Measurement of $H \to b\bar{b}$ 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

February 2021

Author .....
Department of Physics
November 21, 2020

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

Accepted by .....
Nergis Mavalvala
Associate Department Head, Physics

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

by

Daniel Robert Abercrombie

Submitted to the Department of Physics
on November 21, 2020, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

We measured $VH \to b\bar{b}$ with the CMS Detector.

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

Acknowledgments

Thanks.

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 an electron and a positron approaching each other will annihilate and produce 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 speeds, 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, given the total energy of the initial state. The sum of probabilities from different possible field interactions with particular initial conditions is the only thing we can measure. This is also only possible 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 "produc-

tion cross sections" which describe the probabilities of particular intermediate states "occurring" (even though intermediate states never exist in reality).

The main point is that if there exists some interesting particle, and it interacts with other particles, you can see an increased probability of certain initial states resulting in certain final states. This can teach the observer about the role of the interesting particle, without ever directly seeing it.

1.1 Measurement of the Higgs Cross Section

The purpose of the following document is to present the methods and 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 probability contribution the Higgs field has on processes involving 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. The vector boson is in turn produced by the collision of high energy protons.

The measurement is not complete once the intermediate state is generated. The Higgs boson can decay into a number of different particles, with bottom quarks being only one type. The bottom quarks must be identified. 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.

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 before a Higgs boson cross

section measurement can be undertaken.

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 for the better part of a century. The Standard Model has known gaps, such as missing explanations for neutrino mass, as well as the origin 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 [2] and Z [3] bosons and the Higgs boson [4, 5] 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 precise parameters cause excessive tension in that they cannot exist assuming the Standard Model is true, the discrepancies would need to be explained by a different or ammended model.

The Higgs decaying to $b\bar{b}$, or a bottom quark and bottom anti-quark, was observed in 2018 [6,7]. The measurement outlined in this document goes further in that it measures the contribution of $H \to b\bar{b}$ in associated production to final states of different energy. This is called a differential cross section, and places greater constraints on the parameters of the Standard Model. These constraints lead to more precise measurements of the parameters, and have the potential to discover discrepancies that have hitherto been missed.

1.3 Using the CMS Detector at the LHC

This measurement is only possible due to massive efforts by the scientific community. The Higgs boson is not typically generated in conditions on Earth. The Large Hadron Collider (LHC) at the European Organization for Nuclear Research (CERN) was constructed due to the efforts of thousands of scientists and engineers, as well as the funding from countries distributed all around the globe. The LHC performs the proton collisions needed.

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 FRASER. The CMS detector is a general purpose detector, which was used in the discovery of the Higgs boson. Its detection capabilities make it instrumental in a number of state of the art measurements of the Standard Model as well as the potential 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 [8].

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 are

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

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. 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.

In associated production, a Higgs boson is created through radiation from the Z and W bosons, which are also known as vector bosons. The measurement of $H \to b\bar{b}$ in associated production 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 [9]. 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.

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

Gauge bosons are produced by symmetries of the Lagrangian. The Standard Model has SU(3) $\times$ SU(2)$_L$ $\times$ U(1)$_Y$ symmetry. The SU(3) symmetry produces gluons which mediate the strong force between quarks [10]. The SU(2)$_L$ $\times$ 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 [11]. 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 [12–14]. 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) $\times$ 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 [15], 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) $\times$ 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) $\times$ U(1) symmetry back onto the Lagrangian in Equation (2.2). The derivatives must be replaced.

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

To simplify the expansion of Equation (2.2), a particular gauge, or particular doublet state, is chosen. ϕ is written to satisfy the vacuum expectation value of Equation (2.3) in the gauge that will give us the massless neutral boson known as a photon.

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

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

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

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 the 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) $\times$ 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 will produce the masses of fermions due to their coupling to the electroweak force.

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.

$$W^{\pm} = \frac{1}{\sqrt{2}} \left(W^{(1)} \mp i W^{(2)} \right) \tag{2.11}$$

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

$$\frac{1}{4} g_W^2 W_\mu^- W^{+\mu} (v+h)^2 = \frac{1}{4} g_W^2 v^2 W_\mu^- W^{+\mu} + \frac{1}{2} g_W^2 v W_\mu^- W^{+\mu} h + \frac{1}{4} g_W^2 W_\mu^- W^{+\mu} h^2 \tag{2.12}$$

The second term on the right hand side of Equation (2.12) gives us the coupling strength of a vertex with a Higgs and two W bosons.

$$g_{HWW} = \frac{1}{2} g_W^2 v = g_W m_W \tag{2.13}$$

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 $h Z_\mu Z^\mu$.

$$g_{HZZ} = \frac{1}{2} \left(g_W^2 + g'^2 \right) v = \sqrt{g_W^2 + g'^2} m_Z \tag{2.14}$$

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.

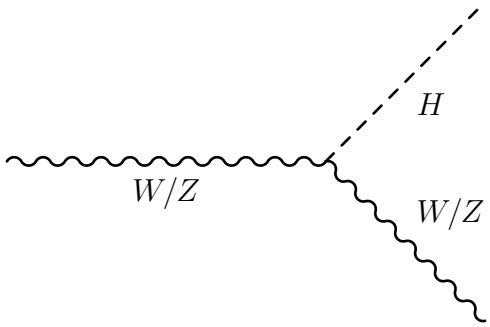

Figure 2-1: Above is the Feynman diagram for associated production. The W or Z boson radiates a Higgs boson. Both bosons later decay into particles detected by CMS.

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 interacts with stable fermions. Since the LHC is a hadron collider, considering the vector bosons' couplings with quarks would be most relevant.

Quarks couple to each other through the strong force, resulting from a SU(3) 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 parametrized by the Cabibbo-Kobayashi-Maskawa (CKM) matrix.

$$\begin{pmatrix} d' \\ s' \\ b' \end{pmatrix} = \begin{pmatrix} V_{ud} & V_{us} & V_{ub} \\ V_{cd} & V_{cs} & V_{cb} \\ V_{td} & V_{ts} & V_{tb} \end{pmatrix} \begin{pmatrix} d \\ s \\ b \end{pmatrix} \tag{2.15}$$

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

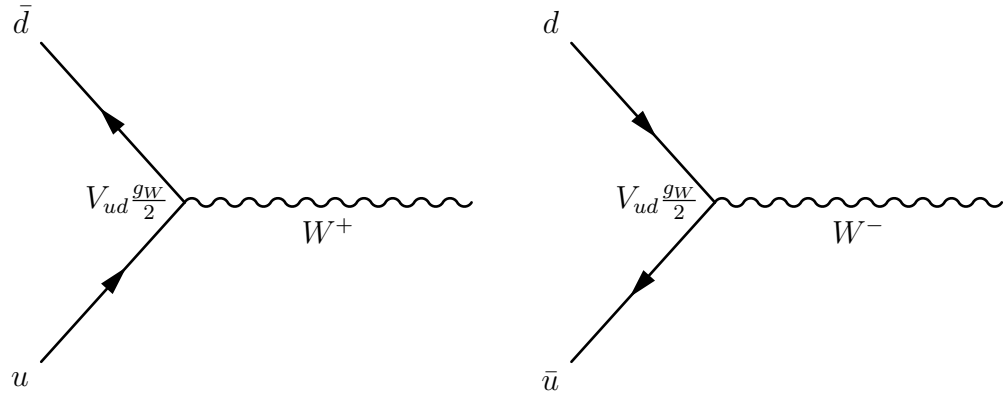

Figure 2-2: Above are diagrams for generating W^+ and W^- bosons. The u and d quarks in the diagram can be replaced with any up-type or down-type quark, respectively. The CKM matrix element would in the vertex element would be changed accordingly.

current vertex interaction.

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

The vertices for this interaction is shown in Figure 2-2 arranged in a way to show 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 labelled $\mathrm{SU}(2)_L$. From Equation (2.11), the $W^\pm$ bosons are completely made up of the $W^{(1)}$ and $W^{(2)}$ components of the $\mathrm{SU}(2)_L$, so they also only interact with left-handed fermions and right-handed anti-fermions.

Both the photon and the Z boson mix the $\mathrm{SU}(2)_L$ and $\mathrm{U}(1)_Y$ components of the Standard Model. Production of the Z boson needs to be directly understood for this measurement, but it is more straightforward to determine the strength of the Z boson couplings to left- and right-handed fermions by exploiting the symmetry of photon interactions. That is, 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. From the mixing in Equation (2.8), the photon and Z fields can be expressed as the following.

$$A_\mu = B_\mu \cos\theta_W + W_\mu^{(3)} \sin\theta_W \tag{2.16}$$

$$Z_\mu = -B_\mu \sin\theta_W + W_\mu^{(3)} \cos\theta_W \tag{2.17}$$

θ_W is known as the weak mixing angle. The relative strengths of the B and $W^{(3)}$ couplings are determined directly through lepton electro-magnetic characteristics, keeping in mind that $W^{(3)}$ only interacts with left handed particles. The following are the electro-magnetic interaction strengths of left- and right-handed electrons and neutrinos.

$$e_L: \qquad Qe = \frac{1}{2} g' Y_{e_L} \cos\theta_W - \frac{1}{2} g_W \sin\theta_W \tag{2.18}$$

$$\nu_L: \qquad 0 = \frac{1}{2} g' Y_{\nu_L} \cos\theta_W - \frac{1}{2} g_W \sin\theta_W \tag{2.19}$$

$$e_R: \qquad Qe = \frac{1}{2} g' Y_{e_R} \cos\theta_W \tag{2.20}$$

$$\nu_R: \qquad 0 = \frac{1}{2} g' Y_{\nu_R} \cos\theta_W \tag{2.21}$$

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

$$Y = 2\left(Q - I_W^{(3)}\right) \tag{2.22}$$

The following relationship also arises from these experimental constraints.

$$e = g_W \sin\theta_W = g' \cos\theta_W \tag{2.23}$$

Returning to the Z boson, from Equation (2.17), and defining

$$g_Z = \frac{e}{\sin\theta_W \cos\theta_W}, \tag{2.24}$$

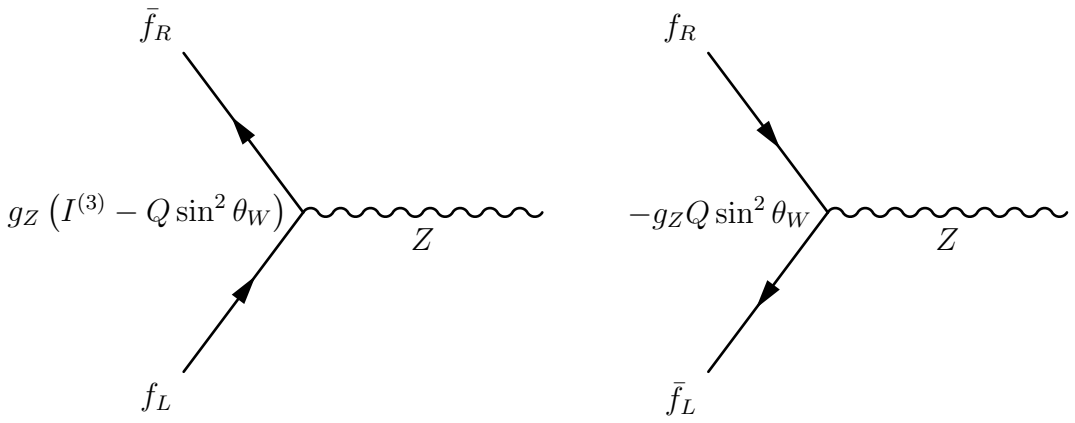

Figure 2-3: Above are diagrams for generating Z bosons. Left- and right-handed fermions are both coupled to, but with different coupling strengths.

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

$$-\frac{1}{2}g'\sin\theta_W(Y_{f_L}\bar{u}_L\gamma^\mu u_L + Y_{f_R}\bar{u}_R\gamma^\mu u_R) + I_W^{(3)}g_W\cos\theta_W\left(\bar{u}_L\gamma^\mu u_L\right) = g_Z\left(\left(I^{(3)} - Q\sin^2\theta_W\right)\bar{u}_L\gamma^\mu u_L - Q\sin^2\theta_W\bar{u}_R\gamma^\mu u_R\right) \quad (2.25)$$

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 labelled red, green, and blue, or r, g, and b. There are also anti-states for each color state, labelled $\bar{r}$, $\bar{g}$, 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

$$\psi(q\bar{q}) = \frac{1}{\sqrt{3}}(r\bar{r} + g\bar{g} + b\bar{b}), \tag{2.26}$$

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

$$\psi(qqq) = \frac{1}{\sqrt{6}}(rgb - rbg + gbrgrb + brg - bgr) \tag{2.27}$$

Baryons can also be composed of three anti-quarks, which has a state corresponding to Equation (2.27), but with anti-color. The resulting spray of hadronic particles generated from the vacuum to restore color singlets 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 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

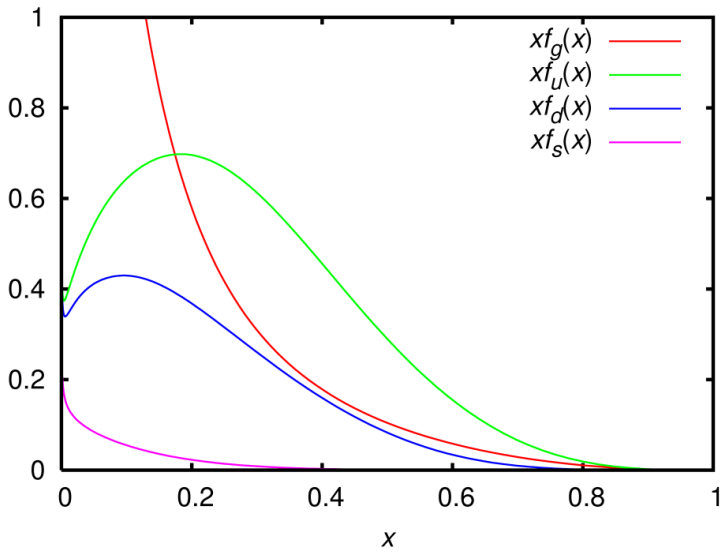

Figure 2-4: The Parton Distribution Function for protons is shown above. Most of the proton's momentum is carried by u and d quarks, but virtual s quarks as well as gluons can also interact with particles passing through the proton.

of the relativistic Breit-Wigner formula.

$$f(E) = \frac{k}{(E^2 - M^2)^2 + M^2\Gamma^2} \tag{2.28}$$

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 into quarks. However, in the hadronic environment produced at the LHC these are not the best indicators of a vector boson intermediate state. This measurement uses leptonic decays in the final state since they are easier to identify and separate from background processes.

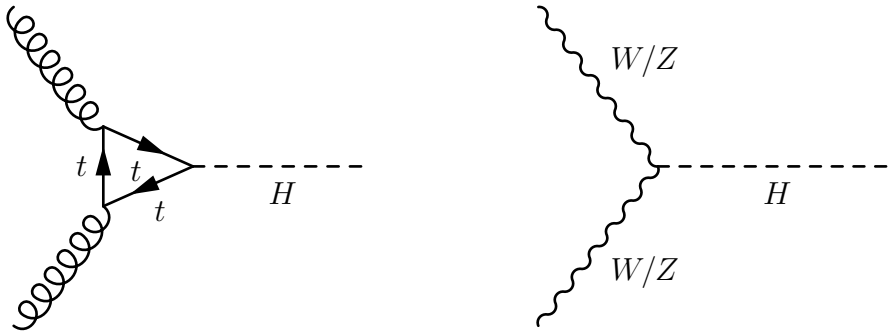

Figure 2-5: Above are the Feynman diagrams for other production mechanisms of the Higgs boson. Gluon fusion is shown on the left, and vector boson fusion is shown on the right.

There are three generations of leptons. 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 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.

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 $b\bar{b}$. Since the Higgs is a $SU(2)_L$ doublet of scalar fields, the term $-g_f(\bar{L}\phi R + \bar{R}\phi^\dagger L)$ in the Standard Model Lagrangian is invariant under $SU(2)_L \times SU(1)_Y$, where L is a left-handed fermion doublet, and

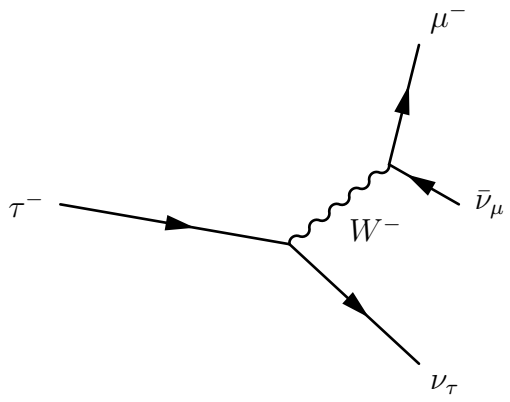

Figure 2-6: Heavier leptons can decay to lighter leptons while emitting two neutrinos. Above is an example of a decay of $\tau \to \nu_\tau \mu \bar{\nu}_\mu$.

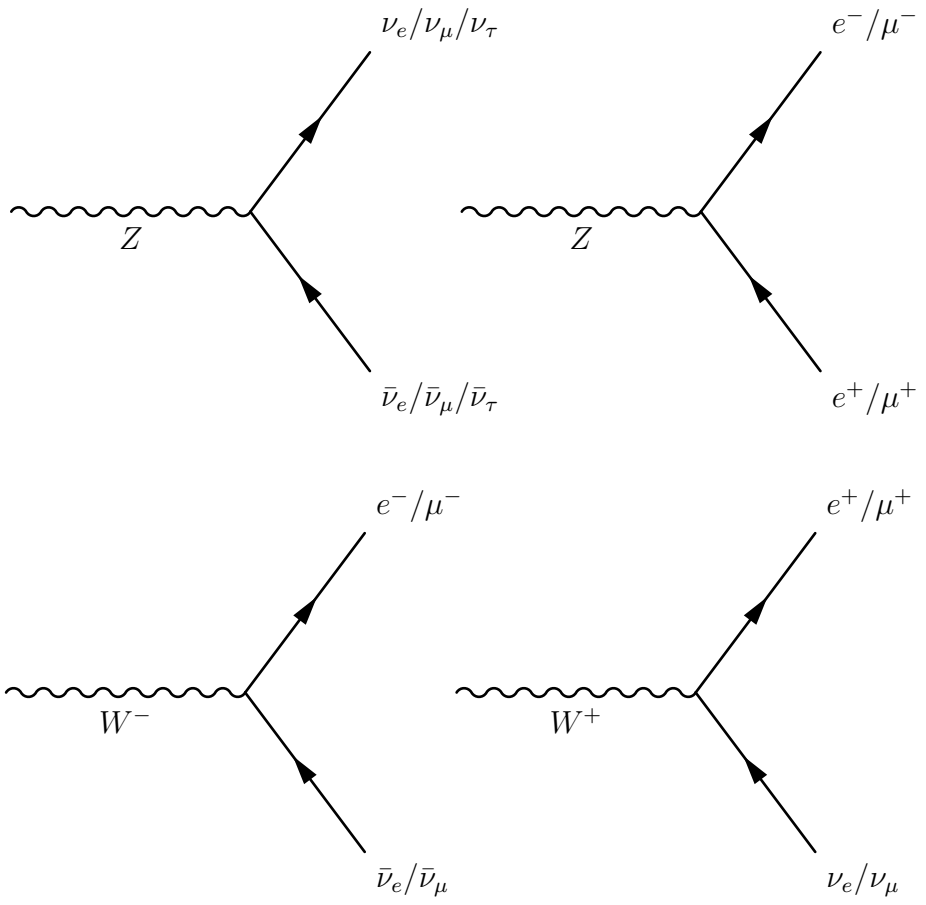

Figure 2-7: Above are the three different vector boson decays we are interested in. τ decays do also contribute to the charged lepton final states as seen by a detector, but the energy carried away by neutrinos significantly reduces those decay modes' contribution to the accepted states.

R is a right-handed singlet. If the Higgs doublet is expanded around the vacuum expectation value, as Equation (2.5), the Lagrangian term becomes the following.

$$\mathcal{L}_f = -\frac{g_f}{\sqrt{2}} v \left(\bar{f}_L f_R + \bar{f}_R f_L\right) - \frac{g_f}{\sqrt{2}} h \left(\bar{f}_L f_R + \bar{f}_R f_L\right) \tag{2.29}$$

In Equation (2.29), f refers to the lower field of the fermion's $\mathrm{SU}(2)_L$ doublet. The Lagrangian also includes terms for the upper field since the conjugate of ϕ has the same symmetries as ϕ.

The Lagrangian showing fermion-Higgs interactions in Equation (2.29) consists of two terms. Since v is constant, the first term is consistent with a fermion's mass, assuming an appropriate coupling constant.

$$g_f = \sqrt{2}\frac{m_f}{v} \tag{2.30}$$

The second term is the coupling of the fermion to the Higgs field with the same coupling constant. This is the mechanism by which the Higgs give fermions their masses, and also why the Higgs couples more strongly to massive particles. 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 GeV, and the cross section for this production drops off as the tail of a relativistic Breit-Wigner function in Equation (2.28) with a mass M of 125 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 (2.28).

As the decay mode with the highest coupling requiring no off-shell particles, the 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

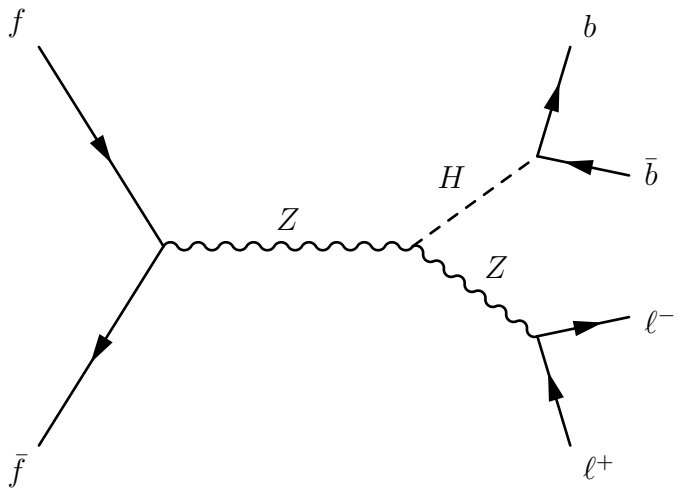

Figure 2-8: Above is the full Feynman diagram for $ZH \to \ell^+\ell^- b\bar{b}$.

diagrams for the processes being measured in this analysis. One such full diagram is shown in Figure 2-8.

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 [16]. 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 [17], 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 [18].

3.2 Detector Requirements

One configuration of possible final state particles was shown previously in Figure 2-8. There, two oppositely charged leptons and two b quarks are the end decay products. The b quarks also hadronize 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 2.15 quantifies the mixing between the different quark flavors. The value of V_{tb} is close to unity, and since the CKM matrix is unitary, V_{cb} and V_{ub} are small. This means the matrix element weak decays of the b hadrons is small. This is the only decay channel available to the lightest b hadrons, so their lifetimes are relatively long. 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 labelled 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 reconstruct 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 pixel 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 pixel detector is that it is non-destructive. Particles it detects pass through to the rest of the detector for additional measurement and identifica-

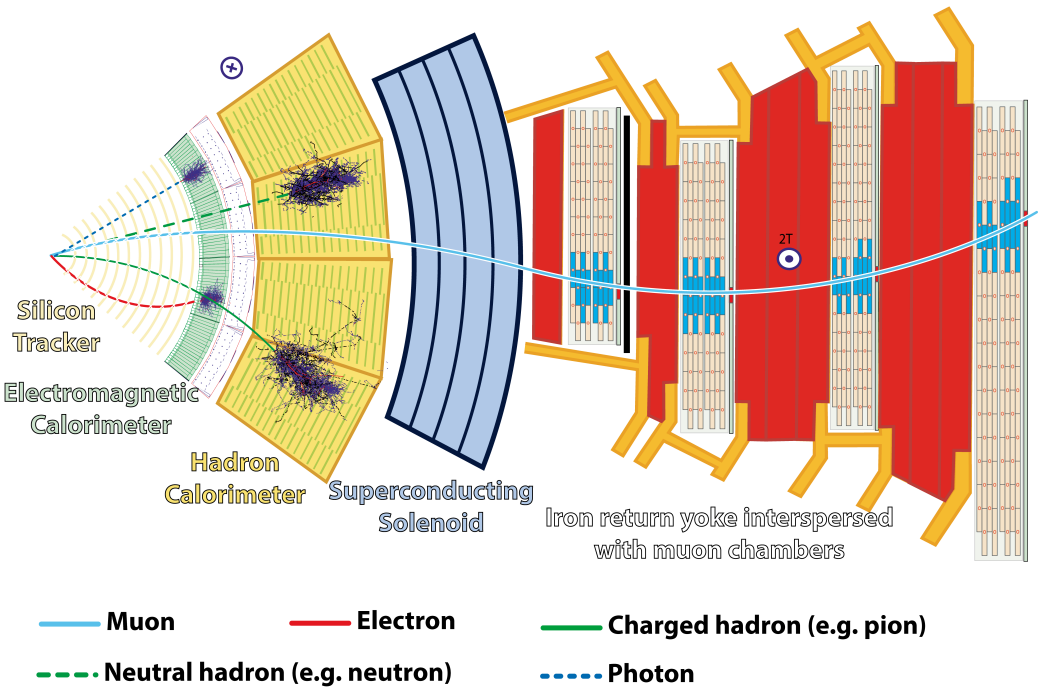

Figure 3-1: A slice of the CMS detector is shown above [19]. The four detector layers are labelled and show the penetration depths of various particles stable enough to travel a measurable distance.

tion. The next sub-detector encountered by most particles is the Electromagnetic Calorimeter, which is designed to measure the energies of photons and electrons. The next sub-detector, the Hadronic Calorimeter, measures the energies of both charged and neutral hadrons. The two calorimeters are destructive because 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

Part of the CMS acronym acknowledges the role of the solenoid magnet. 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 TeV of energy to bend enough to have a momentum resolution of 10%. Inside the solenoid, the magnetic field operates at 3.8 T, with the solenoid design being capable of achieving 4 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 kA is sent through 2168 turns over 12.9 m. The magnetic field stores 2.7 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 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 $r = 4$ cm from the interaction point, are made of hybrid pixel detectors. Each pixel has dimensions of 100×150 µm in order to achieve fine resolution of where the detected particles originiated. 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, silicon strip detectors are used. These are placed in the region that is $20 < r < 55\,\text{cm}$ from the beamline. Strip dimensions give a cell size of approximately $10\,\text{cm} \times 80\,\mu\text{m}$. 2–3% of cells are activiated during a typical bunch crossing. The outermost layers are made of larger strips with cell sizes of $25\,\text{cm} \times 180\,\mu\text{m}$. About 1% of these pixels are triggered each bunch crossing.

The active material of the silicon pixel 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 pixel 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 [20]. 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 [21].

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 $22 \times 22\,\text{mm}$. The length of each crystal is 230 mm, and the far face is slightly larger at $26 \times 26\,\text{mm}$. $PbWO_4$ has a radiation length of $\chi_0 = 8.9\,\text{mm}$ and a Moliere radius of 21 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 desireable 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 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 [22]. Resolution is measured by looking at $Z \to ee$ events. For Run 2, the barrel region of the ECAL performed with 1.6% resolution, and the other regions had a 5% resolution [23].

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 Calorimeter (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 $|\eta| < 2$ [24].

3.3.5 Muon Chambers

Muons are the most penetrative particles that CMS detects. Through the calorimeters, muons act as minimum-ionizing particles [25]. 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 [26, 27], allowing the momentum of the muons to be extracted from the curvature of their trajectory. Layers of muon chambers act much like the pixel 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 [28].

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 is taken largely from reference [29].

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 pixel 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 travelling 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 pixel detector. 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 pixel 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 elec-

tromagnetic 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 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 pixel 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 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 pixel 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 uninteresting

data. 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 [30].

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 [31].

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 [32].

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 2-8 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 $VH \to b\bar{b}$.

- 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 2-8, 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 enviroment 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 will 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 D. 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 [33].

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

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 calcuation. 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 [34] and MadGraph5 [35].

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 [36] as implemented in `PYTHIA8` [37].

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` [38]. The full CMS detector geometry is maintained within CMSSW [39] using a framework written in the Unified Modelling Language [40]. 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 [41]. Together they also provide around 100 Petabytes of space [42]. 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. These n-tuples are created and read by the CMS Software (CMSSW) [43].

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 decription of physics processes described thus far. First, objects are defined in terms of variables and particle candidates provided by the detector reconstruction algorithms. Then selection requirements based on these objects used to reject events entirely from the analysis are given. After that, selection requirements used to classify events into different decay channels of the vector boson are specified. 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. For objects, passing the loose selection often means that the object is defined well enough to veto events for inclusion in categories 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.

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.

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

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.

$$\Delta R = \sqrt{\Delta\eta^2 + \Delta\phi^2} \tag{4.2}$$

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

$$p_T^{\text{PU}} = 0.5 \times \sum p_T^{\text{PU,charged}} \tag{4.3}$$

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 [44], using the closest approach of tracks to the beamline [45]. 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 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.

- The muon must have a relatively high energy 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 cut 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 cuts 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 $Z \to \mu\mu$, one muon only needs to satisfy the slightly looser selection of $p_T > 15\,\text{GeV}$, as long as the other has $p_T > 25\,\text{GeV}$.
- 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 one hit on a pixel.
- 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 accepted by all members of 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 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 cut of $I < 0.4$.

To optimize the electron selection, electrons are identified with the aid of an MVA [46]. 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 also must pass the following cuts.

- 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 either have $|\eta| < 1.4442$ or $|\eta| > 1.5660$.
- For electrons with $|\eta| < 1.4442$:
  - 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 $|\eta| > 1.5660$:
  - 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 [47] using the jet clustering parameter $R = 0.4$. Due to factors like pileup and imperfect detector response, the energy of the reconstructed jets are corrected [48].

Loose jet cuts, based on the constituents, are applied to remove jets constructed from detector noise. Jets that get a significant fraction of their energy from pileup

Table 4.1: The minimal cut 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 mis-tag rates.

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

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 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 consituents. The jets must satisfy $p_T > 25\,\text{GeV}$ for the zero and one lepton signatures. The two lepton signature from $Z(\ell\ell)H$ is cleaner, and looser jet selection criteria of $p_T > 20\,\text{GeV}$ 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 can be considered in a deep neural network (DNN) called Deep Combined Secondary Vertex (DeepCSV) designed to identify b jets [49]. The output of DeepCSV has three working points that are defined based on the amount of false positives that can be expected in a collection of jets passing the cut. The specific values are different for each year of operation. The detector and collision conditions change, and a separate model is trained each 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 prediction on the amount of energy carried away by undetected neutrinos is made. A Deep Neural Network (DNN) is trained in Tensorflow [50], which is designed to improve the energy measurement and resolution of all b jets for CMS [51]. 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 consistuents
- the highest p_T of charged hadron consituents
- 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.

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 transistion from resolved jets to boosted single jets, larger jets, labelled 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 the jet clustering parameter $R = 0.8$. 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 $p_T > 250\,\text{GeV}$.

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 particle whose daughters make up most of the jet constituents. A number of grooming algorithms, which are designed to remove particles from pileup from the jets, were considered within CMS [52]. The soft drop algorithm [53] was chosen as the standard in the experiment and is used in this analysis. 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, which is why a second requirement on fat jets considered for the analysis is $m_{\text{SD}} > 50\,\text{GeV}$.

4.1.7 Missing Transverse Energy

Missing transverse energy, which is actually the missing transverse momentum, also labelled E_T^{miss} or MET, is a vector that takes advantage of the fact that momentum transverse to the beamline is conserved. 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 [54]. The resulting magnitude and direction is a proxy for the transverse momentum of any neutrinos in the event. However, 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.

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 [47] 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 can be reconstructed precisely. A kinematic fit is performed to constrain the momenta of the two leptons by requiring their combined mass to match

Table 4.2: The value of σ/μ for each fitted di-jet mass peak is shown below. Figure 4-1 shows the mass peaks that were fit to fill the inclusive column.

	Low p_T	High p_T	incl.
No R.	0.157	0.133	0.150
Reg.	0.134	0.121	0.130
Fit	0.129	0.112	0.124

the Z boson mass. After this constraint is applied, the tranvserve momentum of the dijet system, along with ISR jets, is balanced with the transverse momentum of the dilepton system. The fit is performed by minimizing the chi-squared of the kinematic system that meets these 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 recoil jets is the same as that used by the $HH \to b\bar{b}b\bar{b}$ analysis from CMS [55]. The mass of the Z boson is given a Gaussian uncertainty of 5 GeV, and it is assumed that the MET in the event is 0. 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` [56]. Only the resulting energies of the b-jets, which are otherwise relatively loosely constrained, are used for analyzing the two-lepton regions. The improvement of the di-jet mass in the signal sample as a result of the kinematic fit are shown in Figure 4-1 and Table 4.2.

4.2 Backgrounds to the Analysis

In order to effectively measure Higgs production, we need to be able to accurately estimate other events that end up in our selection. For example, the final state for the two lepton decay in Figure 2-8 can also be achieved by a Drell-Yan process radiating 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,

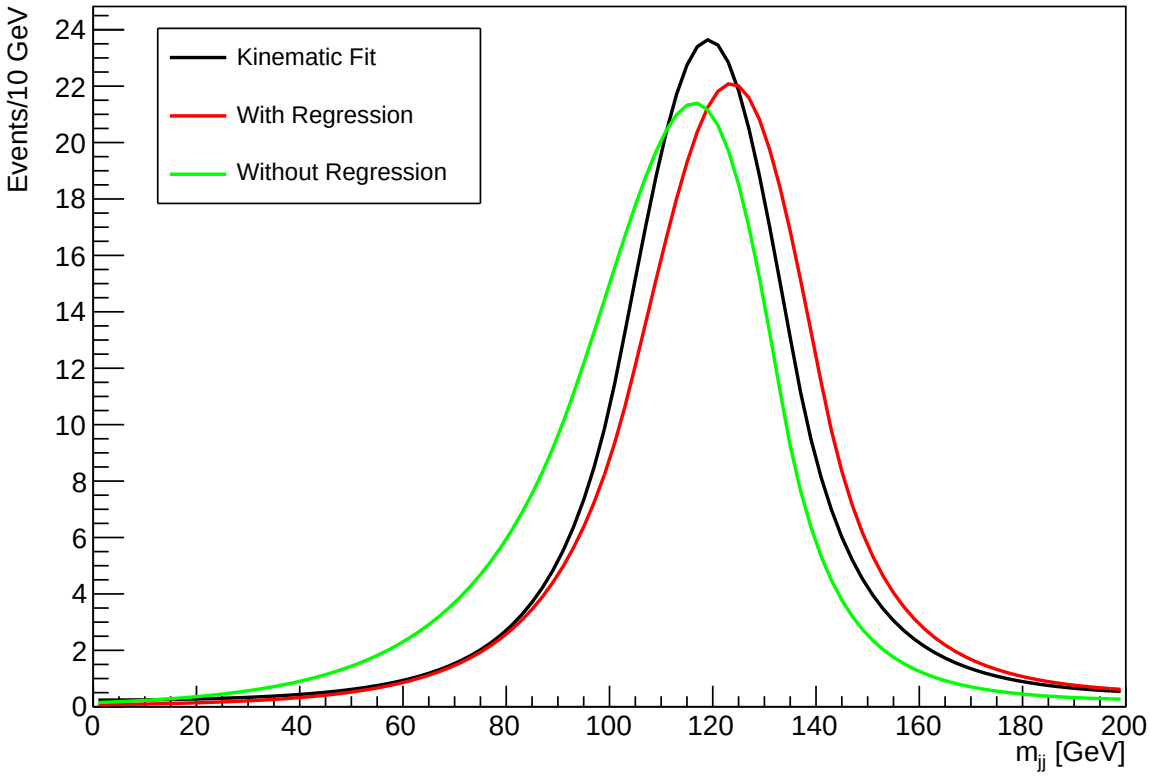

Figure 4-1: The Higgs di-jet mass in the 2-lepton signal samples is shown above. Peaks from the raw jet, the regressed jet energy, and the kinematic fit are compared.

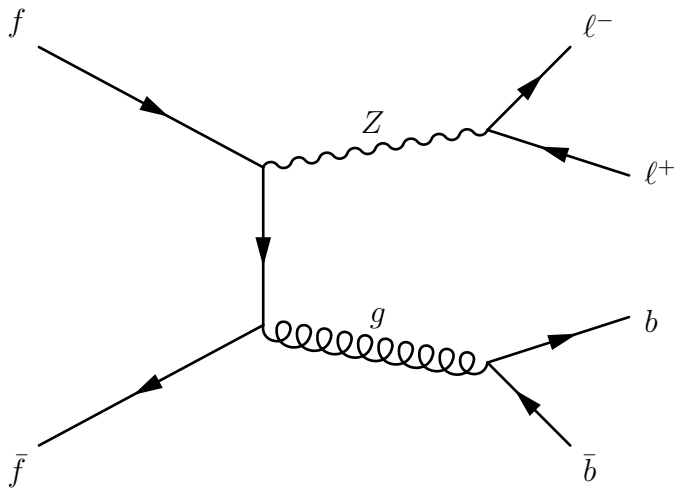

Figure 4-2: Above is the Feynman diagram matching the two lepton final state coming from Drell-Yan and jets.

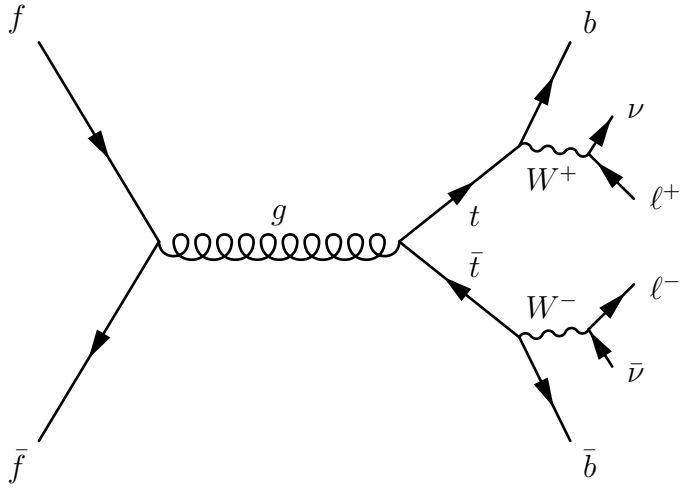

Figure 4-3: Above is the Feynman diagram matching the two lepton final state coming from fully leptonic $t\bar{t}$ decay. In events with little energy carried away by the neutrinos, this can appear to be the same as the two-lepton signal process.

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 signal 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 travelling out of the detector without being

observed. For the zero-lepton channel, the Drell Yan process is instead replaced with $Z \to \nu\bar{\nu}$. 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, di-boson, QCD, and single top backgrounds can also contribute. This simplifies the methods needed to generate Monte Carlo samples, since the decay mode of each intermediate particle can be randomly selected for each trial.

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 can be made for the phase space.

4.3 Simplified Template Cross Section Bins

The measurement performed in this analysis is a differential cross section of Higgs production. This is done with a Simplified Template Cross Section (STXS) measurement [57], where the Higgs boson production cross sections are measured in multiple kinematic bins. This allowed for sensitivity to new physics with reduced model dependence. 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 $p_{T,V} > 250\,\text{GeV}$.

4.4 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. A summary of cuts for each region in each channel is given in Table 4.3. A few channel- or region-specific cuts are left out, but are described in the appropriate sections.

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 events with large values of MET. Slightly different triggers are used for each year of

Table 4.3: Below is a summary of common cuts 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.

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	–	–

Run 2 of the LHC, and they are listed in Table 4.4. The efficiency of each trigger or combination of triggers increases as a function of MET, and that efficiency plateaus at 170 GeV. Therefore, the MET for the event must be larger than 170 GeV. 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 4-4. To handle this, all events with $-1.86 < \phi_{\text{MET}} < -0.7$ 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 cuts are similar for all of the channels. For the Higgs decay, Two jets are selected: the jet with the highest b-tag score with $p_T > 60$ GeV, and the jet with the highest or second

Table 4.4: Below are the trigger paths used for all of 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

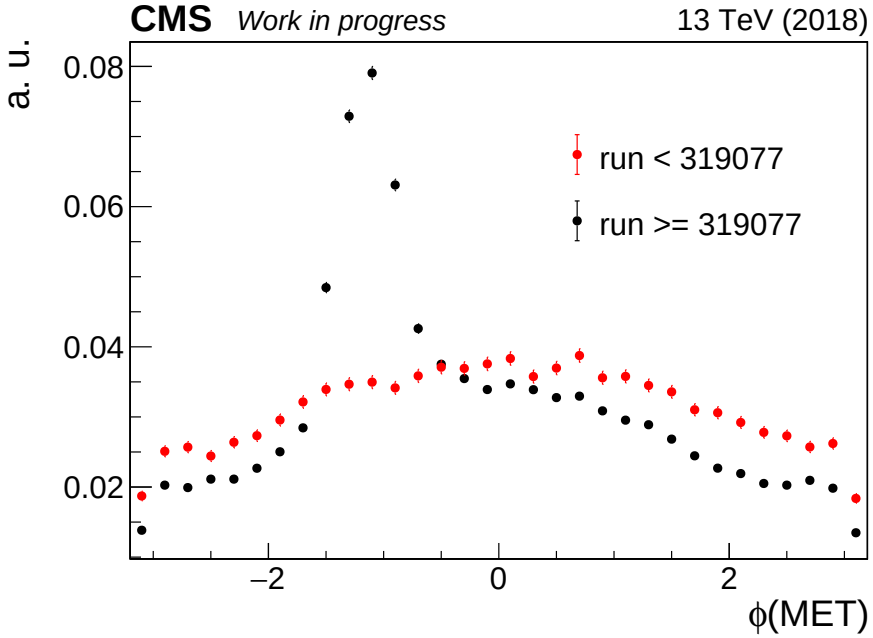

Figure 4-4: Above compares the MET ϕ distributions before and after shutting off problematic HCAL modules during for run 319077. The excess of events in the region $-1.86 < \phi_{\text{MET}} < -0.7$ is caused by mis-measuring the momenta of forward jets in that region.

highest (if the first jet is the highest) b-tag score with $p_T > 35\,\text{GeV}$. Both of these p_T values are evaluated after applying the b-jet energy regression. The di-jet mass is required to be less than 500 GeV and to satisfy $p_T > 120\,\text{GeV}$. Also the di-jet system is selected to be back-to-back with the MET by requiring $\Delta\phi(\text{MET}, jj) > 2.0$. To reduce additional backgrounds, 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 $\Delta\phi(\text{MET}, j) > 0.5$.

The signal region is additionally defined by requiring 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 for the appropriate year, 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 only one jet with $p_T > 30\,\text{GeV}$ is allows in addition to the selected b-jets. Finally, the PF MET and Track MET must have good agreement with $\Delta\phi(\text{MET}, \text{trkMET}) < 0.5$.

There are three background processes that are present in the signal selection. Two processes that need to be treated separately are both when a $Z \to \nu\bar{\nu}$ occurs recoiling off of jets. In one processes, the recoiling jets are b-jets, and in the other, the jets are light jets. Since the fraction of actual $Z \to \nu\bar{\nu}$ events containing b-jets is not well known, we want to scale them separately. The third background process that needs to be separately measured in this phase space is a $t\bar{t}$ semi-leptonic decay where the lepton falls outside of the detector acceptance. This type of background would also result 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. 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 $\Delta\phi(\text{MET}, \text{trkMET})$ 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. 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 cuts 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 0-lepton channel. There is a slightly tighter cut on the di-jet direction of $\Delta\phi(\text{MET}, jj) > 2.5$.

In the signal region, the additional cuts 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 aside from the adjusted common cuts listed above is a lack of dependence on the Track MET. 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 cut 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. They must have an invariant mass satisfying $75\,\text{GeV} < m_{\ell\ell} < 105\,\text{GeV}$. This process is clean enough to relax the kinematic cuts on the selected b-jets even further than the relaxed cuts of the 1-lepton channel. The selected b-jets only need to have a regressed $p_T > 20\,\text{GeV}$, and the di-jet system only needs $p_T > 50\,\text{GeV}$. There are no cuts on the number of additional, outside of categorization for STXS bins. The di-jet system still needs to satisfy the tighter cut of $\Delta\phi(jj, V) > 2.5$.

The signal region uses the same b-tag and mass window cuts as the other two channels. For the Z + heavy flavor control region, the di-lepton mass cut is narrowed to $85\,\text{GeV} < m_{\ell\ell} < 97\,\text{GeV}$ in order to cut out $t\bar{t}$, and the usual Higgs mass veto is applied. The MET is also required to be low with MET $< 60\,\text{GeV}$ for the Z + heavy region, but no other. For the Z + light flavor region, purity is achieved by making the b tagging requirement that both selected jets fail the loose working point. 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 $10\,\text{GeV} < m_{\ell\ell} < 75\,\text{GeV}$ or $m_{\ell\ell} > 120\,\text{GeV}$.

4.5 Boosted Analysis Selection

When the Higgs has very high p_T, the jet clustering algorithms can find both daughter particles as being part of a single jet. The boosted analysis only targets the STXS bins with $p_{T,V} > 250\,\text{GeV}$. The selection differs primarily in the fact that a single fat jet which passes a double b-tag cut [49] 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. The output is decorrelated with mass, allowing for the jet mass to be used to generate control regions, as is done in the resolved analysis.

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 addi-

tional b-jet, the jet must pass the DeepCSV medium working point, have a regressed $p_T > 25\,\text{GeV}$, and be outside of the selected fat jet so that $\Delta R(j, fj) > 0.8$.

4.5.1 0 Leptons

As in Section 4.4.1, the lack of measured leptons is caused by the Z boson decaying to neutrinos, so the 0-lepton channel has high MET. A requirement of MET $> 250\,\text{GeV}$ is applied, balancing out the p_T requirement of the fat jet. As for the resolved analysis, any extra leptons leads to the event not being considered for the 0-lepton channel. To remove QCD background for all regions, the same cut from the resolved analysis of $\Delta\phi(\text{MET}, j) > 0.5$ for all jets with $p_T > 30\,\text{GeV}$ 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 $90 < m_{\text{SD}} < 150\,\text{GeV}$. No additional jets outside of the fatjet are allowed in the event. The control regions are the same as for the resovled analysis: Z + heavy flavor, Z + light flavor, and $t\bar{t}$. For the Z + heavy flavor control region, the mass cut is changed to instead veto the Higgs mass window. For the Z + light flavor, there is no mass requirement outside of the $m_{\text{SD}} > 50\,\text{GeV}$ 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 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 an invariant mass near the Z boson mass for the signal region and the two Z + jets regions. 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 section and the next is orthogonal to all other regions. To prevent any statistical bias in the analysis, both simulated and measured events must not be double counted. This is most relevant when comparing the resolved and boosted channels since the same PF Candidates are reused to define two different kinds of jet collections, making it harder to enforce orthogonality with a single cut. The overlapping events were given priority to boosted or resolved depending on a study done to optimize each STXS bin. Any events that are in both selections are assigned to the resolved analysis, unless the event is in a resolved control region and a boosted signal region.

Chapter 5

Analysis Results

5.1 Run 2 Data Collection

The CMS detector collected proton-proton collision data at $\sqrt{s} = 13\,\text{TeV}$ over three years during Run 2 of the LHC. In 2016, CMS collected $37.80\,\text{fb}^{-1}$ of data [58]. In 2017, $44.98\,\text{fb}^{-1}$ of data was collected [59]. In 2018, CMS collected $63.67\,\text{fb}^{-1}$ of collision data [60].

5.2 Corrections and Uncertainties

Despite efforts to simulate LHC collisions as accurately as possible, a number of differences in the distributions predicted by MC and present in data arise. This is a result of not being able to predict the beam conditions exactly and not being able to predict the calibration accurately. This is made more difficult since the detector degrades in its high radiation environment.

Corrections are made to the simulation by re-weighting based on the pileup and by scaling the predicted energies based on particle type. Most of the corrections are derived by dedicated groups that provide the entire CMS collaboration with proper calibrations. This analysis does depend in particular on predictions of b-jet energies, which are not corrected centrally. Therefore this section includes a description of how that correction is derived.

5.2.1 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. These are measured via the Tag and Probe method, selecting muons that reconstruct the Z boson resonance. A scale factor is then applied to MC to match the data efficiency.

Each of the efficiency measurements and scale factors are binned in muon p_T and η. The p_T bins are $[20, 25, 30, 40, 50, 60, \infty)$ in GeV. Bins in $|\eta|$ are delimited at $[0, 0.9, 1.2, 2.1, 2.4]$.

5.2.2 Electrons

5.2.3 Jets and MET

5.2.4 *b*-Jet Energy Correction

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 $p_T = 0\,\text{GeV}$.

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
  $p_{T,\ell\ell} > 100\,\text{GeV}$ and $71\,\text{GeV} < m_{\ell\ell} < 111\,\text{GeV}$.
- 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 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. 5-1. From each plot, the mean (μ) and the standard deviation (σ) are extracted. σ/μ is fit as a function of α.

$$f(\alpha) = (m \times \alpha) \oplus b \times (1 + c_k \times \alpha) \tag{5.1}$$

c_k 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. 5-2. Smearing is done by scaling difference between $p_{T,reco}$ and $p_{T,gen}$ by b_{data}/b_{MC}. This causes the post-smearing fits

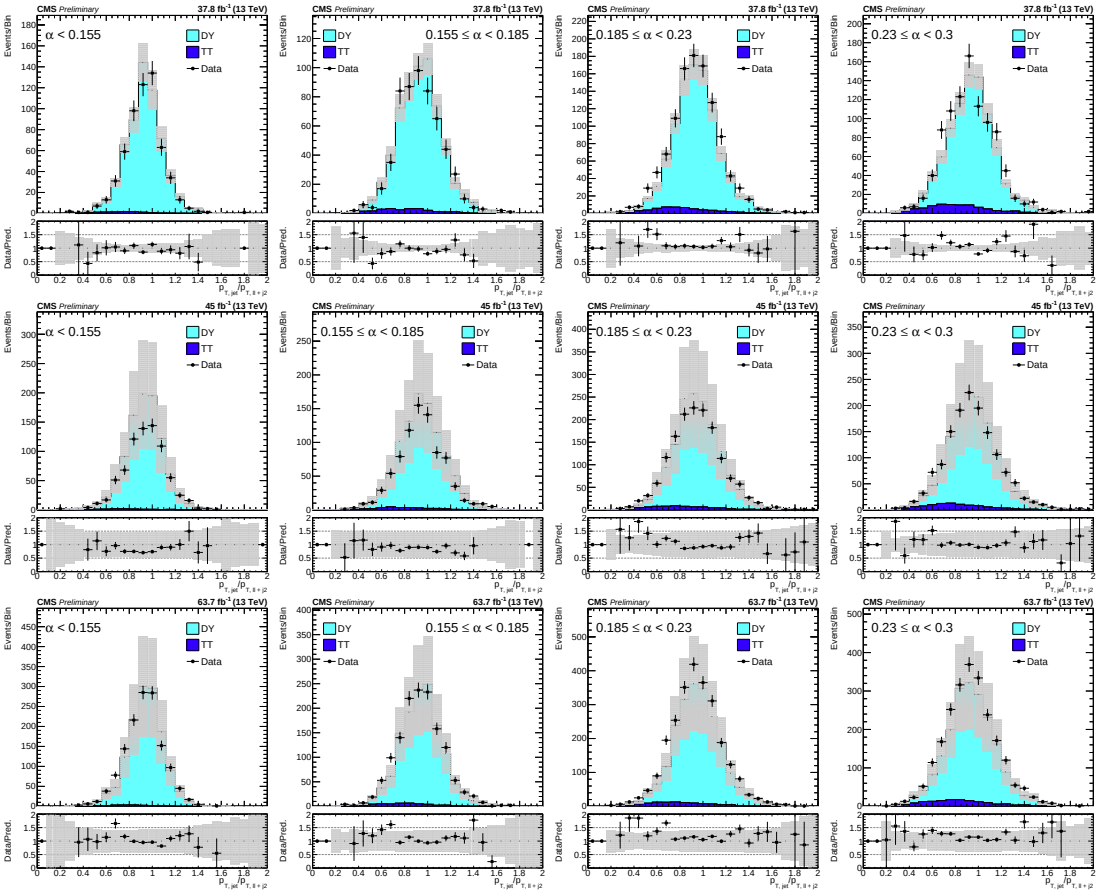

Figure 5-1: The histograms of reponse for each event are shown above. The top row shows 2016, the middle shows 2017, and the bottom row shows 2018 histograms.

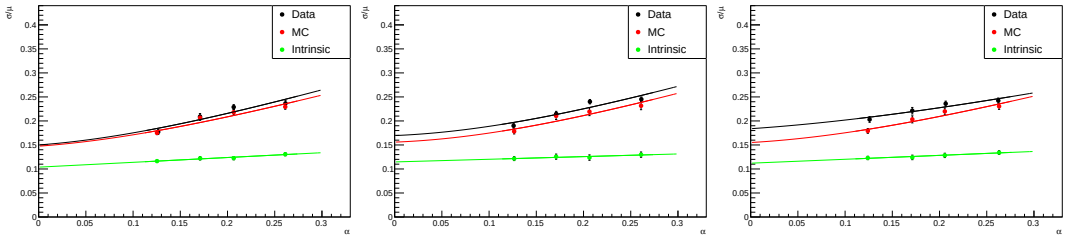

Figure 5-2: The fits to Data, MC, and intrinsic resolutions are shown. From left to right are the fits for 2016, 2017, and 2018.

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

Year	Scaling	Smearing
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

to agree at $\alpha = 0$. Uncertainties are extracted from the fit uncertainties of b for data and MC. The resulting smearings are in Table 5.1.

5.3 Theoretical Uncertainties

5.4 Multivariate Discriminator

In each STXS bin, a mutlivariate discriminator is plotted 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. Using fewer discriminating variables in the boosted events leads to this difference in architecture.

5.4.1 Resolved DNN

The DNN classifier for distinguishing background and signal events is prepared using Keras with a Tensorflow backend 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

64. The final layer is a softmax layer with the target of predicting the probability of each event belonging to a particular class.

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 5.2. All variables that are affected by the kinematic fit in the 2-lepton region use the values calculated by the fit.

5.4.2 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 multijet 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 recontructed fat jet
- Transverse momentum of the fat jet
- Transverse momentum of the recontructed vector boson
- Number of soft-track jets with $p_T > 5\,\text{GeV}$
- Double b-tagger output node for boosted jets

All of these variables were all used in the 0-, 1-, and 2-lepton regions, even though the regions were trained separately.

5.5 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 [61] as

Table 5.2: 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$_{\text{max}}$ WP	1, 2, or 3 if higher b-tag discriminate meets the tight, medium, or loose working point respectively	✓	✓	✓
b-tag$_{\text{min}}$ 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	✓	✓	
$p_{T,\text{lead}}$	Leading jet p_T	✓	✓	✓
$p_{T,\text{trail}}$	Trailing jet p_T	✓	✓	✓
SA5	Number of soft jets, $p_T > 5\,\text{GeV}$	✓	✓	✓
N_{aj}	Number of additional jets	✓	✓	
b-tag$_{\text{add}}$	Maximum b-tag of additional jets	✓		
$p_{T,\text{add}}$	Maximum p_T of additional jets	✓		
$\Delta\phi(\text{add},\text{MET})$	Azimuthal angle between additional jet and MET	✓		
$\Delta\phi(\ell,\text{MET})$	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			✓

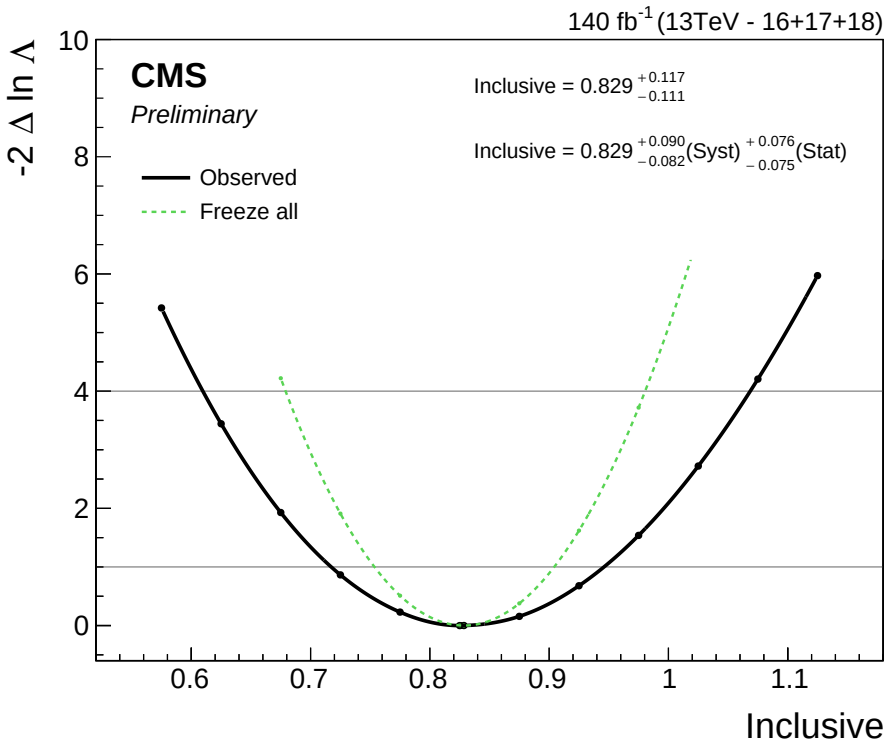

Figure 5-3: The likelihood scan of the inclusive signal strength of the VZ cross check analysis.

part of `CMSSW_10_2_13` [56]. It is commonly used in searches of new physics processes within the CMS collaboration. Part of the method for discovery involves finding the most likely values for all nuisance parameters and signal strength simultaneously. This maximum likelihood feature is used for the measurement of signal strength in the STXS bins.

In addition to the most likely nuisance parameter and scale factor values, the output of `combine` includes likelihood values at multiple signal strengths with and without variation of nuisance parameters from their most likely values. This allows separation of systematic and statistical uncertainties from the global fit. The `combine` tool also evaulated the impacts of each individual nuisance parameter on the likelihood by varying them individually.

5.5.1 *VZ* Cross Check Analysis

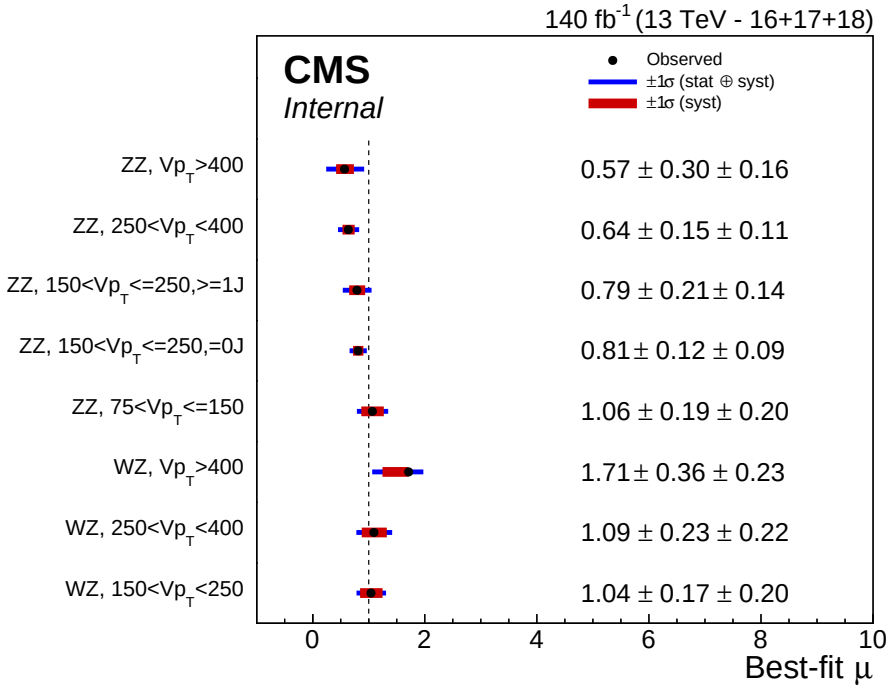

Figure 5-4: The measured most likely values of all STXS bins in the VZ cross check analysis.

Chapter 6

Conclusions

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 [62] 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 [63, 64] and available through the Python Package Index (PyPI) as `dynamo-consistency` and `workflowwebtools`.

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 chosen incorrectly for production jobs that assume the presence of a local file. Last disk copies may also be missing, causing a 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 behaviour is controlled via a simple JSON configuration file, with options for site selection, passed via command line arguments. This allows Dynamo to run separate schedules for differing site architectures.

Because Dynamo runs in a heterogenous computing environment, different sites need to be listed remotely using different methods. Currently implemented are listings using XRootD Python bindings, the `gfal-ls` CLI [65], and a `xrdfs` subshell. These listers are easily extensible in Python, allowing for new site architectures to be checked by 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, with the exception of a few configurable filters. 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

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 webpage. 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 [66].

A.1.1 Installation

Dynamo Consistency requires the XRootD [67] 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 [68]. 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.

Dynamo Consistency only interacts with Dynamo at two points during a check. First, it gets a listing of what should be at a site. The next time it interacts with Dynamo is at the end when it reports results.

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 deletion requests.

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 schemas 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 [**?**], 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. `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 subdirectories too. This is to speed up the file tree comparison, shown in Figure A-1

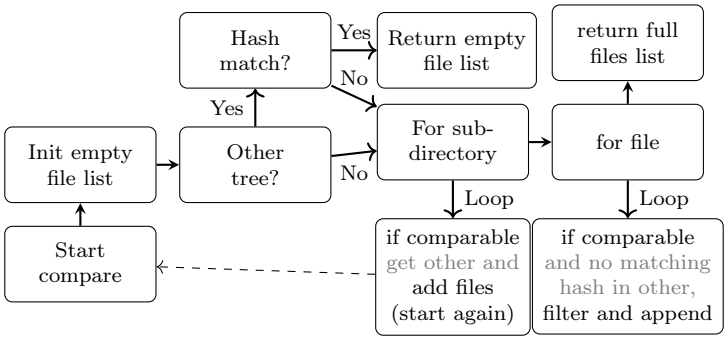

Figure A-1: Comparison algorithm

A.1.3 Remote Listing

The remote listing is equally flexible. The factory function `dynamo_consistency.backend.get_listers()` reads the configuration file to determine the type of lister for a 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 listed transparently:

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

This takes much longer than the std,std-refInventory Listing, since every directory of the site needs to be queried. The 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 each queue is shown below.

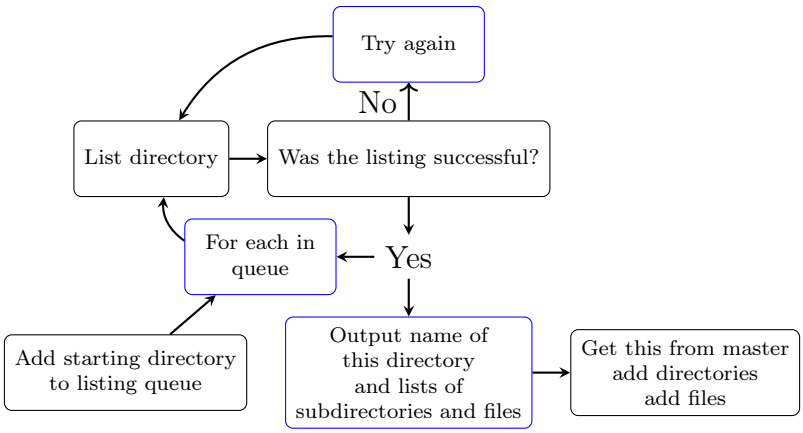

Figure A-2: Listing algorithm. TODO: Make better colors and words and stuff

A.1.4 Executables

A list of some of the executables installed with the package is given below. The usages for each executable are also given in order to show the flexibility that the overall system has.

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.
```

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's 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.

set-status

This script changes the status of a site on the summary webpage. 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 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

A configuration file should be created before pointing to it, like above. The configuration file for Site 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 before `dynamo-consistency` tries to run a check. 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 [69].
- **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. If the prefix is not set for a site, and it fails to list **RootPath**, it uses the default `/cms`.
- **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 live

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 production script, located at `dynamo_consistency/prod/compare.py` at the time of writing, goes through the following steps for each site.

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 webpage'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

Appendix B

Physics Calculations

Appendix C

Data Format

Appendix D

Generator Parameters

Appendix E

Data Card

Bibliography

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

[2] 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.

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

[4] 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.

[5] G. Aad *et al.*, "Combined search for the Standard Model Higgs boson in pp collisions at $\sqrt{s}$=7 TeV with the ATLAS detector,"

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

[7] 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.

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

[9] 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.

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

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

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

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

[14] 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.

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

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

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

[18] 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.

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

[20] 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.

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

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

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

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

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

[26] 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.

[27] 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.

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

[29] 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.

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

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

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

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

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

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

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

[37] 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.

[38] 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.

[39] 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.

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

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

[42] 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.

[43] 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.

[44] 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.

[45] 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.

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

[47] 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.

[48] 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.

[49] 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.

[50] 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.

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

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

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

[54] 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.

[55] 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.

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

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

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

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

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

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

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

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

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

[65] 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.

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

[67] 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.

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

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