgs boson and vector bosons, as well as the coupling between the Higgs boson and bottom quarks is measured using the kappa framework [83]. Results of this analysis alone are shown in Figure 6-22. Standard Model values of $\kappa_V = \kappa_b = 1$ fall within the 68% confidence level. The results of this κ scan can be additionally constrained by combining this analysis with others that focus on using either vector bosons for Higgs production or the $b\bar{b}$ decay path.

The last piece of the analysis shows where future work can improve measurement sensitivity. Nuisance parameters are gathered into groups of nuisances, and the relative effect of each group on the overall systematic uncertainty is evaluated. The results for each of the STXS bins is shown in Table 6.3. The first two rows of each table, "Signal" and "Background", are the theoretical uncertainties applied to each type of simulation. These uncertainties can be reduced slightly by using a lepton collider instead of a hadron collider since there would be no PDF uncertainties in that case.
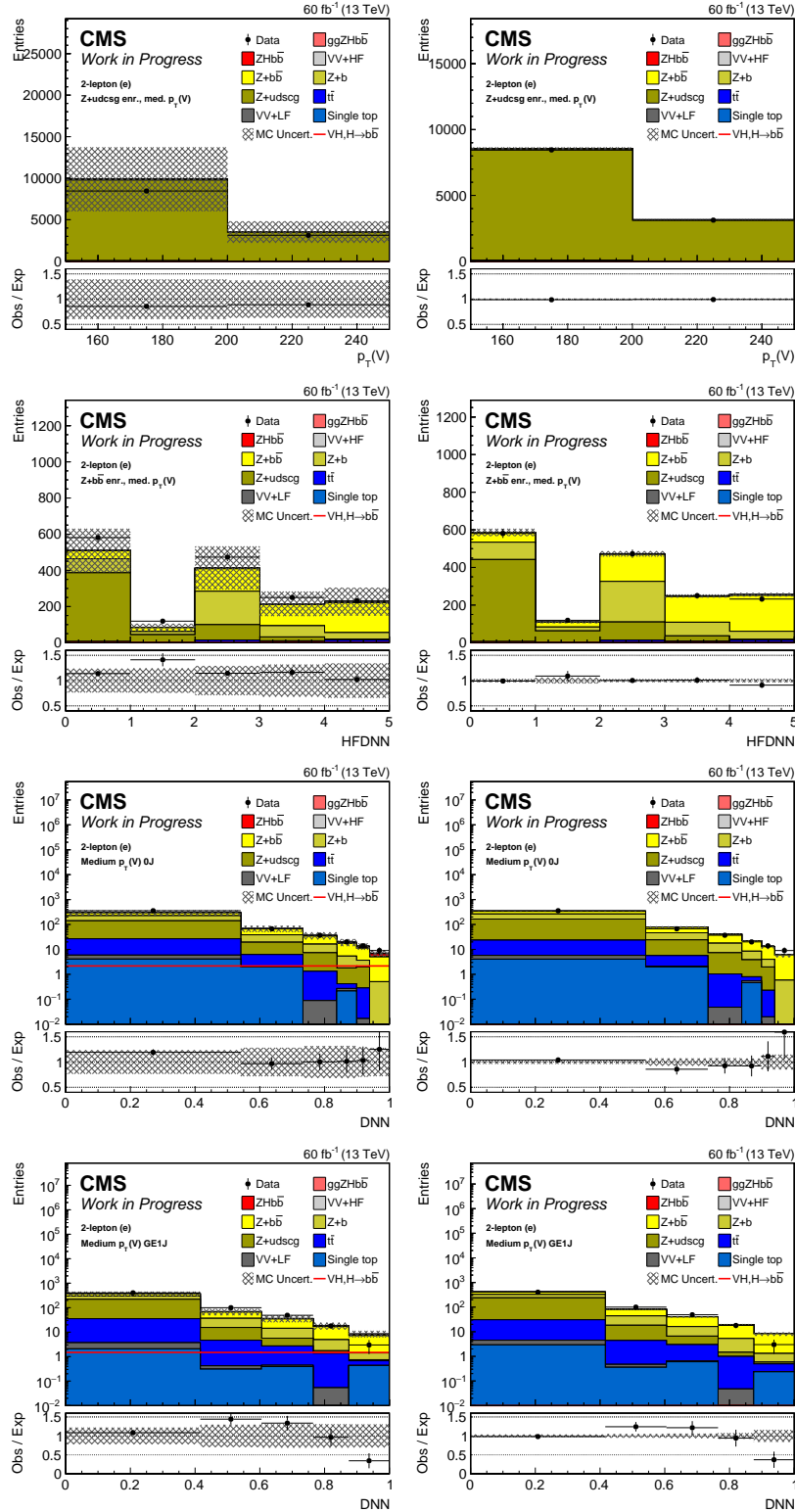
Figure 6-17: Above are $Z \to ee$ pre-fit (left) and post-fit distributions (right) for the medium p_T STXS bins in 2018 in the *VH* measurement. The top row shows the $Z +$ light jets control region, and the second row shows the $Z +$ heavy jets control region. The third row shows the signal region with no additional jets, and the bottom row shows the signal region with additional jets.
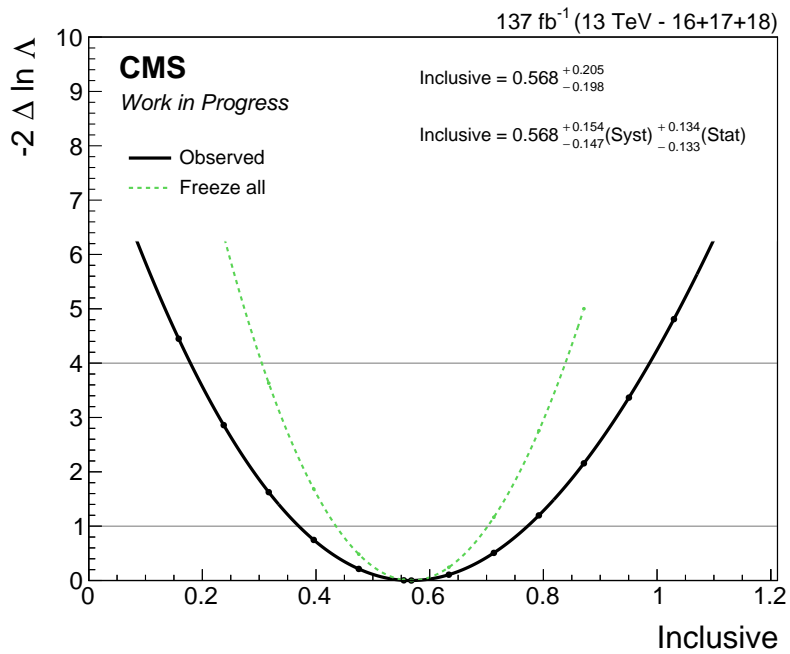
Figure 6-18: The likelihood scan of the inclusive signal strength of the *VH* analysis.
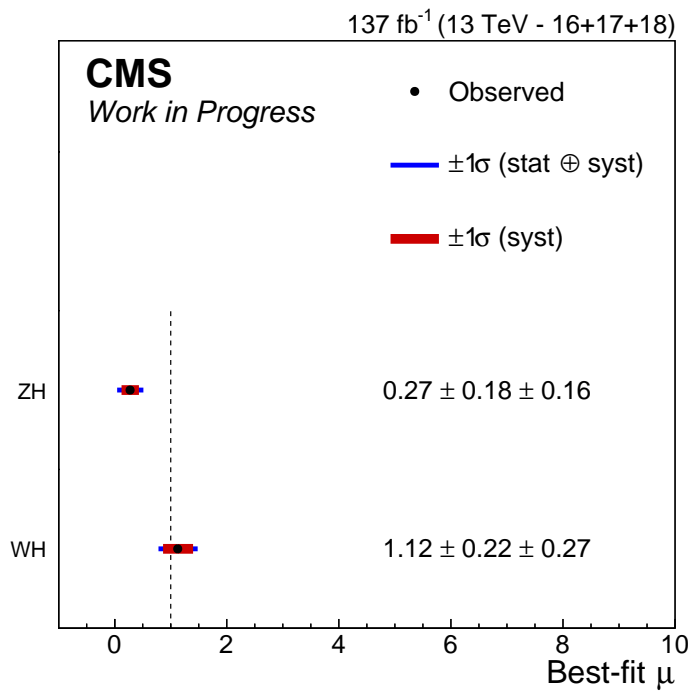
Figure 6-19: The measured most likely values of *WH* and *ZH* in the *VH* analysis.
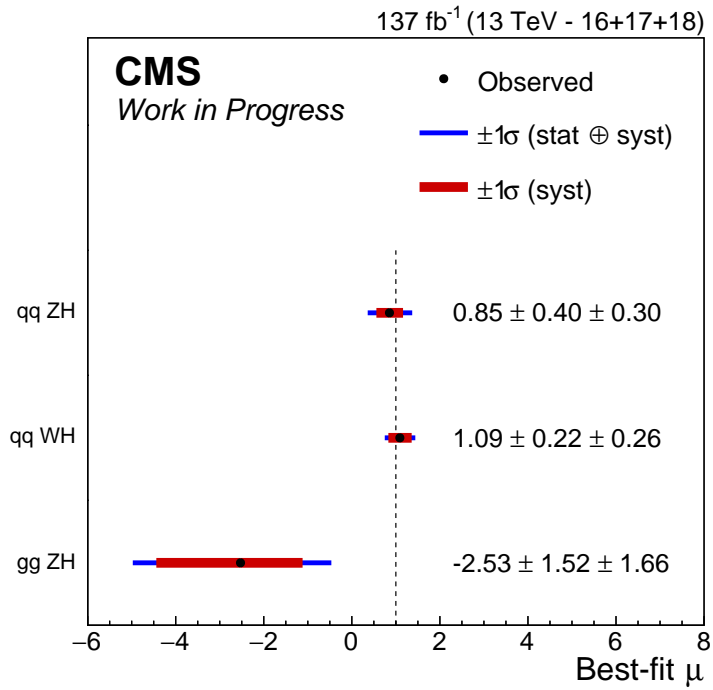
Figure 6-20: The measured most likely values of with the *VH* process split into *qqWH*, *qqZH*, and *ggZH*.
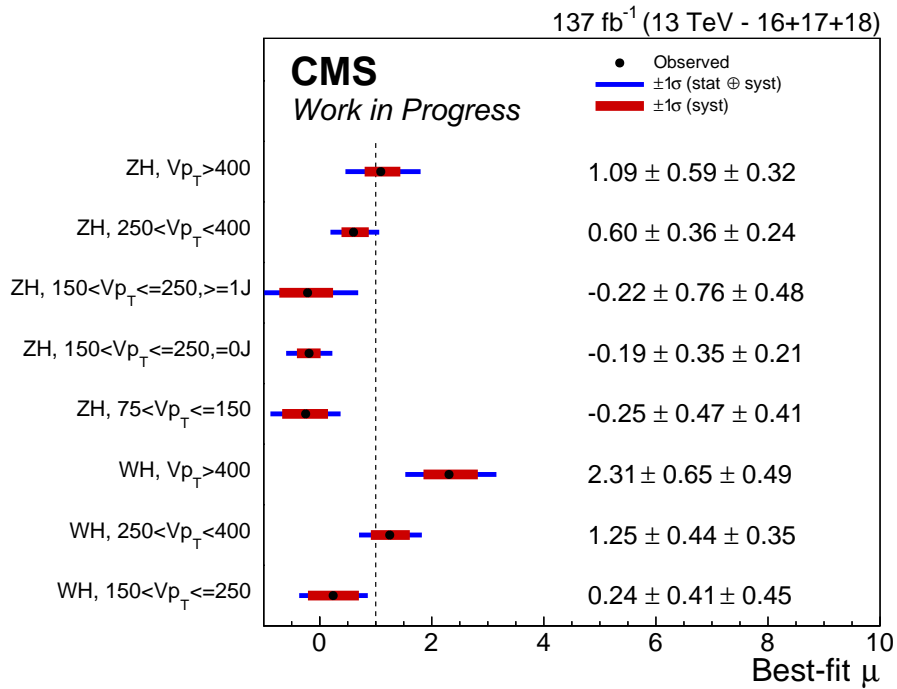
Figure 6-21: The measured most likely values of all STXS bins in the *VH* analysis.
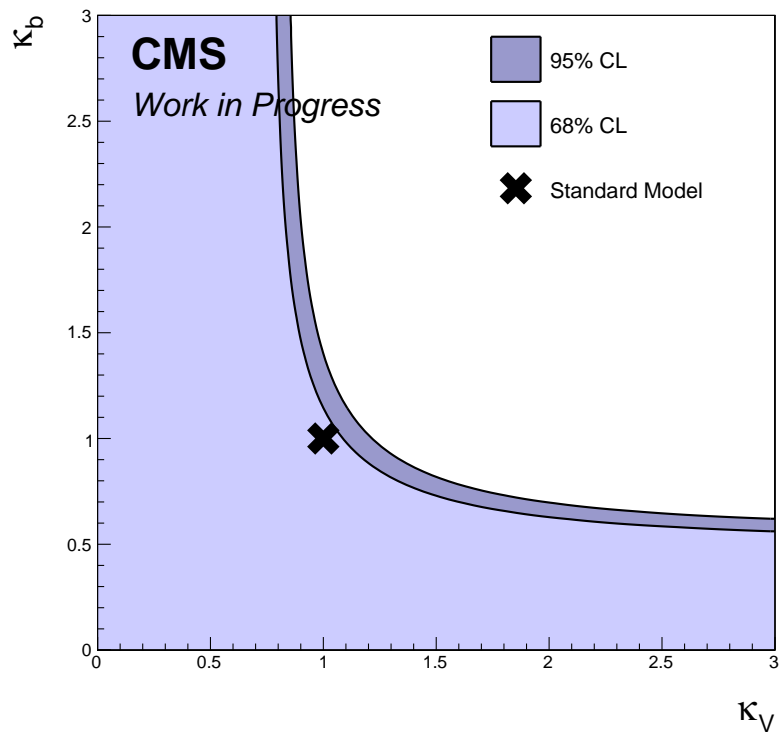
Figure 6-22: The possible values of κ_V and κ_b are shown above. Values of 1.0 reproduce Standard Model predictions.

However, these theoretical uncertainties have a generally smaller effect than some of the experimental systematic uncertainties. In particular, the $p_T(V)$ background migration systematic introduced in Section 5.4.3 has the largest effect on the measurement sensitivity. This systematic effect was assumed to be large since there was no thorough study on this correction. Future work focusing on this correction can improve the measurement sensitivity in all STXS bins. Beyond that, the jet energy has a large effect, which is why there was a dedicated study as outlined in Section 5.2. Finally, the b-tagging systematics were more significant than the jet energy for some bins. Progress on simulation of jets, in particular simulation of heavy-flavor jets, would be able to reduce both of these systematics as well.

Table 6.3: The contribution of different groups of systematic uncertainties are given below for each of the STXS bins. The "Signal" and "Background" groups are theoretical systematic uncertainties. The rest of the groups are results of the experimental apparatus or analysis technique. The "Meas. XS" uncertainties are related to the measured cross sections of the single-top [76] and di-boson [77] processes that are used to normalize backgrounds in this analysis. The total systematic uncertainty as well as the statistical uncertainty for each bin are also included for comparison.

	WH med.	WH high	WH highest
Signal	+0.060 -0.000	+0.117 -0.034	+0.223 -0.117
Background	+0.052 -0.047	+0.053 -0.025	+0.084 -0.060
b-tagging	+0.128 -0.118	+0.084 -0.069	+0.092 -0.077
Jet energy	+0.225 -0.203	+0.090 -0.081	+0.116 -0.076
Lepton ID	+0.022 -0.000	+0.054 -0.042	+0.023 -0.047
LO to NLO	+0.047 -0.038	+0.035 -0.005	+0.050 -0.050
Luminosity	+0.024 -0.000	+0.045 -0.000	+0.066 -0.043
Meas. XS	+0.034 -0.000	+0.039 -0.000	+0.047 -0.057
Triggers	+0.011 -0.000	+0.022 -0.035	+0.034 -0.048
$p_T(V)$ Mig.	+0.165 -0.156	+0.133 -0.123	+0.305 -0.293
Total Sys.	+0.460 -0.450	+0.362 -0.333	+0.518 -0.458
Stat	+0.413 -0.406	+0.447 -0.430	+0.672 -0.633

	ZH low	ZH no J	ZH with J	ZH high	ZH highest
Signal	+0.049 -0.164	+0.040 -0.103	+0.113 -0.280	+0.142 -0.031	+0.130 -0.025
Background	+0.038 -0.050	+0.000 -0.015	+0.069 -0.065	+0.027 -0.027	+0.038 -0.025
b-tagging	+0.131 -0.135	+0.038 -0.036	+0.098 -0.093	+0.052 -0.046	+0.058 -0.043
Jet energy	+0.050 -0.048	+0.019 -0.030	+0.101 -0.093	+0.045 -0.040	+0.070 -0.069
Lepton ID	+0.032 -0.032	+0.007 -0.021	+0.000 -0.067	+0.018 -0.012	+0.028 -0.022
LO to NLO	+0.154 -0.162	+0.035 -0.039	+0.113 -0.083	+0.044 -0.040	+0.045 -0.025
Luminosity	+0.016 -0.020	+0.000 -0.012	+0.017 -0.014	+0.023 -0.023	+0.039 -0.012
Meas. XS	+0.000 -0.016	+0.000 -0.000	+0.000 -0.000	+0.026 -0.016	+0.004 -0.000
Triggers	+0.024 -0.010	+0.000 -0.000	+0.029 -0.000	+0.000 -0.010	+0.023 -0.013
$p_T(V)$ Mig.	+0.362 -0.387	+0.091 -0.079	+0.232 -0.210	+0.133 -0.121	+0.252 -0.235
Total Sys.	+0.529 -0.553	+0.205 -0.216	+0.464 -0.506	+0.274 -0.220	+0.355 -0.306
Stat	+0.479 -0.469	+0.360 -0.346	+0.781 -0.738	+0.369 -0.352	+0.620 -0.563

Chapter 7

Conclusions

Thanks to the theory, technologies, and techniques described, an inclusive cross section is measured that is 2.1 standard deviations below the Standard Model. The binned momentum spectrum of the associated vector bosons has a p-value of 9.3% given the inclusive cross section. These measurements confirm the Standard Model.

7.1   Comparison with ATLAS

After the discovery of $VHb\bar{b}$, announced simultaneously by both CMS and ATLAS, a joint effort produced the Simplified Template Cross Section (STXS) framework for study of the differential cross section of the Higgs boson [85]. ATLAS published a first STXS measurement using only the first two years of the LHC's Run 2 data [86]. This publication, as well as the first publication with all three years of Run 2 data [87] only look at the two highest vector boson transverse momentum bins of $250 < p_T(V) < 400\,\text{GeV}$ and $400\,\text{GeV} < p_T(V)$ for WH and ZH. Lower momentum results using all three years were later published by ATLAS separately [88]. With the result of this thesis, we can compare results between CMS and these two papers. This comparison is shown in Table 7.1, along with Standard Model predictions for each STXS bin [89]. The uncertainty bands for each analysis overlap in every STXS bin, except for ZH events with $150 < p_T(V) < 250\,\text{GeV}$. The analysis presented in this thesis has better sensitivity than the ATLAS result in most STXS bins. Though

Table 7.1: Below are the STXS measurements compared between the ATLAS collaboration and this analysis, along with cross section predictions given by the Standard Model. Note, in the analysis published by ATLAS the *ZH* events with $150 < p_T(V) < 250\,\mathrm{GeV}$ are not split between events with and without additional jets, as is done by CMS. The systematic uncertainty applied to the thesis result in that phase space assumes that the uncertainties for the jet multiplicity bins are uncorrelated. A dedicated analysis would have greater precision.

Process	$p_T(V)$	Prediction [fb]	ATLAS [fb]	This Work [fb]
$W(\ell\nu)H(b\bar{b})$	150—250 GeV	24.0 ± 1.1	19.0 ± 12.1	5.8 ± 14.4
$W(\ell\nu)H(b\bar{b})$	250—400 GeV	5.83 ± 0.26	3.3 ± 4.7	7.3 ± 3.2
$W(\ell\nu)H(b\bar{b})$	400 GeV—∞	1.25 ± 0.06	2.1 ± 1.2	2.9 ± 1.0
$Z(\ell\ell/\nu\nu)H(b\bar{b})$	75—150 GeV	50.6 ± 4.1	42.5 ± 35.9	-12.6 ± 30.9
$Z(\ell\ell/\nu\nu)H(b\bar{b})$	150—250 GeV	18.8 ± 2.4	20.5 ± 6.2	-3.8 ± 7.7
$Z(\ell\ell/\nu\nu)H(b\bar{b})$	250—400 GeV	4.12 ± 0.45	1.4 ± 3.0	2.5 ± 1.8
$Z(\ell\ell/\nu\nu)H(b\bar{b})$	400 GeV—∞	0.72 ± 0.05	0.2 ± 0.7	0.78 ± 0.48

overall the two analyses have similar sensitivity.

The sensitivity on this measurement will continue to improve. A combination result of these two independent experiments will reduce the statistical uncertainties as the amount of data available for analysis increases by a factor of two compared to the data gathered by a single experiment. Both experiments will also produce measurements with decreased systematic uncertainties as physicists gain experience with the LHC data, measurement technique is refined, and detector technology is improved.

7.2 The Future of HEP

Study of the Higgs boson is far from finished. Many more analyses of the Higgs boson will follow this one, just as many preceded it. The first published searches of the Higgs boson in associated production by CMS were given in 2011 [84]. As mentioned in the beginning of this document, the search concluded successfully in 2018. This analysis itself used the techniques used to obtain that previous result as a starting point.

There were some adjustments to the method used before, and researchers will use lessons learned from both studies for future measurements at CMS and beyond.

While it is impossible to predict the future precisely, there are multiple projects underway with designs to continue to study the Higgs after CMS and ATLAS. The LHC itself is receiving an upgrade, in a project titled the High Luminosity LHC (HL-LHC). New technologies will be used to create a machine that will gather an order of magnitude more data suitable for Higgs studies and other searches beyond the Standard Model [90]. In addition, there are plans for other new colliders. For example, the Circular Electron Positron Collider (CEPC) is currently being designed for use as a Higgs factory for precision studies [91]. There are also plans for the Future Circular Collider (FCC), which is foreseen to go through phases of providing electron-positron collisions, and follow with hadronic collisions after an upgrade [92], much like the LHC used the same tunnels as LEP.

Over time, experiments on future colliders will achieve significantly more precise measurements of the Higgs differential cross section, giving scientists the potential to uncover smaller discrepancies from the Standard Model. We do not know what other particles and interactions might exist beyond the Standard Model, but we do know that the Standard Model is incomplete. It does not explain gravity, dark matter, or the matter/anti-matter asymmetry that brought galaxies and ourselves into existence. Our approach at CMS must be combined with results of other physics experiments using entirely different techniques. For example, the LHCb collaboration recently measured a violation of lepton universality at 3.1 standard deviations outside of the Standard Model through measuring b hadron decays alone [93]. Of course, along with the excitement generated by the LHCb results are calls for even more precise measurements of b decays. Additional understanding of this process may reveal much about the full $VHb\bar{b}$ process as well. We will likely pick up hints of deviation from the Standard Model, if they are there to be found, in many different experiments as a new level of understanding is gained from each new result.

Appendix A

Detector Projects

Each collaborator must contribute to the operation of the CMS detector before his or her name is added to the author list. The operation of the detector is distinct from analyzing the data generated by the detector, so all collaborators must adopt some role outside of being a physicist.

This appendix details projects I completed in order to contribute to the operation of the CMS detector. The first project presented is the Dynamo Consistency project. It is a plugin for the dynamic data management system Dynamo [94] that compares the inventory of files Dynamo expects at a site with the files that are actually at a site. The other project described is known as Workflow Web Tools. This is a dynamic web server that displays errors reported by the CMS computing infrastructure to operators, and allows those operators to perform corrective actions through the web page. Workflow Web Tools also tracks operator actions for future use in training various machine learning models. Both projects are published as software packages written in Python [95,96] and are available through the Python Package Index (PyPI) as `dynamo-consistency` and `workflowwebtools`, respectively.

A.1 Dynamo Consistency

Dynamo Consistency is a plugin for Dynamo Dynamic Data Management System that checks consistency between Dynamo's inventory and files actually located at managed

sites. Even though Dynamo controls and tracks the history of file transfers between computing sites, a separate check is needed to ensure files are not lost or accumulated due to user or system errors. For example, sites that can no longer access some files after a power outage can cause problems for many related activities. File transfers requested from a inconsistent site to another site will fail when files are missing. Sites will be also be chosen incorrectly for production jobs that assume the presence of a local file. Last disk copies may also be missing, causing a significant delay when a user requests data. Another type of inconsistency arises when files thought to be deleted are still on disk. This leads to wasted disk space for files that are not accessed, except by accident. Dynamo Consistency regularly checks consistency by listing each remote site and comparing the listed contents to Dynamo's inventory database. The results are reported back to Dynamo, which can then take corrective measures.

A single executable `dynamo-consistency` is provided to run the consistency check. This executable can be used directly in Dynamo's scheduling system. Most of the behavior is controlled via a single JSON configuration file, with options for site selection, passed via command line arguments. Differing command line arguments allows Dynamo to run separate schedules for differing site architectures.

Because Dynamo runs in a heterogeneous computing environment, different sites need to be listed remotely using different methods. Currently implemented are listings using XRootD Python bindings, the `gfal-ls` CLI [97], and a `xrdfs` subshell. These listers are easily extensible in Python, allowing for new site architectures to be added to Dynamo Consistency as well.

The default executable performs the check as expected, listing files that are not tracked by Dynamo as orphans and listing files that are not found at sites as missing. A few configurable filters can be added to modify these lists. Dynamo Consistency avoids listing orphan files that have a modification time that is recent. Paths to avoid deleting can also be set. Deletion and transfer requests that are queued are also used to filter the final report to avoid redundant actions from Dynamo.

In addition to tracking the consistency between Dynamo's inventory and physical site storage, Dynamo Consistency can report all remote files older than a certain age

138

in general directories. These files can also be filtered with path patterns, just as the regular consistency check. The time-based only reporting allows for cleaning of directories that Dynamo does not track. This is a setting recommended for large file systems that are written to with a high frequency.

Summaries of check results, as well as the statuses of running checks, are displayed in a web page. The page consists of a table that includes links to logs and lists of orphan and missing files. Cells are color coded to allow operators to quickly identify problematic sites. Historic summary data for each site is also accessible through this page.

If the available configuration options and listers are not enough, advanced users can also directly use the Python API to run a custom consistency check. For more details on the Dynamo Consistency package, see [98].

A.1.1 Installation

Dynamo Consistency requires the XRootD [99] Python module to be installed separately. In addition, it uses the Dynamo Dynamic Data Management package to get inventory listings and to report results of the consistency check. Any other needed packages are installed with Dynamo Consistency during installation.

The simplest way to install is through pip:

```
pip install dynamo-consistency
```

The source code is maintained on GitHub [100]. Other typical `setuptools` methods are supported by the repository's `setup.py`.

A.1.2 Inventory Listing

Two listings must be done to compare. One is the Inventory Listing, and the other is the Remote Listing. This section describes the inventory listing, and the next describes remote listing.

The inventory is queried before the site is listed remotely due to possible race conditions. It is not uncommon for a site listing to take multiple days. In the

meanwhile, two things can change in the inventory. A file can be deleted from a site or it can be added to a site. An added file is ignored by setting **IgnoreAge** in the configuration to a large enough value. Files that are deleted during the remote listing are filtered out by checking recent deletion requests after the remote listing.

There are currently multiple ways to get the site contents from Dynamo. One is to access the MySQL database use for Dynamo storage directly. This will work as long as the schema does not change. A more reliable way to keep up with major changes in Dyanmo is to use the Dynamo inventory object. This method is less optimized when working with the MySQL storage plugin, but will work for different schema and any different storage types that are added in the future.

The type of inventory lister is selected via command line options, or by setting `dynamo_consistency.opt.V1` to `True` or `False` before importing any modules that rely on the backend. By implementing the three modules `inventory`, `registry`, and `siteinfo`, described in the full documentation [98], any other method of communicating with an inventory can be added.

After selecting the backend, the inventory can be listed transparently using the following method:

```
from dynamo_consistency import inventorylister
listing = inventorylister.listing(sitename)
```

Here, `listing` is a `dynamo_consistency.datatypes.DirectoryInfo` object that is the root node of the full directory tree. Each node of `DirectoryInfo` contains meta data about a directory, such as its modification timestamp and name. It also holds a list of sub-directories, in the form of `DirectoryInfo` objects, and a list of files. The files are represented as dictionaries containing the name, size, and modification time of the file. Each file and `DirectoryInfo` also stores a hash of the meta data. The `DirectoryInfo` hash includes information from the object's files and sub-directories too. This is to speed up the file tree comparison, shown in Figure A-1

Hash match? — Yes → Return empty file list

return full files list

No

Yes

Init empty file list → Other tree? — Yes (up) → For sub-directory → for file

No

Start compare

Loop

Loop

if comparable get other and add files (start again)

if comparable and no matching hash in other, filter and append
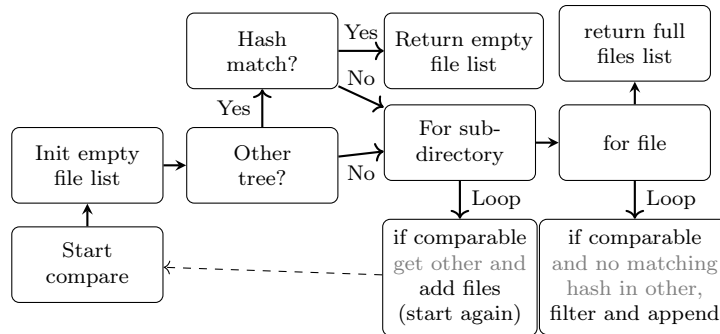
Figure A-1: Comparison algorithm

A.1.3 Remote Listing

The remote listing is equally flexible in terms of having multiple implementations used for the listing. The factory function `dynamo_consistency.backend.get_listers()` reads the configuration file to determine the type of lister for each site. There are currently three different classes implemented, and more can be added by extending the `dynamo_consistency.backend.listers.Lister` class and implementing its `ls_directory` method. The three current listers are the following:

- `dynamo_consistency.backend.listers.XRootDLister` - This listing object uses the `XRootD` Python module to connect to and query each site.

- `dynamo_consistency.backend.listers.GFalLister` - This listing object uses the `gfal-ls` command line tool to list remote sites.

- `dynamo_consistency.backend.listers.XRootDLister` - This listing object opens a subshell using the `xrdfs` command line tool and queries the remote site.

Once the type of lister is set in the configuration the contents of the remote site can be gathered with a simple interface:

```
from dynamo_consistency import remotelister
listing = remotelister.listing(sitename)
```

This takes much longer than the Inventory Listing, since every directory of the site needs to be queried over the network. A software layer between the listing class and

the final output creates multiple connections and works on two queues with multiple threads. There is the input queue, which is a list of directories that still need to be listed, and an output queue which holds the result of each directory listed so far. The workflow of the listing is shown in Figure A-2.
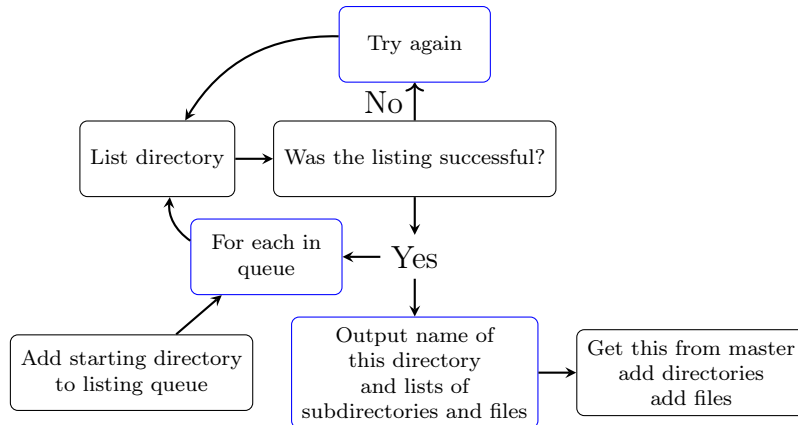
Figure A-2: Listing algorithm with retries

A.1.4 Executables

Other tools are available as part of `dynamo-consistency` in order to simplify operating the system. Many of these come as separate executables. A list of some of the executables installed with the package is given below.

dynamo-consistency

This program runs the Site Consistency Check for Dynamo Dynamic Data Management System.

```
Usage: dynamo-consistency [options]
```

```
Options:
  --version             show program's version number and exit
  -h, --help            show this help message and exit
  --config=FILE         Sets the location of the configuration file to read.
```

```
Selection Options:

  --site=PATTERN       Sets the pattern used to select a site to run on next.

  --lock=NAME          Sets the lock name that should be used for this run.

  --date-string=YYYYMMDD

                       Set the datestring to pull for RAL-Reader listers


Logging Options:

  --update-summary     Forces the update of the summary table, even if

                       loading trees

  --email              Send an email on uncaught exception.

  --info               Displays logs down to info level.

  --debug              Displays logs down to debug level.


Behavior Options:

  These options will change the backend loaded and actions taken


  --no-orphan          Do not delete any orphan files.

  --cms                Run actions specific to CMS collaboration data.

  --no-sam             Disables the SAM readiness check.

  --more-logs          Clean any "AdditionalLogDeletions" directories.

  --no-inventory       Do not connect the inventory. Used to test unmerged

  --unmerged           Run actions on "/store/unmerged".

  --v1                 Connect to Dynamo database directly

  --v1-reporting       Connect to Dynamo database directly for registry only.

  --cnf=FILE           Point to a non-default location of a ``my.cnf`` file.

  --test               Run with a test instance of backend module.
```

set-status

This script changes the status of a site on the summary web page. It can be used to
unlock from a dead process, disable sites from running, and change whether or not
to act on the site. This script can take a `-config <FILE>` parameter to point to a
configuration file, a la `dynamo-consistency`. For the last two arguments, `SITE` will

Table A.1: Valid statuses for sites as tracked by `dynamo-consistency` are described below.

Action	Description
ready	This sets the site status back to idle. This means the site is ready to run. Should be used on a site that is disabled.
halt	This stops a currently running or locked site. This site is still eligible to run.
disable	Can be applied to a site that is either running or ready. It halts the site and also prevents it from running until set to `ready` again.
act	Marks a site as one to report results to the registry.
dry	Opposite of `act`, this action prevents this site from making entries into the registry in future runs.

match the name of the site to change. `ACTION` can be one of the entries in Table A.1

```
Usage: set-status [options] SITE ACTION
```

```
Options:
  --version      show program's version number and exit
  -h, --help     show this help message and exit
  --config=FILE  Sets the location of the configuration file to read.


  Logging Options:
    --info       Displays logs down to info level.
    --debug      Displays logs down to debug level.
```

consistency-dump-tree

Dumps the `dynamo_consistency.datatypes.DirectoryInfo` tree into the cache directory. By default, it dumps the tree that would be read from the inventory.

If the `[NAME]` argument is not given, defaults to `inventory.pkl` or `remote.pkl` when using the `--remote` option.

```
Usage: consistency-dump-tree [options] [NAME]
```

```
Options:
  --version             show program's version number and exit
  -h, --help            show this help message and exit
  --config=FILE         Sets the location of the configuration file to read.


  Selection Options:
    --site=PATTERN      Sets the pattern used to select a site to run on next.
    --remote            Dump the remote site listing instead of the inventory
    --date-string=YYYYMMDD

                        Set the datestring to pull for RAL-Reader listers


  Logging Options:
    --info              Displays logs down to info level.
    --debug             Displays logs down to debug level.


  Behavior Options:
    These options will change the backend loaded and actions taken


    --unmerged          Run actions on "/store/unmerged".
    --v1                Connect to Dynamo database directly
    --test              Run with a test instance of backend module.
```

check-phedex

This program is only useful for double-checking CMS sites. This program checks a site's orphan files against PhEDEx. If any of the datasets are supposed to be at the site, this gives a non-zero exit code.

```
Usage: check-phedex [options] SITE


Options:
  --version       show program's version number and exit
```

```
-h, --help     show this help message and exit
--config=FILE  Sets the location of the configuration file to read.


Logging Options:
  --info         Displays logs down to info level.
  --debug        Displays logs down to debug level.
```

A.1.5   Configuration

The configuration file for `dynamo-consistency` is a JSON or YAML file with the following keys.

- **AccessMethod** - A dictionary of access methods for sites. Sites default to XRootD, but setting a value of `SRM` causes the site to be listed by `gfal-ls` commands.

- **AdditionalLogDeletions** - A dictionary that lists which directories have logs to be cleaned for different sites. These log directories are treated the same as log directories in `/store/unmerged`. This means they use the **UnmergedLogsAge** parameter to determine cleaning policy.

- **DirectoryList** - A list of directories inside of **RootPath** to check consistency.

- **DeleteOrphans** - By default, is true. If set to false, orphan files will all be filtered out so that none are deleted.

- **FreeMem** - The amount of free memory that is required for a check to run. The memory is given in GBs.

- **GFALThreads** - The number of threads used by the GFAL listers

- **GlobalRedirectors** - The redirectors to start all locate calls from, unless looking for a site that is listed in the **Redirectors** configuration.

- **IgnoreAge** - Ignore any files or directories with an age less than this, in days.

- **IgnoreDirectories** - The check ignores any paths that contain any of the strings in this list.

- **InventoryAge** - The age, in days, of how old the information from the inventory can be

- **ListAge** - The age, in days, of how old the list of files directly from the site can be

- **ListDeletable** - Configuration for unmerged cleaning "listdeletable" module. Details on some of the configuration parameters are documented online [101].

- **MaxMissing** - If more files than this number are missing, then there will be no automatic entry into the register.

- **MaxOrphan** - If more than files than this number are orphan files at a site, then there will be no automatic entry into the register.

- **NumThreads** - The number of threads used by the XRootD listers

- **PathPrefix** - A dictionary of prefixes to place before **RootPath** in the XRootD call. This allows for different paths for different sites. If the prefix is not set for a site, and it fails to list **RootPath**, it falls back to a default `/cms` before giving up.

- **RedirectorAge** - The age, in days, of how old the information on doors from redirectors can be. If this value is set to zero, the redirector information is never refreshed.

- **Redirectors** - A dictionary with keys of sites with hard-coded redirector locations. If a site is not listed in this way, the redirector is found by matching domains from `CMSToolBox.siteinfo.get_domain()` to redirectors found in a generic `xrdfs locate` call.

- **Retries** - Number of retries after timeouts to attempt

- **RootPath** - The directory where all of the listed subdirectories will be under. For CMS sites, this will be `"/store"`

- **SaveCache** - If set and evaluates to True, copies old cached directory trees instead of overwriting

- **Timeout** - This gives the amount of time, in seconds, that you want the listing to try to run on a single directory before it times out.

- **Unmerged** - A list of sites to handle cleaning of `/store/unmerged` on. If the list is empty, all the sites are managed centrally

- **UnmergedLogsAge** - The minimum age of the unmerged logs to be deleted, in days

- **UseLoadBalancer** - A list of sites where the main redirector of the site is used

- **UseTransferQueue** - If true, put missing files into tranfer queue table when using `--v1` for reporting. Defaults to true value.

- **VarLocation** - The location for the varying directory. Inside this directory will be:

  - Logs

  - Redirector lists

  - Cached trees

  - Lock files

- **WebDir** - The directory where text files and the sqlite3 database are stored

Configuration parameters can also be quickly overwritten for a given run by setting an environment variable of the same name.

A.1.6 Comparison Script

The full set of operations in a typical run of the `dynamo-consistency` is enumerated below.

1. Points `config.py` to the local `consistency_config.json` file

2. Notes the time, and if it's daylight savings time for entry into the summary database

3. Reads the list of previous missing files, since it requires a file to be missing on multiple runs before registering it to be copied

4. It gathers the inventory tree by calling
   `dynamo_consistency.getinventorycontents.get_db_listing()`.

5. Creates a list of datasets to not report missing files in. This list consists of deletion requests fetched from PhEDEx by
   `dynamo_consistency.checkphedex.set_of_deletions()`

6. It creates a list of datasets to not report orphans in. This list consists of the following.

   - Datasets that have any files on the site, as listed by the dynamo MySQL database

   - Deletion requests fetched from PhEDEx (same list as datasets to skip in missing)

   - Any datasets that have the status flag set to '`IGNORED`' in the dynamo database

   - Merging datasets that are protected by Unified

7. It gathers the site tree by calling
   `dynamo_consistency.getsitecontents.get_site_tree()`. The list of orphans is used during the running to filter out empty directories that are reported to the registry during the run.

8. Does the comparison between the two trees made, using the configuration options concerning file age.

9. If the number of missing files is less than **MaxMissing**, the number of orphans is less than **MaxOrphan**, and the site is under the web page's "Debugged sites" tab, connects to a dynamo registry to report the following errors:

   - Every orphan file and every empty directory that is not too new nor should contain missing files is entered in the deletion queue.

   - For each missing file, every possible source site as listed by the dynamo database, (not counting the site where missing), is entered in the transfer queue. Creates a text file full of files that only exist elsewhere on tape.

10. Creates a text file that contains the missing blocks and groups.

11. `.txt` file lists and details of orphan and missing files are moved to the web space

12. If the site is listed in the configuration under the **Unmerged** list, the unmerged cleaner is run over the site:

    - `dynamo_consistency.getsitecontents.get_site_tree()` is run again, this time only over `/store/unmerged`

    - Empty directories that are not too new nor protected by Unified are entered into the deletion queue

    - The list of files is passed through the unmerged cleaner

    - The list of files to delete from unmerged cleaner are entered in the deletion queue

13. The summary database is updated to show the last update on the website

A.2   Workflow Web Tools

Datasets used by the CMS collaboration are produced centrally by the Workflow Team. During production of these datasets, errors can arise from things such as

150

missing files, too much memory usage, or loss of functionality at a particular site. For each failed dataset, Workflow Team operators must gather information from error reports and decide how to recover in order to complete the workflow.

The Workflow Web Tools has the dual purpose of displaying these errors as well as allowing operators to submit actions directly to the web page, instead of requiring them to open up separate shells to script actions. The existence of this layer, which is aware of both the error reports and the submitted actions, also allows for investigation into automatically attempting to recover simple errors.

A.2.1 Error Information

A matrix of errors is built for each task of each workflow. Each row of the matrix corresponds to a different error code that was reported back as a job failed to complete. Each column corresponds to which computing site that the job was running at. The cell simply contains the number of jobs that threw the particular error code while running at a particular site. At the time when the error report is gathered, Workflow Web Tools also gathers the site readiness reports of each computing site. This can be correlated with the error reports to see if jobs failed due to site problems, rather then problems with the workflow submission itself. An example of the display for this error information is given in Figure A-3.

A.2.2 Possible Actions

Workflows are typically completed after throwing errors by either rerunning the failed jobs, called ACDC, or by discarding the entire workflow, and restarting it. Both options often include adjustment of parameters such as memory requested and the number of events run per job. These actions are all set with a number of radio buttons, meaning that there is a finite set of actions that an operator can take. This makes it easier for an automated system to learn and take over some operations.

Each action submitted by an operator is stored in a database and exposed via an API for Unified, the backend manager of workflows, to read the actions from the

151

	T0_CH_CERN	T1_DE_KIT	T1_ES_PIC	T1_FR_CCIN2P3	T1_IT_CNAF	T1_RU_JINR	T1_UK_RAL	T1_US_FNAL	T2_CH_CERN	T2_CH_CERNBOX	T2_CH_CERN_HLT	T2_DE_DESY	T2_DE_RWTH	T2_ES_IFCA	T2_FR_GRIF_IRFU	T2_FR_GRIF_LLR	T2_IT_Bari	T2_IT_Legnaro	T2_UK_London_Brunel	T2_UK_London_IC	T2_UK_SGrid_RALPP	T2_US_Florida	T2_US_MIT	T2_US_UCSD	T2_US_Wisconsin	T3_US_FNALLPC	null
-1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0
85	0	0	0	1	0	1	11	0	0	0	6	0	1	0	1	2	1	0	1	1	1	1	0	1	0	0	0
86	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
92	0	0	2	1	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0
132	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
134	0	0	0	0	0	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0
139	0	0	0	0	0	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
8004	0	1	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
50110	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
50660	0	0	4	0	1	0	0	2	1	0	59	1	3	0	0	0	1	2	1	1	0	3	1	8	3	0	0
50664	0	0	7	5	3	0	1	1	0	0	33	0	0	0	2	0	4	0	0	0	9	2	0	0	1	0	0
71304	0	0	0	0	0	0	0	0	0	0	88	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
99109	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
99305	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	96

Figure A-3: An example error matrix is given above. The site names are color coded to show their site readiness status. The -1 error code is used to report jobs that did not complete successfully, but also did not return an error code.

web tool. Even after acting on the workflows, these actions are persistently stored for future study. Workflow and task names can be matched and displayed to show both the error reports and the eventual action taken in a machine-readable format.

A.2.3 Automatic Actions

The stored data on error information and submitted actions were used in several machine learning studies to try to predict actions using the error matrix and site statuses in a Deep Neural Network. The unbalanced nature of the submitted actions (the vast majority being ACDC) means that significantly more data will need to be gathered for a naïve approach to work.

A more successful approach has been to create static models that correlate most frequent error codes and statuses of sites relevant to the workflow to automatically predict actions. This separate project, Artificial Intelligence Error Handling (AIEH), has been able to reduce operator loads significantly. The Workflow Web Tools periodically queries AIEH for predictions of all promising workflows, and automatically presents AIEH's predictions to Unified without operator intervention.

Appendix B

Simulation Details

The datasets use for each year are given in Table B.1, Table B.2, and Table B.3.

Sample Name	Xsec (pb)	KFactor	Total Events	Weight
DYBJetsToLL_M-50_Zpt-100to200_TuneCUETP8M1_13TeV-madgraphMLM-pythia8/	3.086	1.23	8.38e+06	1.63e-02
DYBJetsToLL_M-50_Zpt-200toInf_TuneCUETP8M1_13TeV-madgraphMLM-pythia8/	0.3167	1.23	1.28e+06	1.10e-02
DYJetsToLL_BGenFilter_Zpt-100to200_M-50_TuneCUETP8M1_13TeV-madgraphMLM-pythia8/	3.414	1.23	2.67e+05	5.65e-01
DYJetsToLL_BGenFilter_Zpt-200toInf_M-50_TuneCUETP8M1_13TeV-madgraphMLM-pythia8/	0.5084	1.23	1.29e+05	1.75e-01
DYJetsToLL_M-50_HT-100to200_TuneCUETP8M1_13TeV-madgraphMLM-pythia8/	147.3	1.23	9.37e+06	6.96e-01
DYJetsToLL_M-50_HT-1200to2500_TuneCUETP8M1_13TeV-madgraphMLM-pythia8/	0.1512	1.23	5.96e+05	1.12e-02
DYJetsToLL_M-50_HT-200to400_TuneCUETP8M1_13TeV-madgraphMLM-pythia8/	41.04	1.23	8.65e+06	2.10e-01
DYJetsToLL_M-50_HT-2500toInf_TuneCUETP8M1_13TeV-madgraphMLM-pythia8/	0.003659	1.23	3.99e+05	4.06e-04
DYJetsToLL_M-50_HT-400to600_TuneCUETP8M1_13TeV-madgraphMLM-pythia8/	5.676	1.23	7.66e+06	3.28e-02
DYJetsToLL_M-50_HT-600to800_TuneCUETP8M1_13TeV-madgraphMLM-pythia8/	1.359	1.23	8.29e+06	7.26e-03
DYJetsToLL_M-50_HT-800to1200_TuneCUETP8M1_13TeV-madgraphMLM-pythia8/	0.623	1.23	2.67e+06	1.03e-02
DYJetsToLL_M-50_TuneCUETP8M1_13TeV-madgraphMLM-pythia8/	4954.0	1.0	1.46e+08	1.22e+00
QCD_HT1000to1500_TuneCUETP8M1_13TeV-madgraphMLM-pythia8/	1208.0	1.0	1.52e+07	2.86e+00
QCD_HT100to200_TuneCUETP8M1_13TeV-madgraphMLM-pythia8/	27940000.0	1.0	8.23e+07	1.22e+04
QCD_HT1500to2000_TuneCUETP8M1_13TeV-madgraphMLM-pythia8/	120.4	1.0	1.18e+07	3.66e-01
QCD_HT2000toInf_TuneCUETP8M1_13TeV-madgraphMLM-pythia8/	25.27	1.0	5.97e+06	1.52e-01
QCD_HT200to300_TuneCUETP8M1_13TeV-madgraphMLM-pythia8/	1712000.0	1.0	5.76e+07	1.07e+03
QCD_HT300to500_TuneCUETP8M1_13TeV-madgraphMLM-pythia8/	345900.0	1.0	5.46e+07	2.28e+02
QCD_HT500to700_TuneCUETP8M1_13TeV-madgraphMLM-pythia8/	31630.0	1.0	6.26e+07	1.82e+01
QCD_HT700to1000_TuneCUETP8M1_13TeV-madgraphMLM-pythia8/	6828.0	1.0	3.72e+07	6.61e+00
ST_t-channel_antitop_4f_inclusiveDecays_13TeV-powhegV2-madspin-pythia8_TuneCUETP8M1/	80.95	1.0	3.88e+07	7.51e-02
TT_TuneCUETP8M2T4_13TeV-powheg-pythia8/	831.76	1.0	7.69e+07	3.89e-01
ST_s-channel_4f_InclusiveDecays_13TeV-amcatnlo-pythia8/	3.36	1.0	2.95e+07	4.10e-03
ST_t-channel_top_4f_inclusiveDecays_13TeV-powheg-madspin-pythia8_TuneCUETP8M1/	136.02	1.0	6.71e+07	7.30e-02
ST_tW_top_5f_inclusiveDecays_13TeV-powheg-pythia8_TuneCUETP8M1/	35.6	1.0	6.95e+06	1.84e-01
ST_tW_antitop_5f_inclusiveDecays_13TeV-powheg-pythia8_TuneCUETP8M1/	35.6	1.0	6.93e+06	1.85e-01
WBJetsToLNu_Wpt-100to200_TuneCUETP8M1_13TeV-madgraphMLM-pythia8/	5.969	1.22	3.98e+06	6.59e-02
WBJetsToLNu_Wpt-200toInf_TuneCUETP8M1_13TeV-madgraphMLM-pythia8/	0.8744	1.22	5.43e+06	7.07e-02
WJetsToLNu_HT-100To200_TuneCUETP8M1_13TeV-madgraphMLM-pythia8/	1349.0	1.22	7.60e+07	7.80e-01
WJetsToLNu_HT-1200To2500_TuneCUETP8M1_13TeV-madgraphMLM-pythia8/	1.329	1.22	3.17e+06	1.84e-02
WJetsToLNu_HT-200To400_TuneCUETP8M1_13TeV-madgraphMLM-pythia8/	359.3	1.22	3.90e+07	4.05e-01
WJetsToLNu_HT-2500ToInf_TuneCUETP8M1_13TeV-madgraphMLM-pythia8/	0.03216	1.22	2.64e+06	5.35e-04
WJetsToLNu_HT-400To600_TuneCUETP8M1_13TeV-madgraphMLM-pythia8/	48.86	1.22	6.07e+06	3.54e-01
WJetsToLNu_HT-600To800_TuneCUETP8M1_13TeV-madgraphMLM-pythia8/	12.03	1.22	1.87e+07	2.83e-02
WJetsToLNu_HT-70To100_TuneCUETP8M1_13TeV-madgraphMLM-pythia8/	1352.0	1.22	3.26e+07	1.82e+00
WJetsToLNu_HT-800To1200_TuneCUETP8M1_13TeV-madgraphMLM-pythia8/	5.482	1.22	6.29e+06	3.83e-02
WJetsToLNu_BGenFilter_Wpt-100to200_TuneCUETP8M1_13TeV-madgraphMLM-pythia8/	26.89	1.22	4.19e+05	2.82e+00
WJetsToLNu_BGenFilter_Wpt-200toInf_TuneCUETP8M1_13TeV-madgraphMLM-pythia8/	3.786	1.22	2.15e+05	7.72e-01
WJetsToLNu_TuneCUETP8M1_13TeV-madgraphMLM-pythia8/	61526.7	1.0	8.69e+07	2.55e+01
WW_TuneCUETP8M1_13TeV-pythia8/	117.6	1.0	7.98e+06	5.30e-01
WWTo1L1Nu2Q_13TeV_amcatnloFXFX_madspin_pythia8/	51.87	1.0	4.41e+08	4.23e-03
WZ_TuneCUETP8M1_13TeV-pythia8/	48.1	1.0	4.00e+06	4.33e-01
WZTo1L1Nu2Q_13TeV_amcatnloFXFX_madspin_pythia8/	10.875	1.0	1.94e+08	2.01e-03
WminusH_HToBB_WToLNu_M125_13TeV_powheg_pythia8/	0.10899	1.0	4.21e+05	9.32e-03
WplusH_HToBB_WToLNu_M125_13TeV_powheg_pythia8/	0.17202	1.0	6.90e+05	8.97e-03
ZBJetsToNuNu_Zpt-100to200_TuneCUETP8M1_13TeV-madgraphMLM-pythia8	5.93	1.23	6.79e+06	3.87e-02
ZBJetsToNuNu_Zpt-200toInf_TuneCUETP8M1_13TeV-madgraphMLM-pythia8	0.6075	1.23	4.44e+05	6.06e-02
ZH_HToBB_ZToLL_M125_13TeV_powheg_pythia8/	0.04718	1.0	2.24e+05	7.59e-03
ZH_HToBB_ZToNuNu_M125_13TeV_powheg_pythia8/	0.09322	1.0	4.35e+05	7.72e-03
ZJetsToNuNu_BGenFilter_Zpt-100to200_TuneCUETP8M1_13TeV-madgraphMLM-pythia8	2.162	1.23	1.50e+05	6.37e-01
ZJetsToNuNu_BGenFilter_Zpt-200toInf_TuneCUETP8M1_13TeV-madgraphMLM-pythia8	0.3151	1.23	9.78e+04	1.43e-01
ZJetsToNuNu_HT-100To200_13TeV-madgraph	280.59	1.23	2.42e+07	5.12e-01
ZJetsToNuNu_HT-1200To2500_13TeV-madgraph	0.28629	1.23	5.13e+05	2.47e-02
ZJetsToNuNu_HT-200To400_13TeV-madgraph	77.79	1.23	2.48e+07	1.39e-01
ZJetsToNuNu_HT-2500ToInf_13TeV-madgraph	0.006912	1.23	4.05e+05	7.56e-04
ZJetsToNuNu_HT-400To600_13TeV-madgraph	10.755	1.23	8.84e+06	5.38e-02
ZJetsToNuNu_HT-600To800_13TeV-madgraph	2.5644	1.23	5.77e+06	1.97e-02
ZJetsToNuNu_HT-800To1200_13TeV-madgraph	1.1802	1.23	2.17e+06	2.41e-02
ZZ_TuneCUETP8M1_13TeV-pythia8/	17.2	1.0	1.69e+06	3.66e-01
ZZTo2L2Q_13TeV_amcatnloFXFX_madspin_pythia8/	2.387	1.0	7.34e+07	1.17e-03
ZZTo2Q2Nu_13TeV_amcatnloFXFX_madspin_pythia8/	4.726	1.0	1.99e+08	8.55e-04
ggZH_HToBB_ZToLL_M125_13TeV_powheg_pythia8/	0.0072	1.0	3.00e+06	8.65e-05
ggZH_HToBB_ZToNuNu_M125_13TeV_powheg_pythia8/	0.01437	1.0	3.00e+06	1.72e-04

Table B.1: The MC datasets used for 2016 are listed above.

Sample Name	Xsec (pb)	KFactor	Total Events	Weight
DYBJetsToLL_M-50_Zpt-100to200_TuneCP5_13TeV-madgraphMLM-pythia8_newgridpack/	3.224	1.23	2.72e+06	6.09e-02
DYBJetsToLL_M-50_Zpt-200toInf_TuneCP5_13TeV-madgraphMLM-pythia8_newgridpack/	0.3298	1.23	2.40e+05	7.05e-02
DYJetsToLL_BGenFilter_Zpt-100to200_M-50_TuneCP5_13TeV-madgraphMLM-pythia8_newgridpack/	2.671	1.23	1.91e+06	7.16e-02
DYJetsToLL_BGenFilter_Zpt-200toInf_M-50_TuneCP5_13TeV-madgraphMLM-pythia8_newgridpack/	0.3934	1.23	3.21e+05	6.28e-02
DYJetsToLL_M-50_HT-100to200_TuneCP5_13TeV-madgraphMLM-pythia8/	161.1	1.23	1.51e+07	5.45e-01
DYJetsToLL_M-50_HT-1200to2500_TuneCP5_13TeV-madgraphMLM-pythia8/	0.1933	1.23	6.25e+05	1.59e-02
DYJetsToLL_M-50_HT-200to400_TuneCP5_13TeV-madgraphMLM-pythia8/	48.66	1.23	9.74e+06	2.56e-01
DYJetsToLL_M-50_HT-2500toInf_TuneCP5_13TeV-madgraphMLM-pythia8/	0.003468	1.23	4.19e+05	4.24e-04
DYJetsToLL_M-50_HT-400to600_TuneCP5_13TeV-madgraphMLM-pythia8/	6.968	1.23	7.05e+06	5.07e-02
DYJetsToLL_M-50_HT-600to800_TuneCP5_13TeV-madgraphMLM-pythia8/	1.743	1.23	6.22e+06	1.44e-02
DYJetsToLL_M-50_HT-800to1200_TuneCP5_13TeV-madgraphMLM-pythia8/	0.8052	1.23	2.10e+06	1.96e-02
DYJetsToLL_M-4to50_HT-100to200_TuneCP5_13TeV-madgraphMLM-pythia8/	204.0	1.23	9.88e+06	1.06e+00
DYJetsToLL_M-4to50_HT-200to400_TuneCP5_13TeV-madgraphMLM-pythia8/	54.39	1.23	4.02e+06	6.94e-01
DYJetsToLL_M-4to50_HT-400to600_TuneCP5_13TeV-madgraphMLM-pythia8/	5.697	1.23	2.98e+06	9.81e-02
DYJetsToLL_M-4to50_HT-600toInf_TuneCP5_13TeV-madgraphMLM-pythia8/	1.85	1.23	2.93e+06	3.24e-02
DYJetsToLL_M-50_TuneCP5_13TeV-madgraphMLM-pythia8/	5343.0	1.23	9.70e+07	2.83e+00
QCD_HT1000to1500_TuneCP5_13TeV-madgraph-pythia8/	1088.0	1.0	1.66e+07	2.73e+00
QCD_HT1500to2000_TuneCP5_13TeV-madgraph-pythia8/	99.11	1.0	1.16e+07	3.55e-01
QCD_HT2000toInf_TuneCP5_13TeV-madgraph-pythia8/	20.23	1.0	5.94e+06	1.42e-01
QCD_HT200to300_TuneCP5_13TeV-madgraph-pythia8/	1547000.0	1.0	5.94e+07	1.09e+03
QCD_HT300to500_TuneCP5_13TeV-madgraph-pythia8/	322600.0	1.0	6.02e+07	2.23e+02
QCD_HT500to700_TuneCP5_13TeV-madgraph-pythia8/	29980.0	1.0	5.62e+07	2.22e+01
QCD_HT700to1000_TuneCP5_13TeV-madgraph-pythia8/	6334.0	1.0	7.53e+07	3.51e+00
ST_s-channel_4f_leptonDecays_TuneCP5_PSweights_13TeV-amcatnlo-pythia8/	3.74	1.0	9.91e+06	1.57e-02
ST_t-channel_antitop_4f_InclusiveDecays_TuneCP5_PSweights_13TeV-powheg-pythia8/	80.95	1.0	6.26e+07	5.39e-02
ST_t-channel_top_4f_InclusiveDecays_TuneCP5_PSweights_13TeV-powheg-pythia8/	136.02	1.0	1.75e+08	3.23e-02
ST_tW_antitop_5f_inclusiveDecays_TuneCP5_PSweights_13TeV-powheg-pythia8/	35.85	1.0	7.75e+06	1.93e-01
ST_tW_top_5f_inclusiveDecays_TuneCP5_PSweights_13TeV-powheg-pythia8/	35.85	1.0	7.95e+06	1.88e-01
TTToHadronic_TuneCP5_PSweights_13TeV-powheg-pythia8/	377.96	1.0	1.30e+08	1.21e-01
TTTo2L2Nu_TuneCP5_PSweights_13TeV-powheg-pythia8/	88.29	1.0	3.78e+07	9.73e-02
TTToSemiLeptonic_TuneCP5_PSweights_13TeV-powheg-pythia8/	365.34	1.0	8.41e+07	1.81e-01
WBJetsToLNu_Wpt-100to200_TuneCP5_13TeV-madgraphMLM-pythia8_newgridpack/	5.542	1.21	4.15e+06	6.74e-02
WBJetsToLNu_Wpt-200toInf_TuneCP5_13TeV-madgraphMLM-pythia8_newgridpack/	0.801	1.21	7.72e+05	5.24e-02
WJetsToLNu_HT-100To200_TuneCP5_13TeV-madgraphMLM-pythia8/	1395.0	1.21	3.45e+07	2.04e+00
WJetsToLNu_HT-1200To2500_TuneCP5_13TeV-madgraphMLM-pythia8/	1.074	1.21	1.24e+07	4.36e-03
WJetsToLNu_HT-200To400_TuneCP5_13TeV-madgraphMLM-pythia8/	407.9	1.21	2.13e+07	9.69e-01
WJetsToLNu_HT-2500ToInf_TuneCP5_13TeV-madgraphMLM-pythia8/	0.03216	1.21	1.28e+07	1.26e-04
WJetsToLNu_HT-400To600_TuneCP5_13TeV-madgraphMLM-pythia8/	57.48	1.21	5.63e+06	5.15e-01
WJetsToLNu_HT-600To800_TuneCP5_13TeV-madgraphMLM-pythia8/	12.87	1.21	1.88e+07	3.45e-02
WJetsToLNu_HT-800To1200_TuneCP5_13TeV-madgraphMLM-pythia8/	5.366	1.21	8.15e+06	3.32e-02
WJetsToLNu_BGenFilter_Wpt-100to200_TuneCP5_13TeV-madgraphMLM-pythia8_newgridpack/	20.56	1.21	1.34e+07	7.76e-02
WJetsToLNu_BGenFilter_Wpt-200toInf_TuneCP5_13TeV-madgraphMLM-pythia8_newgridpack/	2.936	1.21	2.48e+06	5.96e-02
WJetsToLNu_TuneCP5_13TeV-madgraphMLM-pythia8/	52940.0	1.0	7.77e+07	2.84e+01
WW_TuneCP5_13TeV-pythia8/	115.3	1.0	7.71e+06	6.24e-01
WWTo1L1Nu2Q_13TeV_amcatnloFXFX_madspin_pythia8/	50.858	1.0	5.05e+06	4.20e-01
WWToLNuQQ_NNPDF31_TuneCP5_PSweights_13TeV-powheg-pythia8/	50.858	1.0	8.79e+06	2.41e-01
WZ_TuneCP5_13TeV-pythia8/	48.1	1.0	3.93e+06	5.11e-01
WminusH_HToBB_WToLNu_M125_13TeV_powheg_pythia8/	0.10899	1.0	2.38e+06	1.91e-03
WplusH_HToBB_WToLNu_M125_13TeV_powheg_pythia8/	0.17202	1.0	4.84e+06	1.48e-03
ZBJetsToNuNu_Zpt-100to200_TuneCP5_13TeV-madgraphMLM-pythia8_newgridpack/	6.209	1.23	4.98e+06	6.40e-02
ZBJetsToNuNu_Zpt-200toInf_TuneCP5_13TeV-madgraphMLM-pythia8_newgridpack/	0.6286	1.23	5.14e+05	6.27e-02
ZH_HToBB_ZToNuNu_M125_13TeV_powheg_pythia8/	0.09322	1.0	2.58e+06	1.51e-03
ZJetsToNuNu_BGenFilter_Zpt-100to200_TuneCP5_13TeV-madgraphMLM-pythia8_newgridpack/	1.689	1.23	1.37e+06	6.31e-02
ZJetsToNuNu_BGenFilter_Zpt-200toInf_TuneCP5_13TeV-madgraphMLM-pythia8_newgridpack/	0.2476	1.23	1.89e+05	6.73e-02
ZJetsToNuNu_HT-100To200_13TeV-madgraph/	302.8	1.23	2.27e+07	6.83e-01
ZJetsToNuNu_HT-200To400_13TeV-madgraph/	92.59	1.23	2.17e+07	2.19e-01
ZJetsToNuNu_HT-2500ToInf_13TeV-madgraph/	0.005146	1.23	6.73e+03	3.92e-02
ZJetsToNuNu_HT-400To600_13TeV-madgraph/	13.18	1.23	1.83e+07	3.70e-02
ZJetsToNuNu_HT-600To800_13TeV-madgraph/	3.257	1.23	1.13e+07	1.47e-02
ZJetsToNuNu_HT-800To1200_13TeV-madgraph/	1.496	1.23	2.06e+06	3.73e-02
ZZ_TuneCP5_13TeV-pythia8/	14.6	1.0	1.95e+06	3.12e-01
ZZTo2L2Q_13TeV_amcatnloFXFX_madspin_pythia8/	10.88	1.0	2.45e+07	1.85e-02
ZZTo4L_13TeV_powheg_pythia8/	2.038	1.0	2.50e+07	3.40e-03
ggZH_HToBB_ZToLL_M125_13TeV_powheg_pythia8/	0.0072	1.0	2.85e+06	1.05e-04
ggZH_HToBB_ZToNuNu_M125_13TeV_powheg_pythia8/	0.01437	1.0	2.98e+06	2.01e-04

Table B.2: The MC datasets used for 2017 are listed above.

Sample Name	Xsec (pb)	KFactor	Total Events	Weight
DYBJetsToLL_M-50_Zpt-100to200_TuneCP5_13TeV-madgraphMLM-pythia8/	3.206	1.23	1.99e+07	1.19e-02
DYBJetsToLL_M-50_Zpt-200toInf_TuneCP5_13TeV-madgraphMLM-pythia8/	0.3304	1.23	1.75e+06	1.40e-02
DYJetsToLL_BGenFilter_Zpt-100to200_M-50_TuneCP5_13TeV-madgraphMLM-pythia8/	2.662	1.23	6.54e+05	3.01e-01
DYJetsToLL_BGenFilter_Zpt-200toInf_M-50_TuneCP5_13TeV-madgraphMLM-pythia8/	0.3949	1.23	1.03e+05	2.83e-01
DYJetsToLL_M-50_HT-100to200_TuneCP5_PSweights_13TeV-madgraphMLM-pythia8/	160.8	1.23	1.15e+07	1.03e+00
DYJetsToLL_M-50_HT-1200to2500_TuneCP5_PSweights_13TeV-madgraphMLM-pythia/	0.1931	1.23	5.32e+05	2.68e-02
DYJetsToLL_M-50_HT-200to400_TuneCP5_PSweights_13TeV-madgraphMLM-pythia8/	48.63	1.23	1.07e+07	3.35e-01
DYJetsToLL_M-50_HT-2500toInf_TuneCP5_PSweights_13TeV-madgraphMLM-pythia8/	0.003513	1.23	4.16e+05	6.24e-04
DYJetsToLL_M-50_HT-400to600_TuneCP5_PSweights_13TeV-madgraphMLM-pythia8/	6.982	1.23	3.44e+07	1.50e-02
DYJetsToLL_M-50_HT-600to800_TuneCP5_PSweights_13TeV-madgraphMLM-pythia8/	1.756	1.23	8.83e+06	1.47e-02
DYJetsToLL_M-50_HT-800to1200_TuneCP5_PSweights_13TeV-madgraphMLM-pythia8/	0.8094	1.23	3.12e+06	1.91e-02
DYJetsToLL_M-50_TuneCP5_13TeV-madgraphMLM-pythia8/	5343.0	1.23	9.78e+07	4.03e+00
QCD_HT1000to1500_TuneCP5_13TeV-madgraphMLM-pythia8/	1088.0	1.0	1.54e+07	4.24e+00
QCD_HT1500to2000_TuneCP5_13TeV-madgraphMLM-pythia8/	99.11	1.0	1.01e+07	5.87e-01
QCD_HT2000toInf_TuneCP5_13TeV-madgraphMLM-pythia8/	20.23	1.0	5.41e+06	2.24e-01
QCD_HT200to300_TuneCP5_13TeV-madgraphMLM-pythia8/	1547000.0	1.0	5.42e+07	1.71e+03
QCD_HT300to500_TuneCP5_13TeV-madgraphMLM-pythia8/	322600.0	1.0	5.46e+07	3.55e+02
QCD_HT500to700_TuneCP5_13TeV-madgraphMLM-pythia8/	29980.0	1.0	5.50e+07	3.27e+01
QCD_HT700to1000_TuneCP5_13TeV-madgraphMLM-pythia8/	6334.0	1.0	4.47e+07	8.50e+00
ST_s-channel_4f_leptonDecays_TuneCP5_13TeV-madgraph-pythia8/	3.74	1.0	1.48e+08	1.51e-03
ST_t-channel_antitop_4f_InclusiveDecays_TuneCP5_13TeV-powheg-madspin-pythia8/	80.95	1.0	5.13e+09	9.48e-04
ST_t-channel_top_4f_InclusiveDecays_TuneCP5_13TeV-powheg-madspin-pythia8/	136.02	1.0	1.66e+10	4.92e-04
ST_tW_antitop_5f_inclusiveDecays_TuneCP5_13TeV-powheg-pythia8/	35.85	1.0	2.66e+08	8.07e-03
ST_tW_antitop_5f_NoFullyHadronicDecays_TuneCP5_13TeV-powheg-pythia8/	19.56	1.0	2.42e+08	4.86e-03
ST_tW_top_5f_inclusiveDecays_TuneCP5_13TeV-powheg-pythia8/	35.85	1.0	3.35e+08	6.42e-03
ST_tW_top_5f_NoFullyHadronicDecays_TuneCP5_13TeV-powheg-pythia8/	19.56	1.0	3.04e+08	3.86e-03
TTToHadronic_TuneCP5_13TeV-powheg-pythia8/	377.96	1.0	1.03e+11	2.20e-04
TTTo2L2Nu_TuneCP5_13TeV-powheg-pythia8/	88.29	1.0	4.50e+09	1.18e-03
TTToSemiLeptonic_TuneCP5_13TeV-powheg-pythia8/	365.34	1.0	5.38e+10	4.07e-04
WBJetsToLNu_Wpt-100to200_TuneCP5_13TeV-madgraphMLM-pythia8/	5.527	1.21	7.36e+06	5.45e-02
WBJetsToLNu_Wpt-200toInf_TuneCP5_13TeV-madgraphMLM-pythia8/	0.7996	1.21	3.02e+05	1.92e-01
WJetsToLNu_HT-100To200_TuneCP5_13TeV-madgraphMLM-pythia8/	1392.0	1.21	2.83e+07	3.57e+00
WJetsToLNu_HT-1200To2500_TuneCP5_13TeV-madgraphMLM-pythia8/	1.084	1.21	7.40e+06	1.06e-02
WJetsToLNu_HT-200To400_TuneCP5_13TeV-madgraphMLM-pythia8/	410.3	1.21	2.22e+07	1.34e+00
WJetsToLNu_HT-2500ToInf_TuneCP5_13TeV-madgraphMLM-pythia8/	0.008067	1.21	3.19e+06	1.84e-04
WJetsToLNu_HT-400To600_TuneCP5_13TeV-madgraphMLM-pythia8/	57.85	1.21	5.91e+06	7.10e-01
WJetsToLNu_HT-600To800_TuneCP5_13TeV-madgraphMLM-pythia8/	12.95	1.21	1.68e+07	5.61e-02
WJetsToLNu_HT-800To1200_TuneCP5_13TeV-madgraphMLM-pythia8/	5.45	1.21	8.36e+06	4.73e-02
WJetsToLNu_BGenFilter_Wpt-100to200_TuneCP5_13TeV-madgraphMLM-pythia8/	20.49	1.21	2.36e+05	6.30e+00
WJetsToLNu_BGenFilter_Wpt-200toInf_TuneCP5_13TeV-madgraphMLM-pythia8/	2.935	1.21	8.30e+05	2.57e-01
WJetsToLNu_HT-70To100_TuneCP5_13TeV-madgraphMLM-pythia8/	1353.0	1.21	7.04e+07	1.40e+00
WminusH_HToBB_WToLNu_M125_13TeV_powheg_pythia8/	0.10899	1.0	6.40e+05	1.02e-02
WplusH_HToBB_WToLNu_M125_13TeV_powheg_pythia8/	0.17202	1.0	1.01e+06	1.02e-02
ZBJetsToNuNu_Zpt-100to200_TuneCP5_13TeV-madgraphMLM-pythia8	6.195	1.23	6.65e+07	6.87e-03
ZBJetsToNuNu_Zpt-200toInf_TuneCP5_13TeV-madgraphMLM-pythia8	0.6293	1.23	7.35e+06	6.32e-03
ZH_HToBB_ZToLL_M125_13TeV_powheg_pythia8/	0.04718	1.0	5.53e+05	5.12e-03
ZH_HToBB_ZToNuNu_M125_13TeV_powheg_pythia8/	0.09322	1.0	7.15e+05	7.82e-03
ZJetsToNuNu_BGenFilter_Zpt-100to200_TuneCP5_13TeV-madgraphMLM-pythia8	1.679	1.23	2.01e+05	6.16e-01
ZJetsToNuNu_BGenFilter_Zpt-200toInf_TuneCP5_13TeV-madgraphMLM-pythia8	0.2468	1.23	3.73e+04	4.88e-01
ZJetsToNuNu_HT-100To200_13TeV-madgraph	303.4	1.23	2.37e+07	9.46e-01
ZJetsToNuNu_HT-1200To2500_13TeV-madgraph	0.3425	1.23	3.40e+05	7.43e-02
ZJetsToNuNu_HT-200To400_13TeV-madgraph	91.71	1.23	2.32e+07	2.91e-01
ZJetsToNuNu_HT-2500ToInf_13TeV-madgraph	0.005263	1.23	3.50e+05	1.11e-03
ZJetsToNuNu_HT-400To600_13TeV-madgraph	13.1	1.23	9.48e+06	1.02e-01
ZJetsToNuNu_HT-600To800_13TeV-madgraph	3.248	1.23	5.73e+06	4.19e-02
ZJetsToNuNu_HT-800To1200_13TeV-madgraph	1.496	1.23	2.06e+06	5.37e-02
ggZH_HToBB_ZToLL_M125_13TeV_powheg_pythia8/	0.01437	1.0	2.70e+04	3.20e-02
ggZH_HToBB_ZToNuNu_M125_13TeV_powheg_pythia8/	0.01437	1.0	5.62e+04	1.53e-02

Table B.3: The MC datasets used for 2018 are listed above.

Appendix C

Data Card

The tables in this appendix give the observed yields and pre-fit predictions for each channel and each year.

Table C.1: The observed and predicted yields are given for the signal regions for 0-lepton in 2016.

Process	Medium $p_T(V)$ 0J	Medium $p_T(V)$ GE1J	High $p_T(V)$	High $p_T(V)$, boosted	Highest $p_T(V)$	Highest $p_T(V)$, boosted
Data	3444	4036	1601	2035	152	423
WH 150_250 0J	7.297e+00	1.392e+00	4.894e-01	0	0	0
WH 150_250 GE1J	7.867e-01	1.892e+00	2.302e-01	3.321e-02	0	0
WH 250_400	1.419e+00	9.698e-01	3.503e+00	7.489e-01	4.932e-02	1.924e-02
WH 75_150	6.560e-01	1.731e-01	0	0	0	0
WH GT400	2.859e-02	8.905e-03	3.160e-01	1.040e-01	3.322e-01	3.394e-01
ZH 0_75	0	0	2.550e-03	0	0	0
ZH 150_250 0J	2.299e+01	2.596e+00	2.481e+00	2.669e-02	0	0
ZH 150_250 GE1J	2.809e+00	5.882e+00	1.016e+00	1.134e-01	0	0
ZH 250_400	3.374e+00	1.271e+00	1.613e+01	3.322e+00	5.961e-01	2.165e-01
ZH 75_150	1.814e+00	4.654e-01	6.357e-03	0	0	0
ZH GT400	7.561e-03	6.880e-03	1.040e+00	4.143e-01	3.322e+00	2.615e+00
ggZH 0_75	2.600e-04	0	0	0	0	0
ggZH 150_250 0J	5.120e+00	6.577e-01	4.522e-01	7.052e-03	0	0
ggZH 150_250 GE1J	2.206e+00	3.975e+00	6.435e-01	8.790e-02	0	0
ggZH 250_400	6.495e-01	4.919e-01	3.457e+00	8.091e-01	8.446e-02	4.281e-02
ggZH 75_150	5.832e-01	3.063e-01	1.357e-03	0	0	0
ggZH GT400	7.085e-04	4.001e-04	9.972e-02	6.591e-02	3.183e-01	2.834e-01
TT	1.530e+03	2.557e+03	6.268e+02	7.935e+02	1.888e+01	9.874e+01
VVHF	8.001e+01	4.534e+01	5.194e+01	1.625e+01	6.761e+00	6.037e+00
VVLF	2.920e+01	2.805e+01	2.264e+01	4.700e+01	3.771e+00	1.411e+01
Wj0b_c	8.414e+01	7.524e+01	5.032e+01	7.504e+01	4.481e+00	1.932e+01
Wj0b_udsg	1.088e+02	9.527e+01	6.117e+01	1.265e+02	6.757e+00	2.355e+01
Wj1b	7.093e+01	5.827e+01	3.482e+01	5.805e+01	3.440e+00	1.271e+01
Wj2b	1.423e+02	1.451e+02	9.751e+01	1.211e+02	1.099e+01	3.260e+01
Zj0b_c	1.314e+02	7.419e+01	7.680e+01	1.128e+02	9.821e+00	4.681e+01
Zj0b_udsg	1.027e+02	6.430e+01	6.435e+01	1.522e+02	1.174e+01	6.052e+01
Zj1b	1.607e+02	9.877e+01	7.365e+01	1.124e+02	8.575e+00	3.373e+01
Zj2b	3.791e+02	2.391e+02	1.861e+02	1.273e+02	2.537e+01	6.232e+01
s_Top	1.867e+02	1.846e+02	6.419e+01	7.596e+01	3.774e+00	1.171e+01

Table C.2: The observed and predicted yields are given for the control regions for 0-lepton in 2016.

Process	Z+udcsg enr., med. $p_T(V)$	Z+udcsg enr., high $p_T(V)$	Z+udcsg enr., high $p_T(V)$, boosted
Data	9160	2742	78659
TT	1.367e+03	3.053e+02	3.191e+03
VVHF	1.133e+01	5.163e+00	1.883e+01
VVLF	1.446e+02	6.736e+01	1.889e+03
Wj0b_c	7.463e+02	2.018e+02	2.943e+03
Wj0b_udsg	2.264e+03	5.809e+02	2.496e+04
Wj1b	9.205e+01	2.615e+01	3.510e+02
Wj2b	5.640e+01	1.715e+01	1.879e+02
Zj0b_c	7.793e+02	2.818e+02	4.405e+03
Zj0b_udsg	1.770e+03	7.512e+02	3.749e+04
Zj1b	1.575e+02	5.061e+01	6.923e+02
Zj2b	8.495e+01	3.031e+01	2.533e+02
s_Top	1.699e+02	3.858e+01	5.277e+02
Process	Z+bb enr., med. $p_T(V)$	Z+bb enr., high $p_T(V)$	Z+bb enr., high $p_T(V)$, boosted
Data	13062	2952	2567
TT	5.855e+03	1.058e+03	8.015e+02
VVHF	9.473e+01	3.227e+01	1.869e+01
VVLF	7.388e+01	2.008e+01	8.989e+01
Wj0b_c	4.188e+02	1.023e+02	1.498e+02
Wj0b_udsg	5.343e+02	1.285e+02	2.215e+02
Wj1b	3.060e+02	7.521e+01	9.754e+01
Wj2b	6.184e+02	1.710e+02	2.421e+02
Zj0b_c	4.219e+02	1.581e+02	2.285e+02
Zj0b_udsg	3.478e+02	1.481e+02	3.006e+02
Zj1b	5.748e+02	1.592e+02	2.419e+02
Zj2b	1.294e+03	3.618e+02	3.005e+02
s_Top	6.552e+02	1.414e+02	9.027e+01
Process	$t\bar{t}$ enr., med. $p_T(V)$	$t\bar{t}$ enr., high $p_T(V)$	$t\bar{t}$ enr., high $p_T(V)$, boosted
Data	3395	838	1229
TT	2.846e+03	6.922e+02	1.108e+03
VVHF	3.394e+00	1.482e+00	5.775e+00
VVLF	8.099e+00	1.330e+00	8.263e+00
Wj0b_c	3.301e+01	1.050e+01	1.648e+01
Wj0b_udsg	3.762e+01	1.403e+01	1.502e+01
Wj1b	2.378e+01	7.778e+00	1.714e+01
Wj2b	3.117e+01	1.160e+01	4.850e+01
Zj0b_c	2.130e+01	8.779e+00	1.491e+01
Zj0b_udsg	1.805e+01	8.977e+00	9.209e+00
Zj1b	2.265e+01	8.555e+00	1.992e+01
Zj2b	4.769e+01	2.146e+01	7.903e+01
s_Top	1.281e+02	3.589e+01	7.903e+01

Table C.3: The observed and predicted yields are given for the signal regions for 1-lepton (e) in 2016.

Process	Medium $p_T(V)$	High $p_T(V)$	High $p_T(V)$, boosted	Highest $p_T(V)$	Highest $p_T(V)$, boosted
Data	12006	1563	1921	129	423
WH 150_250 0J	2.357e+01	1.537e+00	1.458e-02	0	0
WH 150_250 GE1J	8.049e+00	5.998e-01	1.080e-01	0	0
WH 250_400	3.431e+00	1.148e+01	2.492e+00	4.427e-01	1.747e-01
WH 75_150	6.580e+00	1.243e-05	0	0	0
WH GT400	7.395e-03	8.824e-01	4.518e-01	2.574e+00	2.214e+00
ZH 0_75	9.896e-04	0	0	0	0
ZH 150_250 0J	6.733e-01	2.646e-02	0	0	0
ZH 150_250 GE1J	2.275e-01	1.549e-02	0	0	0
ZH 250_400	1.887e-01	1.845e-01	3.590e-02	2.067e-03	6.941e-04
ZH 75_150	2.842e-01	0	0	0	0
ZH GT400	1.640e-02	3.799e-02	2.742e-02	2.018e-02	2.326e-02
TT	9.417e+03	1.103e+03	1.440e+03	4.114e+01	2.545e+02
VVHF	6.272e+01	1.928e+01	6.898e+00	3.129e+00	2.331e+00
VVLF	4.461e+01	1.783e+01	3.013e+01	9.757e-01	1.840e+01
Wj0b_c	4.404e+02	8.025e+01	1.247e+02	1.365e+01	4.264e+01
Wj0b_udsg	3.764e+02	7.213e+01	1.116e+02	1.416e+01	4.817e+01
Wj1b	2.422e+02	5.471e+01	6.768e+01	9.447e+00	2.286e+01
Wj2b	3.780e+02	1.223e+02	1.553e+02	2.464e+01	7.352e+01
Zj0b_c	2.214e+01	1.496e+00	3.790e+00	1.217e-01	1.344e+00
Zj0b_udsg	3.634e+01	2.586e+00	5.240e+00	3.025e-01	1.067e+00
Zj1b	4.432e+01	3.186e+00	3.964e+00	1.562e-01	6.485e-01
Zj2b	4.838e+01	5.169e+00	5.912e+00	4.418e-01	1.759e+00
s_Top	1.373e+03	2.444e+02	2.081e+02	1.871e+01	7.334e+01

Table C.4: The observed and predicted yields are given for the control regions for 1-lepton (e) in 2016.

Process	W+udcsg enr., med. $p_T(V)$	W+udcsg enr., high $p_T(V)$	W+udcsg enr., high $p_T(V)$, boosted
Data	105027	24155	39175
TT	1.924e+04	4.547e+03	5.435e+03
VVHF	2.438e+01	8.982e+00	1.229e+01
VVLF	2.289e+03	9.025e+02	1.819e+03
Wj0b_c	1.579e+04	3.539e+03	4.458e+03
Wj0b_udsg	5.239e+04	1.313e+04	2.742e+04
Wj1b	1.134e+03	3.139e+02	4.068e+02
Wj2b	3.572e+02	1.527e+02	2.367e+02
Zj0b_c	5.605e+02	7.261e+01	1.302e+02
Zj0b_udsg	3.489e+03	4.150e+02	1.050e+03
Zj1b	1.292e+02	1.385e+01	2.376e+01
Zj2b	2.497e+01	4.251e+00	8.959e+00
s_Top	3.104e+03	8.494e+02	1.057e+03
Process	W+bb enr., med. $p_T(V)$	W+bb enr., high $p_T(V)$	W+bb enr., high $p_T(V)$, boosted
Data	20605	3470	2024
TT	1.545e+04	2.473e+03	1.547e+03
VVHF	5.574e+01	1.951e+01	6.503e+00
VVLF	5.773e+01	2.360e+01	8.361e+01
Wj0b_c	8.084e+02	1.778e+02	2.028e+02
Wj0b_udsg	7.505e+02	1.958e+02	1.985e+02
Wj1b	5.339e+02	1.292e+02	1.141e+02
Wj2b	8.627e+02	2.425e+02	3.134e+02
Zj0b_c	4.896e+01	5.681e+00	7.085e+00
Zj0b_udsg	6.818e+01	6.542e+00	9.776e+00
Zj1b	7.259e+01	6.353e+00	6.781e+00
Zj2b	7.790e+01	7.958e+00	1.088e+01
s_Top	2.190e+03	5.302e+02	2.807e+02
Process	$t\bar{t}$ enr., med. $p_T(V)$	$t\bar{t}$ enr., high $p_T(V)$	$t\bar{t}$ enr., high $p_T(V)$, boosted
Data	38193	9459	3450
TT	3.708e+04	8.817e+03	4.079e+03
VVHF	6.869e+01	4.074e+01	4.040e+00
VVLF	5.154e+01	3.035e+01	1.061e+01
Wj0b_c	2.374e+02	9.584e+01	1.743e+01
Wj0b_udsg	2.763e+02	9.764e+01	6.360e+00
Wj1b	6.486e+02	2.500e+02	1.625e+01
Wj2b	8.032e+02	3.519e+02	5.856e+01
Zj0b_c	1.436e+01	3.906e+00	1.312e+00
Zj0b_udsg	2.280e+01	4.638e+00	9.908e-01
Zj1b	7.477e+01	1.424e+01	9.796e-01
Zj2b	7.816e+01	1.694e+01	3.997e+00
s_Top	2.540e+03	1.054e+03	3.411e+02

Table C.5: The observed and predicted yields are given for the signal regions for 1-lepton (μ) in 2016.

Process	Medium $p_T(V)$	High $p_T(V)$	High $p_T(V)$, boosted	Highest $p_T(V)$	Highest $p_T(V)$, boosted
Data	16249	2127	2463	154	581
WH 0_75	1.951e-02	0	0	0	0
WH 150_250 0J	2.983e+01	1.981e+00	5.517e-03	0	0
WH 150_250 GE1J	1.037e+01	8.396e-01	1.194e-01	0	0
WH 250_400	4.128e+00	1.381e+01	2.839e+00	5.374e-01	2.125e-01
WH 75_150	8.437e+00	2.732e-04	0	0	0
WH GT400	9.994e-03	8.083e-01	4.759e-01	2.982e+00	2.604e+00
ZH 150_250 0J	9.829e-01	4.906e-02	8.978e-04	0	0
ZH 150_250 GE1J	3.053e-01	1.697e-02	0	0	0
ZH 250_400	1.964e-01	2.314e-01	6.373e-02	4.133e-03	1.195e-03
ZH 75_150	4.484e-01	1.194e-03	0	0	0
ZH GT400	1.158e-02	2.205e-02	2.343e-02	2.572e-02	2.038e-02
TT	1.227e+04	1.362e+03	1.762e+03	5.458e+01	2.972e+02
VVHF	8.050e+01	2.528e+01	5.329e+00	3.062e+00	2.303e+00
VVLF	6.213e+01	1.977e+01	3.065e+01	7.991e-01	2.044e+01
Wj0b_c	5.133e+02	1.044e+02	1.527e+02	1.645e+01	7.126e+01
Wj0b_udsg	5.218e+02	1.039e+02	1.568e+02	2.600e+01	6.334e+01
Wj1b	3.279e+02	6.970e+01	8.666e+01	1.097e+01	2.953e+01
Wj2b	4.827e+02	1.604e+02	1.817e+02	3.198e+01	9.275e+01
Zj0b_c	2.968e+01	3.868e+00	5.701e+00	4.552e-01	1.209e+00
Zj0b_udsg	4.615e+01	3.884e+00	1.021e+01	5.188e-01	1.400e+00
Zj1b	5.880e+01	5.032e+00	5.481e+00	3.946e-01	8.907e-01
Zj2b	6.992e+01	6.915e+00	7.405e+00	6.539e-01	1.927e+00
s_Top	1.724e+03	2.801e+02	2.417e+02	2.087e+01	1.099e+02

Table C.6: The observed and predicted yields are given for the control regions for 1-lepton (μ) in 2016.

Process	W+udcsg enr., med. $p_T(V)$	W+udcsg enr., high $p_T(V)$	W+udcsg enr., high $p_T(V)$, boosted
Data	142862	32273	47879
TT	2.434e+04	5.758e+03	6.666e+03
VVHF	3.255e+01	1.459e+01	1.761e+01
VVLF	3.064e+03	1.098e+03	2.213e+03
Wj0b_c	2.029e+04	4.513e+03	5.586e+03
Wj0b_udsg	7.003e+04	1.729e+04	3.488e+04
Wj1b	1.483e+03	4.007e+02	5.024e+02
Wj2b	4.767e+02	1.902e+02	3.026e+02
Zj0b_c	8.355e+02	1.060e+02	1.540e+02
Zj0b_udsg	5.102e+03	6.679e+02	1.386e+03
Zj1b	1.789e+02	2.034e+01	2.862e+01
Zj2b	3.429e+01	6.691e+00	1.223e+01
s_Top	3.971e+03	1.084e+03	1.319e+03
Process	W+bb enr., med. $p_T(V)$	W+bb enr., high $p_T(V)$	W+bb enr., high $p_T(V)$, boosted
Data	28053	4545	2548
TT	1.985e+04	3.019e+03	1.848e+03
VVHF	7.078e+01	1.932e+01	8.710e+00
VVLF	1.010e+02	2.428e+01	1.191e+02
Wj0b_c	1.117e+03	2.483e+02	2.620e+02
Wj0b_udsg	1.069e+03	2.638e+02	2.446e+02
Wj1b	7.172e+02	1.623e+02	1.406e+02
Wj2b	1.082e+03	2.992e+02	3.850e+02
Zj0b_c	7.770e+01	6.280e+00	9.266e+00
Zj0b_udsg	9.657e+01	1.436e+01	2.218e+01
Zj1b	9.800e+01	8.921e+00	9.081e+00
Zj2b	1.165e+02	1.253e+01	1.402e+01
s_Top	2.764e+03	6.598e+02	3.339e+02
Process	$t\bar{t}$ enr., med. $p_T(V)$	$t\bar{t}$ enr., high $p_T(V)$	$t\bar{t}$ enr., high $p_T(V)$, boosted
Data	50788	12017	4244
TT	4.647e+04	1.103e+04	5.029e+03
VVHF	8.711e+01	4.722e+01	4.302e+00
VVLF	6.077e+01	2.759e+01	9.766e+00
Wj0b_c	3.250e+02	1.275e+02	1.982e+01
Wj0b_udsg	3.234e+02	1.365e+02	9.878e+00
Wj1b	8.308e+02	3.153e+02	2.039e+01
Wj2b	1.030e+03	4.480e+02	7.608e+01
Zj0b_c	2.010e+01	4.969e+00	9.419e-01
Zj0b_udsg	2.642e+01	7.535e+00	1.095e+00
Zj1b	9.908e+01	1.860e+01	1.435e+00
Zj2b	1.073e+02	2.229e+01	5.295e+00
s_Top	3.127e+03	1.264e+03	3.906e+02

Table C.7: The observed and predicted yields are given for the signal regions for 2-lepton (e) in 2016.

Process	Low $p_T(V)$	Medium $p_T(V)$ 0J	Medium $p_T(V)$ GE1J	High $p_T(V)$	High $p_T(V)$, boosted	Highest $p_T(V)$	Highest $p_T(V)$, boosted
Data	2478	234	292	108	52	12	18
ZH 0_75	9.538e-02	0	0	0	0	0	0
ZH 150_250 0J	1.721e-01	4.048e+00	6.058e-01	2.573e-02	1.095e-03	0	0
ZH 150_250 GE1J	7.791e-02	5.432e-01	1.733e+00	9.020e-03	3.283e-03	0	0
ZH 250_400	0	6.703e-02	3.716e-02	2.274e+00	3.555e-01	7.597e-03	4.489e-03
ZH 75_150	1.312e+01	5.066e-02	2.020e-02	0	0	0	0
ZH GT400	0	0	0	2.072e-02	2.095e-03	4.797e-01	2.361e-01
ggZH 0_75	8.620e-03	0	0	0	0	0	0
ggZH 150_250 0J	5.764e-02	1.150e+00	1.833e-01	3.648e-03	0	0	0
ggZH 150_250 GE1J	5.569e-02	5.661e-01	1.371e+00	8.899e-03	5.241e-04	0	0
ggZH 250_400	0	1.471e-02	1.984e-02	6.971e-01	1.250e-01	1.338e-03	5.944e-04
ggZH 75_150	3.695e+00	2.109e-02	1.220e-02	0	0	0	0
ggZH GT400	0	0	0	4.722e-03	2.156e-03	6.754e-02	3.801e-02
TT	5.685e+02	1.746e+01	2.116e+01	2.151e+00	6.171e-01	0	0
VVHF	3.261e+01	7.761e+00	6.748e+00	5.616e+00	1.486e+00	7.302e-01	4.118e-01
VVLF	1.509e+01	2.074e+00	1.875e+00	1.238e+00	2.246e+00	5.842e-02	8.934e-01
Zj0b_c	2.221e+02	2.743e+01	3.311e+01	1.333e+01	8.462e+00	2.051e+00	3.351e+00
Zj0b_udsg	3.374e+02	1.902e+01	4.046e+01	1.639e+01	1.204e+01	2.821e+00	4.843e+00
Zj1b	6.742e+02	5.491e+01	5.389e+01	1.795e+01	1.217e+01	2.492e+00	3.251e+00
Zj2b	5.994e+02	6.390e+01	6.629e+01	2.919e+01	1.791e+01	4.618e+00	8.243e+00
s_Top	2.582e+01	7.407e-01	1.750e+00	3.764e-01	0	0	0

Table C.8: The observed and predicted yields are given for the control regions for 2-lepton (e) in 2016.

Process	Z+udcsg enr., low $p_T(V)$	Z+udcsg enr., med. $p_T(V)$	Z+udcsg enr., high $p_T(V)$	Z+udcsg enr., high $p_T(V)$, boosted
Data	32439	6944	1325	843
TT	6.175e+01	5.378e+00	0	4.557e-01
VVHF	1.472e+00	5.498e-01	2.765e-01	1.024e+00
VVLF	2.073e+02	6.213e+01	1.521e+01	2.033e+01
Zj0b_c	1.764e+03	3.333e+02	8.301e+01	1.220e+02
Zj0b_udsg	3.223e+04	6.009e+03	1.211e+03	7.673e+02
Zj1b	3.540e+02	6.422e+01	1.221e+01	3.903e+01
Zj2b	2.945e+01	6.816e+00	1.885e+00	2.033e+01
s_Top	7.798e+00	1.109e+00	0	2.329e-01
Process	Z+bb enr., low $p_T(V)$	Z+bb enr., med. $p_T(V)$	Z+bb enr., high $p_T(V)$	Z+bb enr., high $p_T(V)$, boosted
Data	3001	795	171	77
TT	3.387e+02	1.956e+01	7.755e-01	8.541e-01
VVHF	3.179e+01	1.263e+01	3.497e+00	1.382e+00
VVLF	1.355e+01	4.871e+00	1.454e+00	3.168e+00
Zj0b_c	3.448e+02	9.059e+01	2.615e+01	1.928e+01
Zj0b_udsg	4.584e+02	1.173e+02	3.146e+01	2.205e+01
Zj1b	9.570e+02	1.822e+02	3.439e+01	2.195e+01
Zj2b	8.844e+02	2.292e+02	5.217e+01	3.521e+01
s_Top	1.531e+01	2.523e+00	5.006e-01	3.363e-01
Process	$t\bar{t}$ enr., low $p_T(V)$	$t\bar{t}$ enr., med. $p_T(V)$	$t\bar{t}$ enr., high $p_T(V)$	$t\bar{t}$ enr., high $p_T(V)$, boosted
Data	9768	1011	84	42
TT	9.714e+03	8.896e+02	7.875e+01	1.396e+01
VVHF	7.954e+00	3.593e+00	1.106e+00	4.372e-01
VVLF	3.282e+00	9.188e-01	2.343e-01	1.889e+00
Zj0b_c	1.847e+01	4.388e+00	1.153e+00	5.458e+00
Zj0b_udsg	3.727e+01	4.997e+00	1.585e+00	5.493e+00
Zj1b	1.154e+02	2.219e+01	4.946e+00	4.049e+00
Zj2b	1.392e+02	3.314e+01	9.068e+00	7.211e+00
s_Top	3.775e+02	7.618e+01	1.605e+01	9.051e+00

Table C.9: The observed and predicted yields are given for the signal regions for 2-lepton (μ) in 2016.

Process	Low $p_T(V)$	Medium $p_T(V)$ 0J	Medium $p_T(V)$ GE1J	High $p_T(V)$	High $p_T(V)$, boosted	Highest $p_T(V)$	Highest $p_T(V)$, boosted
Data	3842	332	374	135	73	12	36
ZH 0_75	1.443e-01	0	0	0	0	0	0
ZH 150_250 0J	2.002e-01	5.410e+00	7.628e-01	5.285e-02	1.260e-04	0	0
ZH 150_250 GE1J	1.102e-01	7.763e-01	2.293e+00	4.025e-02	3.512e-03	0	0
ZH 250_400	0	1.012e-01	5.866e-02	3.050e+00	4.853e-01	2.629e-02	7.796e-03
ZH 75_150	1.872e+01	1.355e-01	5.663e-02	0	0	0	0
ZH GT400	0	0	0	3.975e-02	1.540e-02	6.683e-01	4.154e-01
ggZH 0_75	1.512e-02	0	0	0	0	0	0
ggZH 150_250 0J	7.090e-02	1.566e+00	2.368e-01	1.190e-02	0	0	0
ggZH 150_250 GE1J	7.971e-02	7.678e-01	1.855e+00	2.708e-02	2.043e-03	0	0
ggZH 250_400	0	2.628e-02	3.569e-02	9.327e-01	1.640e-01	4.346e-03	1.996e-03
ggZH 75_150	4.987e+00	4.276e-02	3.347e-02	0	0	0	0
ggZH GT400	0	0	0	5.215e-03	3.877e-03	9.077e-02	5.374e-02
TT	8.450e+02	1.217e+01	2.144e+01	1.735e+00	3.223e+00	0	0
VVHF	4.729e+01	9.589e+00	9.398e+00	4.835e+00	1.620e+00	6.426e-01	6.430e-01
VVLF	2.178e+01	2.814e+00	4.348e+00	1.211e+00	3.049e+00	4.137e-01	1.680e+00
Zj0b_c	3.280e+02	3.884e+01	4.729e+01	1.787e+01	1.355e+01	2.589e+00	4.890e+00
Zj0b_udsg	4.631e+02	3.885e+01	7.390e+01	2.093e+01	1.528e+01	4.095e+00	6.173e+00
Zj1b	9.827e+02	7.556e+01	8.228e+01	2.659e+01	1.734e+01	3.787e+00	5.521e+00
Zj2b	8.415e+02	8.274e+01	9.380e+01	3.872e+01	2.588e+01	7.027e+00	1.208e+01
s_Top	2.933e+01	3.763e+00	4.099e+00	1.484e+00	3.822e-01	0	4.486e-01

Table C.10: The observed and predicted yields are given for the control regions for 2-lepton (μ) in 2016.

Process	Z+udcsg enr., low $p_T(V)$	Z+udcsg enr., med. $p_T(V)$	Z+udcsg enr., high $p_T(V)$	Z+udcsg enr., high $p_T(V)$, boosted
Data	48206	9545	1719	1163
TT	1.027e+02	1.326e+01	9.527e-01	6.617e+00
VVHF	2.069e+00	6.008e-01	1.863e-01	1.635e+00
VVLF	2.793e+02	8.190e+01	2.494e+01	2.857e+01
Zj0b_c	2.344e+03	4.663e+02	1.008e+02	1.809e+02
Zj0b_udsg	4.372e+04	8.170e+03	1.662e+03	1.078e+03
Zj1b	4.638e+02	8.890e+01	1.596e+01	5.479e+01
Zj2b	3.142e+01	8.627e+00	2.572e+00	2.852e+01
s_Top	1.009e+01	1.182e+00	2.228e-01	5.445e-01
Process	Z+bb enr., low $p_T(V)$	Z+bb enr., med. $p_T(V)$	Z+bb enr., high $p_T(V)$	Z+bb enr., high $p_T(V)$, boosted
Data	4737	1125	234	128
TT	5.063e+02	2.992e+01	5.177e-01	1.444e+00
VVHF	4.581e+01	1.710e+01	4.736e+00	2.115e+00
VVLF	1.905e+01	6.396e+00	2.277e+00	3.906e+00
Zj0b_c	4.673e+02	1.288e+02	3.411e+01	2.739e+01
Zj0b_udsg	6.792e+02	1.690e+02	4.515e+01	3.105e+01
Zj1b	1.315e+03	2.466e+02	4.682e+01	3.048e+01
Zj2b	1.227e+03	3.127e+02	6.924e+01	5.268e+01
s_Top	2.254e+01	4.532e+00	3.693e-01	4.003e-01
Process	$t\bar{t}$ enr., low $p_T(V)$	$t\bar{t}$ enr., med. $p_T(V)$	$t\bar{t}$ enr., high $p_T(V)$	$t\bar{t}$ enr., high $p_T(V)$, boosted
Data	15396	1485	154	84
TT	1.362e+04	1.260e+03	1.138e+02	2.973e+01
VVHF	1.269e+01	4.208e+00	1.979e+00	4.603e-01
VVLF	4.902e+00	1.263e+00	1.123e-01	3.373e+00
Zj0b_c	4.336e+01	8.082e+00	2.100e+00	9.983e+00
Zj0b_udsg	4.271e+01	9.574e+00	4.269e+00	1.572e+01
Zj1b	2.004e+02	3.670e+01	8.129e+00	7.660e+00
Zj2b	2.276e+02	6.146e+01	1.810e+01	1.529e+01
s_Top	4.979e+02	1.125e+02	1.752e+01	1.490e+01

Table C.11: The observed and predicted yields are given for the signal regions for 0-lepton in 2017.

Process	Medium $p_T(V)$ 0J	Medium $p_T(V)$ GE1J	High $p_T(V)$	High $p_T(V)$, boosted	Highest $p_T(V)$	Highest $p_T(V)$, boosted
Data	3517	3942	1787	2178	132	493
WH 150_250 0J	6.525e+00	1.151e+00	4.862e-01	1.102e-02	0	0
WH 150_250 GE1J	9.167e-01	1.757e+00	2.232e-01	1.422e-02	0	0
WH 250_400	1.938e+00	1.079e+00	3.604e+00	6.879e-01	4.006e-02	2.263e-02
WH 75_150	6.067e-01	1.454e-01	3.344e-03	0	0	0
WH GT400	4.012e-02	1.552e-02	4.324e-01	2.015e-01	4.269e-01	2.856e-01
ZH 0_75	0	0	7.024e-05	0	0	0
ZH 150_250 0J	2.230e+01	2.264e+00	2.374e+00	1.071e-02	0	0
ZH 150_250 GE1J	3.041e+00	5.752e+00	1.074e+00	7.932e-02	0	0
ZH 250_400	5.077e+00	1.758e+00	1.759e+01	3.106e+00	5.365e-01	1.868e-01
ZH 75_150	1.617e+00	4.150e-01	7.042e-03	0	0	0
ZH GT400	1.911e-02	0	1.375e+00	5.896e-01	3.591e+00	2.809e+00
ggZH 0_75	4.031e-04	0	0	0	0	0
ggZH 150_250 0J	6.024e+00	6.618e-01	5.750e-01	4.992e-03	0	0
ggZH 150_250 GE1J	2.357e+00	4.031e+00	7.207e-01	8.634e-02	1.669e-06	0
ggZH 250_400	1.081e+00	7.692e-01	4.272e+00	8.838e-01	1.023e-01	4.760e-02
ggZH 75_150	6.208e-01	2.435e-01	3.867e-03	4.688e-04	0	0
ggZH GT400	1.303e-03	3.375e-04	1.871e-01	9.133e-02	3.556e-01	2.588e-01
TT	1.572e+03	2.401e+03	7.255e+02	7.945e+02	1.605e+01	8.150e+01
VVHF	9.968e+01	4.924e+01	7.134e+01	2.819e+01	8.717e+00	1.244e+01
VVLF	2.985e+01	2.112e+01	2.473e+01	4.020e+01	1.594e+00	9.143e+00
Wj0b_c	8.633e+01	9.060e+01	6.001e+01	9.490e+01	2.909e+00	1.726e+01
Wj0b_udsg	9.339e+01	7.690e+01	4.821e+01	9.944e+01	3.872e+00	1.943e+01
Wj1b	5.688e+01	5.159e+01	3.295e+01	4.947e+01	2.060e+00	9.091e+00
Wj2b	1.290e+02	1.201e+02	9.334e+01	9.753e+01	1.103e+01	2.187e+01
Zj0b_c	1.475e+02	9.531e+01	8.939e+01	1.253e+02	1.012e+01	4.077e+01
Zj0b_udsg	7.939e+01	6.434e+01	6.476e+01	1.336e+02	8.547e+00	4.703e+01
Zj1b	1.841e+02	9.792e+01	9.285e+01	1.278e+02	8.445e+00	3.107e+01
Zj2b	3.857e+02	2.463e+02	2.301e+02	1.556e+02	2.784e+01	6.055e+01
s_Top	1.840e+02	2.015e+02	7.780e+01	6.451e+01	1.952e+00	1.365e+01

Table C.12: The observed and predicted yields are given for the control regions for 0-lepton in 2017.

Process	Z+udcsg enr., med. $p_T(V)$	Z+udcsg enr., high $p_T(V)$	Z+udcsg enr., high $p_T(V)$, boosted
Data	9341	3286	79789
TT	1.415e+03	3.452e+02	3.154e+03
VVHF	1.691e+01	6.540e+00	3.202e+01
VVLF	1.556e+02	5.183e+01	1.690e+03
Wj0b_c	8.973e+02	2.041e+02	2.596e+03
Wj0b_udsg	2.231e+03	6.510e+02	2.380e+04
Wj1b	7.235e+01	2.001e+01	2.239e+02
Wj2b	4.304e+01	1.673e+01	1.101e+02
Zj0b_c	9.979e+02	3.476e+02	3.796e+03
Zj0b_udsg	1.812e+03	8.137e+02	3.586e+04
Zj1b	1.876e+02	6.596e+01	6.781e+02
Zj2b	9.290e+01	3.952e+01	2.117e+02
s_Top	1.801e+02	4.438e+01	5.380e+02
Process	Z+bb enr., med. $p_T(V)$	Z+bb enr., high $p_T(V)$	Z+bb enr., high $p_T(V)$, boosted
Data	13149	3508	2970
TT	5.779e+03	1.190e+03	8.650e+02
VVHF	1.100e+02	4.172e+01	2.812e+01
VVLF	4.393e+01	8.120e+00	4.627e+01
Wj0b_c	4.654e+02	1.100e+02	1.293e+02
Wj0b_udsg	4.761e+02	1.407e+02	1.832e+02
Wj1b	2.488e+02	6.411e+01	8.236e+01
Wj2b	5.602e+02	1.627e+02	1.927e+02
Zj0b_c	5.447e+02	1.899e+02	2.037e+02
Zj0b_udsg	3.570e+02	1.532e+02	2.578e+02
Zj1b	6.298e+02	1.963e+02	2.632e+02
Zj2b	1.360e+03	4.538e+02	3.382e+02
s_Top	6.969e+02	1.692e+02	9.035e+01
Process	$t\bar{t}$ enr., med. $p_T(V)$	$t\bar{t}$ enr., high $p_T(V)$	$t\bar{t}$ enr., high $p_T(V)$, boosted
Data	3250	915	2250
TT	2.739e+03	6.866e+02	1.947e+03
VVHF	3.258e+00	1.133e+00	6.451e+00
VVLF	2.107e+00	8.234e-01	3.545e+00
Wj0b_c	3.273e+01	1.077e+01	1.672e+01
Wj0b_udsg	3.426e+01	9.288e+00	8.567e+00
Wj1b	1.597e+01	4.720e+00	1.332e+01
Wj2b	2.468e+01	9.192e+00	3.787e+01
Zj0b_c	1.942e+01	9.638e+00	1.380e+01
Zj0b_udsg	1.384e+01	8.046e+00	9.211e+00
Zj1b	1.988e+01	9.141e+00	2.597e+01
Zj2b	4.638e+01	2.100e+01	9.221e+01
s_Top	1.152e+02	3.578e+01	1.218e+02

Table C.13: The observed and predicted yields are given for the signal regions for 1-lepton (e) in 2017.

Process	Medium $p_T(V)$	High $p_T(V)$	High $p_T(V)$, boosted	Highest $p_T(V)$	Highest $p_T(V)$, boosted
Data	13949	1826	2127	106	537
WH 0_75	1.418e-02	0	0	0	0
WH 150_250 0J	2.659e+01	1.625e+00	6.539e-03	0	0
WH 150_250 GE1J	9.725e+00	7.076e-01	1.236e-01	0	0
WH 250_400	5.286e+00	1.265e+01	2.364e+00	4.084e-01	1.532e-01
WH 75_150	7.872e+00	1.541e-03	2.247e-03	0	0
WH GT400	1.401e-02	1.291e+00	6.061e-01	2.990e+00	2.533e+00
ZH 0_75	5.364e-03	0	0	0	0
ZH 150_250 0J	6.486e-01	2.807e-02	4.764e-04	0	0
ZH 150_250 GE1J	2.260e-01	1.036e-02	1.141e-03	0	0
ZH 250_400	2.498e-01	1.672e-01	5.078e-02	2.242e-03	0
ZH 75_150	3.428e-01	3.799e-05	0	0	0
ZH GT400	8.527e-03	3.945e-02	2.457e-02	1.977e-02	2.104e-02
TT	1.158e+04	1.279e+03	1.696e+03	5.609e+01	2.835e+02
VVHF	8.274e+01	2.100e+01	1.055e+01	3.531e+00	2.584e+00
VVLF	5.138e+01	1.148e+01	2.908e+01	6.610e-01	1.252e+01
Wj0b_c	6.022e+02	9.819e+01	1.315e+02	1.500e+01	4.313e+01
Wj0b_udsg	4.274e+02	6.417e+01	1.075e+02	1.461e+01	3.442e+01
Wj1b	2.768e+02	5.659e+01	6.607e+01	6.229e+00	1.759e+01
Wj2b	4.191e+02	1.274e+02	1.445e+02	2.156e+01	5.506e+01
Zj0b_c	3.495e+01	2.973e+00	4.878e+00	3.237e-01	9.950e-01
Zj0b_udsg	3.561e+01	4.847e+00	4.773e+00	2.669e-02	9.319e-01
Zj1b	5.321e+01	2.422e+00	3.704e+00	1.224e-01	5.530e-01
Zj2b	5.692e+01	6.474e+00	5.370e+00	3.893e-01	1.332e+00
s_Top	1.632e+03	2.550e+02	2.426e+02	1.657e+01	8.558e+01

Table C.14: The observed and predicted yields are given for the control regions for 1-lepton (e) in 2017.

Process	W+udcsg enr., med. $p_T(V)$	W+udcsg enr., high $p_T(V)$	W+udcsg enr., high $p_T(V)$, boosted
Data	127417	27731	40956
TT	2.313e+04	5.286e+03	5.695e+03
VVHF	5.414e+01	1.822e+01	1.436e+01
VVLF	2.191e+03	7.324e+02	1.534e+03
Wj0b_c	2.650e+04	4.872e+03	4.599e+03
Wj0b_udsg	6.696e+04	1.498e+04	2.871e+04
Wj1b	1.181e+03	2.810e+02	2.661e+02
Wj2b	3.486e+02	1.339e+02	1.435e+02
Zj0b_c	8.695e+02	9.443e+01	1.049e+02
Zj0b_udsg	4.717e+03	5.236e+02	1.013e+03
Zj1b	1.762e+02	1.670e+01	1.799e+01
Zj2b	2.761e+01	4.001e+00	5.696e+00
s_Top	3.732e+03	9.610e+02	1.049e+03
Process	W+bb enr., med. $p_T(V)$	W+bb enr., high $p_T(V)$	W+bb enr., high $p_T(V)$, boosted
Data	24941	4103	2345
TT	1.943e+04	2.917e+03	1.709e+03
VVHF	7.643e+01	2.173e+01	1.162e+01
VVLF	8.407e+01	1.519e+01	6.308e+01
Wj0b_c	1.244e+03	2.301e+02	1.870e+02
Wj0b_udsg	1.027e+03	1.881e+02	1.798e+02
Wj1b	6.172e+02	1.174e+02	1.002e+02
Wj2b	9.861e+02	2.475e+02	2.542e+02
Zj0b_c	7.991e+01	7.877e+00	6.622e+00
Zj0b_udsg	7.274e+01	1.086e+01	9.457e+00
Zj1b	1.011e+02	8.157e+00	7.164e+00
Zj2b	1.112e+02	1.017e+01	8.770e+00
s_Top	2.716e+03	5.945e+02	2.660e+02
Process	$t\bar{t}$ enr., med. $p_T(V)$	$t\bar{t}$ enr., high $p_T(V)$	$t\bar{t}$ enr., high $p_T(V)$, boosted
Data	45244	10894	4127
TT	4.431e+04	1.039e+04	4.966e+03
VVHF	6.674e+01	3.137e+01	3.626e+00
VVLF	4.447e+01	1.663e+01	4.968e+00
Wj0b_c	3.004e+02	8.725e+01	1.309e+01
Wj0b_udsg	2.719e+02	9.144e+01	7.953e+00
Wj1b	5.947e+02	1.945e+02	1.111e+01
Wj2b	7.838e+02	3.012e+02	4.308e+01
Zj0b_c	1.654e+01	3.480e+00	7.343e-01
Zj0b_udsg	2.267e+01	1.489e+01	2.604e-01
Zj1b	7.937e+01	1.313e+01	9.288e-01
Zj2b	8.508e+01	1.514e+01	2.818e+00
s_Top	2.883e+03	1.060e+03	3.409e+02

Table C.15: The observed and predicted yields are given for the signal regions for 1-lepton (μ) in 2017.

Process	Medium $p_T(V)$	High $p_T(V)$	High $p_T(V)$, boosted	Highest $p_T(V)$	Highest $p_T(V)$, boosted
Data	18798	2464	2732	162	665
WH 0_75	3.694e-02	0	0	0	0
WH 150_250 0J	3.311e+01	2.020e+00	1.831e-02	0	0
WH 150_250 GE1J	1.134e+01	8.578e-01	9.079e-02	0	0
WH 250_400	6.230e+00	1.534e+01	2.909e+00	5.658e-01	1.857e-01
WH 75_150	1.014e+01	1.005e-02	0	0	0
WH GT400	1.061e-02	1.434e+00	5.902e-01	3.441e+00	2.965e+00
ZH 0_75	6.183e-03	0	0	0	0
ZH 150_250 0J	1.040e+00	4.830e-02	0	0	0
ZH 150_250 GE1J	3.461e-01	1.816e-02	1.883e-03	0	0
ZH 250_400	2.535e-01	2.538e-01	5.431e-02	4.154e-03	3.046e-03
ZH 75_150	5.166e-01	9.071e-04	0	0	0
ZH GT400	1.602e-02	3.079e-02	1.761e-02	1.983e-02	1.081e-02
TT	1.439e+04	1.584e+03	2.020e+03	5.080e+01	3.335e+02
VVHF	9.825e+01	2.551e+01	1.165e+01	3.004e+00	5.589e+00
VVLF	6.873e+01	1.513e+01	5.244e+01	6.777e+00	2.489e+01
Wj0b_c	8.628e+02	1.471e+02	1.632e+02	1.006e+01	5.310e+01
Wj0b_udsg	5.355e+02	1.099e+02	1.605e+02	7.281e+00	5.388e+01
Wj1b	3.299e+02	6.156e+01	8.005e+01	8.450e+00	1.986e+01
Wj2b	5.197e+02	1.652e+02	1.739e+02	2.595e+01	6.220e+01
Zj0b_c	4.129e+01	1.208e+00	5.689e+00	1.494e-01	1.688e+00
Zj0b_udsg	4.090e+01	6.197e+00	8.876e+00	8.134e-01	1.861e+00
Zj1b	6.614e+01	4.489e+00	5.209e+00	1.710e-01	8.871e-01
Zj2b	8.463e+01	8.702e+00	6.330e+00	8.084e-01	1.107e+00
s_Top	2.062e+03	3.213e+02	3.128e+02	2.509e+01	1.070e+02

Table C.16: The observed and predicted yields are given for the control regions for 1-lepton (μ) in 2017.

Process	W+udcsg enr., med. $p_T(V)$	W+udcsg enr., high $p_T(V)$	W+udcsg enr., high $p_T(V)$, boosted
Data	170543	36717	50184
TT	2.830e+04	6.347e+03	6.625e+03
VVHF	6.997e+01	2.199e+01	2.028e+01
VVLF	2.882e+03	9.800e+02	1.889e+03
Wj0b_c	3.302e+04	6.040e+03	5.453e+03
Wj0b_udsg	8.513e+04	1.893e+04	3.462e+04
Wj1b	1.518e+03	3.494e+02	3.185e+02
Wj2b	4.375e+02	1.646e+02	1.780e+02
Zj0b_c	1.159e+03	1.339e+02	1.344e+02
Zj0b_udsg	6.451e+03	7.281e+02	1.344e+03
Zj1b	2.369e+02	2.451e+01	2.396e+01
Zj2b	4.248e+01	7.277e+00	7.447e+00
s_Top	4.666e+03	1.153e+03	1.239e+03
Process	W+bb enr., med. $p_T(V)$	W+bb enr., high $p_T(V)$	W+bb enr., high $p_T(V)$, boosted
Data	34502	5370	2934
TT	2.404e+04	3.449e+03	1.916e+03
VVHF	9.896e+01	2.161e+01	1.411e+01
VVLF	1.123e+02	2.232e+01	6.996e+01
Wj0b_c	1.793e+03	2.914e+02	2.609e+02
Wj0b_udsg	1.136e+03	2.547e+02	2.299e+02
Wj1b	7.846e+02	1.467e+02	1.184e+02
Wj2b	1.255e+03	3.051e+02	3.072e+02
Zj0b_c	8.734e+01	7.739e+00	9.965e+00
Zj0b_udsg	1.137e+02	1.615e+01	1.455e+01
Zj1b	1.433e+02	1.006e+01	9.770e+00
Zj2b	1.633e+02	1.690e+01	1.108e+01
s_Top	3.424e+03	7.287e+02	3.197e+02
Process	$t\bar{t}$ enr., med. $p_T(V)$	$t\bar{t}$ enr., high $p_T(V)$	$t\bar{t}$ enr., high $p_T(V)$, boosted
Data	60000	13808	5239
TT	5.412e+04	1.242e+04	5.786e+03
VVHF	8.366e+01	4.016e+01	3.656e+00
VVLF	6.860e+01	3.587e+01	2.761e+00
Wj0b_c	3.763e+02	1.110e+02	1.734e+01
Wj0b_udsg	3.443e+02	1.145e+02	1.058e+01
Wj1b	7.605e+02	2.414e+02	1.520e+01
Wj2b	9.595e+02	3.615e+02	5.273e+01
Zj0b_c	2.195e+01	3.310e+00	6.956e-01
Zj0b_udsg	2.987e+01	6.385e+00	7.106e-01
Zj1b	9.613e+01	1.569e+01	1.038e+00
Zj2b	1.114e+02	2.016e+01	4.148e+00
s_Top	3.508e+03	1.313e+03	4.174e+02

Table C.17: The observed and predicted yields are given for the signal regions for 2-lepton (e) in 2017.

Process	Low $p_T(V)$	Medium $p_T(V)$ 0J	Medium $p_T(V)$ GE1J	High $p_T(V)$	High $p_T(V)$, boosted	Highest $p_T(V)$	Highest $p_T(V)$, boosted
Data	2761	273	327	130	78	13	29
ZH 0_75	1.750e-01	0	0	0	0	0	0
ZH 150_250 0J	1.977e-01	5.088e+00	6.433e-01	3.382e-02	0	0	0
ZH 150_250 GE1J	7.274e-02	7.305e-01	2.136e+00	2.181e-02	1.854e-03	0	0
ZH 250_400	0	6.883e-02	3.994e-02	2.878e+00	3.603e-01	1.202e-02	5.480e-03
ZH 75_150	1.667e+01	1.189e-01	2.853e-02	0	0	0	0
ZH GT400	0	0	0	2.573e-02	1.298e-02	6.193e-01	2.723e-01
ggZH 0_75	1.366e-02	0	0	0	0	0	0
ggZH 150_250 0J	8.057e-02	1.792e+00	2.377e-01	9.407e-03	0	0	0
ggZH 150_250 GE1J	7.154e-02	7.384e-01	1.689e+00	1.463e-02	1.391e-03	0	0
ggZH 250_400	0	1.667e-02	2.716e-02	9.354e-01	1.315e-01	3.634e-03	1.403e-03
ggZH 75_150	5.049e+00	3.751e-02	2.720e-02	0	0	0	0
ggZH GT400	0	0	0	5.431e-03	1.746e-03	8.830e-02	4.035e-02
TT	7.439e+02	1.228e+01	2.073e+01	0	3.314e-01	2.463e-01	0
VVHF	3.299e+01	6.787e+00	4.998e+00	3.336e+00	1.106e+00	7.033e-01	5.497e-01
VVLF	9.130e+00	1.564e+00	1.242e+00	6.538e-01	7.744e-01	1.564e-01	3.506e-01
Zj0b_c	2.957e+02	3.349e+01	3.968e+01	1.401e+01	1.245e+01	2.316e+00	4.094e+00
Zj0b_udsg	2.874e+02	3.464e+01	4.266e+01	2.008e+01	1.413e+01	1.887e+00	2.963e+00
Zj1b	7.915e+02	7.337e+01	6.230e+01	2.016e+01	1.451e+01	2.788e+00	3.169e+00
Zj2b	7.362e+02	8.954e+01	8.706e+01	4.075e+01	1.724e+01	4.978e+00	6.735e+00
s_Top	2.612e+01	1.715e+00	1.465e+00	3.041e-01	1.165e+00	0	0

Table C.18: The observed and predicted yields are given for the control regions for 2-lepton (e) in 2017.

Process	Z+udcsg enr., low $p_T(V)$	Z+udcsg enr., med. $p_T(V)$	Z+udcsg enr., high $p_T(V)$	Z+udcsg enr., high $p_T(V)$, boosted
Data	41833	8612	1717	1018
TT	7.976e+01	5.396e+00	4.966e-01	3.495e+00
VVHF	3.972e+00	1.234e+00	2.484e-01	6.720e-01
VVLF	2.119e+02	6.111e+01	1.494e+01	1.488e+01
Zj0b_c	2.249e+03	4.887e+02	8.662e+01	1.196e+02
Zj0b_udsg	4.349e+04	8.694e+03	1.673e+03	8.001e+02
Zj1b	4.056e+02	8.682e+01	1.695e+01	3.863e+01
Zj2b	2.705e+01	9.459e+00	2.251e+00	1.759e+01
s_Top	8.625e+00	1.129e+00	2.533e-01	4.488e-01
Process	Z+bb enr., low $p_T(V)$	Z+bb enr., med. $p_T(V)$	Z+bb enr., high $p_T(V)$	Z+bb enr., high $p_T(V)$, boosted
Data	3537	932	222	123
TT	4.481e+02	2.518e+01	6.125e-01	1.104e+00
VVHF	2.867e+01	1.103e+01	3.215e+00	1.320e+00
VVLF	6.000e+00	2.292e+00	8.359e-01	9.310e-01
Zj0b_c	4.417e+02	1.406e+02	2.771e+01	1.995e+01
Zj0b_udsg	4.438e+02	1.241e+02	2.905e+01	2.419e+01
Zj1b	1.043e+03	2.227e+02	4.257e+01	2.518e+01
Zj2b	1.124e+03	3.056e+02	6.509e+01	3.870e+01
s_Top	1.726e+01	3.900e+00	0	1.860e-01
Process	$t\bar{t}$ enr., low $p_T(V)$	$t\bar{t}$ enr., med. $p_T(V)$	$t\bar{t}$ enr., high $p_T(V)$	$t\bar{t}$ enr., high $p_T(V)$, boosted
Data	12300	1240	118	87
TT	1.241e+04	1.123e+03	1.030e+02	2.702e+01
VVHF	6.536e+00	2.123e+00	8.447e-01	4.164e-01
VVLF	8.911e+00	7.944e-01	5.264e-01	3.886e+00
Zj0b_c	4.443e+01	1.025e+01	1.647e+00	5.752e+00
Zj0b_udsg	4.109e+01	8.556e+00	1.655e+00	7.094e+00
Zj1b	1.476e+02	2.621e+01	5.805e+00	5.939e+00
Zj2b	2.021e+02	4.878e+01	1.123e+01	8.915e+00
s_Top	4.456e+02	9.877e+01	1.452e+01	1.859e+01

Table C.19: The observed and predicted yields are given for the signal regions for 2-lepton (μ) in 2017.

Process	Low $p_T(V)$	Medium $p_T(V)$ 0J	Medium $p_T(V)$ GE1J	High $p_T(V)$	High $p_T(V)$, boosted	Highest $p_T(V)$	Highest $p_T(V)$, boosted
Data	4133	354	413	174	106	16	31
ZH 0_75	1.790e-01	0	0	0	0	0	0
ZH 150_250 0J	2.847e-01	6.542e+00	8.123e-01	6.978e-02	8.183e-09	0	0
ZH 150_250 GE1J	1.209e-01	9.774e-01	2.605e+00	3.429e-02	3.960e-03	0	0
ZH 250_400	0	1.333e-01	6.798e-02	3.558e+00	3.957e-01	3.647e-02	7.580e-03
ZH 75_150	2.210e+01	1.417e-01	5.190e-02	0	0	0	0
ZH GT400	0	0	0	4.813e-02	1.798e-02	7.756e-01	3.945e-01
ggZH 0_75	2.605e-02	0	0	0	0	0	0
ggZH 150_250 0J	1.103e-01	2.295e+00	3.094e-01	2.010e-02	0	0	0
ggZH 150_250 GE1J	1.014e-01	9.266e-01	2.210e+00	2.956e-02	1.416e-03	0	0
ggZH 250_400	0	3.770e-02	4.257e-02	1.142e+00	1.650e-01	8.721e-03	2.215e-03
ggZH 75_150	6.625e+00	6.998e-02	3.438e-02	0	0	0	0
ggZH GT400	0	0	0	1.232e-02	3.953e-03	9.916e-02	5.060e-02
TT	1.006e+03	3.256e+01	3.746e+01	1.727e+00	1.919e+00	0	0
VVHF	4.349e+01	8.685e+00	6.474e+00	4.948e+00	1.465e+00	9.020e-01	7.292e-01
VVLF	1.046e+01	1.636e+00	1.911e+00	1.109e+00	1.473e+00	5.283e-02	7.281e-01
Zj0b_c	4.050e+02	4.059e+01	5.583e+01	2.061e+01	1.532e+01	2.350e+00	4.069e+00
Zj0b_udsg	4.158e+02	4.041e+01	5.926e+01	2.161e+01	1.635e+01	2.703e+00	4.642e+00
Zj1b	1.021e+03	9.766e+01	8.840e+01	2.952e+01	1.807e+01	3.040e+00	3.511e+00
Zj2b	1.032e+03	1.098e+02	1.082e+02	4.522e+01	2.311e+01	5.873e+00	8.465e+00
s_Top	4.079e+01	3.668e+00	2.479e+00	8.773e-01	2.453e-02	0	0

Table C.20: The observed and predicted yields are given for the control regions for 2-lepton (μ) in 2017.

Process	Z+udcsg enr., low p$_T(V)$	Z+udcsg enr., med. $p_T(V)$	Z+udcsg enr., high $p_T(V)$	Z+udcsg enr., high $p_T(V)$, boosted
Data	59966	11480	2213	1320
TT	1.112e+02	1.603e+01	6.773e-01	7.867e+00
VVHF	4.975e+00	1.565e+00	3.981e-01	8.253e-01
VVLF	2.748e+02	7.337e+01	1.975e+01	1.836e+01
Zj0b_c	2.875e+03	5.916e+02	1.084e+02	1.581e+02
Zj0b_udsg	5.657e+04	1.091e+04	2.105e+03	1.008e+03
Zj1b	5.628e+02	1.039e+02	1.945e+01	5.163e+01
Zj2b	3.505e+01	9.590e+00	2.400e+00	2.366e+01
s_Top	1.266e+01	1.712e+00	4.158e-01	1.176e+00
Process	Z+bb enr., low $p_T(V)$	Z+bb enr., med. $p_T(V)$	Z+bb enr., high $p_T(V)$	Z+bb enr., high $p_T(V)$, boosted
Data	5029	1245	251	175
TT	6.372e+02	3.839e+01	8.654e-01	4.461e+00
VVHF	3.725e+01	1.388e+01	3.959e+00	2.373e+00
VVLF	7.647e+00	2.685e+00	1.054e+00	1.366e+00
Zj0b_c	5.644e+02	1.587e+02	3.810e+01	2.425e+01
Zj0b_udsg	6.259e+02	1.631e+02	4.301e+01	2.669e+01
Zj1b	1.420e+03	2.871e+02	5.238e+01	3.470e+01
Zj2b	1.506e+03	3.996e+02	7.789e+01	5.144e+01
s_Top	2.459e+01	4.882e+00	3.704e-01	1.821e-01
Process	$t\bar{t}$ enr., low $p_T(V)$	$t\bar{t}$ enr., med. $p_T(V)$	$t\bar{t}$ enr., high $p_T(V)$	$t\bar{t}$ enr., high $p_T(V)$, boosted
Data	18333	1717	169	122
TT	1.675e+04	1.533e+03	1.330e+02	5.020e+01
VVHF	9.141e+00	3.197e+00	1.148e+00	7.198e-01
VVLF	3.289e+00	5.260e+00	2.951e+00	4.437e+00
Zj0b_c	5.519e+01	1.730e+01	3.317e+00	1.180e+01
Zj0b_udsg	4.124e+01	1.213e+01	2.502e+00	1.588e+01
Zj1b	2.259e+02	4.693e+01	9.529e+00	1.246e+01
Zj2b	2.814e+02	7.289e+01	2.269e+01	1.586e+01
s_Top	5.785e+02	1.315e+02	2.160e+01	2.727e+01

Table C.21: The observed and predicted yields are given for the signal regions for 0-lepton in 2018.

Process	Medium $p_T(V)$ 0J	Medium $p_T(V)$ GE1J	High $p_T(V)$	High $p_T(V)$, boosted	Highest $p_T(V)$	Highest $p_T(V)$, boosted
Data	4366	4602	2650	1685	215	347
WH 150_250 0J	7.915e+00	1.080e+00	4.422e-01	0	0	0
WH 150_250 GE1J	1.043e+00	2.077e+00	1.978e-01	1.931e-02	0	0
WH 250_400	2.361e+00	1.459e+00	4.706e+00	7.740e-01	3.834e-02	1.119e-02
WH 75_150	3.045e-01	1.219e-01	0	0	0	0
WH GT400	2.061e-02	2.518e-02	6.051e-01	2.477e-01	5.549e-01	3.466e-01
ZH 150_250 0J	2.700e+01	2.604e+00	2.578e+00	9.690e-03	0	0
ZH 150_250 GE1J	3.626e+00	7.010e+00	1.117e+00	9.824e-02	0	0
ZH 250_400	5.745e+00	2.078e+00	2.269e+01	3.465e+00	5.773e-01	1.735e-01
ZH 75_150	1.083e+00	2.965e-01	0	0	0	0
ZH GT400	1.081e-02	4.248e-03	1.897e+00	7.018e-01	5.076e+00	3.233e+00
ggZH 150_250 0J	7.161e+00	8.040e-01	5.466e-01	6.478e-03	0	0
ggZH 150_250 GE1J	2.755e+00	4.897e+00	6.555e-01	8.022e-02	0	0
ggZH 250_400	1.339e+00	9.974e-01	5.723e+00	9.787e-01	1.041e-01	3.854e-02
ggZH 75_150	4.532e-01	2.043e-01	8.407e-04	0	0	0
ggZH GT400	3.370e-04	5.812e-04	2.378e-01	1.046e-01	5.053e-01	3.111e-01
TT	1.836e+03	3.001e+03	8.563e+02	7.961e+02	2.266e+01	8.071e+01
VVHF	1.099e+02	6.280e+01	8.554e+01	2.586e+01	1.202e+01	8.848e+00
VVLF	2.430e+01	1.812e+01	3.538e+01	2.419e+01	3.460e+00	1.202e+01
Wj0b_c	1.730e+02	1.197e+02	9.644e+01	3.831e+01	1.885e+00	8.100e+00
Wj0b_udsg	1.626e+02	1.528e+02	1.372e+02	2.192e+01	1.517e+01	1.382e+00
Wj1b	7.076e+01	5.824e+01	4.559e+01	4.634e+01	3.268e+00	6.576e+00
Wj2b	1.540e+02	1.497e+02	1.260e+02	1.070e+02	1.341e+01	2.332e+01
Zj0b_c	1.767e+02	1.287e+02	1.281e+02	7.226e+01	1.404e+01	1.945e+01
Zj0b_udsg	1.839e+02	1.240e+02	1.310e+02	2.404e+01	2.358e+01	6.320e+00
Zj1b	2.375e+02	1.283e+02	1.280e+02	1.230e+02	1.480e+01	3.053e+01
Zj2b	4.422e+02	2.845e+02	2.818e+02	1.635e+02	3.867e+01	6.779e+01
s_Top	2.330e+02	2.489e+02	9.822e+01	6.799e+01	3.088e+00	1.571e+01

Table C.22: The observed and predicted yields are given for the control regions for 0-lepton in 2018.

Process	Z+udcsg enr., med. $p_T(V)$	Z+udcsg enr., high $p_T(V)$	Z+udcsg enr., high $p_T(V)$, boosted
Data	12379	4990	95195
TT	1.587e+03	4.039e+02	4.111e+03
VVHF	1.466e+01	9.073e+00	4.532e+01
VVLF	2.011e+02	1.101e+02	2.118e+03
Wj0b_c	1.154e+03	2.958e+02	3.420e+03
Wj0b_udsg	3.383e+03	1.011e+03	3.105e+04
Wj1b	8.136e+01	2.965e+01	3.103e+02
Wj2b	4.687e+01	1.791e+01	1.862e+02
Zj0b_c	1.104e+03	4.583e+02	4.859e+03
Zj0b_udsg	2.777e+03	1.398e+03	4.649e+04
Zj1b	2.088e+02	8.279e+01	9.499e+02
Zj2b	9.193e+01	4.756e+01	3.538e+02
s_Top	2.053e+02	5.327e+01	6.468e+02
Process	Z+bb enr., med. $p_T(V)$	Z+bb enr., high $p_T(V)$	Z+bb enr., high $p_T(V)$, boosted
Data	15987	4868	2646
TT	7.015e+03	1.555e+03	9.006e+02
VVHF	1.357e+02	5.211e+01	4.142e+01
VVLF	2.933e+01	4.993e+01	3.335e+01
Wj0b_c	7.221e+02	1.563e+02	1.113e+02
Wj0b_udsg	1.067e+03	3.095e+02	4.882e+01
Wj1b	3.420e+02	9.114e+01	9.119e+01
Wj2b	6.378e+02	1.949e+02	2.111e+02
Zj0b_c	7.108e+02	2.549e+02	1.528e+02
Zj0b_udsg	7.362e+02	3.417e+02	7.935e+01
Zj1b	8.520e+02	2.848e+02	3.014e+02
Zj2b	1.542e+03	5.560e+02	3.863e+02
s_Top	8.532e+02	2.150e+02	9.127e+01
Process	$t\bar{t}$ enr., med. $p_T(V)$	$t\bar{t}$ enr., high $p_T(V)$	$t\bar{t}$ enr., high $p_T(V)$, boosted
Data	3831	1152	2434
TT	3.488e+03	8.786e+02	2.212e+03
VVHF	4.540e+00	2.031e+00	7.929e+00
VVLF	3.735e+00	4.366e+00	3.684e+00
Wj0b_c	5.369e+01	1.926e+01	1.224e+01
Wj0b_udsg	7.749e+01	3.150e+01	3.779e+00
Wj1b	2.112e+01	7.183e+00	1.436e+01
Wj2b	2.933e+01	1.315e+01	4.750e+01
Zj0b_c	3.192e+01	1.352e+01	1.055e+01
Zj0b_udsg	3.032e+01	1.835e+01	3.877e+00
Zj1b	3.061e+01	1.428e+01	2.507e+01
Zj2b	5.670e+01	2.657e+01	1.103e+02
s_Top	1.634e+02	5.835e+01	1.153e+02

Table C.23: The observed and predicted yields are given for the signal regions for 1-lepton (e) in 2018.

Process	Medium $p_T(V)$	High $p_T(V)$	High $p_T(V)$, boosted	Highest $p_T(V)$	Highest $p_T(V)$, boosted
Data	23105	2999	2362	235	526
WH 0_75	1.643e-02	0	0	0	0
WH 150_250 0J	4.007e+01	2.263e+00	1.707e-02	0	0
WH 150_250 GE1J	1.485e+01	9.443e-01	1.037e-01	0	0
WH 250_400	7.017e+00	1.968e+01	3.252e+00	5.046e-01	1.492e-01
WH 75_150	9.732e+00	1.254e-03	0	0	0
WH GT400	4.116e-03	1.601e+00	5.061e-01	4.142e+00	3.202e+00
ZH 0_75	2.667e-03	0	0	0	0
ZH 150_250 0J	1.064e+00	3.187e-02	0	0	0
ZH 150_250 GE1J	3.764e-01	1.110e-02	2.182e-03	0	0
ZH 250_400	3.517e-01	2.618e-01	6.275e-02	1.922e-03	9.850e-04
ZH 75_150	4.193e-01	0	0	0	0
ZH GT400	1.888e-02	7.269e-02	3.960e-02	4.452e-02	3.022e-02
TT	1.720e+04	1.930e+03	1.885e+03	8.391e+01	3.460e+02
VVHF	1.283e+02	3.503e+01	9.498e+00	1.983e+00	2.262e+00
VVLF	6.941e+01	1.886e+01	1.829e+01	7.554e+00	1.191e+01
Wj0b_c	1.043e+03	2.080e+02	8.942e+01	2.483e+01	2.735e+01
Wj0b_udsg	8.461e+02	2.037e+02	2.526e+01	3.201e+01	6.577e+00
Wj1b	4.374e+02	8.371e+01	6.410e+01	1.018e+01	1.773e+01
Wj2b	6.222e+02	2.029e+02	1.624e+02	3.434e+01	6.436e+01
Zj0b_c	4.593e+01	8.185e+00	2.330e+00	3.867e-01	6.397e-01
Zj0b_udsg	9.769e+01	1.024e+01	7.990e-01	5.467e-01	2.590e-01
Zj1b	9.190e+01	5.413e+00	4.972e+00	5.891e-01	7.025e-01
Zj2b	8.387e+01	8.816e+00	5.940e+00	7.114e-01	1.708e+00
s_Top	2.429e+03	4.121e+02	2.601e+02	3.611e+01	1.119e+02

Table C.24: The observed and predicted yields are given for the control regions for 1-lepton (e) in 2018.

Process	W+udcsg enr., med. $p_T(V)$	W+udcsg enr., high $p_T(V)$	W+udcsg enr., high $p_T(V)$, boosted
Data	193770	41572	57462
TT	2.880e+04	6.670e+03	8.374e+03
VVHF	6.427e+01	2.215e+01	2.635e+01
VVLF	3.082e+03	1.176e+03	2.157e+03
Wj0b_c	3.520e+04	6.690e+03	6.644e+03
Wj0b_udsg	1.040e+05	2.452e+04	4.129e+04
Wj1b	1.482e+03	3.708e+02	4.111e+02
Wj2b	4.340e+02	1.822e+02	2.725e+02
Zj0b_c	1.062e+03	1.163e+02	1.662e+02
Zj0b_udsg	6.537e+03	7.505e+02	1.606e+03
Zj1b	2.113e+02	2.135e+01	3.017e+01
Zj2b	2.967e+01	5.638e+00	1.102e+01
s_Top	4.710e+03	1.258e+03	1.570e+03
Process	W+bb enr., med. $p_T(V)$	W+bb enr., high $p_T(V)$	W+bb enr., high $p_T(V)$, boosted
Data	39928	6587	2565
TT	2.817e+04	4.350e+03	1.880e+03
VVHF	1.187e+02	3.467e+01	1.789e+01
VVLF	1.409e+02	5.294e+01	5.436e+01
Wj0b_c	2.453e+03	4.259e+02	1.857e+02
Wj0b_udsg	2.138e+03	4.823e+02	5.293e+01
Wj1b	9.698e+02	1.885e+02	1.152e+02
Wj2b	1.385e+03	3.772e+02	3.070e+02
Zj0b_c	9.179e+01	6.610e+00	9.481e+00
Zj0b_udsg	1.441e+02	2.227e+01	3.278e+00
Zj1b	1.454e+02	1.271e+01	1.042e+01
Zj2b	1.333e+02	1.608e+01	1.189e+01
s_Top	3.971e+03	8.775e+02	3.134e+02
Process	$t\bar{t}$ enr., med. $p_T(V)$	$t\bar{t}$ enr., high $p_T(V)$	$t\bar{t}$ enr., high $p_T(V)$, boosted
Data	66941	15880	5432
TT	6.491e+04	1.531e+04	6.251e+03
VVHF	9.074e+01	4.928e+01	3.667e+00
VVLF	6.806e+01	4.098e+01	3.396e+00
Wj0b_c	4.724e+02	1.643e+02	1.100e+01
Wj0b_udsg	5.317e+02	1.664e+02	7.421e+00
Wj1b	9.104e+02	3.050e+02	1.347e+01
Wj2b	1.175e+03	4.533e+02	5.924e+01
Zj0b_c	3.027e+01	4.695e+00	6.660e-01
Zj0b_udsg	4.163e+01	8.545e+00	2.443e-01
Zj1b	1.101e+02	1.897e+01	1.634e+00
Zj2b	1.169e+02	2.323e+01	4.284e+00
s_Top	4.292e+03	1.636e+03	4.395e+02

Table C.25: The observed and predicted yields are given for the signal regions for 1-lepton (μ) in 2018.

Process	Medium $p_T(V)$	High $p_T(V)$	High $p_T(V)$, boosted	Highest $p_T(V)$	Highest $p_T(V)$, boosted
Data	31579	4021	2960	294	599
WH 0_75	2.262e-02	0	0	0	0
WH 150_250 0J	5.175e+01	2.568e+00	1.132e-02	0	0
WH 150_250 GE1J	1.781e+01	1.101e+00	1.364e-01	0	0
WH 250_400	8.456e+00	2.355e+01	3.771e+00	7.492e-01	1.761e-01
WH 75_150	1.357e+01	3.760e-03	0	0	0
WH GT400	1.499e-03	1.937e+00	7.714e-01	5.590e+00	3.732e+00
ZH 0_75	2.724e-03	0	0	0	0
ZH 150_250 0J	1.726e+00	5.302e-02	7.135e-04	0	0
ZH 150_250 GE1J	5.691e-01	2.210e-02	4.960e-03	0	0
ZH 250_400	3.605e-01	3.943e-01	7.821e-02	4.068e-03	3.366e-04
ZH 75_150	7.081e-01	0	0	0	0
ZH GT400	1.129e-02	4.281e-02	2.207e-02	3.700e-02	2.988e-02
TT	2.256e+04	2.427e+03	2.312e+03	9.527e+01	4.177e+02
VVHF	1.619e+02	4.066e+01	1.123e+01	7.950e+00	8.388e+00
VVLF	1.393e+02	2.313e+01	8.864e+00	5.578e+00	7.667e+00
Wj0b_c	1.384e+03	2.563e+02	1.049e+02	2.844e+01	2.673e+01
Wj0b_udsg	1.348e+03	2.757e+02	1.980e+01	3.341e+01	1.150e+01
Wj1b	5.783e+02	1.126e+02	8.645e+01	1.518e+01	2.484e+01
Wj2b	7.780e+02	2.410e+02	2.095e+02	4.294e+01	8.545e+01
Zj0b_c	5.316e+01	7.564e+00	4.040e+00	4.491e-01	4.339e-01
Zj0b_udsg	1.592e+02	8.038e+00	1.700e+00	1.958e+00	1.576e-01
Zj1b	1.119e+02	7.940e+00	6.659e+00	4.733e-01	6.843e-01
Zj2b	1.115e+02	1.542e+01	9.168e+00	8.659e-01	1.446e+00
s_Top	3.203e+03	5.017e+02	3.558e+02	4.362e+01	1.380e+02

Table C.26: The observed and predicted yields are given for the control regions for 1-lepton (μ) in 2018.

Process	W+udcsg enr., med. $p_T(V)$	W+udcsg enr., high $p_T(V)$	W+udcsg enr., high $p_T(V)$, boosted
Data	260398	55094	69980
TT	3.728e+04	8.479e+03	1.010e+04
VVHF	8.490e+01	2.915e+01	2.662e+01
VVLF	4.182e+03	1.351e+03	2.737e+03
Wj0b_c	4.624e+04	8.475e+03	8.310e+03
Wj0b_udsg	1.414e+05	3.232e+04	5.274e+04
Wj1b	1.930e+03	4.689e+02	5.258e+02
Wj2b	5.624e+02	2.182e+02	3.405e+02
Zj0b_c	1.519e+03	1.611e+02	2.152e+02
Zj0b_udsg	9.581e+03	1.210e+03	2.167e+03
Zj1b	2.806e+02	3.073e+01	3.963e+01
Zj2b	5.045e+01	7.584e+00	1.454e+01
s_Top	6.177e+03	1.574e+03	1.927e+03

Process	W+bb enr., med. $p_T(V)$	W+bb enr., high $p_T(V)$	W+bb enr., high $p_T(V)$, boosted
Data	55450	8895	3283
TT	3.702e+04	5.489e+03	2.252e+03
VVHF	1.509e+02	4.565e+01	2.048e+01
VVLF	2.055e+02	5.641e+01	2.483e+01
Wj0b_c	3.287e+03	6.385e+02	2.670e+02
Wj0b_udsg	3.118e+03	6.437e+02	9.683e+01
Wj1b	1.317e+03	2.517e+02	1.480e+02
Wj2b	1.855e+03	4.605e+02	3.761e+02
Zj0b_c	1.474e+02	1.912e+01	7.336e+00
Zj0b_udsg	2.803e+02	3.444e+01	3.636e+00
Zj1b	1.958e+02	2.054e+01	1.189e+01
Zj2b	2.158e+02	2.033e+01	1.548e+01
s_Top	5.168e+03	1.149e+03	3.692e+02

Process	$t\bar{t}$ enr., med. $p_T(V)$	$t\bar{t}$ enr., high $p_T(V)$	$t\bar{t}$ enr., high $p_T(V)$, boosted
Data	89639	20932	6748
TT	8.353e+04	1.917e+04	7.629e+03
VVHF	1.187e+02	6.113e+01	5.071e+00
VVLF	7.179e+01	4.845e+01	1.603e+00
Wj0b_c	5.961e+02	1.833e+02	1.732e+01
Wj0b_udsg	6.820e+02	2.045e+02	5.317e+00
Wj1b	1.166e+03	3.775e+02	1.699e+01
Wj2b	1.482e+03	5.833e+02	7.215e+01
Zj0b_c	3.751e+01	6.439e+00	6.920e-01
Zj0b_udsg	5.452e+01	1.295e+01	3.370e-01
Zj1b	1.458e+02	2.531e+01	1.742e+00
Zj2b	1.676e+02	3.272e+01	5.100e+00
s_Top	5.606e+03	2.052e+03	5.378e+02

Table C.27: The observed and predicted yields are given for the signal regions for 2-lepton (e) in 2018.

Process	Low $p_T(V)$	Medium $p_T(V)$ 0J	Medium $p_T(V)$ GE1J	High $p_T(V)$	High $p_T(V)$, boosted	Highest $p_T(V)$	Highest $p_T(V)$, boosted
Data	4999	504	575	245	64	31	19
ZH 0_75	2.289e-01	0	0	0	0	0	0
ZH 150_250 0J	3.373e-01	7.785e+00	9.678e-01	4.255e-02	8.011e-04	0	0
ZH 150_250 GE1J	1.493e-01	1.111e+00	3.269e+00	2.580e-02	3.602e-03	0	0
ZH 250_400	0	1.249e-01	6.117e-02	4.417e+00	4.647e-01	1.842e-02	4.898e-03
ZH 75_150	2.555e+01	1.220e-01	4.382e-02	0	0	0	0
ZH GT400	0	0	0	3.598e-02	8.566e-03	8.884e-01	3.661e-01
ggZH 0_75	2.594e-02	0	0	0	0	0	0
ggZH 150_250 0J	1.379e-01	2.646e+00	3.597e-01	9.620e-03	0	0	0
ggZH 150_250 GE1J	1.155e-01	1.113e+00	2.648e+00	1.633e-02	1.324e-03	0	0
ggZH 250_400	0	3.244e-02	4.043e-02	1.483e+00	1.766e-01	2.691e-03	5.990e-04
ggZH 75_150	7.824e+00	4.744e-02	3.488e-02	0	0	0	0
ggZH GT400	0	0	0	8.888e-03	2.601e-03	1.293e-01	6.461e-02
TT	1.150e+03	2.658e+01	3.997e+01	1.226e+00	6.795e-01	0	2.631e-01
VVHF	5.039e+01	1.014e+01	8.163e+00	5.592e+00	6.454e-01	1.208e+00	6.631e-01
VVLF	1.260e+01	2.754e+00	2.359e+00	1.569e+00	1.225e+00	2.223e-01	4.583e-01
Zj0b_c	5.340e+02	6.238e+01	7.837e+01	2.303e+01	7.862e+00	3.571e+00	1.798e+00
Zj0b_udsg	7.475e+02	7.387e+01	1.212e+02	4.420e+01	1.816e+00	5.508e+00	6.458e-01
Zj1b	1.259e+03	1.188e+02	1.116e+02	3.898e+01	1.423e+01	3.843e+00	3.455e+00
Zj2b	1.167e+03	1.354e+02	1.368e+02	5.830e+01	2.181e+01	8.041e+00	7.796e+00
s_Top	4.040e+01	6.273e+00	3.225e+00	6.880e-01	0	0	0

Table C.28: The observed and predicted yields are given for the control regions for 2-lepton (e) in 2018.

Process	Z+udcsg enr., low $p_T(V)$	Z+udcsg enr., med. $p_T(V)$	Z+udcsg enr., high $p_T(V)$	Z+udcsg enr., high $p_T(V)$, boosted
Data	57826	11576	2368	1509
TT	9.427e+01	9.152e+00	2.976e-01	6.309e+00
VVHF	4.364e+00	1.404e+00	4.813e-01	1.243e+00
VVLF	2.698e+02	7.876e+01	2.135e+01	2.603e+01
Zj0b_c	3.314e+03	6.992e+02	1.307e+02	1.898e+02
Zj0b_udsg	6.430e+04	1.248e+04	2.380e+03	1.248e+03
Zj1b	5.998e+02	1.201e+02	2.142e+01	6.234e+01
Zj2b	3.336e+01	1.058e+01	3.179e+00	3.254e+01
s_Top	1.419e+01	2.919e+00	2.442e-01	1.097e+00
Process	Z+bb enr., low $p_T(V)$	Z+bb enr., med. $p_T(V)$	Z+bb enr., high $p_T(V)$	Z+bb enr., high $p_T(V)$, boosted
Data	6249	1655	420	119
TT	7.093e+02	3.974e+01	7.926e-01	1.231e+00
VVHF	4.632e+01	1.621e+01	4.900e+00	2.484e+00
VVLF	9.093e+00	4.180e+00	1.246e+00	9.493e-01
Zj0b_c	8.230e+02	2.183e+02	5.327e+01	1.710e+01
Zj0b_udsg	1.066e+03	3.136e+02	8.774e+01	7.434e+00
Zj1b	1.748e+03	3.791e+02	7.215e+01	3.246e+01
Zj2b	1.655e+03	4.718e+02	1.008e+02	5.050e+01
s_Top	3.355e+01	6.414e+00	0	4.955e-01
Process	$t\bar{t}$ enr., low $p_T(V)$	$t\bar{t}$ enr., med. $p_T(V)$	$t\bar{t}$ enr., high $p_T(V)$	$t\bar{t}$ enr., high $p_T(V)$, boosted
Data	19584	1953	199	68
TT	1.964e+04	1.878e+03	1.694e+02	3.396e+01
VVHF	1.091e+01	4.021e+00	1.113e+00	6.643e-01
VVLF	6.835e+00	1.387e+00	3.596e+00	2.395e-01
Zj0b_c	7.226e+01	2.006e+01	2.272e+00	5.338e+00
Zj0b_udsg	6.501e+01	2.096e+01	2.974e+00	1.732e+00
Zj1b	2.739e+02	4.585e+01	8.711e+00	5.616e+00
Zj2b	2.836e+02	6.862e+01	1.788e+01	1.034e+01
s_Top	7.100e+02	1.608e+02	2.780e+01	1.576e+01

Table C.29: The observed and predicted yields are given for the signal regions for 2-lepton (μ) in 2018.

Process	Low $p_T(V)$	Medium $p_T(V)$ 0J	Medium $p_T(V)$ GE1J	High $p_T(V)$	High $p_T(V)$, boosted	Highest $p_T(V)$	Highest $p_T(V)$, boosted
Data	7331	705	783	295	70	31	35
ZH 0_75	2.889e-01	0	0	0	0	0	0
ZH 150_250 0J	4.431e-01	1.005e+01	1.257e+00	1.034e-01	2.004e-04	0	0
ZH 150_250 GE1J	1.992e-01	1.453e+00	4.195e+00	7.381e-02	3.357e-03	0	0
ZH 250_400	0	1.758e-01	1.022e-01	5.524e+00	5.845e-01	5.370e-02	1.236e-02
ZH 75_150	3.444e+01	2.071e-01	8.425e-02	0	0	0	0
ZH GT400	0	0	0	6.876e-02	2.680e-02	1.108e+00	5.466e-01
ggZH 0_75	3.529e-02	0	2.525e-04	0	0	0	0
ggZH 150_250 0J	1.599e-01	3.375e+00	4.952e-01	2.391e-02	3.038e-04	0	0
ggZH 150_250 GE1J	1.533e-01	1.427e+00	3.348e+00	4.348e-02	3.056e-03	0	0
ggZH 250_400	0	5.089e-02	6.001e-02	1.751e+00	2.287e-01	1.206e-02	1.561e-03
ggZH 75_150	1.022e+01	8.581e-02	5.159e-02	0	0	0	0
ggZH GT400	0	0	0	1.904e-02	3.315e-03	1.617e-01	7.121e-02
TT	1.591e+03	4.376e+01	5.171e+01	3.347e+00	2.998e+00	0	0
VVHF	6.999e+01	1.453e+01	1.083e+01	6.557e+00	1.277e+00	1.198e+00	8.875e-01
VVLF	2.145e+01	2.783e+00	2.816e+00	1.973e+00	1.236e+00	1.780e-01	7.260e-01
Zj0b_c	7.340e+02	8.511e+01	8.904e+01	4.224e+01	8.671e+00	4.209e+00	3.089e+00
Zj0b_udsg	8.845e+02	9.877e+01	1.409e+02	5.813e+01	2.451e+00	8.230e+00	1.108e+00
Zj1b	1.815e+03	1.606e+02	1.462e+02	4.960e+01	1.742e+01	5.261e+00	4.598e+00
Zj2b	1.570e+03	1.734e+02	1.710e+02	7.151e+01	2.678e+01	1.001e+01	1.118e+01
s_Top	6.838e+01	5.039e+00	3.626e+00	0	0	0	0

Table C.30: The observed and predicted yields are given for the control regions for 2-lepton (μ) in 2018.

Process	Z+udcsg enr., low $p_T(V)$	Z+udcsg enr., med. $p_T(V)$	Z+udcsg enr., high $p_T(V)$	Z+udcsg enr., high $p_T(V)$, boosted
Data	81615	15434	3057	1976
TT	1.535e+02	1.872e+01	1.412e+00	1.317e+01
VVHF	6.382e+00	1.728e+00	2.966e-01	1.749e+00
VVLF	3.682e+02	1.020e+02	2.735e+01	2.809e+01
Zj0b_c	4.466e+03	8.822e+02	1.574e+02	2.460e+02
Zj0b_udsg	8.488e+04	1.594e+04	3.004e+03	1.597e+03
Zj1b	8.033e+02	1.509e+02	2.763e+01	7.948e+01
Zj2b	4.713e+01	1.298e+01	3.389e+00	4.048e+01
s_Top	2.289e+01	1.760e+00	7.158e-01	1.158e+00
Process	Z+bb enr., low $p_T(V)$	Z+bb enr., med. $p_T(V)$	Z+bb enr., high $p_T(V)$	Z+bb enr., high $p_T(V)$, boosted
Data	9330	2258	495	149
TT	9.774e+02	6.429e+01	3.166e+00	3.249e+00
VVHF	6.060e+01	1.973e+01	6.293e+00	3.539e+00
VVLF	1.244e+01	4.716e+00	1.767e+00	1.391e+00
Zj0b_c	1.058e+03	2.945e+02	6.315e+01	2.417e+01
Zj0b_udsg	1.455e+03	3.920e+02	1.054e+02	9.642e+00
Zj1b	2.452e+03	4.951e+02	9.026e+01	4.420e+01
Zj2b	2.276e+03	6.077e+02	1.276e+02	6.493e+01
s_Top	4.292e+01	7.726e+00	2.201e-01	0
Process	$t\bar{t}$ enr., low $p_T(V)$	$t\bar{t}$ enr., med. $p_T(V)$	$t\bar{t}$ enr., high $p_T(V)$	$t\bar{t}$ enr., high $p_T(V)$, boosted
Data	28843	2891	333	112
TT	2.677e+04	2.490e+03	2.344e+02	3.955e+01
VVHF	1.455e+01	5.380e+00	1.617e+00	1.020e+00
VVLF	1.352e+01	4.143e-01	2.265e+00	6.160e-01
Zj0b_c	1.083e+02	2.056e+01	7.434e+00	8.186e+00
Zj0b_udsg	1.037e+02	2.296e+01	6.708e+00	5.635e+00
Zj1b	4.156e+02	7.519e+01	1.594e+01	1.082e+01
Zj2b	4.631e+02	1.288e+02	3.064e+01	2.091e+01
s_Top	9.704e+02	2.096e+02	3.764e+01	2.076e+01

Appendix D

Nuisance Parameter Impacts

The following figures show the impacts of the 60 most influential nuisance parameters on each STXS bin. There are 1878 nuisance parameters, but the influence of each on the final results quickly drops off. Notice the difference in the scale of each canvas in the figures.
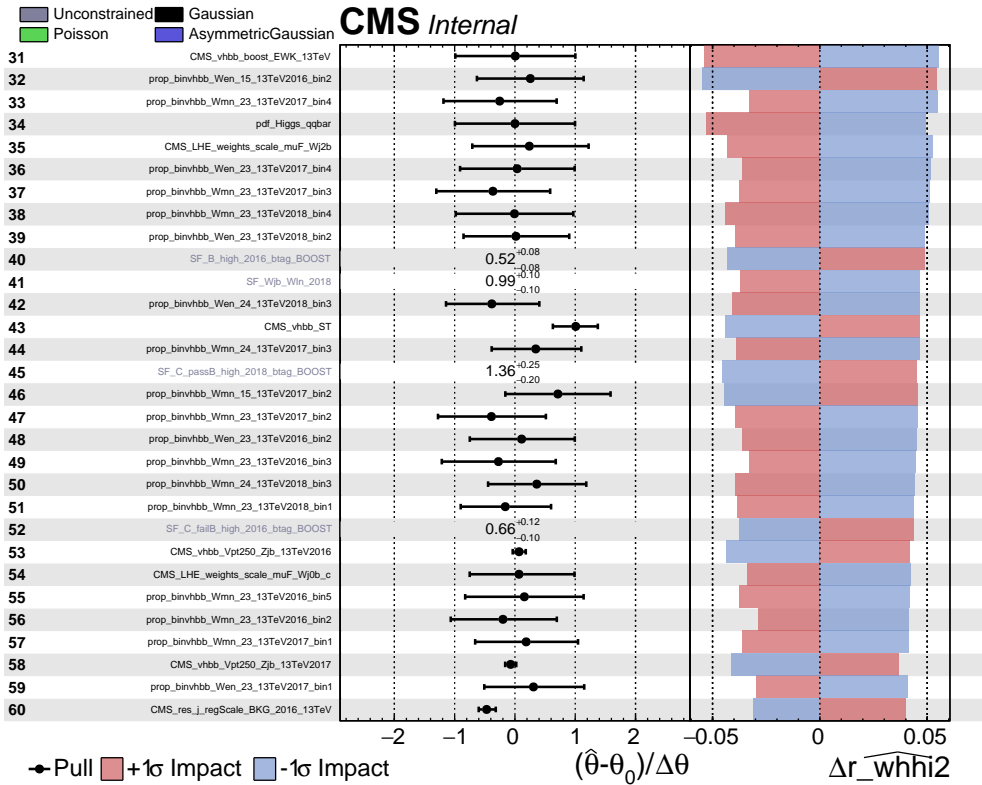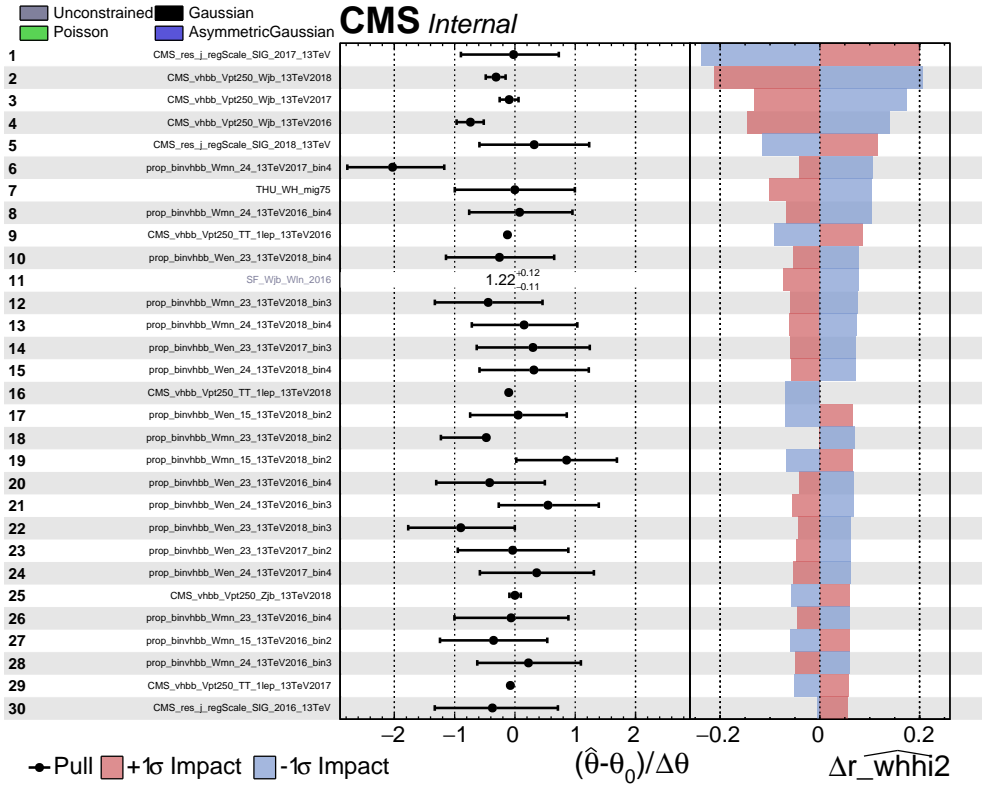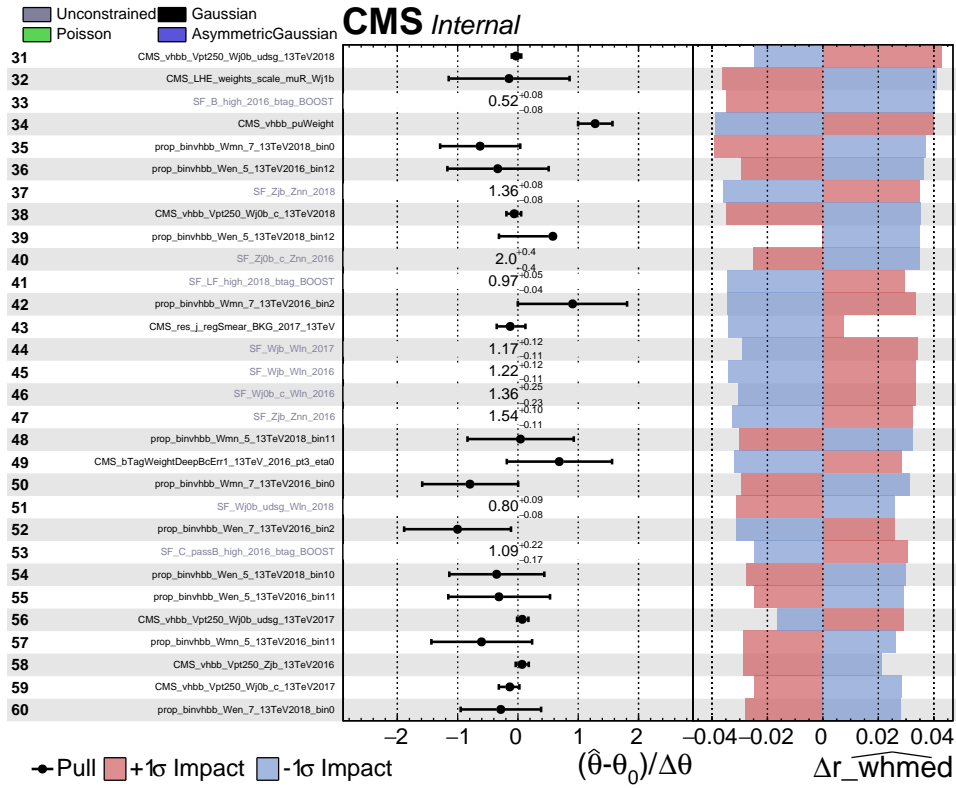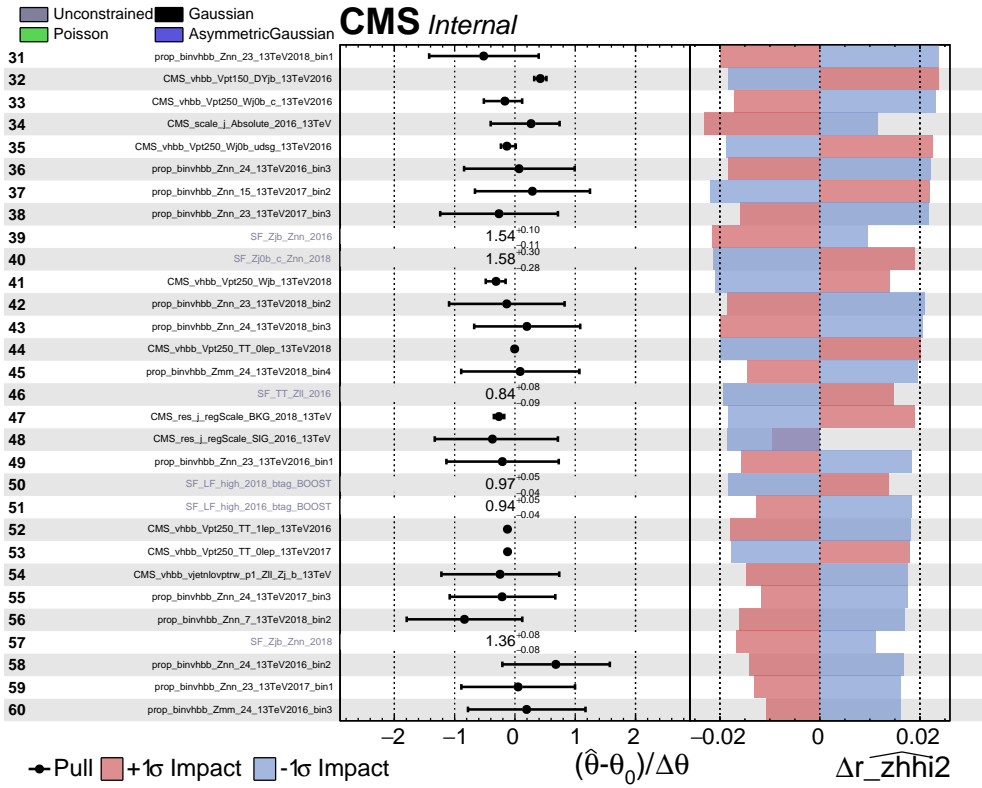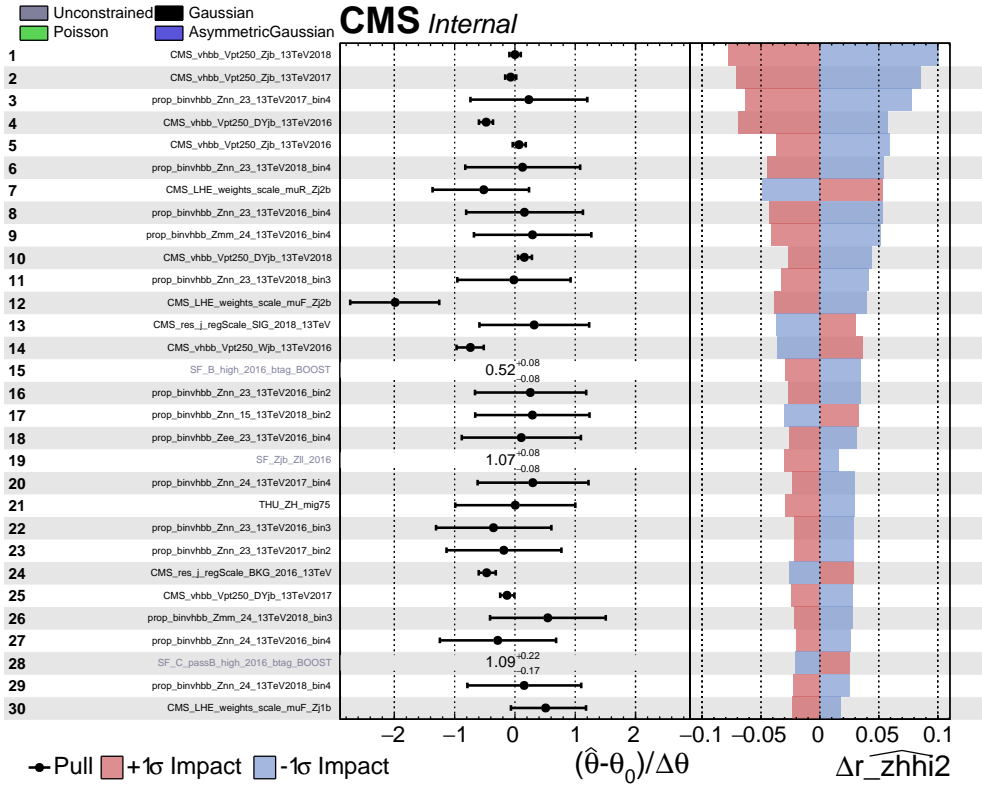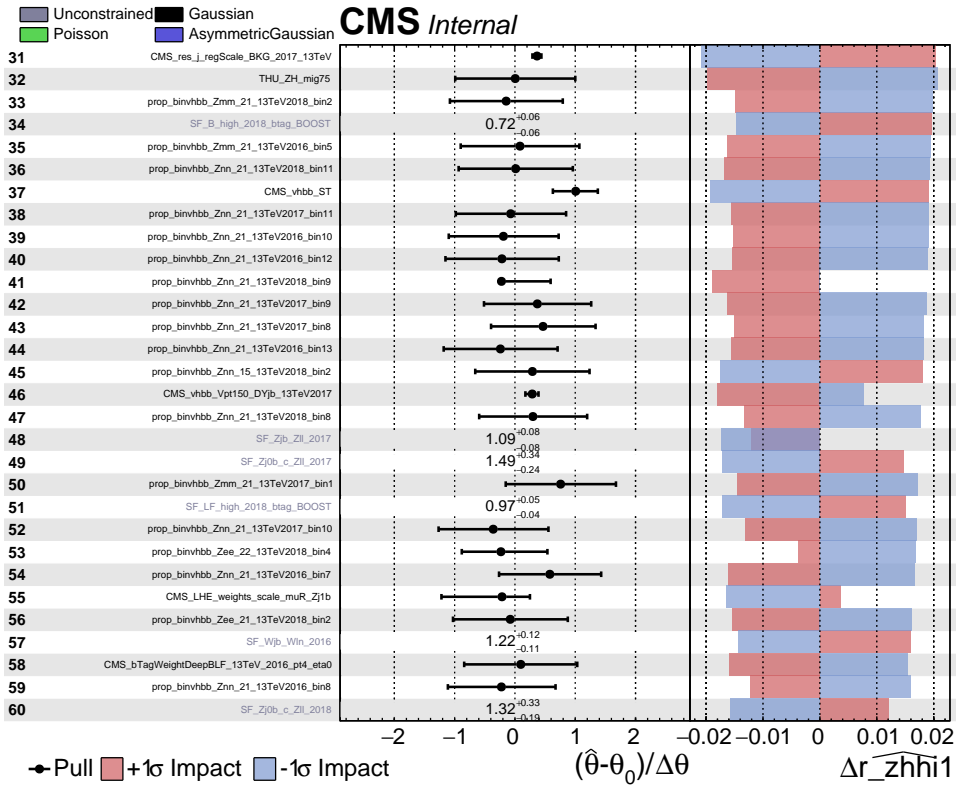
Figure D-1: Impacts on WH $p_T(V) > 400\,\mathrm{GeV}$ by the 60 most influential nuisance parameters.
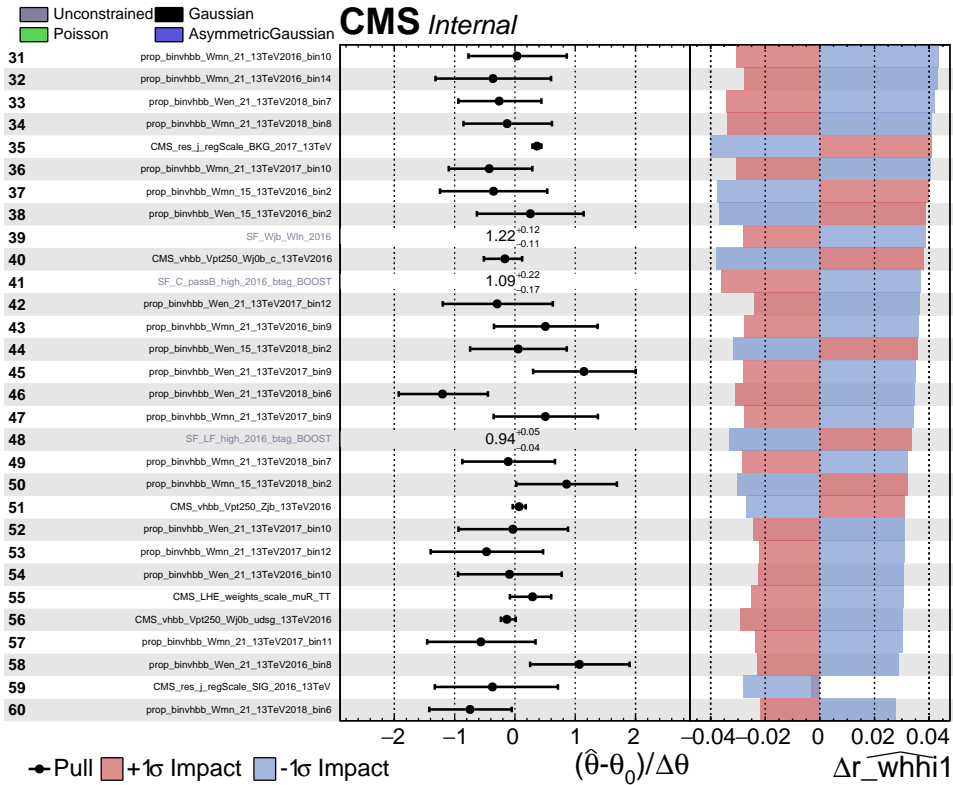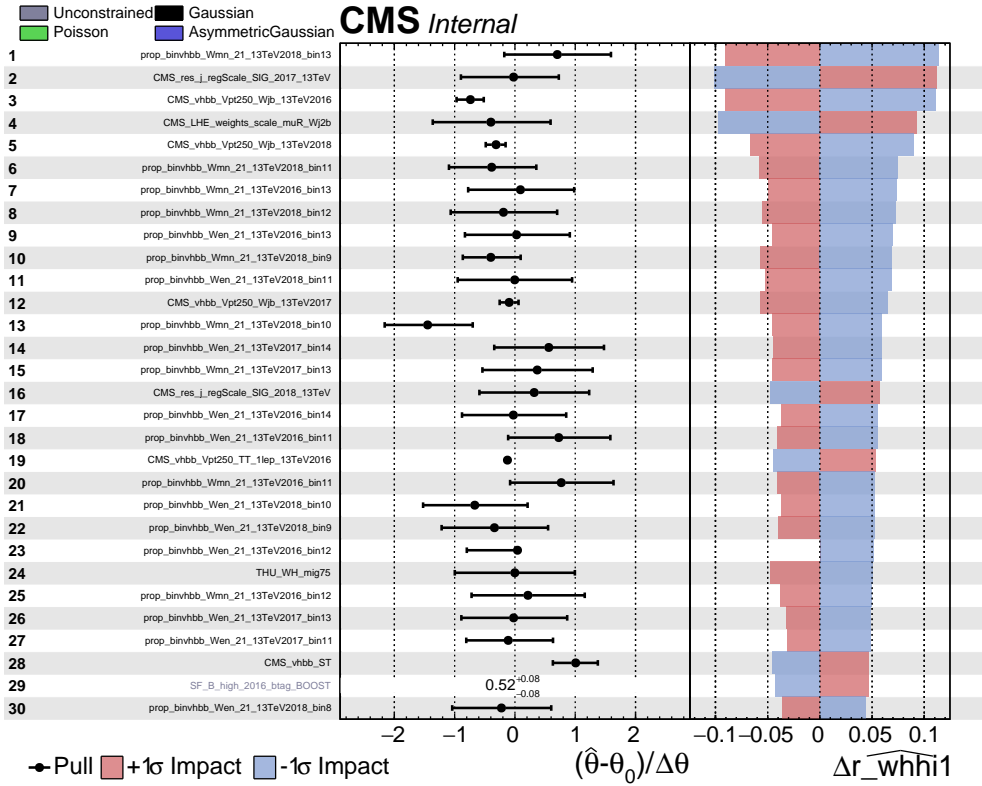
Figure D-2: Impacts on WH $250 < p_T(V) < 400\,\mathrm{GeV}$ by the 60 most influential nuisance parameters.
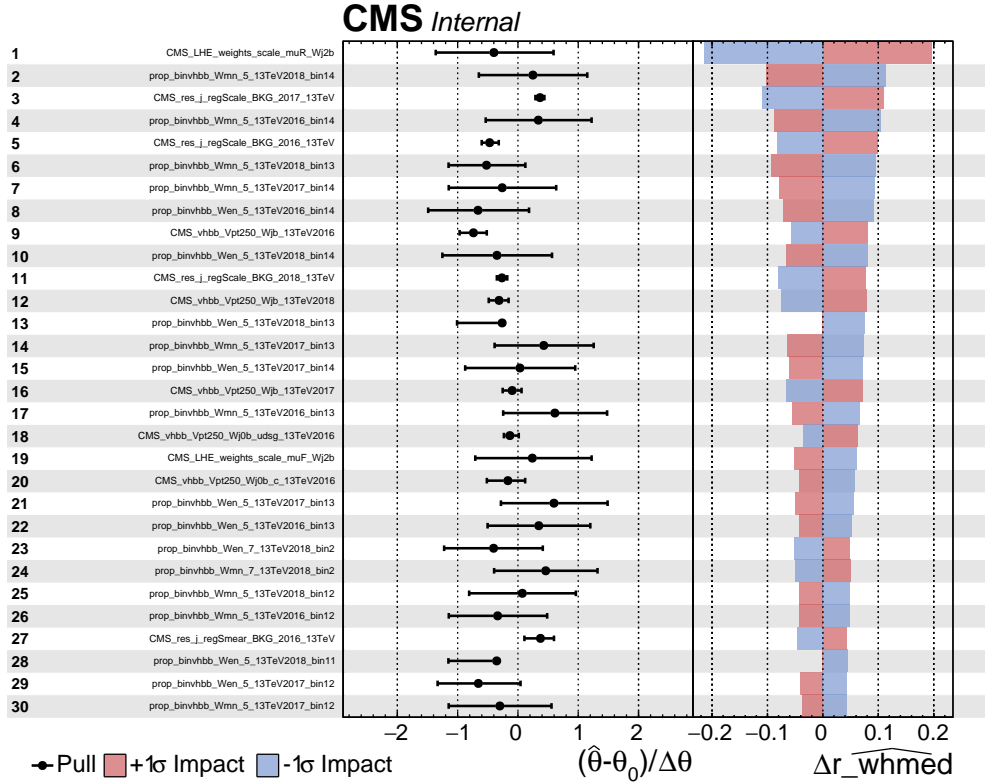
Figure D-3: Impacts on WH $150 < p_T(V) < 250\,\mathrm{GeV}$ by the 60 most influential nuisance parameters.

Figure D-4: Impacts on ZH $p_T(V) > 400\,\mathrm{GeV}$ by the 60 most influential nuisance parameters.
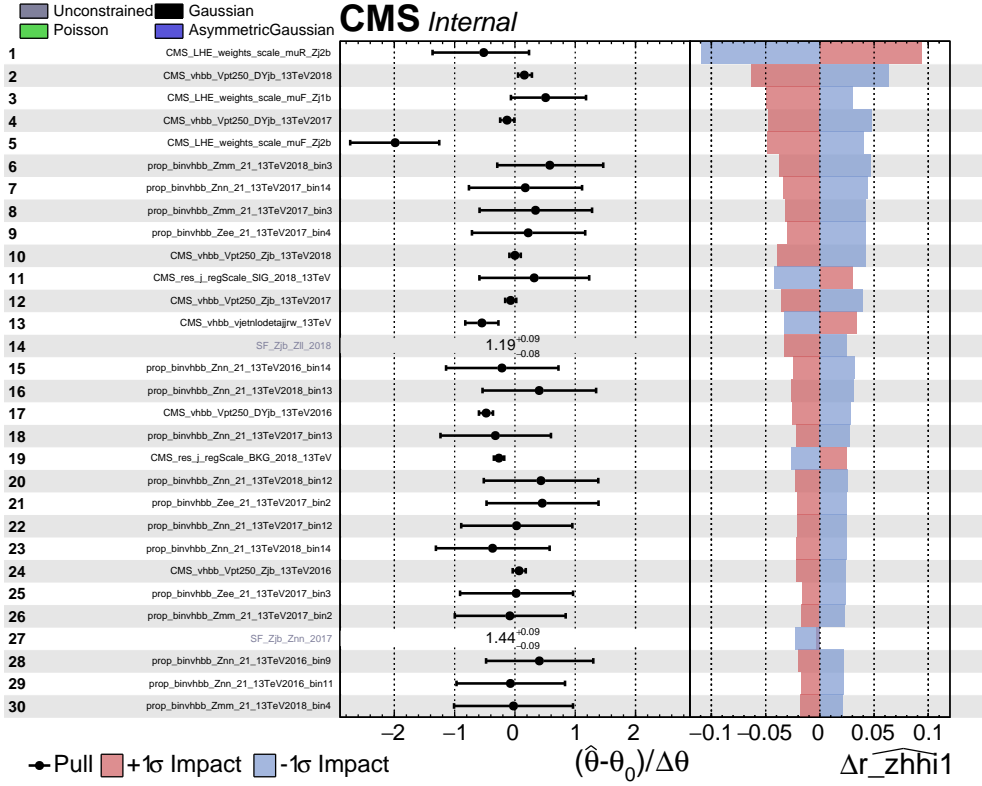
Figure D-5: Impacts on ZH $250 < p_T(V) < 400\,\mathrm{GeV}$ by the 60 most influential nuisance parameters.
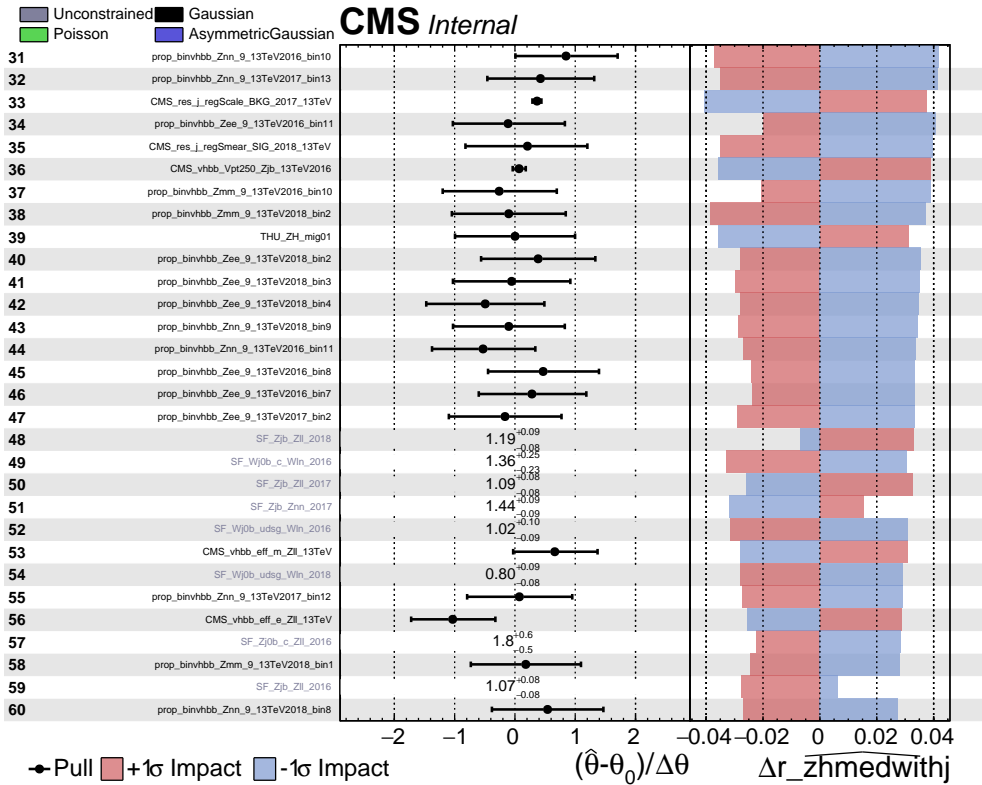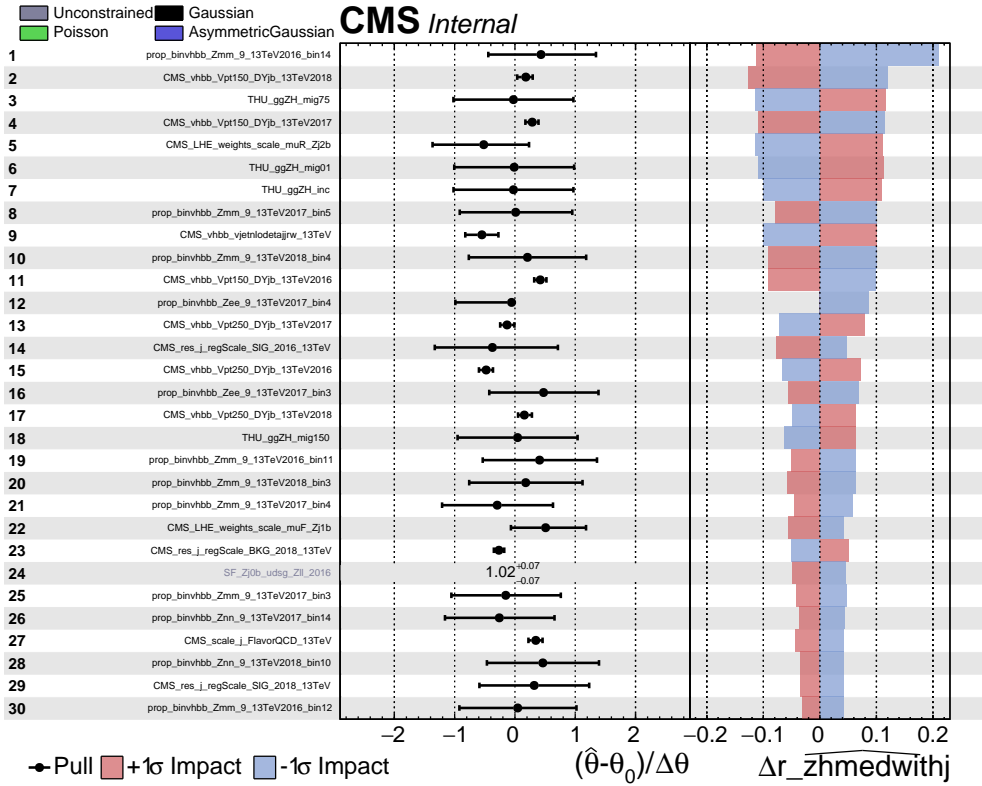
Figure D-6: Impacts on ZH $150 < p_T(V) < 250\,\mathrm{GeV}$ with additional jets by the 60 most influential nuisance parameters.
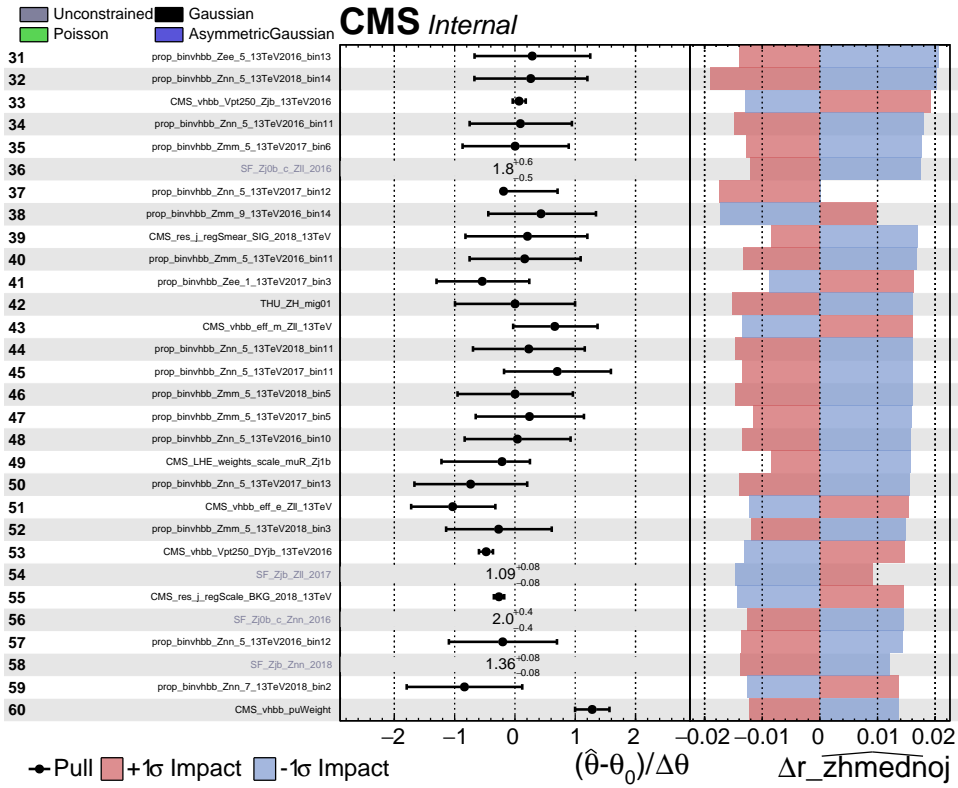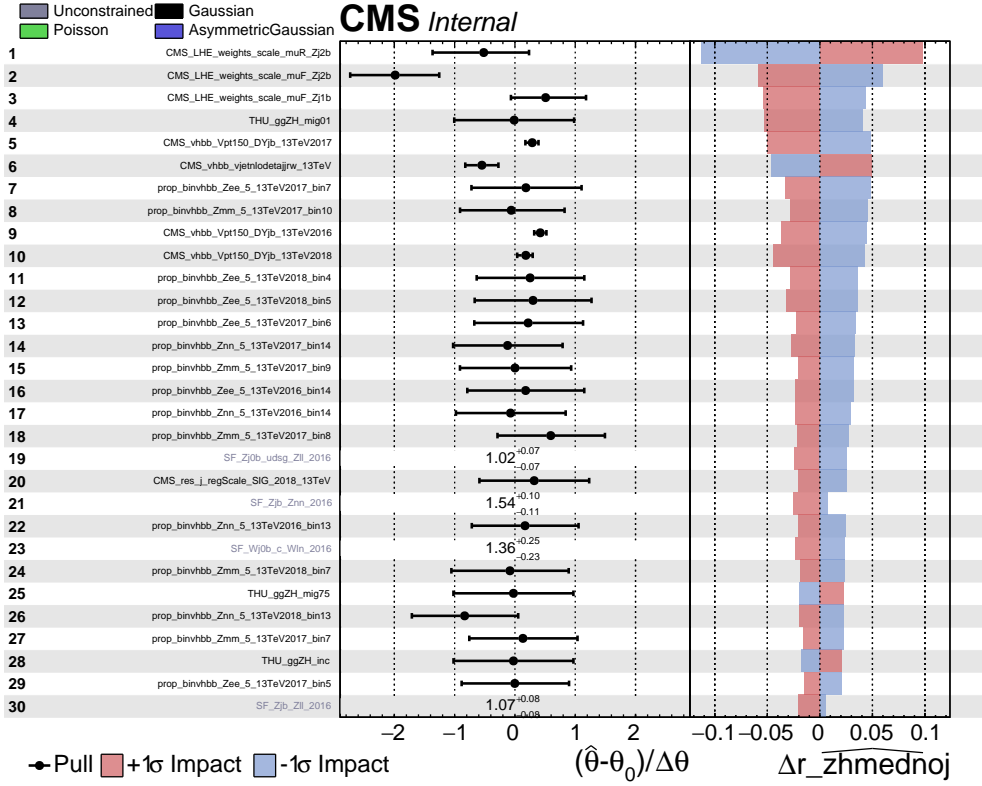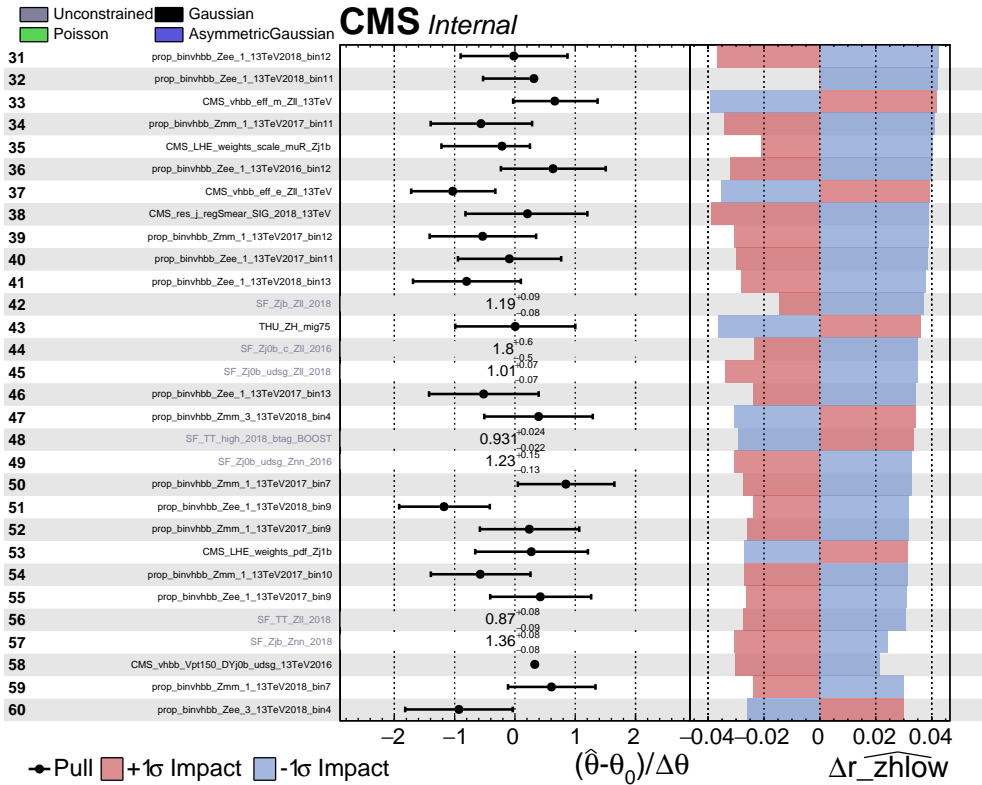
Figure D-7: Impacts on ZH $150 < p_T(V) < 250\,\mathrm{GeV}$ without additional jets by the 60 most influential nuisance parameters.
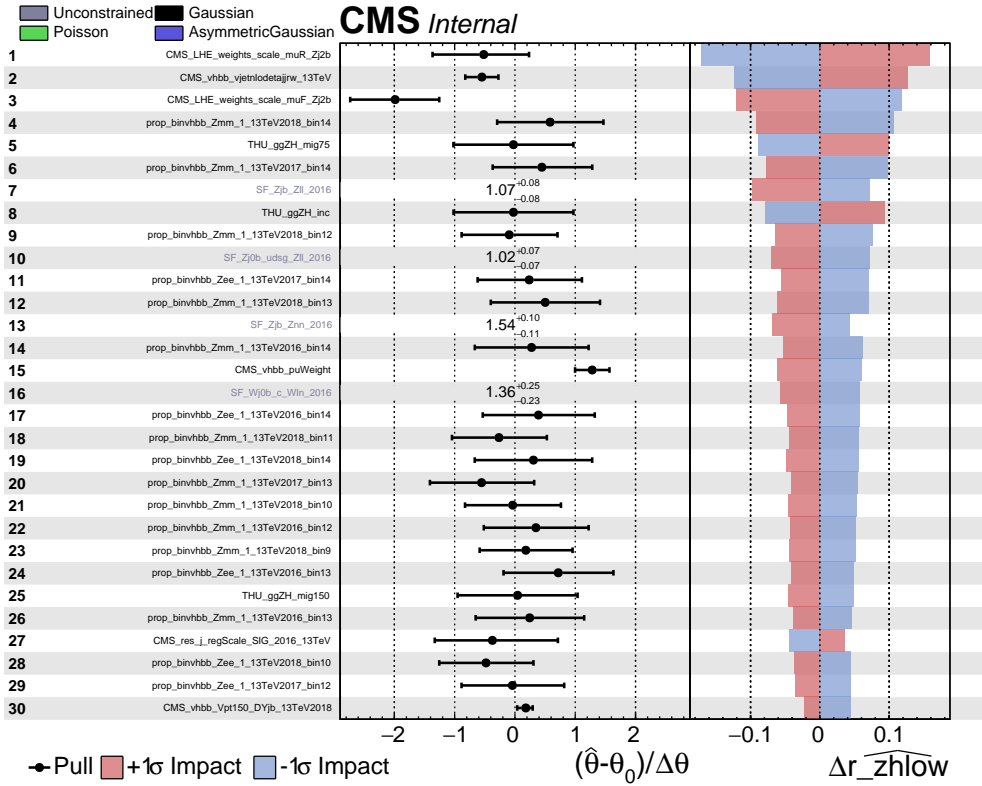
Figure D-8: Impacts on ZH $75 < p_T(V) < 150\,\mathrm{GeV}$ by the 60 most influential nuisance parameters.

Bibliography

[1] T. S. Kuhn, *The structure of scientific revolutions*. Chicago: University of Chicago Press, 1970.

[2] S. Weinberg, "A model of leptons," *Phys. Rev. Lett.*, vol. 19, pp. 1264–1266, Nov 1967.

[3] J. Smith, W. L. van Neerven, and J. A. M. Vermaseren, "Transverse mass and width of the w boson," *Phys. Rev. Lett.*, vol. 50, pp. 1738–1740, May 1983.

[4] W. Dau, "Ua1 results from $p\overline{p}$ collisions at $\sqrt{s} = 540$ gev," tech. rep., 198401126, 1983.

[5] S. Chatrchyan, V. Khachatryan, A. Sirunyan, A. Tumasyan, W. Adam, E. Aguilo, T. Bergauer, M. Dragicevic, J. Erö, C. Fabjan, *et al.*, "Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC," *Physics Letters B*, vol. 716, p. 30–61, Sep 2012.

[6] G. Aad *et al.*, "Combined search for the Standard Model Higgs boson in pp collisions at $\sqrt{s}$=7 TeV with the ATLAS detector," *Phys. Rev. D*, vol. 86, p. 032003, Aug 2012.

[7] CMS Collaboration, "Observation of Higgs Boson Decay to Bottom Quarks," *Physical Review Letters*, vol. 121, Sep 2018.

[8] M. Aaboud, G. Aad, B. Abbott, O. Abdinov, B. Abeloos, D. Abhayasinghe, S. Abidi, O. AbouZeid, N. Abraham, H. Abramowicz, and et al., "Observation of $h \to b\bar{b}$ decays and vh production with the atlas detector," *Physics Letters B*, vol. 786, p. 59–86, Nov 2018.

[9] S. Holmes, R. S. Moore, and V. Shiltsev, "Overview of the tevatron collider complex: goals, operations and performance," *Journal of Instrumentation*, vol. 6, pp. T08001–T08001, aug 2011.

[10] S. Myers, *The LEP Collider, from design to approval and commissioning*. John Adams' memorial lecture, Geneva: CERN, 1991. Delivered at CERN, 26 Nov 1990.

[11] M. Thomson, *Modern particle physics*. Cambridge: Cambridge University Press, 2013.

[12] S. Brandt, *Measurement of W and Z boson production cross sections in proton-proton collisions at $\sqrt{s} = 5.02$ TeV and $\sqrt{s} = 13$ TeV*. PhD thesis, Massachusetts Institute of Technology, 2020.

[13] D. J. Gross and F. Wilczek, "Ultraviolet behavior of non-abelian gauge theories," *Phys. Rev. Lett.*, vol. 30, pp. 1343–1346, Jun 1973.

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

[15] P. W. Higgs, "Broken symmetries and the masses of gauge bosons," *Phys. Rev. Lett.*, vol. 13, pp. 508–509, Oct 1964.

[16] G. S. Guralnik, C. R. Hagen, and T. W. B. Kibble, "Global conservation laws and massless particles," *Phys. Rev. Lett.*, vol. 13, pp. 585–587, Nov 1964.

[17] V. Fock, "Bemerkung zum Virialsatz," *Zeitschrift fur Physik*, vol. 63, pp. 855–858, Nov. 1930.

[18] S. Weinberg, "General theory of broken local symmetries," *Phys. Rev. D*, vol. 7, pp. 1068–1082, Feb 1973.

[19] V. Khachatryan, A. Sirunyan, A. Tumasyan, W. Adam, T. Bergauer, M. Dragicevic, J. Erö, C. Fabjan, M. Friedl, R. Frühwirth, and et al., "Measurement of the ratiob(t→wb)/b(t→wq)in pp collisions ats=8tev," *Physics Letters B*, vol. 736, p. 33–57, Sep 2014.

[20] S. Dittmaier, C. Mariotti, G. Passarino, R. Tanaka, S. Alekhin, J. Alwall, E. Bagnaschi, A. Banfi, J. Blumlein, S. Bolognesi, *et al.*, "Handbook of lhc higgs cross sections: 2. differential distributions," *arXiv preprint arXiv:1201.3084*, 2012.

[21] CMS Collaboration, *CMS Physics: Technical Design Report Volume 1: Detector Performance and Software*. Technical Design Report CMS, Geneva: CERN, 2006.

[22] L. Evans and P. Bryant, "LHC machine," *Journal of Instrumentation*, vol. 3, pp. S08001–S08001, aug 2008.

[23] R. Bruce, G. Arduini, H. Bartosik, R. De Maria, M. Giovannozzi, G. Iadarola, J. Jowett, K. S. B. Li, M. Lamont, A. Lechner, E. Metral, D. Mirarchi, T. Pieloni, S. Redaelli, G. Rumolo, B. Salvant, R. Tomas Garcia, and J. Wenninger, "LHC Run 2: Results and challenges," p. MOAM5P50. 7 p, Jul 2016.

[24] D. Barney, "CMS Detector Slice." CMS Collection., Jan 2016.

[25] A. Dominguez, D. Abbaneo, K. Arndt, N. Bacchetta, A. Ball, E. Bartz, W. Bertl, G. M. Bilei, G. Bolla, H. W. K. Cheung, M. Chertok, S. Costa, N. Demaria, D. D. Vazquez, K. Ecklund, W. Erdmann, K. Gill, G. Hall,

K. Harder, F. Hartmann, R. Horisberger, W. Johns, H. C. Kaestli, K. Klein, D. Kotlinski, S. Kwan, M. Pesaresi, H. Postema, T. Rohe, C. Schäfer, A. Starodumov, S. Streuli, A. Tricomi, P. Tropea, J. Troska, F. Vasey, and W. Zeuner, "CMS Technical Design Report for the Pixel Detector Upgrade," Tech. Rep. CERN-LHCC-2012-016. CMS-TDR-11, Sep 2012. Additional contacts: Jeffrey Spalding, Fermilab, Jeffrey.Spalding@cern.ch Didier Contardo, Universite Claude Bernard-Lyon I, didier.claude.contardo@cern.ch.

[26] A. Modak, "CMS Phase-1 Pixel Detector: Operational Experience, Performance and Lessons Learned," Tech. Rep. CMS-CR-2019-283, CERN, Geneva, Nov 2019.

[27] F. Monti, "CMS ECAL monitoring and calibration in LHC Run 2," Tech. Rep. CMS-CR-2018-171, CERN, Geneva, Aug 2018.

[28] N. Bartosik, "Performance of the CMS Electromagnetic Calorimeter in LHC Run2," *PoS*, vol. LeptonPhoton2019, p. 126. 4 p, 2019.

[29] M. Chadeeva and N. Lychkovskaya, "Calibration of the CMS hadron calorimeter in run 2," *Journal of Instrumentation*, vol. 13, pp. C03025–C03025, mar 2018.

[30] E. James, Y. Maravin, M. Mulders, and N. Neumeister, "Muon Identification in CMS," Tech. Rep. CMS-NOTE-2006-010, CERN, Geneva, Jan 2006.

[31] V. I. Klioukhine, D. Campi, B. Cure, A. Desirelli, S. Farinon, H. Gerwig, D. Green, J. P. Grillet, A. Herve, F. Kircher, B. Levesy, R. Lovelesss, and R. P. Smith, "3d magnetic analysis of the cms magnet," *IEEE Transactions on Applied Superconductivity*, vol. 10, no. 1, pp. 428–431, 2000.

[32] V. Klyukhin, N. Amapane, A. Ball, B. Curé, A. Gaddi, H. Gerwig, M. Mulders, V. Calvelli, A. Hervé, and R. Loveless, "Validation of the CMS Magnetic Field Map. Validation of the CMS Magnetic Field Map," *J. Supercond. Novel Magn.*, vol. 28, pp. 701–704. 7 p, May 2016.

[33] N. Pozzobon, "The CMS muon system: performance during the LHC run-2," *Journal of Instrumentation*, vol. 14, pp. C11031–C11031, nov 2019.

[34] A. Sirunyan, A. Tumasyan, W. Adam, E. Asilar, T. Bergauer, J. Brandstetter, E. Brondolin, M. Dragicevic, J. Erö, M. Flechl, and et al., "Particle-flow reconstruction and global event description with the cms detector," *Journal of Instrumentation*, vol. 12, p. P10003–P10003, Oct 2017.

[35] M. Tosi, "The CMS trigger in Run 2," Tech. Rep. CMS-CR-2017-340, CERN, Geneva, Oct 2017.

[36] L. Cadamuro, "The CMS level-1 trigger system for LHC run II," *Journal of Instrumentation*, vol. 12, pp. C03021–C03021, mar 2017.

[37] H. Sert, "CMS High Level Trigger Performance in Run 2," Tech. Rep. CMS-CR-2019-278, CERN, Geneva, Nov 2019.

[38] M. E. Peskin and D. V. Schroeder, *An introduction to quantum field theory*. Boulder, CO: Westview, 1995. Includes exercises.

[39] S. Frixione, Z. Kunszt, and A. Signer, "Three-jet cross sections to next-to-leading order," *Nuclear Physics B*, vol. 467, p. 399–442, May 1996.

[40] C. Oleari, "The powheg box," *Nuclear Physics B - Proceedings Supplements*, vol. 205-206, p. 36–41, Aug 2010.

[41] V. Hirschi and O. Mattelaer, "Automated event generation for loop-induced processes," 2015.

[42] B. Andersson, G. Gustafson, G. Ingelman, and T. Sjöstrand, "Parton fragmentation and string dynamics," *Physics Reports*, vol. 97, no. 2, pp. 31 – 145, 1983.

[43] T. Sjöstrand, S. Ask, J. R. Christiansen, R. Corke, N. Desai, P. Ilten, S. Mrenna, S. Prestel, C. O. Rasmussen, and P. Z. Skands, "An introduction to pythia 8.2," *Computer Physics Communications*, vol. 191, pp. 159 – 177, 2015.

[44] S. Agostinelli *et al.*, "Geant4—a simulation toolkit," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 506, no. 3, pp. 250 – 303, 2003.

[45] M. Hildreth, V. N. Ivanchenko, D. J. Lange, and M. J. Kortelainen, "CMS full simulation for run-2," *Journal of Physics: Conference Series*, vol. 664, p. 072022, dec 2015.

[46] V. Lefébure, S. Banerjee, and I. González, "CMS Simulation Software Using Geant4," Tech. Rep. CMS-NOTE-1999-072, CERN, Geneva, Dec 1999.

[47] K. Bloom, "Cms software and computing for lhc run 2," *Proceedings of 38th International Conference on High Energy Physics — PoS(ICHEP2016)*, Feb 2017.

[48] N. Ratnikova, C.-H. Huang, A. Sanchez-Hernandez, T. Wildish, and X. Zhang, "CMS space monitoring," *Journal of Physics: Conference Series*, vol. 513, p. 042036, jun 2014.

[49] J. Blomer, "A quantitative review of data formats for HEP analyses," *Journal of Physics: Conference Series*, vol. 1085, p. 032020, sep 2018.

[50] R. Brun and F. Rademakers, "ROOT—an object oriented data analysis framework," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 389, no. 1-2, pp. 81–86, 1997.

[51] V. Innocente, L. Silvestris, and D. Stickland, "Cms software architecture: Software framework, services and persistency in high level trigger, reconstruction and analysis," *Computer Physics Communications*, vol. 140, no. 1, pp. 31 – 44, 2001. CHEP2000.

[52] K. Rose, "Deterministic annealing for clustering, compression, classification, regression, and related optimization problems," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2210–2239, 1998.

[53] CMS Collaboration, "Description and performance of track and primary-vertex reconstruction with the CMS tracker," *Journal of Instrumentation*, vol. 9, pp. P10009–P10009, oct 2014.

[54] J. R. and, "CMS electron and photon performance at 13 TeV," *Journal of Physics: Conference Series*, vol. 1162, p. 012008, jan 2019.

[55] M. Cacciari, G. P. Salam, and G. Soyez, "The anti-ktjet clustering algorithm," *Journal of High Energy Physics*, vol. 2008, p. 063–063, Apr 2008.

[56] V. Khachatryan, A. Sirunyan, A. Tumasyan, W. Adam, E. Asilar, T. Bergauer, J. Brandstetter, E. Brondolin, M. Dragicevic, J. Erö, and et al., "Jet energy scale and resolution in the cms experiment in pp collisions at 8 tev," *Journal of Instrumentation*, vol. 12, p. P02014–P02014, Feb 2017.

[57] A. Sirunyan, A. Tumasyan, W. Adam, F. Ambrogi, E. Asilar, T. Bergauer, J. Brandstetter, E. Brondolin, M. Dragicevic, J. Erö, and et al., "Identification of heavy-flavour jets with the cms detector in pp collisions at 13 tev," *Journal of Instrumentation*, vol. 13, p. P05011–P05011, May 2018.

[58] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. J. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Józefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. G. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. A. Tucker, V. Vanhoucke, V. Vasudevan, F. B. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, "Tensorflow: Large-scale machine learning on heterogeneous distributed systems," *CoRR*, vol. abs/1603.04467, 2016.

[59] CMS Collaboration, "A deep neural network for simultaneous estimation of b jet energy and resolution," 2019.

[60] D. Bertolini, P. Harris, M. Low, and N. Tran, "Pileup per particle identification," *Journal of High Energy Physics*, vol. 2014, no. 10, p. 59, 2014.

[61] D. Abercrombie, "A search for dark matter via higgs decay using quark jet substructure," 2014.

[62] A. J. Larkoski, S. Marzani, G. Soyez, and J. Thaler, "Soft drop," *Journal of High Energy Physics*, vol. 2014, May 2014.

[63] The CMS collaboration, "Performance of the CMS missing transverse momentum reconstruction in pp data at $\sqrt{s} = 8$ TeV," *Journal of Instrumentation*, vol. 10, pp. P02006–P02006, feb 2015.

[64] A. M. Sirunyan, A. Tumasyan, W. Adam, F. Ambrogi, E. Asilar, T. Bergauer, J. Brandstetter, M. Dragicevic, J. Erö, and et al., "Search for production of higgs boson pairs in the four b quark final state using large-area jets in proton-proton collisions at s = 13 $\sqrt{s} = 13$ tev," *Journal of High Energy Physics*, vol. 2019, Jan 2019.

[65] The CMS Collaboration, *CMSSW Reference Manual*.

[66] C. Kato, "Status and prospects of STXS measurements in ATLAS and CMS," Tech. Rep. ATL-PHYS-PROC-2019-080, CERN, Geneva, Aug 2019.

[67] "Muon reconstruction performance during Run II," Jul 2019.

[68] T. C. collaboration, "Identification of b-quark jets with the cms experiment," *Journal of Instrumentation*, vol. 8, p. P04013–P04013, Apr 2013.

[69] L. Scodellaro, "b tagging in atlas and cms," 2017.

[70] J. Alwall, R. Frederix, S. Frixione, V. Hirschi, F. Maltoni, O. Mattelaer, H. S. Shao, T. Stelzer, P. Torrielli, and M. Zaro, "The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations," *JHEP*, vol. 07, p. 079, 2014.

[71] S. Kallweit, J. M. Lindert, P. Maierhöfer, S. Pozzorini, and M. Schönherr, "Nlo qcd+ew predictions for v + jets including off-shell vector-boson decays and multijet merging," *Journal of High Energy Physics*, vol. 2016, p. 1–51, Apr 2016.

[72] D. E. De Florian Sabaris, C. Grojean, F. Maltoni, C. Mariotti, A. Nikitenko, M. Pieri, P. Savard, M. Schumacher, and R. Tanaka, "Handbook of LHC Higgs cross sections: 4. Deciphering the nature of the Higgs sector," Sep 2016.

[73] "CMS Luminosity Measurements for the 2016 Data Taking Period," Tech. Rep. CMS-PAS-LUM-17-001, CERN, Geneva, 2017.

[74] "CMS luminosity measurement for the 2017 data-taking period at $\sqrt{s} = 13$ TeV," Tech. Rep. CMS-PAS-LUM-17-004, CERN, Geneva, 2018.

[75] "CMS luminosity measurement for the 2018 data-taking period at $\sqrt{s} = 13$ TeV," Tech. Rep. CMS-PAS-LUM-18-002, CERN, Geneva, 2019.

[76] A. Sirunyan *et al.*, "Cross section measurement of t-channel single top quark production in pp collisions at s=13tev," *Physics Letters B*, vol. 772, pp. 752 – 776, 2017.

[77] A. Sirunyan *et al.*, "Combination of searches for heavy resonances decaying to ww, wz, zz, wh, and zh boson pairs in proton–proton collisions at s=8 and 13 tev," *Physics Letters B*, vol. 774, pp. 533 – 558, 2017.

[78] L. Mijovic, L. Scodellaro, and M. zur Nedden, "Treatment of the correlations in b-tagging systematics in atlas and cms," tech. rep.

[79] "Performance of Deep Tagging Algorithms for Boosted Double Quark Jet Topology in Proton-Proton Collisions at 13 TeV with the Phase-0 CMS Detector," Jul 2018.

[80] R. Barlow and C. Beeston, "Fitting using finite monte carlo samples," *Computer Physics Communications*, vol. 77, no. 2, pp. 219 – 228, 1993.

[81] CMS Collaboration and others, "Documentation for the higgs combine tool."

[82] R. D. Cousins, "Generalization of chisquare goodness-of-fit test for binned data using saturated models, with application to histograms," *URL http://www. physics. ucla. edu/~ cousins/stats/cousins_ saturated. pdf*, 2013.

[83] J. de Blas, M. Cepeda, J. D'Hondt, R. Ellis, C. Grojean, B. Heinemann, F. Maltoni, A. Nisati, E. Petit, R. Rattazzi, and et al., "Higgs boson studies at future particle colliders," *Journal of High Energy Physics*, vol. 2020, Jan 2020.

[84] "Search for the Standard Model Higgs Boson Decaying to Bottom Quarks and Produced in Association with a W or a Z Boson," Tech. Rep. CMS-PAS-HIG-11-012, CERN, Geneva, 2011.

[85] N. Berger, C. Bertella, T. P. Calvet, M. Calvetti, V. Dao, M. Delmastro, M. Duehrssen-Debling, P. Francavilla, Y. Haddad, O. Kivernyk, J. M. Langford, C. Li, G. Marchiori, P. Milenovic, C. E. Pandini, E. Scott, F. J. Tackmann, K. Tackmann, L. Viliani, and M. Xiao, "Simplified template cross sections - stage 1.1," 2019.

[86] M. Aaboud, G. Aad, B. Abbott, D. C. Abbott, O. Abdinov, A. Abed Abud, D. K. Abhayasinghe, S. H. Abidi, O. S. AbouZeid, and et al., "Measurement of VH, $H \to b\bar{b}$ production as a function of the vector-boson transverse momentum in 13 TeV pp collisions with the ATLAS detector," *Journal of High Energy Physics*, vol. 2019, May 2019.

[87] G. Aad, B. Abbott, D. C. Abbott, *et al.*, "Measurement of the associated production of a Higgs boson decaying into b-quarks with a vector boson at high transverse momentum in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector," *Phys. Lett. B*, vol. 816, p. 136204. 49 p, Aug 2020.

[88] G. Aad, B. Abbott, D. C. Abbott, *et al.*, "Measurements of WH and ZH production in the $H \to b\bar{b}$ decay channel in pp collisions at 13 TeV with the ATLAS detector," *Eur. Phys. J. C*, vol. 81, p. 178. 41 p, Jul 2020.

[89] D. de Florian, C. Grojean, F. Maltoni, C. Mariotti, A. Nikitenko, M. Pieri, P. Savard, M. Schumacher, R. Tanaka, R. Aggleton, *et al.*, "Handbook of lhc higgs cross sections: 4. deciphering the nature of the higgs sector," *arXiv. org*, 2016.

[90] G. Apollinari, I. Béjar Alonso, O. Brüning, M. Lamont, and L. Rossi, "High-luminosity large hadron collider (hl-lhc) : Preliminary design report," 12 2015.

[91] T. C. S. Group, "Cepc conceptual design report: Volume 1 - accelerator," 2018.

[92] M. Benedikt, A. Blondel, P. Janot, M. Mangano, and F. Zimmermann, "Future circular colliders succeeding the lhc," *Nature Physics*, vol. 16, no. 4, pp. 402–407, 2020.

[93] L. collaboration, "Test of lepton universality in beauty-quark decays," 2021.

[94] Y. Iiyama, B. Maier, D. Abercrombie, M. Goncharov, and C. Paus, "Dynamo – handling scientific data across sites and storage media," 2020.

[95] G. Van Rossum and F. L. Drake Jr, *Python reference manual*. Centrum voor Wiskunde en Informatica Amsterdam, 1995.

[96] G. Van Rossum and F. L. Drake, *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace, 2009.

[97] E. Laure, A. Edlund, F. Pacini, P. Buncic, M. Barroso, A. Di Meglio, F. Prelz, A. Frohner, O. Mulmo, A. Krenek, *et al.*, "Programming the grid with glite," tech. rep., CERN, Geneva, Switzerland, 2006.

[98] D. Abercrombie, *Dynamo Consistency*.

[99] A. Dorigo, P. Elmer, F. Furano, and A. Hanushevsky, "XROOTD-A Highly scalable architecture for data access," *WSEAS Transactions on Computers*, vol. 1, no. 4.3, pp. 348–353, 2005.

[100] D. Abercrombie and et al., *SmartDataProjects/dynamo-consistency*.

[101] D. Abercrombie, *Site Admin Toolkit*.