SEARCH FOR THE 125 GeV STANDARD MODEL HIGGS BOSON DECAYING VIA $H \rightarrow WW \rightarrow l\nu jj$ AT $\sqrt{s} = 8$ TeV

A Dissertation

by

ALEXX S. PERLOFF

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Ricardo Eusebi
Committee Members,	Alexei Safonov
	Bhaskar Dutta
	Sherry J. Yennello
Head of Department,	Grigory Rogachev

May 2018

Major Subject: Physics

ABSTRACT

The Higgs boson discovery was announced on July 4th, 2012. It was measured to have a mass of 125.7 ± 0.3 (stat) ± 0.3 (syst) GeV and since then boson has been seen in many decay paths, including the $\mathrm{H}\rightarrow\gamma\gamma$, $\mathrm{H}\rightarrow\mathrm{ZZ}\rightarrow4\mathrm{l}$, $\mathrm{H}\rightarrow\tau\tau$, and $\mathrm{H}\rightarrow\mathrm{W}^+\mathrm{W}^-\rightarrow\mathrm{l}\nu\mathrm{l}\nu$ channels. However, no one has looked for the boson at this mass using the $\mathrm{H}\rightarrow\mathrm{W}^+\mathrm{W}^-\rightarrow\mathrm{l}\nu\mathrm{jj}$ decay channel. This dissertation presents a search for the $\sim$125 GeV Higgs in semi-leptonic W decays using both traditional kinematically discriminating variables as well as a matrix element technique. The data for this analysis was collected in 2012 by the Compact Muon Solenoid (CMS) experiment at the Large Hadron Collider (LHC) and amounts to 19.7 fb^{-1} of proton-proton collisions at a center of mass energy of 8 TeV. Although this analysis presents a step forward in complexity, we were still not able to see a significant excess above the standard model background prediction. However, we were able to set an upper limit of 5.4 on $\sigma/\sigma_{\mathrm{SM}}$ at the 95% confidence level for the semi-leptonic W decay of the Higgs boson. These represent some of the first such limits recorded.

DEDICATION

To my parents and my clone (brother).

ACKNOWLEDGMENTS

It's an interesting thing taking stock of seven years of one's life. To be so grateful to so many people, but be utterly unable to convey the depth of one's gratitude. Nevertheless, that is just how I feel about my time in graduate school.

It has been a true honor to work with my advisor, Professor Ricardo Eusebi. I remember meeting Ricardo at Fermilab when I was an undergraduate summer student, not realizing that this would be a relationship that would enrich my life forever. I have learned an innumerable amount from Ricardo about an ever widening range of topics, from the intricacies of high energy physics, to the navigation of Texas A&M University regulation, to the benefits of making a solid plan before execution. I am a better student and physicist because of him.

I must also thank the wider TAMU collider physics group for their support and advise. Whether it was a question on physics, career advice, or time on a computing cluster, this group was always there for me. I have spent a significant amount of time at Texas A&M, so this list is too long to enumerate, but I want each and every one of these people to know that I am forever grateful to them. I am also thankful for our analysis collaborators Professor Chris Neu and Joseph Goodell from the University of Virginia for their invaluable contributions to this analysis. It is a better and more complete analysis because of them. I am also thankful for my fellow physics graduate students. This group of amazing individuals made my time in College Station that much more enjoyable and helped me to grow as a person outside of physics.

Besides my dissertation analysis I have spent many years working for the JetMET/JERC group within CMS. In that time I have worked with countless individuals who should know what a pleasure it was to work alongside them. There are to many names to list here, but there are some which are deserving of explicit recognition: Mikko Voutilainen, Ia Iashvili, Viola Sordini, Hartmut Stadie, Francesco Pandolfi, Salvatore Rappoccio, Philip Harris, Konstantinos Kousouris, and Robert Schoefbeck.

The CMS experiment is a nearly unbelievable undertaking requiring the hard work and dedication of thousands of scientists, engineers, and students. Like everyone else, I owe some portion of my success and results to each and every one of these people. However, I would be remiss if I didn't single out my collaborators and friends from the Fermilab LPC. John Stupak and Ben Kreis, without you I would be a lesser physicist and weaker rock climber than I am today. Kevin Pedro and Lindsay Grey, I can only hope that when I grow up I understand programming languages like you. Nahn Tran, Jim Dolen, and Justin Pilot, thank you for your support on my JEC/JMAR endeavors. Nadja Strobbe, Hansjorge Webber, Scarlet Norberg, Joe Pastika, Jamie Antonelli, and Doug Berry, thank you for indulging my never ending and sometimes random questions. I promise they hade a purpose. To everyone else at the LPC, you my everlasting gratitude for your support.

My parents always told me that we collect some relationships from each stage in our lives; that the ones that remain are meant to be. I met Breann Sitarski as undergraduates at UCLA and we remain friends to this day. Each day I am thankful for her friendship and support. And while this may seem like an odd acknowledgment, I know that Breann and I have talked through many analysis challenges and strategies. She is my sounding board.

Finally, I thank my family, whom I love and appreciate more than words can ever describe. My brother Spenser is and always has been my go-to guy. Whether he knows it or not, he is my inspiration and I can only strive to be as engaged and dedicated as he is. My parents Laura and Gregg have supported me throughout this entire graduate process, even when they couldn't understand the title of my dissertation. I am the person I am today...I am here today because of them.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supported by a dissertation committee consisting of Professors Ricardo Eusebi (advisor), Alexei Safonov, and Bhaskar Dutta of the Department of Physics and Professor Sherry Yennello of the Department of Chemistry.

The data analyzed in Chapters 5-7 was collected by the CMS collaboration and is used with their permission. The event selection as well as the boosted decision tree training and optimization in chapter 5 was done in consultation with Professor Christopher Neu of the Department of Physics, University of Virginia and his former graduate student Dr. Joseph Goodell.

All other work conducted for the dissertation was completed by the student independently.

Funding Sources

This work was made possible in part by the Department of Energy under grant numbers DE-FOA-0001414, DE-FOA-0000733, and DE-FOA-0001358. The contents are solely the responsibility of the authors and do not necessarily represent the official views of the Department of Energy.

NOMENCLATURE

AOD	Analysis Object Data
APD	Avalanche Photodiode
ASIC	Application-Specific Integrated Circuits
ATLAS	A Toroidal LHC ApparatuS
AVF	Adaptive Vertex Fitter
BDT	Boosted Decision Tree
BSM	Beyond Standard Model
CERN	European Organization for Nuclear Research
CH	Charged Hadron
CKM	Cabibbo-Kobayashi-Maskawa
CL	Confidence Level
CMS	Compact Muon Solenoid
CMSSW	CMS Software Framework
CP	Charge-Parity
CPU	Central Processing Unit
CSC	Cathode Strip Chamber
CSCTF	Cathode Strip Chamber Track Finder
CSV	Combined Secondary Vertex
CTEQ	Coordinated Theoretical-Experimental Project on QCD
CTF	Combinatorial Track Finder
DA	Deterministic Annealing

DAQ	Data Aquisition
DT	Drift Tube
DTTF	Drift Tube Track Finder
EB	ECAL Barrel
ECAL	Electromagnetic Calorimeter
EE	ECAL Endcap
EM	Electromagnetic
ES	ECAL Preshower
EWSB	Electroweak Symmetry Breaking
FPGA	Field-Programmable Gate Array
FSR	Final-State Radiation
HB	HCAL Barrel
HCAL	Hadronic Calorimeter
HE	HCAL Endcap
HF	HCAL Forward
HLT	High-Level Trigger
HO	HCAL Outer
HPD	Hybrid Photodiode
IP	Interaction Point
ISR	Initial-State Radiation
JEC	Jet Energy Correction
JER	Jet Energy Resolution
L1	Level 1
L1A	Level-1 Accept
L2	Level 2

L3	Level 3
LEP	Large Electron-Positron Collider
LHC	Large Hadron Collider
LHCb	Large Hadron Collider beauty
LL	Leading Log
LO	Leading Order
LSP	Lightest Supersymmetric Particle
MB	Muon Barrel
MC	Monte Carlo
ME	Muon Endcap
MEM	Matrix Element Method
MET	Missing Transverse Energy
MIP	Minimum Ionizing Particle
MVA	Multivariate
NDF	Number of Degrees of Freedom
NH	Neutral Hadron
NLO	Next-to-Leading Order
NNLL	Next-to-Next-to-Leading Logarithmic
NNLO	Next-to-Next-to-Leading Order
NPV	Number of Primary Vertices
PAG	Physics Analysis Group
PD	Primary Dataset
PDF	Parton Distribution Function
PF	Particle Flow
PMT	Photomultiplier Tube

POG	Physics Object Group
PS	Proton Synchrotron
PSB	Proton Synchrotron Booster
PU	Pileup
QCD	Quantum Chromodynamics
QED	Quantum Electrodynamics
RF	Radio Frequency
RMS	Root Mean Square
RPC	Resistive Plate Chamber
SM	Standard Model
SPS	Super Proton Synchrotron
SUSY	Supersymmetry
TAMU	Texas A&M University
TCS	Trigger Control System
TEC	Tracker End Cap
TIB	Tracker Inner Barrel
TID	Tracker Inner Disks
TOB	Tracker Outer Barrel
TPG	Trigger Primitive Generator
TTC	Timing, Trigger and Control
UE	Underlying Event
VEV	Vacuum Expectation Value
VPT	Vacuum Phototriode
WIMP	Weakly Interacting Massive Particle
WLS	Wavelength-Shifting

TABLE OF CONTENTS

Page

LIST OF FIGURES

FIGURE Page

LIST OF TABLES

TABLE Page

1. INTRODUCTION

Particle physicists seek to understand the building blocks of the universe and how they interact. An understated search to characterize the fundamental constituents of nature which can be built up into the world we see. In this quest there has been no better tool than the synchrotron, a circular accelerator which collides particles at speeds approaching that of light. As the accelerators reach higher and higher energies, physicists are able to probe smaller distance scales and even create heavy, short lived particles which are otherwise inaccessible. The standard model (SM) of particle physics is the codification of such constituents over a century of study. It describes all of the observed elementary particles, their properties, and the electromagnetic, weak, and strong forces through which they interact. The standard model, a specific framework born out of quantum field theory (QFT), has predicted quantities and been proven accurate time and time again. Yet until recently it remained an incomplete model, at least experimentally.

One of the primary missions of the Large Hadron Collider (LHC), the worlds highest energy particle accelerator located at the European Organization for Nuclear Research (CERN), was to search for a long theorized missing piece to the SM. On July 4th, 2012 the ATLAS (A Toroidal LHC Apparatus) and CMS (Compact Muon Solenoid) collaborations at the LHC simultaneously confirmed the discovery of a new boson [30, 26]. Since its discovery, the particle has been shown to be consistent with the hypothesized scalar Higgs boson, said to give mass to itself and all of the other fundamentally massive particles through the process of electroweak symmetry breaking. It took almost 50 years for experimentalists to confirm the existence of the boson first proposed in 1964 as the spin zero mediator to the standard models only scalar field.

Using 19.7 fb^{-1} of 8 TeV data from the CMS experiment at CERN, the Higgs boson mass was measured to be 125.7 ± 0.3 (stat) ± 0.3 (syst) GeV[1][2] by five major decay modes: $\mathrm{H} \to \gamma\gamma$, $\mathrm{H} \to \tau\tau$, $\mathrm{H} \to \mathrm{b\overline{b}}$, $\mathrm{H} \to \mathrm{ZZ}{\to}4\mathrm{l}$, and $\mathrm{H} \to \mathrm{WW} \to \mathrm{l}\nu\mathrm{l}\nu$ [32]. Since then, the experiment has

[1] Unless otherwise indicated this document will use natural units, where $c = \hbar = 1$.

[2] This measurement has subsequently been improved by combining the ATLAS and CMS measurements. The measured Higgs mass as of 2015 was 125.09 ± 0.21 (stat) ± 0.11 (syst) GeV [31].

entered a phase of intense study of the new particle. Every property of the new boson and all of its decay channels must be studied in great detail to confirm that it is indeed the SM Higgs boson and not a different particle with similar characteristics. Currently the properties of the new boson are consistent with those predicted by the SM, but any deviation from the SM predictions could point to some new, as yet unexplored physics.

This dissertation will present a search for the 125 GeV Higgs boson in the the $H\to WW\to l\nu jj$ decay channel using 8 TeV proton-proton data collected by the CMS detector. Although the $H\to WW\to l\nu jj$ channel was used in the original combined limit, the previous search was not sensitive to the "low mass" Higgs, but only to $M_{\mathrm{H}} > 2M_{\mathrm{W}}$ [33].[3] Because the Higgs mass is less than two times the mass of the W boson, at least one of the W bosons must be created "off-shell", meaning that its measured mass is not $\sim 80\,\mathrm{GeV}$. On top of that, the presence of a neutrino makes it a challenge to fully reconstruct the initiating particle. For these reasons the $WW \to l\nu l\nu$ decay channel was the most sensitive of the WW channels during the 2012 combination. Nevertheless this analysis will search for the low mass Higgs boson in the semi-leptonic channel using a matrix element (ME) technique to boost the signal extraction sensitivity.

This dissertation will be organized in the following way. Section 2 will present an overview of the standard model, the Higgs mechanism, and a brief introduction to how the Higgs can point to physics beyond the standard model (BSM). The LHC and CMS will be described in section 3. Section 4 describes the reconstruction of an event at CMS and all of the final physics objects. Section 5 discusses the analysis work-flow from data samples used to signal extraction techniques while the results are presented in section 6. Section 7 gives my concluding remarks.

[3] The lowest search mass was $M_{\mathrm{H}} = 170\,\mathrm{GeV}$.

2. THEORETICAL FRAMEWORK

Since the mid-1970s, the Standard Model (SM) of particle physics has been the leading theory describing three of the four known fundamental forces (not including gravity) as well as classifying all of the known elementary particles. Even during it's formative years, the SM's success at predicting new particles (i.e. the top quark in 1995) and describing the properties of known particles (i.e. $W^{\pm}$ to Z^{0} mass ratio) was undeniable. The model's roots can be traced back to 1930 when Herman Weyl was able to describe electromagnetism as a local symmetry represented by the Lie group $U\left(1\right)$ [34]. In 1954 Yang and Mills created a theory which tried to extend the idea of gauge theory to non-abelian groups [35]. This laid the ground work for Sheldon Glashow to combine the electromagnetic and weak interactions in 1961 [36]. This combined interaction is described by the $SU\left(2\right)\times U\left(1\right)$ group. In 1967 Steven Weinberg and Abdus Salam [37, 38] continued this work by adding in the Higgs mechanism first proposed by Robert Brout and Francois Englert [39], Peter Higgs [40, 41], and Gerald Guralnik, Carl. R. Hagen, and Tom Kibble [42, 43]. Although all of these theorist contributed to this advancement, the mechanism eventually became known as the Brout-Englert-Higgs (BEH) mechanism. The model entered its current form around 1964 with the introduction of the strong force and quantum chromodynamics (QCD) [44, 45, 46, 47, 48]. The initial theory by Gell-Man and Zweig only included the up, down, and strange quarks and was incomplete until the introduction of the color charge by Greenberg [49]. The full theory is described by the symmetry group

$$SU\left(3\right)_{C}\otimes SU\left(2\right)_{L}\otimes U\left(1\right)_{Y} \tag{2.1}$$

where $SU\left(2\right)_{L}\otimes U\left(1\right)_{Y}$ is the electroweak (EW) symmetry group describing both the electromagnetic and weak interactions and $SU\left(3\right)_{C}$ is the symmetry group describing the strong interaction [50, 51].

The rest of this chapter will discuss the standard model, both its structure and some of its mathematical underpinnings, in more detail. Section 2.1 will introduce the particle content of the SM. The QFTs that govern the SM interactions will be discussed in sections 2.2 to 2.5. In section 2.7 we will briefly reference how Higgs physics can relate to physics beyond the SM. More information about the history of the standard model can be found in appendix A.

2.1 The Standard Model

The standard model is a locally gauge-invariant quantum field theory (QFT) in four-dimensional Minkowski space [50, 52]. The structure and particle content of the SM can be found in fig. 2.1. The SM is composed of 12 fermions, the particles that make up matter, and 4 gauge bosons, the force-carrying particles which mediate the electromagnetic, weak, and strong interactions. On its own, the basic symmetries of the standard model require that the gauge bosons (W$^{\pm}$,Z,γ,gluons) be massless. However, we know that this is not true as experiments have shown that the W and Z bosons have relatively large masses. The aforementioned Higgs mechanism takes care of this by spontaneously breaking the electroweak symmetry, giving mass to the quarks, the leptons, and the W and Z bosons [37, 38, 53].

Fermions are particles which obey Fermi-Dirac statistics and the Pauli exclusion principle, meaning that no two fermions may occupy the same quantum state within a given quantum system. These particles have half-integer spin, often denoted as spin-1/2, which means that their intrinsic angular momentum is $\hbar/2$. For every fermion f in the SM there exists an anti-fermion $\bar{f}$, which has oppositely signed quantum numbers, but the same mass. The fermions in the SM are separated into six leptons and six quarks with these further separated into 3 generations of pairs of particles. Each subsequent generation is ostensibly a heavier version of the previous generation, with the same quantum numbers.[1]

Each generation of lepton can be broken down into a charged and neutral lepton. For instance, the first generation is composed of the electron (e), with charge $-e$, and the electron neutrino (ν_e). The second and third generations contain the muon (μ) and tau (τ) along with their associated neutrinos. Although the SM specifies that the neutrinos are massless, experiments have shown that this is not true. While their exact masses are still unknown, upper bounds have been places on these

[1] The neutrinos may have a different mass ordering.

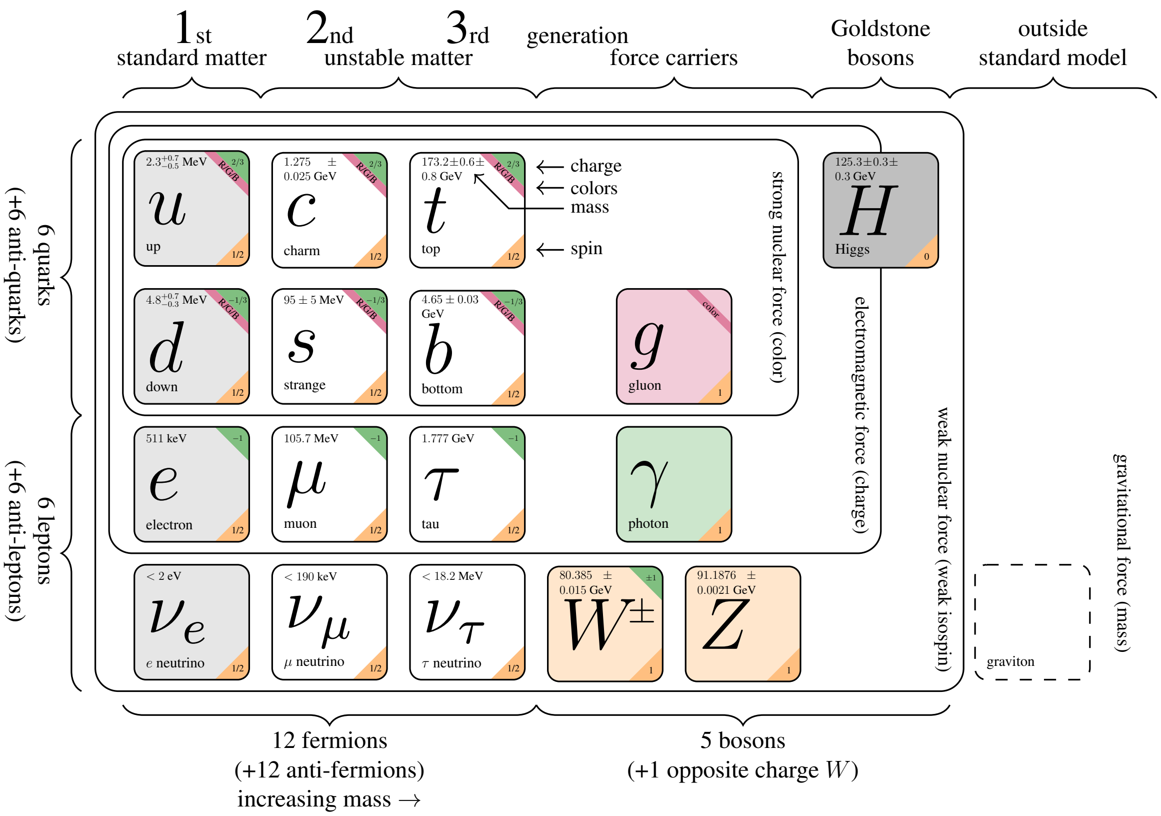


Figure 2.1: The Standard Model of particle physics. The model includes three generations of matter particles (leptons and quarks) as well as the gauge and Higgs bosons. Included in this drawing are the particle names, symbols, masses, spin, electric charge, and color charge, if applicable.

and can be seen in fig. 2.1. Each generation of lepton has an associated quantum number, called the lepton number, defined as $L_\ell = n_\ell - n_{\bar{\ell}}$. First generation leptons have quantum numbers $L_e = +1$ and $L_\mu = L_\tau = 0$ while the second and third generations have value $+1$ for their associated lepton number and zero otherwise. The antileptons have oppositely signed lepton numbers. The lepton numbers are a conserved quantity in the SM, which means that only lepton-antilepton pairs can be created or destroyed. That being said, neutrino oscillations, the phenomena of neutrinos changing flavor from one generation to the next, has been observed [54]. While this violates the conservation of lepton numbers within a generation, the total lepton number $L \equiv L_e + L_\mu + L_\tau$ may still be conserved. All leptons interact through the weak interaction, but only the charged leptons interact using the electromagnetic interaction. Because leptons lack the color charge they do not interact using the strong force.

Like the leptons, the three generation of quarks can be broken into one up-type quark and one down-type quark, categories which gain their name through the content of the first generation containing the up (u) and down (d) quarks. The second generation is made up of the charm (c) and strange (s) quarks while the third is made up of the top (t) and bottom (b) quarks. The up-type quarks have fractional electric charge of $Q = +2e/3$ and the bottom-type quarks have electric charge $Q = -e/3$. As in the case of the leptons, the quarks have an associated baryon quantum number, B. This quantity is conserved in all SM interactions and no exception has every been seen. This means that only quark-antiquark pairs may be created or destroyed and also results in the stability of the lightest baryon, the proton. Baryon number is defined as $B = \frac{1}{3}\left(n_q - n_{\bar{q}}\right)$, where, for example, the baryon number for a quark is $+1/3$ and $-1/3$ for an antiquark. Quarks may interact through the electromagnetic and weak interactions, but unlike the lepton, quarks can also interact via the strong force. This is because quarks also have color charge, which can have three values referred to as red, green, or blue. Antiquarks may contain charges of anti-red, anti-green, or anti-blue. In the SM colorless particles are forbidden from existing on their own, which means that individual quarks, often referred to as bare quarks, have never been seen in nature. Rather, quarks are always found as constituents of bound states called hadrons. This group of composite particles may be further divided into mesons, bound states of a quark-antiquark pair, and baryons, bound states of three quarks and antiquarks. The hadrons contain quark and antiquark combinations such that the bound state is a color singlet, often referred to as being colorless. Mesons contain color-anticolor pairs while baryons consist of red, green, and blue charged quarks. The masses of the quarks are hard to measure due to their confinement in hadrons, however, global averages have been made.

So far the particle content of the SM has been introduced along with the various force carriers. The next few sections will go into greater detail about the specifics of the particle-particle interactions. It will be helpful to keep in mind fig. 2.2, which shows all of the leading order SM interactions.

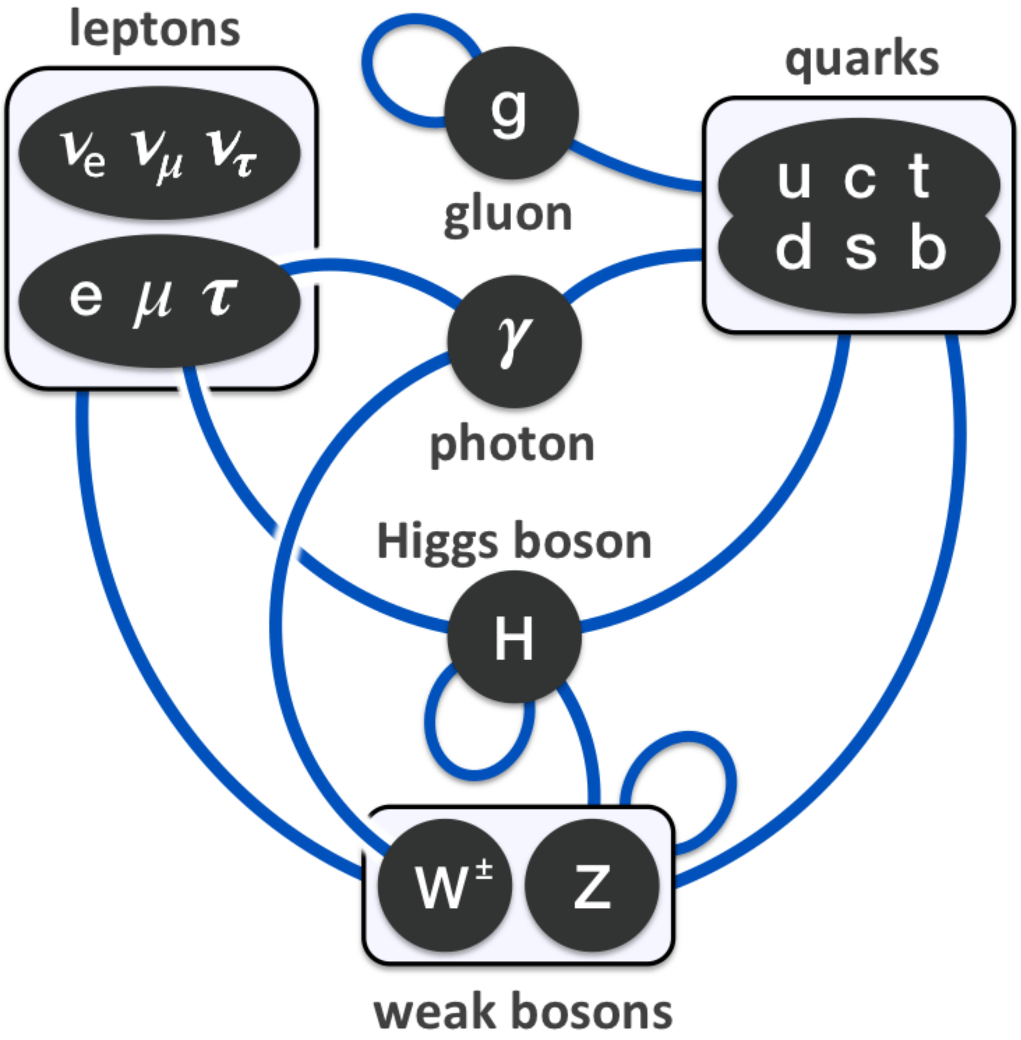


Figure 2.2: A diagram illustrating the leading order interactions between particles in the standard model, including self-interactions. Reprinted from [1].

2.2 Quantum Electrodynamics & the Electromagnetic Interaction

Quantum electrodynamic (QED) is a quantum field theory which describes the dynamics of the electromagnetic interaction and corresponds to the $U_{EM}(1)$ group. In a QFT, particles are represented by fields, which are in turn represented mathematically by Lagrangian densities $\mathcal{L}$. QED was formulated to described the interactions of spin-1/2 particles, namely leptons and quarks.

Like a classical field theory, the interactions and equations of motion of a quantum system are described by a Lagrangian. QED is described by the Dirac Lagrangian density

$$\mathcal{L} = i\bar{\psi}\gamma^{\mu}\partial_{\mu}\psi - m\bar{\psi}\psi \tag{2.2}$$

where ψ a four-component column vector representing the wave function of a spin-1/2 particle[2], γ^{μ} are the four Dirac gamma matrices, $\bar{\psi} \equiv \psi^{\dagger}\gamma^{0}$, and m is the mass of the particle.

In order for QED to be gauge invariant it must be invariant under both local and global gauge transformations. Let there exist a global $U\left(1\right)$ transformation

$$\psi \rightarrow \psi' = e^{-i\alpha}\psi \tag{2.3}$$

with constant α. Then ψ in the Lagrangian 2.2 can be replaced by equation 2.3, which means that $\mathcal{L} \rightarrow \mathcal{L}' = \mathcal{L}$. Therefore QED is invariant under this type of transformation. If instead we have $\alpha \rightarrow \alpha\left(x\right)$ where α is allowed to vary as a function of space-time, then equation 2.3 becomes a local $U\left(1\right)$ transformation. Therefore equation 2.2 becomes

$$\mathcal{L} \rightarrow \mathcal{L}' = \mathcal{L} + \bar{\psi}\gamma^{\mu}\left(\partial_{\mu}\alpha\left(x\right)\right)\psi \tag{2.4}$$

and is thus not invariant under the local transformation as is. To return the gauge invariance we can replace the partial derivative in the Lagrangian density by a covariant derivative

$$D_{\mu} = \partial_{\mu} + iqA_{\mu} \tag{2.5}$$

, where $q = -e$ is the electron charge, in case of an electron, and A_{μ} is a new gauge field representing the photon, the mediator of electromagnetic interactions. This new gauge field transforms as

$$A_{\mu}{\rightarrow}A'_{\mu} = A_{\mu} + \partial_{\mu}\chi\left(x\right) \tag{2.6}$$

[2] ψ is a field known as a Dirac spinor.

, where $\chi\left(x\right)$ is an arbitrary function of space-time. By applying the transformation in equation 2.3 to a lepton field, the photon field transforms as in equation 2.6, and $\chi\left(x\right)=\alpha\left(x\right)/q$, the covariant derivative transforms in the same way as $\psi\left(x\right)$, namely $D_\mu\psi\rightarrow\left(D_\mu\psi\right)'=e^{-i\alpha}D_\mu\psi$. After the changes listed above, equation 2.2 will be locally gauge invariant and take the form

$$\mathcal{L}=\bar{\psi}\left(i\gamma^\mu D_\mu-m\right)\psi-\frac{1}{4}F^{\mu\nu}F_{\mu\nu} \tag{2.7}$$

where

$$F^{\mu\nu}=\left(\partial^\mu A^\nu-\partial^\nu A^\mu\right) \tag{2.8}$$

is the electromagnetic field strength tensor.

Notice that in equation 2.7 does not contain a $m^2A_\mu A^\mu$ term, which would be the mass of the gauge field. This fits with experimental observations given that the photon is massless and thus the electromagnetic interaction has an infinite range. Lagrangian 2.7 does introduce lepton-photon interactions and does contain an $\ell^+\ell^-\gamma$ interaction and a term quadratic in the field strength tensor, which is the photon kinetic energy. The complete QED Lagrangian can be created by generalizing to all leptons by $\psi\rightarrow\psi_i$ and summing over all leptons $i=e,\mu,\tau,u,d,c,s,t,b$ as in equation 2.9.

$$\mathcal{L}=\sum_i\left[\bar{\psi}_i\left(i\gamma^\mu D_\mu-m_i\right)\psi_i\right]-\frac{1}{4}F_{\mu\nu}F^{\mu\nu} \tag{2.9}$$

2.3 Electroweak Interaction

As mentioned in sec 2, the electromagnetic and weak interactions can be unified into a single, non-abelian gauge theory, work started by Yang & Mills and then completed by Glashow, Weinberg, and Salam [53]. In order to explain this unification, we will first work with a fermionic doublet representing and $SU\left(2\right)$ symmetry. A doublet of Dirac fields can be represented as

$$\psi=\begin{pmatrix}\psi_1\left(x\right)\\ \psi_2\left(x\right)\end{pmatrix} \tag{2.10}$$

The doublet will transform under the three dimensional rotation

$$\psi \rightarrow \exp \left\langle i\alpha^{i} \frac{\sigma_{i}}{2} \right\rangle \psi \tag{2.11}$$

This is again a global transformation, but note that by generalizing to higher order interaction we must use matrices instead of a local $\alpha\left(x\right)$ function to describe dynamics. These matrices, σ^{i}, are the Pauli sigma matrices shown in equation 2.12 and satisfy the identity $\sigma^{i}\sigma^{j} = \delta^{ij} + i\epsilon^{ijk}\sigma^{k}$ where $\epsilon^{ijk} = +1$ and where ϵ is an antisymmetric tensor.

$$\sigma^{1} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \ \sigma^{2} = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \ \sigma^{3} = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \tag{2.12}$$

As in sec 2.2 we can turn equation 2.11 into a local transformation by having $\alpha \rightarrow \alpha^{i}\left(x\right)$ and thus

$$\psi\left(x\right) \rightarrow V\left(x\right)\psi\left(x\right), \text{ where } V\left(x\right) = \exp\left(i\alpha^{i}\left(x\right)\frac{\sigma^{i}}{2}\right) \tag{2.13}$$

Still, the Lagrangian must be invariant under this transformation and in order to do this we introduce three vector fields $A_{\mu}^{i}\left(x\right)$, where $i = 1,\ 2,\ 3$. We once again use a covariant derivative

$$D_{\mu} = \partial_{\mu} - igA_{\mu}^{i}\frac{\sigma^{i}}{2} \tag{2.14}$$

, which means that the newly introduced fields transform as

$$A_{\mu}^{i}\left(x\right)\frac{\sigma^{i}}{2} \rightarrow V\left(x\right)\left(A_{\mu}^{i}\left(x\right)\frac{\sigma^{i}}{2} + \frac{i}{g}\partial_{\mu}\right)V^{\dagger}\left(x\right) \tag{2.15}$$

Unfortunately, this transformation is not trivial to calculate given that the Pauli matrices do not commute. By assuming infinitesimally small transformations and expanding $V\left(x\right)$ to first order in

α we obtain the simpler form

$$A^i_\mu \frac{\sigma^i}{2} \rightarrow A^i_\mu \frac{\sigma^i}{2} + \frac{1}{g}\left(\partial_\mu \alpha^i\right)\frac{\sigma^i}{2} + i\left[\alpha^i \frac{\sigma^i}{2}, A^i_\mu \frac{\sigma^i}{2}\right] + ... \tag{2.16}$$

With the above ingredients the covariant derivative will transform as

$$D_\mu \psi \rightarrow \left(1 + i\alpha^i \frac{\sigma^i}{2}\right) D_\mu \psi \tag{2.17}$$

and the field strength tensor will be

$$F^i_{\mu\nu} = \partial_\mu A^i_\nu - \partial_\nu A^i_\mu + g\epsilon^{ijk} A^j_\mu A^k_\nu \tag{2.18}$$

Given all of the above, the Yang-Mills Lagrangian will be

$$\mathcal{L} = -\frac{1}{4}\left(F^i_{\mu\nu}\right)^2 + \bar{\psi}\left(i\gamma^\mu \partial_\mu - igA^i_\mu \frac{\sigma^i}{2}\right)\psi \tag{2.19}$$

Given the above process from Yang-Mills theory, we can now show how to obtain the electroweak interaction, which is based on a local $SU\left(2\right)_L \times U\left(1\right)_Y$ gauge symmetry. This process will follow what was done in section 2.2 in that requiring a local invariance will lead to the introduction of new gauge fields and determine their interactions. It is also important to note that SM fermions can be grouped based on their chirality, which is a fundamental property of a particle and describes how the particles wave function will behave under rotation. Spin-1/2 particles will pick up a minus sign under a 2π rotation, but left-chiral (left-handed) particles will go one way around the complex plane while right-chiral (right-handed) particles will go the opposite direction. In the SM, the left-handed up- and down-type quarks form a weak doublet q_L and the left-handed charged leptons and neutrinos form a separate weak doublet ℓ_L. The right-handed particles form weak singlets, but right-handed neutrinos and left-handed antineutrinos don't exist in the SM.

Given the prerequisites, an explanation of electroweak unification can now be made. This explanation will start by using the first generation of leptons as an example, but will then generalize to more particles. The SM contains an $SU\left(2\right)$ doublet of the left-handed components of the electron neutrino and electron. The $SU\left(2\right)$ invariant right-handed component of the electron is placed in a singlet.

$$L_e = \begin{pmatrix} \nu_L \\ e_L \end{pmatrix},\ e_R \tag{2.20}$$

The kinetic energy term of electroweak Lagrangian for the first generation leptons takes the form

$$\mathcal{L}^e_{KE} = L_e^\dagger \tilde{\sigma}^\mu i \partial_\mu L_e + e_R^\dagger \sigma^\mu i \partial_\mu e_R \tag{2.21}$$

where $\sigma = \left(\sigma^0, \sigma^1, \sigma^2, \sigma^3\right)$, $\tilde{\sigma} = \left(\sigma^0, -\sigma^1, -\sigma^2, -\sigma^3\right)$, σ^0 is an identity matrix, and σ^i are again the Pauli matrices. Equation 2.21 is invariant under the global $SU\left(2\right)_L \times U\left(1\right)_Y$ transformation given by

$$L {\rightarrow} L' = e^{i\theta} U L \quad \forall \quad \theta \in \mathbb{R} \tag{2.22}$$

$$e_R {\rightarrow} e'_R = e^{2i\theta} e_R \quad \forall \quad \theta \in \mathbb{R} \tag{2.23}$$

where $U = e^{-i\alpha^k \sigma^k}$ and α^k is a real number. However, if θ and α^k are allowed to vary as a function of space-time, then the Lagrangian will not be invariant under a local $SU\left(2\right)_L \times U\left(1\right)_Y$ transformation.

To make Lagrangian 2.21 invariant we construct a $U\left(1\right)$ gauge field $B_\mu\left(x\right)$ and three $SU\left(2\right)$ gauge fields $W_\mu\left(x\right) = W_\mu^k\left(x\right)\sigma_k$ which transform as

$$B_\mu\left(x\right) {\rightarrow} B'_\mu\left(x\right) = B_\mu\left(x\right) + \frac{2}{g_1}\partial_\mu \theta\left(x\right) \tag{2.24}$$

$$W_\mu\left(x\right) {\rightarrow} W'_\mu\left(x\right) = U\left(x\right) W_\mu\left(x\right) U^\dagger\left(x\right) + \frac{2i}{g_2}\left(\partial_\mu U\left(x\right)\right) U^\dagger\left(x\right) \tag{2.25}$$

where g_1 and g_2 are dimensionless coupling strengths of the interactions. The covariant derivatives are then

$$D_\mu L_e = \left(\partial_\mu + i\frac{g_1}{2}YB_\mu + i\frac{g_2}{2}YW_\mu\right) L_e \tag{2.26}$$

$$D_\mu e_R = \left(\partial_\mu + i\frac{g_1}{2}YB_\mu\right) e_R \tag{2.27}$$

where Y is the hypercharge operator. The weak hypercharge can be calculated as $Y = 2\left(Q - T_3\right)$, where T_3 is the third component of the weak isospin quantum number T. A notable property of the weak interaction is that it only acts on particles with weak isospin T and that T_3 is conserved in all interactions. The SM gauge fields and their associated electric and hypercharge values can be found in table 2.1. Combining the kinetic and gauge interaction terms of the Lagrangian yields

$$\mathcal{L} = \mathcal{L}_{KE} + \mathcal{L}_{gauge} = L_e^\dagger \tilde{\sigma}^\mu i D_\mu L_e + e_R^\dagger \sigma^\mu i D_\mu e_R - \frac{1}{4}B_{\mu\nu}B^{\mu\nu} - \sum_{i=1}^{3}\frac{1}{4}W^i_{\mu\nu}W^{i\mu\nu} \tag{2.28}$$

where $B_{\mu\nu} = \partial_\mu B_\nu - \partial_\nu B_\mu$ and $W_{\mu\nu} = \left[\partial_\mu + \left(i\frac{g_2}{2}\right)W_\mu\right]W_\nu - \left[\partial_\nu + \left(i\frac{g_2}{2}\right)W_\nu\right]W_\mu$ are the field strength tensors. This Lagrangian, without any mass terms, is now locally invariant. The addition of the mass terms and electroweak symmetry breaking (EWSB) will be covered in section 2.5, but given that the mediators of the weak force are massive, its range is limited to about 10^{-18} m.

The observed electroweak gauge bosons are actually combinations of the B and W fields as shown in equation 2.29

$$W_\mu^\pm = \frac{W_\mu^1 \mp i W_\mu^2}{\sqrt{2}} \tag{2.29}$$

$$Z_\mu = \frac{g_1 W_\mu^3 - g_2 B_\mu}{\sqrt{g_1^2 + g_2^2}} = W_\mu^3 \cos\left(\theta_W\right) - B_\mu \sin\left(\theta_W\right) \tag{2.30}$$

$$A_\mu = \frac{g_1 W_\mu^3 + g_2 B_\mu}{\sqrt{g_1^2 + g_2^2}} = W_\mu^3 \sin\left(\theta_W\right) - B_\mu \cos\left(\theta_W\right) \tag{2.31}$$

, where θ_W is the Weinberg angle defined as $\sin\left(\theta_W\right) = g_1/\sqrt{g_1^2 + g_2^2}$. Note that W_1 and W_2 are electrically charged while W_3 and B are electrically neutral. Given equation 2.28, the $W^\pm$ will only couple to the left-handed doublets while the Z and photon (A) will couple to both the left- and right-handed leptons in the SM. Lagrangian 2.28 can be generalized to include the other generations by appropriately summing over all leptons as in equation 2.32.

$$\mathcal{L}^{\ell} = \sum_{leptons} \left(L_e^{\dagger} \tilde{\sigma}^{\mu} i D_{\mu} L_e + e_R^{\dagger} \sigma^{\mu} i D_{\mu} e_R \right) - \frac{1}{4} B_{\mu\nu} B^{\mu\nu} - \sum_{i=1}^{3} \frac{1}{4} W_{\mu\nu}^{i} W^{i\mu\nu} \tag{2.32}$$

These ideas can be extended to the quarks by making a doublets out of the left-handed up- and down-type quarks and singlets out of the right handed components, as in 2.33.

$$Q_u = \begin{pmatrix} u_L \\ d_L \end{pmatrix}, \; u_R, \; d_R \tag{2.33}$$

A similar kinetic component to the lepton Lagrangian in 2.21 can also be formed

$$\mathcal{L}_{KE}^{quark} = Q_u^{\dagger} \tilde{\sigma}^{\mu} i D_{\mu} Q_u + u_R^{\dagger} \sigma^{\mu} i D_{\mu} u_R + d_R^{\dagger} \sigma^{\mu} i D_{\mu} d_R \tag{2.34}$$

As we saw with the leptons, the $W^{\pm}$ will only couple to the left-handed quark doublets while the Z and photon will couple to both the left- and right-handed quarks.

	Particle-Type	Q	T_3	Y	B	L
Quarks	$\mathrm{q}_L = \begin{pmatrix} \mathrm{u} \\ \mathrm{d} \end{pmatrix}_L$	$\begin{pmatrix} 2/3 \\ -1/3 \end{pmatrix}$	$\begin{pmatrix} 1/2 \\ -1/2 \end{pmatrix}$	$1/3$	$1/3$	0
	u_R	$2/3$	0	$4/3$	$1/3$	0
	d_R	$-1/3$	0	$-2/3$	$1/3$	0
Leptons	$\ell_L = \begin{pmatrix} \nu_e \\ e \end{pmatrix}_L$	$\begin{pmatrix} 0 \\ -1 \end{pmatrix}$	$\begin{pmatrix} 1/2 \\ -1/2 \end{pmatrix}$	-1	0	1
	e_R	-1	0	-2	0	1

Table 2.1: The quantum numbers of the SM fermions grouped by chirality and particle-type, independent of generation. The various particle-types in the SM are up-type quarks, down-type quarks, charged leptons, and neutrinos.

Adding equation 2.34 to equation 2.32 gives the full electroweak Lagrangian as

$$\mathcal{L}^{EW} = \mathcal{L}_{KE}^{lepton} + \mathcal{L}_{KE}^{quark} + \mathcal{L}_{gauge} \tag{2.35}$$

This Lagrangian exhibits an invariance to the $U(1)$ transformation $L_e \rightarrow e^{i\alpha} L_e,\ e_R \rightarrow e^{i\alpha} e_R$, which leads to conservation of electron number. There is a similar invariance to transformations using the muon and tau fields. The Lagrangian is also invariant to another $U(1)$ transformation where all negatively (positively) charged fields are multiplied by $e^{i\alpha}$ $(e^{-i\alpha})$. This invariance leads to the conservation of electric charge. However, the electroweak Lagrangian is not invariant under charge conjugation or a parity transformation. Charge conjugation is when the sign of all quantum numbers is changed, which can also be thought of as exchanging all particles (antiparticles) for antiparticles (particles). Parity transformations occur when the sign of the spacial coordinates are flipped as in $\boldsymbol{r} \rightarrow -\boldsymbol{r}$. Interactions mediated by the photon and Z boson, also known as neutral current interactions, are invariant under the combination of charge and parity transformations, known as CP invariance. On the other hand, interactions involving quarks which are mediated by the $W^{\pm}$ bosons are not invariant under a CP transformation [55].

2.4 Strong Interaction

Quantum Chromodynamics (QCD) is the theory that described the interaction between quarks, the strong interaction, and is represented by a local $SU(3)_C$ gauge symmetry. As described in section 2.1, quarks contain any one of three color charges (C); red, green, or blue. Only color neutral (colorless) hadrons are allowed in nature, which requires that a baryon contain equal parts of each color and that a meson contains a color-anticolor pair. Because of this each quark is represented as a color triplet

$$q_u = \begin{pmatrix} u_r \\ u_g \\ u_b \end{pmatrix} \tag{2.36}$$

The mediator of the strong force, the electrically neutral gluon, must then contain two color charges in order to conserve color. The eight known color combination for the gluon will represented by the eight gauge fields introduced below.

A QCD Lagrangian which is globally $SU\left(3\right)$ invariant can be represented as

$$\mathcal{L}^q_{QCD} = \sum_{i=1}^{6} \bar{q}_i i\gamma^\mu \partial_\mu q_i \tag{2.37}$$

where q_i represents one of the six quark flavors. This Lagrangian will be invariant under a transformation of the form $q_i \rightarrow q'_i = Uq_i$ where U is a member of $SU\left(3\right)$. When using a local $SU\left(3\right)$ transformation where $U \rightarrow U\left(x\right)$, Lagrangian 2.37 is no longer invariant. To return invariance, we must introduce eight gauge fields ($G_\mu\left(x\right)$), which represent the gluons, and the appropriate covariant derivative. The transformation of the gauge fields and the covariant derivative will take the form

$$G_\mu \rightarrow G'_\mu = UG_\mu U^\dagger + \frac{i}{g_s}\left(\partial_\mu U\right)U^\dagger \tag{2.38}$$

$$D_\mu q_i = \left(\partial_\mu + ig_s G_\mu\right) q_i \tag{2.39}$$

where g_s is the dimensionless coupling strength of the color interaction and whose value can be seen in fig. 2.3 where $g_s = \alpha_s$. The field strength tensor for QCD is

$$G_{\mu\nu} = \partial_\mu G_\nu - \partial_\nu G_\mu + ig_s\left(G_\mu G_\nu - G_\nu G_\mu\right) \tag{2.40}$$

and the locally $SU\left(3\right)$ gauge invariant QCD Lagrangian is given as

$$\mathcal{L}^q_{QCD} = \sum_{i=1}^{6}\left(\bar{q}_i i\gamma^\mu D_\mu q_i\right) - \frac{1}{4}\sum_{i=1}^{8} G^i_{\mu\nu} G^{i\mu\nu} \tag{2.41}$$

There are a few interesting facts about the strong interaction which must be noted. In contrast to the electroweak interaction C, P, and T are all conserved. Additionally, the strong force has a

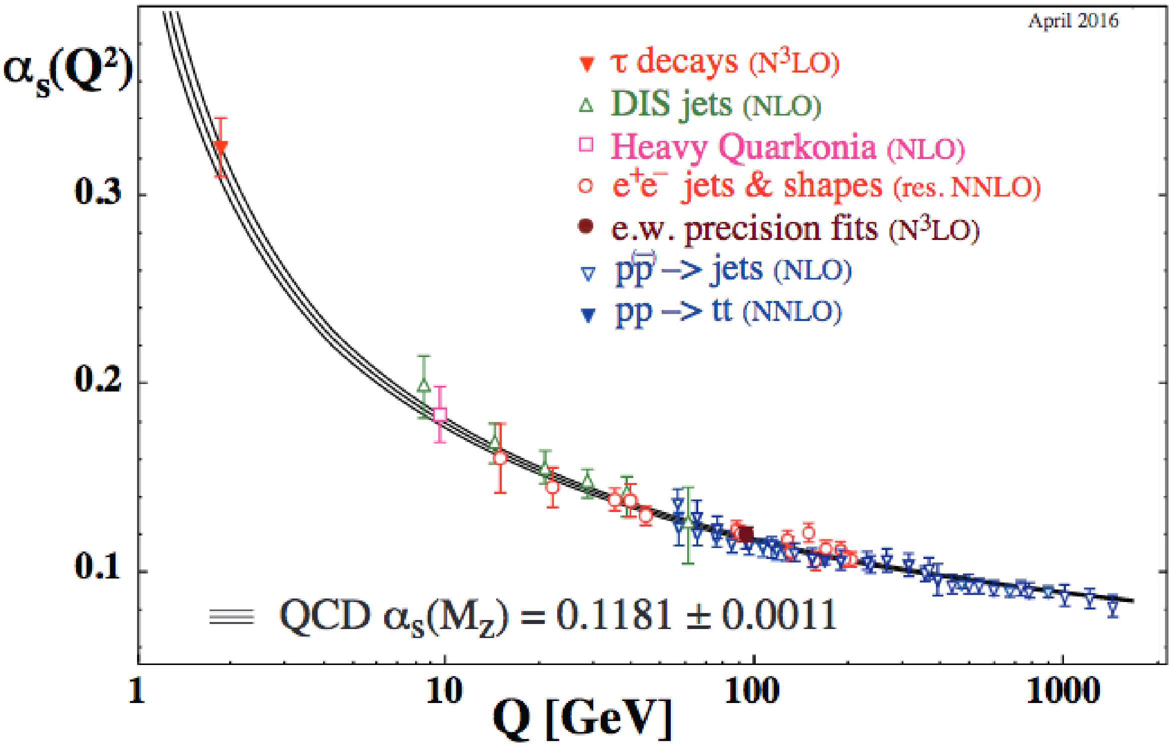


Figure 2.3: Summary of measurements of α_s as a function of the energy scale Q. The respective degree of QCD perturbation theory used in the extraction of α_s is indicated in brackets (NLO: next-to-leading order; NNLO: next-to-next-to leading order; res. NNLO: NNLO matched with re-summed next-to-leading logs; N^{3}LO: next-to-NNLO). Figure and caption reprinted from [2].

range of about 10^{-15} m, which is enough to act on nucleons, i.e. protons and neutrons, to form atomic nuclei. Lastly, QCD is a strongly coupled theory at low energies and large distance scales and weakly interacting at high energies and small distance scales. Quarks are confined particles, meaning that the attractive force between them does not decrease as they move farther apart. Instead the force decreases as the particles move closer and increases as they move farther apart, a behavior called asymptotic freedom [56]. When in the high energy regime the typical perturbative calculation can be made[3], but in the low energy regime theorists must use more advanced techniques such as lattice gauge theory [57].

[3] The leading order (LO) terms can be calculated perturbatively. Corrections must be added to account for the next-to-leading order (NLO) effects, with further corrections for the next-to-next-to leading order (NNLO) effects, and so on.

2.5 Brout-Englert-Higgs Mechanism & The Higgs Boson

The EW and QCD Lagrangians covered in sections 2.3 and 2.4 contain no mass terms, which means the bosons within the SM should be massless. However, we know from experiments at CERN that the $W^{\pm}$ [58] and Z [59] bosons do indeed have mass. The method by which mass is added to the SM while maintaining the necessary gauge invariance is the BEH mechanism [39, 40]. This is accomplished by adding one or more complex scalar fields, the Higgs field(s), to the SM Lagrangian. These fields will acquire a vacuum expectation value (vev) which will spontaneously break the symmetry of the Lagrangian. The Goldstone theorem tells us that for every spontaneously broken continuous symmetry there will be a new massive scalar "Goldstone" boson. So the number of Goldstone bosons will be equal to the number of broken generators of the symmetry group. The massless standard model bosons then acquire mass by absorbing these Goldstone bosons. So the number of massive SM bosons will be equal to the number of broken generators.

Remember from section 2.3 there are four massless electroweak gauge bosons, W^1, W^2, W^3, and B^0. The experimentally observed bosons, however, are the massless photon (γ) and three massive bosons ($W^{\pm}$, Z). We also know that the electric charge Q is conserved in electroweak interactions. This means that the $SU\left(2\right)_L \times U\left(1\right)_Y$ electroweak theory is broken such that a new $U\left(1\right)_{EM}$ symmetry group is formed which corresponds to electromagnetism. In order for three gauge bosons to acquire mass they must absorb three Goldstone bosons. The simplest method to accomplish this is to introduce a complex, scalar $SU\left(2\right)$ doublet Φ with hypercharge $Y = 1$.

$$\Phi = \begin{pmatrix} \phi^+ \\ \phi^0 \end{pmatrix} \tag{2.42}$$

The part of the SM Lagrangian which includes the electroweak gauge bosons and the leptons can be written as

$$\mathcal{L}_{SM} = -\frac{1}{4} W^a_{\mu\nu} W^{\mu\nu}_a - \frac{1}{4} B_{\mu\nu} B^{\mu\nu} + \bar{L}_i \left(i D_\mu \gamma^\mu\right) L_i + \bar{e}_{R,i} \left(i D_\mu \gamma^\mu\right) e_{R,i} \tag{2.43}$$

where i runs over the three generations, μ and ν are Lorentz indices, and a runs over the generators in the gauge group. The field strengths are given by

$$W^a_{\mu\nu} = \partial_\mu W^a_\nu - \partial_\nu W^a_\mu + g_2 \epsilon^{abc} W^b_\mu W^c_\nu \tag{2.44}$$

$$B_{\mu\nu} = \partial_\mu B_\nu - \partial_\nu B_\mu \tag{2.45}$$

and the covariant derivatives for the left- and right-handed leptons are

$$D_\mu L_L = \left(\partial_\mu - i g_2 T_a W^a_\mu - i g_1 Y B_\mu\right) L_L \tag{2.46}$$

$$D_\mu e_R = \left(\partial_\mu - i g_1 Y B_\mu\right) e_R \tag{2.47}$$

where T_a are the generators of the $SU\left(2\right)_L$ gauge group and g_1, g_2 are the coupling constants for the electroweak interaction.

By adding the scalar field in equation 2.42 we must add an additional scalar part to the Lagrangian

$$\mathcal{L}_S = \left(D^\mu \Phi\right)^\dagger \left(D_\mu \Phi\right) - V\left(\Phi\right) \tag{2.48}$$

where the first term is the kinetic term and the second term is the scalar potential, also known as the "Mexican Hat" potential. While the form of the scalar potential is not known from first principles, we can make the assumption that it takes the simplest form possible which has the desired properties of spontaneous symmetry breaking and the ability to be renormalized

$$V\left(\Phi\right) = \mu^2 \Phi^\dagger \Phi + \lambda \left(\Phi^\dagger \Phi\right)^2 \tag{2.49}$$

The value of λ must be positive in order for the vacuum to be stable. The sign of μ^2 specified one of two cases for the potential, both of which are illustrated in fig. 2.4. When $\mu^2 > 0$, the potential

$V\left(\Phi\right)$ is always positive and has a minimum at

$$\left\langle 0\right|\Phi\left|0\right\rangle \equiv \Phi_0 = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \tag{2.50}$$

where no spontaneous symmetry breaking can occur. In contrast, when $\mu^2 < 0$ the potential takes its namesake "Mexican hat" shape with a minimum value not located at the origin. In this case, the neutral component of the scalar field can acquire a vacuum expectation value (vev) v, a process known as electroweak symmetry breaking (EWSB).

$$\left\langle 0\right|\Phi\left|0\right\rangle = \Phi_0 = \frac{1}{\sqrt{2}}\begin{pmatrix} 0 \\ v \end{pmatrix},\ v = \sqrt{\frac{-\mu^2}{\lambda}} \tag{2.51}$$

By only adding a vev to the neutral component of the scalar field electromagnetism is unbroken and the $U\left(1\right)_{EM}$ symmetry keeps a conserved electric charge of $Q = T_3 + \frac{Y}{2}$.

At this point we can expand the scalar field Φ around the minimum Φ_0 to get

$$\Phi\left(x\right) = \frac{1}{\sqrt{2}}\begin{pmatrix} 0 \\ v + h\left(x\right) \end{pmatrix} \tag{2.52}$$

where $h\left(x\right)$ is a new scalar field. Next we insert this field into the kinetic part of the Lagrangian 2.48 and redefine the gauge fields as

$$W_\mu^{\pm} = \frac{1}{\sqrt{2}}\left(W_\mu^1 \mp i W_\mu^2\right) \tag{2.53}$$

$$Z_\mu = \frac{1}{\sqrt{g_1^2 + g_2^2}}\left(g_2 W_\mu^3 - g_1 B_\mu\right) \tag{2.54}$$

$$A_\mu = \frac{1}{\sqrt{g_1^2 + g_2^2}}\left(g_2 W_\mu^3 + g_1 B_\mu\right) \tag{2.55}$$

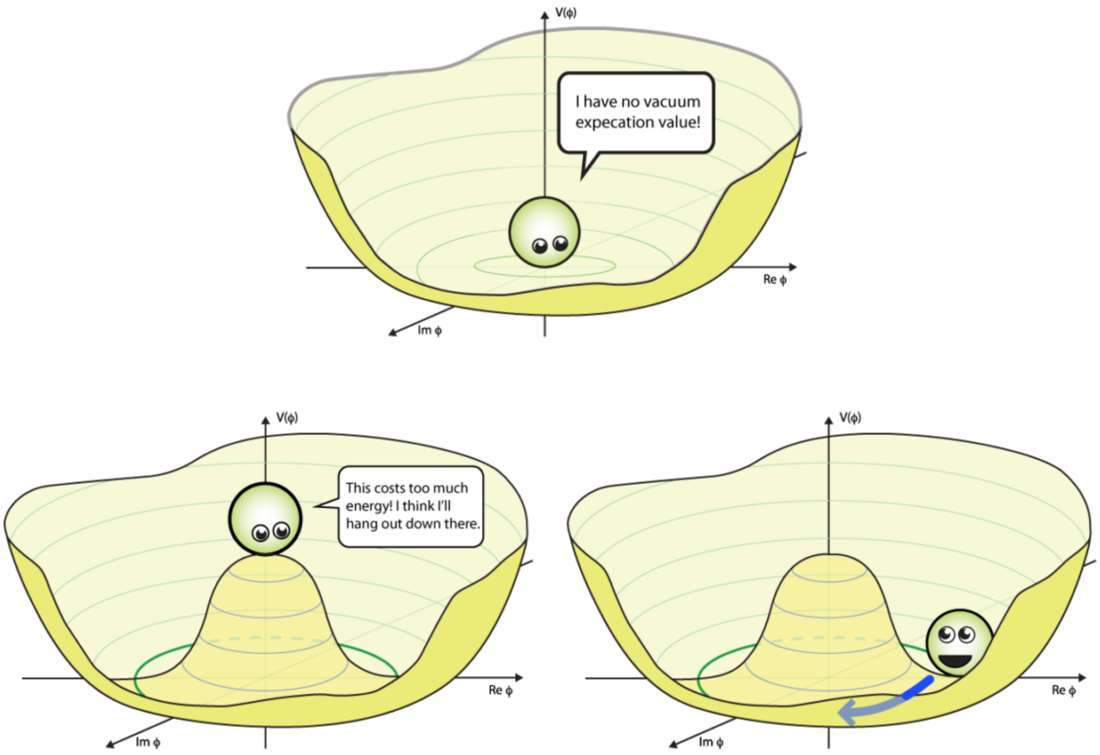

V(φ)
I have no vacuum expecation value!
Re φ
Im φ
V(φ)
This costs too much energy! I think I'll hang out down there.
Re φ
Im φ
V(φ)
Re φ
Im φ

Figure 2.4: (Top) The scalar potential when $\mu^2 > 0$. In this case the potential will always be positive and its minimum value will be at the origin. The vacuum expectation value for this potential is zero. (Bottom) When $\mu^2 < 0$ the potential will take the shape of a "Mexican Hat" with its minimum value being in a degenerate ring around the origin. As soon as the scalar field has moves away from the origin and closer to the minimum the symmetry has been spontaneously broken and will acquire a non-zero vev. Because the scalar field picked a particular direction when falling towards the minimum, it is no longer invariant under a rotation. Reprinted from [3].

which correspond to the observed gauge bosons. After this the covariant derivative becomes

$$|D_\mu\Phi|^2 = \frac{1}{2}\left(\partial_\mu H\right)^2 + \frac{1}{2}g_2^2\left(v+H\right)^2 W_\mu^+ W^{\mu-} + \frac{1}{8}\left(v+H\right)^2\left(g_1^2+g_2^2\right) Z_\mu Z^\mu \tag{2.56}$$

From this we see that the photon A_μ remains massless, but that the mass terms for the W and Z bosons take the general forms $M_W^2 W_\mu W^\mu$ and $\frac{1}{2}M_Z^2 Z_\mu Z^\mu$ respectively. Thus the masses of the electroweak gauge bosons are

$$M_W = \frac{1}{2} v g_2 \tag{2.57}$$

$$M_Z = \frac{1}{2} v \sqrt{g_1^2 + g_2^2} \tag{2.58}$$

$$M_A = 0 \tag{2.59}$$

Three of the degrees of freedom from the scalar field, which would have been two charged and one neutral Goldstone boson, have been absorbed by the gauge bosons in order to give them mass. These appear as *flat* directions in the scalar potential. There is one remaining degree of freedom, an oscillation in the radial direction, which corresponds to the neutral Higgs boson and can be seen in fig. 2.5 where the potential is concave.

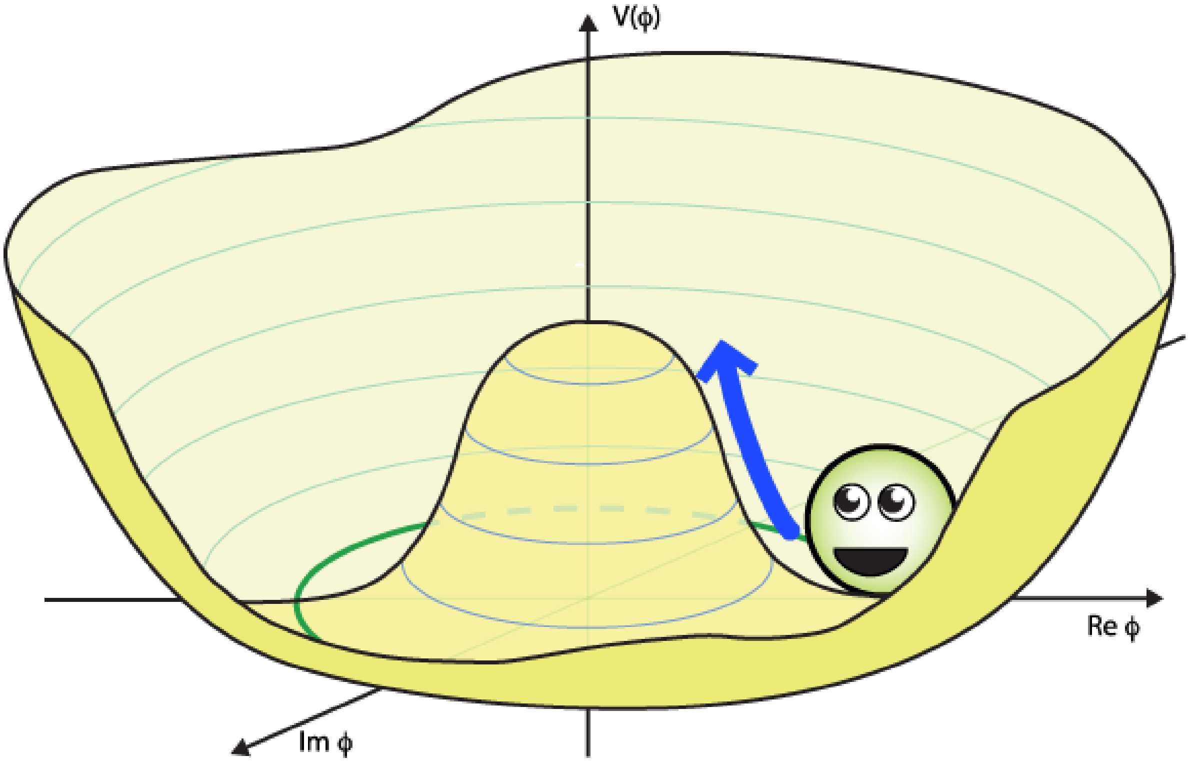


Figure 2.5: The Higgs boson corresponds to an oscillation of the scalar field in the radial direction. Reprinted from [3]

Several relationships can be formed between the various bosons. The Weinberg angle also known as the weak mixing angle θ_W, defined as $\sin\theta_W = \frac{g_1}{\sqrt{g_1^2+g_2^2}}$, can be used to describe the photon an Z as

$$A_\mu = \cos\theta_W B_\mu + \sin\theta_W W_\mu^3 \tag{2.60}$$

$$Z_\mu = -\sin\theta_W B_\mu + \cos\theta_W W_\mu^3 \tag{2.61}$$

Equation 2.62 shows a relationship between the masses of the W and Z at tree level, which is one reason why measurements of their masses are so important.

$$\frac{M_W}{M_Z} = \frac{g_2}{\sqrt{g_1^2+g_2^2}} = \cos\theta_W \tag{2.62}$$

There also exists a relationship between the coupling strength of the weak and electromagnetic interactions which makes use of the weak mixing angle,

$$e = g_2 \sin\theta_W \tag{2.63}$$

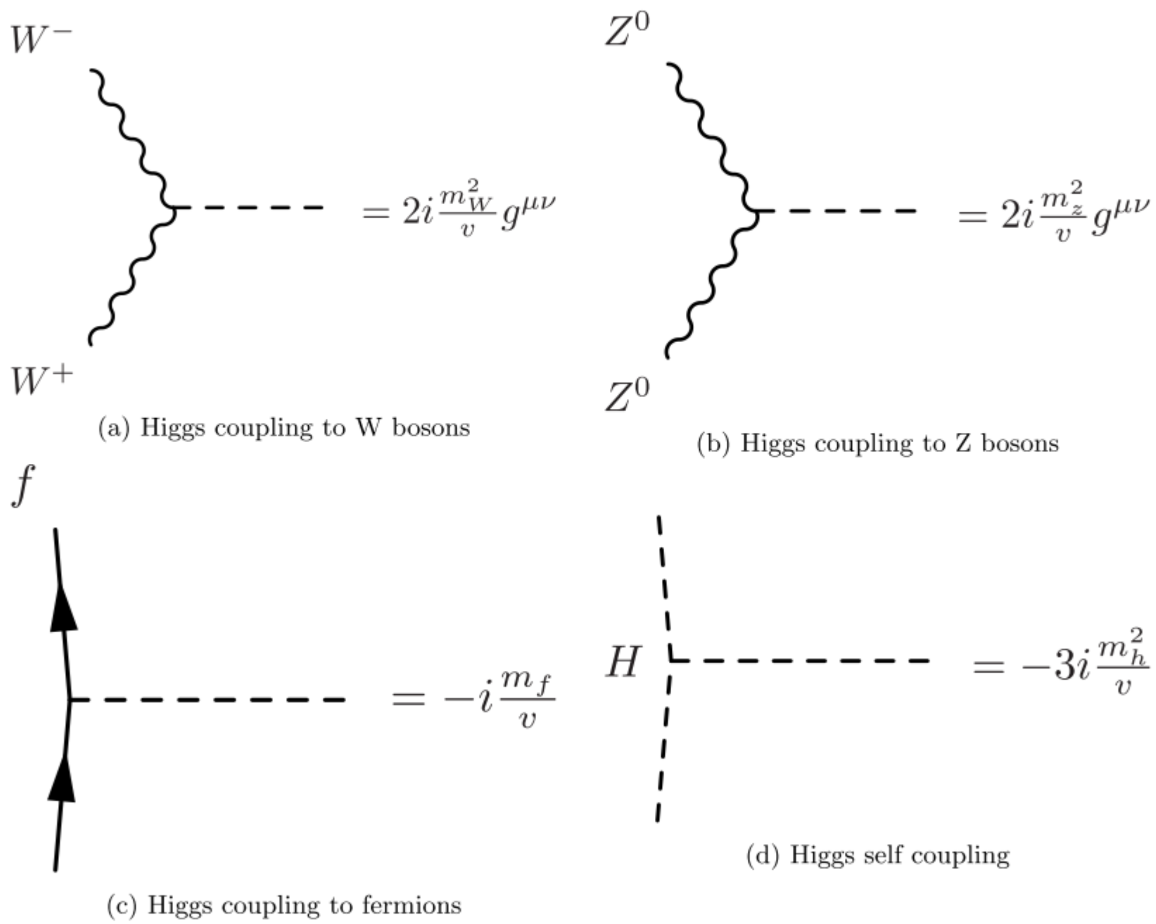
By substituting equation 2.52 into Lagrangian 2.48, using $v^2 = -\frac{\mu^2}{\lambda}$, and looking at only the pieces involving the Higgs we can study the mass and couplings of the Higgs itself. This section of the Lagrangian will take the form

$$\mathcal{L}_H = \frac{1}{2}\left(\partial_\mu H\right)\left(\partial^\mu H\right) - \lambda v^2 H^2 - \lambda v H^3 - \frac{\lambda}{4} H^4 \tag{2.64}$$

Since scalar masses have the general form $\frac{1}{2}m\phi^2$ we find that the Higgs boson mass is

$$m_H = 2\lambda v^2 = -2\mu^2, \tag{2.65}$$

where λ, and thus the Higgs mass, needs to be determined experimentally. We can also see that the Higgs couples to vector bosons, fermions, and itself, all interactions which are shown in fig. 2.6.



(a) Higgs coupling to W bosons

(b) Higgs coupling to Z bosons

(c) Higgs coupling to fermions

(d) Higgs self coupling

Figure 2.6: Tree level Feynman diagrams showing how the Higgs couples to vector bosons (a,b), fermions (c), and to itself (d).

Besides having massive bosons, the SM also has a whole host of massive fermions. These particles can be shown to acquire mass by adding Yukawa couplings between the fermion fields and the scalar field to the SM Lagrangian. The part of the Lagrangian that corresponds to the first generation fermions is given by

$$\mathcal{L}_F = -G_e\bar{L}\Phi e_R - G_d\bar{Q}\Phi d_R - G_u\bar{Q}\tilde{\Phi}u_R + h.c. \tag{2.66}$$

where $\tilde{\Phi} = i\tau_2\Phi^*$ is the conjugate of Φ with negative hypercharge. There are additional terms added to the full Lagrangian which correspond to the second and third generations which are not shown here. By substituting equation 2.52 into Lagrangian 2.66 we find

$$\begin{aligned}\mathcal{L}_F = & -\frac{1}{\sqrt{2}}\left[G_e\begin{pmatrix}\bar{\nu}\,\bar{e}\end{pmatrix}_L\begin{pmatrix}0\\ v+H\end{pmatrix}e_R + G_d\begin{pmatrix}\bar{u}\,\bar{d}\end{pmatrix}_L\begin{pmatrix}0\\ v+h\end{pmatrix}d_R\right.\\ & \left.+G_u\begin{pmatrix}\bar{u}\,\bar{d}\end{pmatrix}_L\begin{pmatrix}v+h\\ 0\end{pmatrix}u_R\right] + h.c.\end{aligned} \tag{2.67}$$

$$= -\frac{1}{\sqrt{2}}\left(v+h\right)\left(G_e\bar{e}_L e_R + G_d\bar{d}_L d_R + G_u\bar{u}_L u_R\right) + h.c. \tag{2.68}$$

where $h.c.$ is a placeholder for the hermitian conjugate terms. The fermion masses take the form $m\bar{f}_L f_R + h.c.$, which means that the fermion masses for the first generation are

$$m_e = \frac{G_e v}{\sqrt{2}}, \qquad m_u = \frac{G_u v}{\sqrt{2}}, \qquad m_d = \frac{G_d v}{\sqrt{2}} \tag{2.69}$$

The second and third generations have similar mass terms. Since there is no right handed neutrino in the SM the neutrinos that do exist remain massless. As the coupling constants, G, and the fermion masses are not predicted by the SM they must be measured and added to the model.

2.6 Higgs Production in a Proton-Proton Collider

The Higgs boson has several accessible productions mechanisms at a proton-proton collider. Fig. 2.7 shows the 8 TeV production cross sections for the five production modes at the LHC. The production mode with the highest rate, by far, is the gluon-gluon fusion process shown in the blue curve, often abbreviated as ggH. Since gluons are massless so they can't couple directly to the Higgs boson. Instead, this production mode proceeds through a fermion loop as shown in 2.8a. The Higgs couplings to fermions goes as $g_{Hf\bar{f}} = \frac{m_f}{v}$, where v is the vacuum expectation value for the Higgs field, $v = \left(\sqrt{2}G_F\right)^{1/2} \approx 246\,\text{GeV}$, where G_F is the Fermi coupling determined by muon decay measurements [60]. This means that the coupling is directly dependent upon the fermion mass and because of this the fermion loop in the gluon-gluon fusion diagram is dominated by top quarks, the heaviest of the fermions in the standard model. The cross section for this production mechanism at $\sqrt{s} = 8$ TeV and assuming a 125 GeV Higgs is

$$\sigma_{\text{ggF}} = 19.27^{+7.2\%}_{-7.8\%}\,(\text{QCD Scale Unc.})^{+7.4\%}_{-6.9\%}\,(\text{PDF} + \alpha_S\ \text{Unc.})\ \text{pb}^{-1} \tag{2.70}$$

where QCD Scale uncertainty refers to the next to next to leading order (NNLO) radiative corrections and PDF + α_S uncertainty refers to the uncertainties on the parton distribution function and strong coupling parameters.

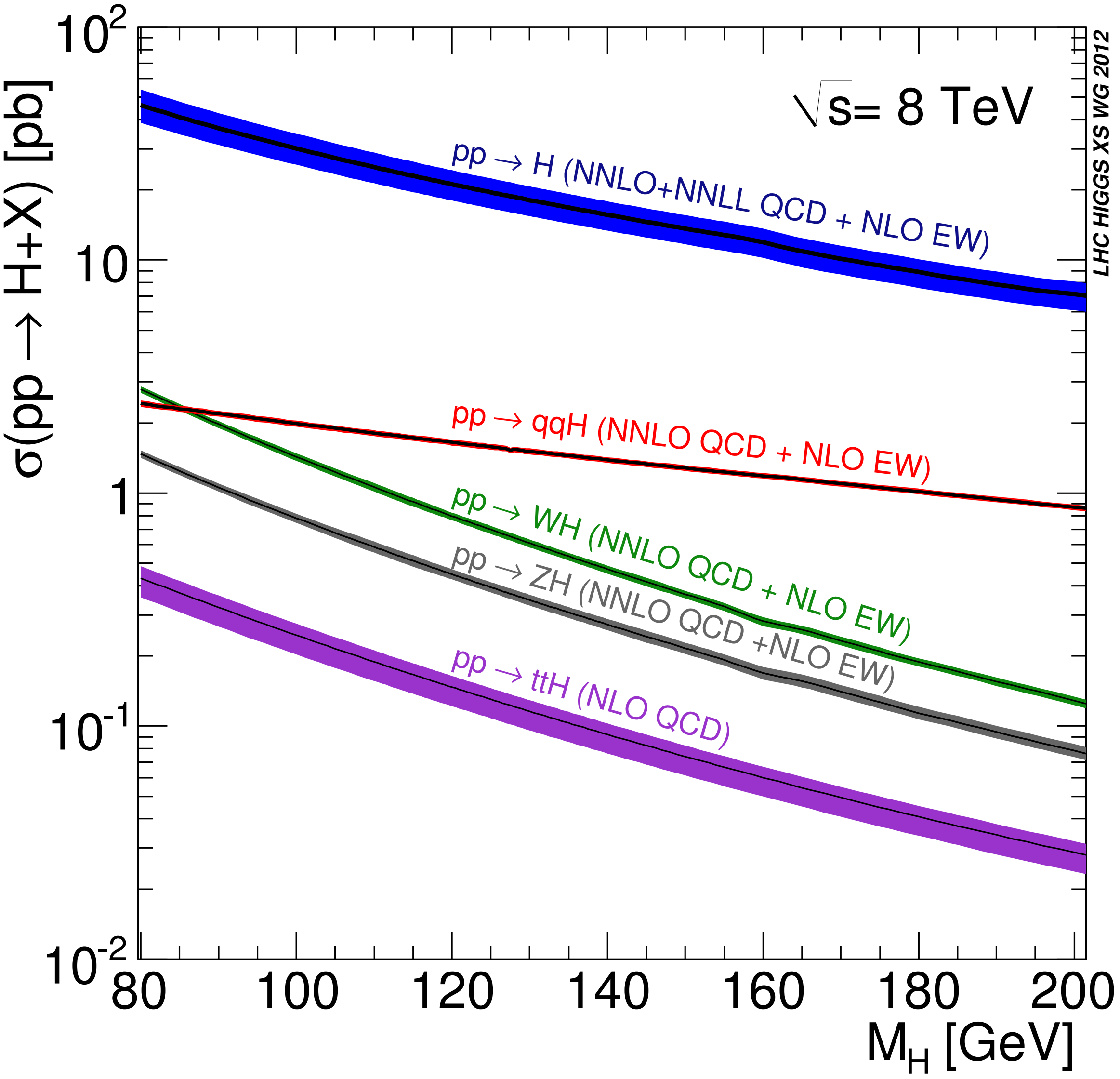


Figure 2.7: Higgs production cross-sections at the LHC for 8 TeV proton-proton collisions.

The production mechanism with the next highest cross section is the vector boson fusion (VBF) process (fig. 2.8b) where either two oppositely charged W bosons or two Z bosons merge and produce a Higgs boson. The final state particles for this process are those from the Higgs decay as well as the two initial quarks, which will preferentially be found in the forward regions of the detector, which is why this process is often abbreviated as qqH. The production cross section in this case is

$$\sigma_{\text{VBF}} = 1.653^{+4.5\%}_{-4.5\%}\,(\text{EW Unc.})^{+0.2\%}_{-0.2\%}\,(\text{QCD Scale Unc.})^{+2.6\%}_{-2.8\%}\,(\text{PDF} + \alpha_S\text{ Unc.})\ \mathrm{pb}^{-1} \tag{2.71}$$

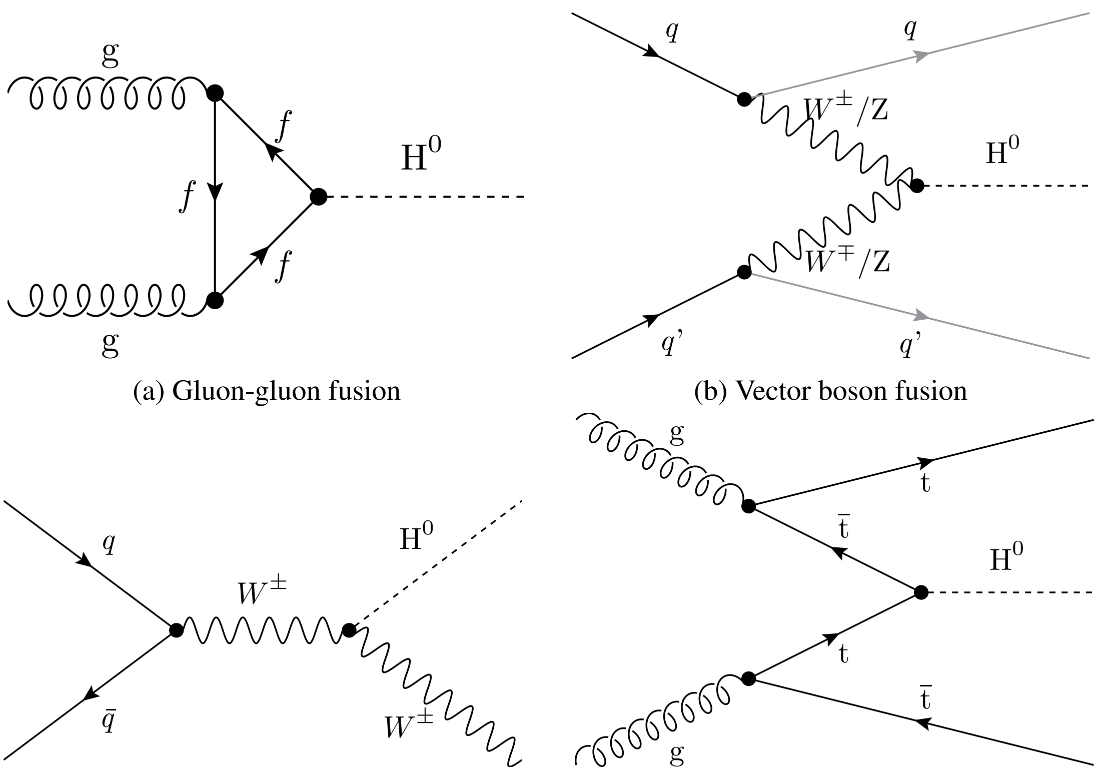
where the electroweak uncertainty is calculated at next to leading order (NLO).

The other processes found in fig. 2.7 can all be grouped as associated production mechanisms. The Higgs is produced along with either a $\mathrm{W}^{\pm}$ boson, Z^0 boson, or a $\mathrm{t\bar{t}}$ pair, often abbreviated as WH, ZH, or ttH. The first two cases, seen in fig. 2.8c, are also referred to as “Higgsstralung” because the Higgs can be seen as being radiated from the vector bosons, similar to how a photon is radiated by an electron during bremsstrahlung. The latter case is seen in fig. 2.8d. The associated production cross sections are

$$\begin{aligned}
\sigma_{\text{WH}} &= 0.7046^{+1.0\%}_{-1.0\%}\,(\text{QCD Scale Unc.})^{+2.3\%}_{-2.3\%}\,(\text{PDF} + \alpha_S\text{ Unc.})\ \mathrm{pb}^{-1}\\
\sigma_{\text{ZH}} &= 0.4153^{+3.1\%}_{-3.1\%}\,(\text{QCD Scale Unc.})^{+2.5\%}_{-2.5\%}\,(\text{PDF} + \alpha_S\text{ Unc.})\ \mathrm{pb}^{-1}\\
\sigma_{\mathrm{t\bar{t}H}} &= 0.1293^{+3.8\%}_{-9.3\%}\,(\text{QCD Scale Unc.})^{+8.1\%}_{-8.1\%}\,(\text{PDF} + \alpha_S\text{ Unc.})\ \mathrm{pb}^{-1}
\end{aligned} \tag{2.72}$$

Just as the Higgs boson can be produced in several ways it can also decay in many ways. Fig. 2.9 shows the Higgs decay branching ratios (BR) as well as $\sigma \times$ BR for final states containing four fermions. It is clear from fig. 2.9a that the WW decay has one of the highest branching ratios and from fig. 2.9b that the $\mathrm{l\nu qq}$ final state has the highest $\sigma \times$ BR.

Given the production cross sections and branching ratios discussed above, fig. 2.10 shows the dominant Feynman diagram searched for in this analysis, the gluon-gluon fusion production and semi-leptonic W decay mode. Nevertheless, we search for a given final state and not an exact



(a) Gluon-gluon fusion

(b) Vector boson fusion

(c) Associated production with a vector boson

(d) Associated production of a Higgs with a pair of t quarks

Figure 2.8: Feynman diagrams for the four Higgs production mechanisms at the LHC.

production and decay chain, so there are several branching ratios which are useful to this analysis and are listed in table 2.2. The $\mathrm{H} \rightarrow \mathrm{ZZ}$ and $\mathrm{H} \rightarrow \mathrm{b\overline{b}}$ BR are included because they can produce a $\mathrm{l}\nu\mathrm{qq}$ final state given a mis-identification or mis-reconstruction issue. The signal cross sections used in this analysis are listed in table 2.3 and present a couple of insights into our signal makeup. First is that the gluon-gluon fusion process is indeed dominant with $\sim$10 times higher of a cross section than the other channels. Additionally, the WH channel where $\mathrm{H} \rightarrow \mathrm{b\overline{b}}$ is non-negligible and comparable in size to the VBF production mode, even though this is not the decay channel we are looking for. By using some cuts to remove b-jets I will later show how to remove this signal contamination.

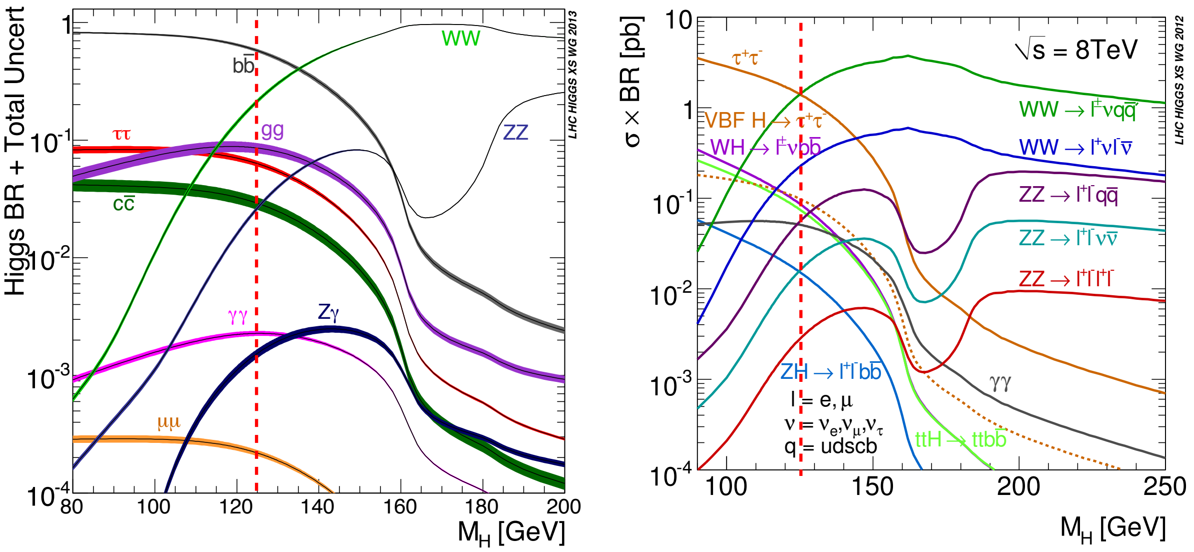
In addition to the true signal events, volunteer signal events (i.e. $\mathrm{H} \rightarrow \mathrm{b\overline{b}}$), this analysis must content with several other standard model processes which can produce a $\mathrm{l}\nu\mathrm{qq}$ final state. These

Decay	BR
H→WW	$0.215^{+4.26\%}_{-4.20\%}$
W→$l\nu$	0.3257
W→qq	0.676
WW → lνqq	0.2203
H → ZZ	$0.0246^{+4.28\%}_{-4.21\%}$
H → b$\overline{\text{b}}$	$0.577^{+3.21\%}_{-3.27\%}$

Table 2.2: Useful Higgs and W branching ratios

Channel	$\sigma \times$ BR
ggH, where H → WW → lνqq	1.823 pb^{-1}
qqH, where H → WW → lνqq	0.1493 pb^{-1}
WH, where H → WW	0.1515 pb^{-1}
ZH, where H → WW	0.08929 pb^{-1}
ttH, where H → WW	0.0278 pb^{-1}
WH, where H → b$\overline{\text{b}}$ → lνqq	0.1324 pb^{-1}
ttH, where H → b$\overline{\text{b}}$ → lνqq	0.0746 pb^{-1}
WH, where H → ZZ	0.01860 pb^{-1}
ZH, where H → ZZ	0.01096 pb^{-1}
ttH, where H → ZZ	0.00341 pb^{-1}

Table 2.3: A table of $\sigma \times$ BR for the lνqq final state resulting from any Higgs production mode and several decay channels.



(a) Higgs branching ratios

(b) Higgs $\sigma \times$ BR for four fermion final states

Figure 2.9: The dominant Higgs decay modes at the LHC. The vertical, dashed red line indicates a Higgs mass of 125 GeV.

background events can even have rates several orders of magnitude higher than that of our signal. There are two varieties of backgrounds which will be encountered, reducible and irreducible. The irreducible backgrounds, like the SM WW process, will exactly produce the $\mathrm{l}\nu\mathrm{qq}$ final state. On the other hand, reducible backgrounds produce slightly different final states, but may still enter the signal region for a variety of reasons. An example of a reducible background is the $\mathrm{t\bar{t}}$ process, which will have extra (b-)jets that may be removed through additional cuts.

The backgrounds considered in this analysis are as follows:

- **W+jets**: This is the production of a single $\mathrm{W}^{\pm}$ boson in association with final state quarks or gluons. If the $\mathrm{W}^{\pm}$ decays leptonically then the final state will match that of our signal. This process has an extremely high cross section and is thus the dominant background in the analysis.

- **Drell-Yan Z/γ^*+ jets**: In this case a Z or γ boson is produced in association with final state quarks or gluons. In order for this process to mimic the signal one lepton from the boson decay must be lost due to being outside the acceptance region or due to some reconstruction

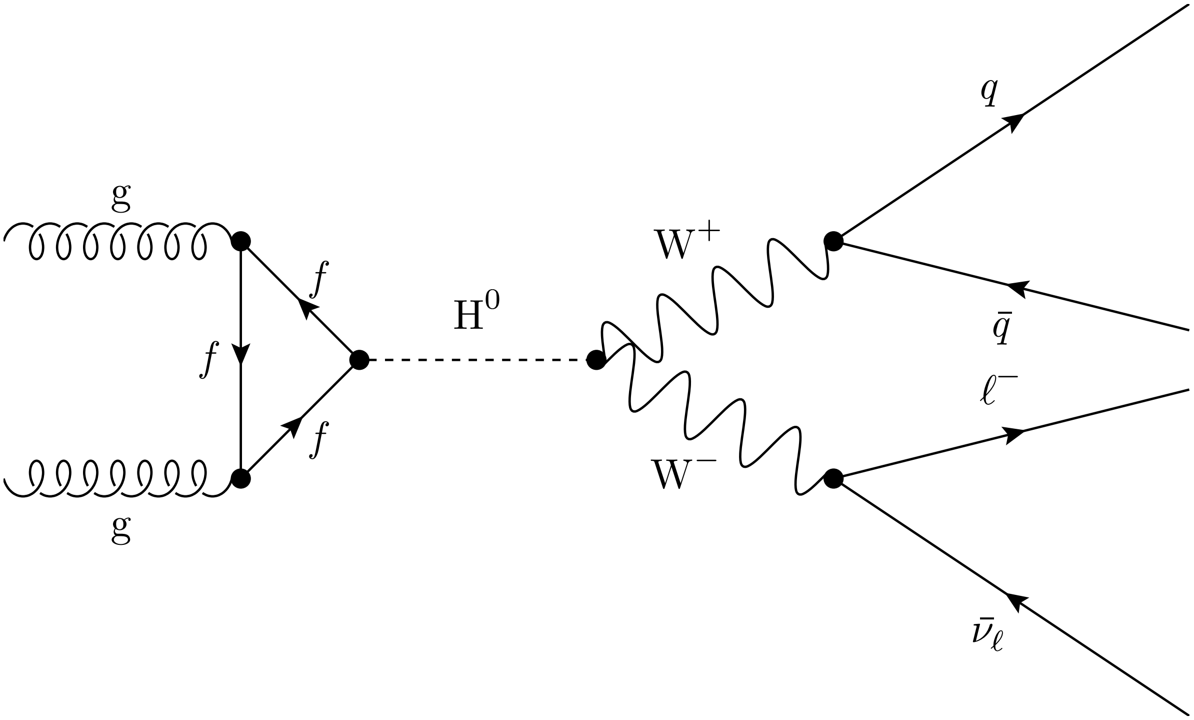
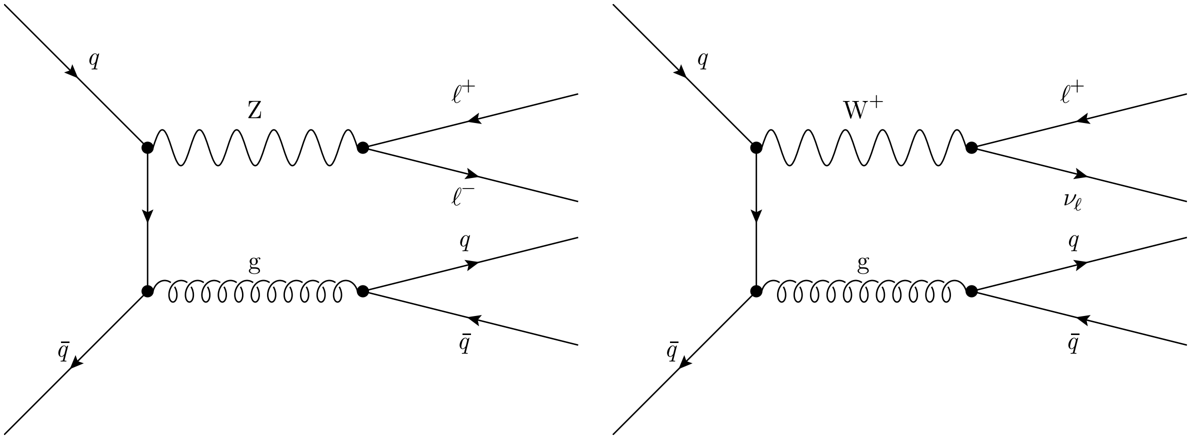


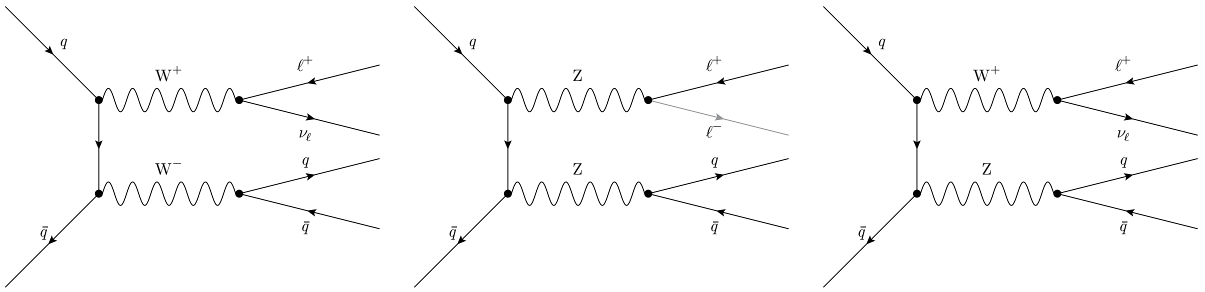
Figure 2.10: Feynman diagram for the gluon-gluon fusion SM Higgs production process where the Higgs decays semi-leptonically to two quarks, one lepton, and one neutrino.



(a) Z^0 production in association with jets (b) W^+ production in association with jets

Figure 2.11: Example Feynman diagrams for the standard model V + jets process decaying to the $\ell\nu jj$ final state.

inefficiency. Although this process also has a high cross section, the requirement of having only one lepton reduces the prevalence of these processes in our signal region.

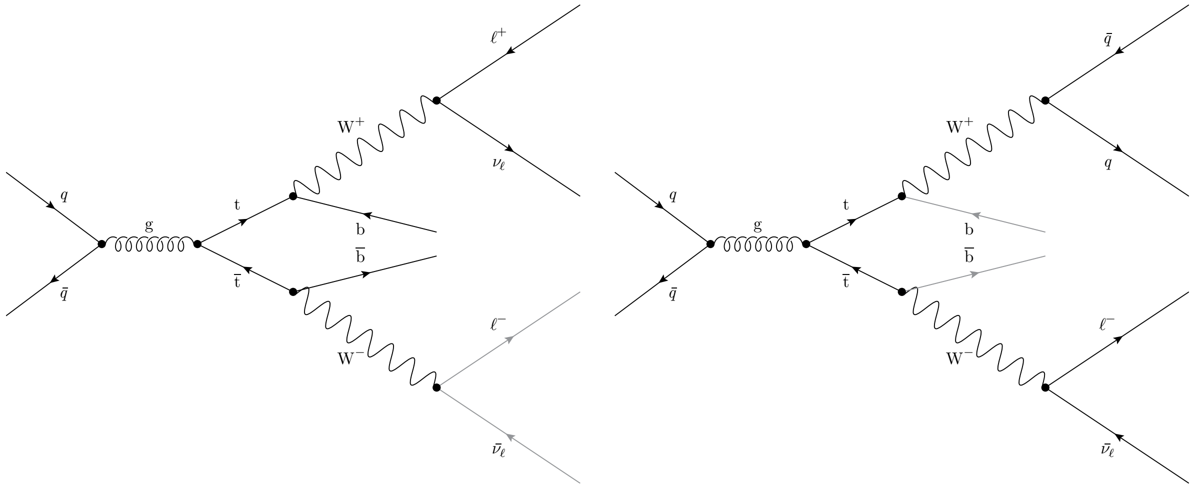


(a) W^+W^- pair production

(b) Z boson pair production

(c) Production of a W^+ and Z boson

Figure 2.12: Example Feynman diagrams for the standard model diboson processes decaying to the $\ell\nu jj$ final state.

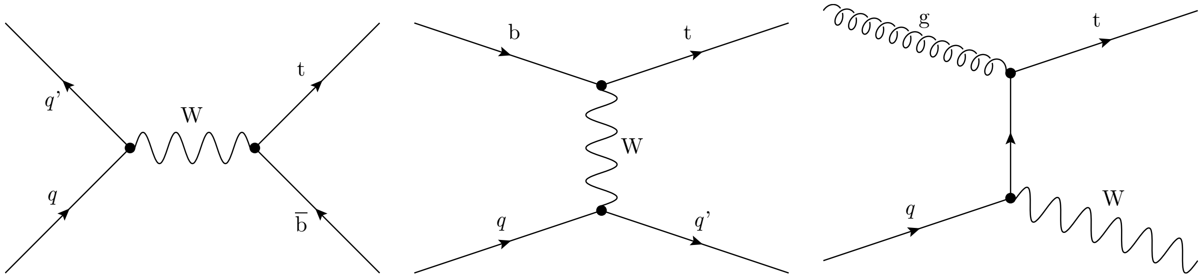


(a)

(b)

Figure 2.13: Two possible $\mathrm{t\bar{t}}$ Feynman diagrams which could have final states similar to the Higgs signal. Lines in gray are either mis-reconstructed or missing.

- **Diboson**: It is possible to mimic the final state signature with decays from several non-resonant diboson processes. The WW process is an irreducible background as it can exactly mimic our signal. The WZ process can produce the $\mathrm{l}\nu\mathrm{qq}$ final state in two ways: either the W decays leptonically and the Z decays hadronically or the W decays hadronically and one of the leptons from the Z decay is lost. The ZZ process is similar in that one lepton from the leptonic Z must be lost in order for the event to make it into the signal region.



(a) Production of a single top quark via the s-channel

(b) Production of a single top quark via the t-channel

(c) Production of a single top quark via the tW-channel

Figure 2.14: Example Feynman diagrams for the standard model single top processes. The final state particles are not pictured here.

- $\mathrm{t\bar{t}}$: The tops will each decays to a $\mathrm{b}$quark and a W boson via the weak interaction. If the W bosons decay semi-leptonically then the final state will be very similar, save for the presence of two additional b-quarks. If the b-jets can be identified then the events can be removed. Still, due to inefficiencies in identifying the b-quarks some $\mathrm{t\bar{t}}$ may still pass all selection requirements.

- **Single Top**: There are three production channels for this type of process: s-channel, t-channel, and the tW-channel. These processes have low cross sections and can produce reducible signatures.

- **Multi-jet**: This is the production of n jets where one jet is mistakenly identified as a lepton and the jet energies are mis-reconstructed enough to produce a sufficient imbalance in the event to mimic the neutrino. While this might seem improbable, the QCD cross section is quite large and thus this become a non-negligible background for this analysis.

The Feynman diagrams for all of these backgrounds, except for QCD, can be found in figs. 2.11, 2.12, 2.13, and 2.14.

2.7 Beyond the Standard Model

While the standard model has been an incredibly successful theory (see appendix A), it too has limitations. These shortcoming manifest themselves as either observations which are not covered by SM or characteristics of SM for which there is no fundamental explanation. In order to combat these shortcomings, a plethora of new theories have been created with the guiding principle that the new theories must be a superset of the standard model. That is, they must be able to reproduce all of the SM observations that have been so thoroughly tested. The following is a non-exhaustive list of shortcomings.

- **Gravity** is not included as either a field or particle within the Standard Model. In addition, there is no explanation as to why gravity is a much weaker force when compared to the electroweak or strong forces. Nevertheless, we expect that quantum gravity effects will become important at the Planck scale, $m_P \sim 10^{19}$ GeV. There have been attempts to create supergravity theories [61, 62, 63], but these have not yet been unified with the rest of the Standard Model. Most of these theories include a particle called the graviton, which is the quantum of a spin-2 field.

- According to cosmological experiments such as Planck, the universe is made of only about 5% ordinary, visible matter. Part of this remainder, about 26%, is made of what is termed **dark matter** (DM) [64, 65]. We know that this gravitationally interacting substance must exist because of astrophysical measurements of galactic rotation curves and galaxy cluster collisions [66, 67]. Still, the Standard Models does not provide any particle candidate. While the exact nature of DM is unknown, we do know that any DM particle must be stable, electrically neutral[4], weakly interacting, and a have a reasonably large mass. While this may sound like the SM neutrino, we already know that neutrino masses are too small [68]. This type of particle has been termed the WIMP or weakly interacting massive particle [66], but other candidates have been proposed as well [60].

- Besides visible and dark matter, the universe contains 69% of something else which has been termed **dark energy** and is not included in the Standard Model. Scientists know very little about dark energy other than that it seems to be causing the acceleration of universal expan-

[4] The term "dark" comes from the fact that DM does not interact with photons and therefore is not visible to the human eye.

sion, an action which could not come from any of the SM particles. Planck measurements indicate that dark energy is consistent with the theory of a cosmological constant. However, when there have been attempts to calculate the cosmological constant in terms of vacuum energy there have been mismatches of 100 orders of magnitude.

- Physicists expect that equal amounts of **matter and antimatter** were created during the Big Bang. Nevertheless the visible universe is filled with matter, but contains very little antimatter. The Standard Model offers no explanation for this discrepancy unless some of the symmetries were violated (i.e. baryon number conservation, CP invariance, and C conservation) [69, 70].

- As explained in section 2.5, Standard Model neutrinos are massless because they have no chiral right-handed counterparts and no Yukawa coupling with the scalar Higgs field. At the same time, there have been observations of **neutrinos oscillating** between flavors, which can only occur if at least two of the three neutrino types have mass [71, 72, 73, 74, 75]. To complicate matters, the physical neutrino eigenstates are mixtures of mass eigenstates (ν_1, ν_2, ν_3), which cannot be measured directly. There have been no direct measurements of the neutrino masses to date, but there have been upper limits placed on the masses and the squared mass differences are known.

- Studies of the Z boson have shown that no fourth generation of fermions with light neutrinos exists [76]. However, there is nothing in the Standard Model which forbids a **fourth generation**. Could there be a fourth family of fermions with heavy neutrinos?

- Is there a reason for the Standard Model fermion couplings to the Higgs boson? In other words, why do the fermion masses vary over five orders of magnitude from 0.511 MeV for the electron to 173 GeV for the top quark? This is sometimes called the **fermion mass hierarchy problem**.

- **Baryon and lepton conservation** are accidental symmetries without enforcement by a local gauge symmetry. Are these really conserved quantities?

- Why is the μ^2 from the Higgs potential negative? It needs to be negative to ensure EWSB, but there is no other compelling reason.

- We know that there are different mass scales in the universe. The Standard Model is effective at the electroweak scale of $\mathcal{O}\left(100\,\text{GeV}\right)$. However, at the Planck scale, $\mathcal{O}\left(10^{19}\,\text{GeV}\right)$, the model starts to break down and requires quantum gravity effects to be valid. As a consequence of the different scales, the bare parameters of the SM can differ from their renormalized values by several orders of magnitude. This in and of itself will not invalidate the model, but one would need to accept some amount of "fine tuning". These types of problems are called "hierarchy problems". There is one problem in particular, however, which is known colloquially as **the hierarchy problem**. Observed particle masses are a combination of the "bare" mass at tree level and the radiative corrections from loop diagrams. The problem comes from loop corrections to the Higgs mass parameter $\mu = m_h/\sqrt{2}$ introduced in section 2.5. The Higgs mass can be written in terms of the bare mass parameter μ_0 and radiative corrections $\delta\mu$

$$\mu^2 = \mu_0^2 + \delta\mu^2 \tag{2.73}$$

  The largest correction comes from the one-loop diagrams dealing with the top quark, the heaviest particle in the Standard Model. The one loop corrections to Higgs are shown in fig. 2.15. While fermions and bosons are protected from these divergences, scalars like the Higgs have a large dependence on the ultraviolet (UV) cutoff. This means that while the observed Higgs mass is $\sim 125\,\text{GeV}$, radiative corrections should drive μ^2 and thus the mass up to very large values. If the bare mass and radiative corrections happened to cancel at such a precise level as to lead to the observed mass it would be and *unnatural* amount of fine tuning [77]. Fine tuning problems like this have traditionally been interpreted as the existence of new physics [66].

In order to answer the open questions or provide a more complete theory, there have been

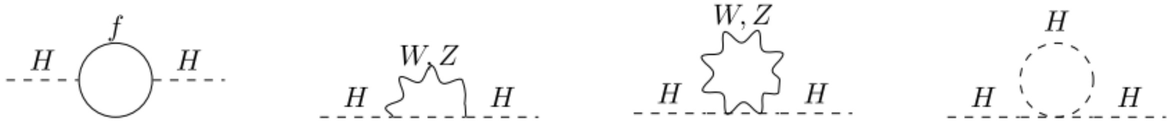


Figure 2.15: Feynman diagrams for the one-loop corrections to the Higgs boson mass. From left to right: contribution from the Yukawa interaction; two contributions from the gauge interaction; contribution from the Higgs self-interaction.

numerous models of beyond-the-SM (BSM) physics developed. While some of these models are still being tested by the LHC and other experiments, some previous BSM models were ruled out by Higgs discovery [78]. Below I list a small selection of BSM models which have been proposed as extensions to the SM.

- There are a whole host of little Higgs theories proposed [78, 79, 80]. In these models the Higgs boson is seen as the pseudo-Goldstone boson of a global symmetry broken around 10 TeV. In addition to the current array of SM particles, a little Higgs model would include new particles with the same spin as the SM particles. An additional symmetry called T-parity would be introduced, which says that particles must be introduced in pairs. This implies that the additions to the SM would only impact observables at the loop-level.

- Models of extra spatial dimensions say that the electroweak scale is the only fundamentally short distance scale and that loop corrections to the Higgs mass cut off at the electroweak scale and not the Planck scale [81]. The reduction in the cutoff scale leads to less fine tuning. A key feature of these models is that gravity, but not the other gauge interactions, permeates the new dimensions, which is why it is seen as being much weaker than the other forces. The Planck scale in $(4 + n)$ dimensions is assumed to be on the order of the electroweak scale. For $n \geq 2$ the size of the new dimensions is sub-millimeter, which is a scale where gravity has not been thoroughly tested.

- Supersymmetry (SUSY) was first proposed by Miyazawa in 1966 in order to relate mesons and baryons for hadronic physics [82, 83]. In the 1970s it was rediscovered as a QFT by several groups. In short SUSY introduces a new space-time symmetry that relates fermions to bosons and immediately provides a solution to the hierarchy problem [84, 85, 86, 87, 88, 89, 90, 91]. Each SM particle has a SUSY partner that differs in spin by 1/2. The coupling of the particles are chosen so that the Higgs mass does not diverge due to loop corrections. Additionally, SUSY causes EWSB in a different way that does not require negative μ^2, answering another question left by the SM. Although SUSY models can cause protons to decay, many introduce R-parity to prevent this. If R-parity conserved then the lightest supersymmetric particle (LSP) is stable (i.e. the LSP can't decay without violating R-parity). SUSY is a appealing model because the LSP also provides a DM candidate as the particle would be heavy and weakly interacting. While the theory has significant promise, SUSY has not yet been observed. This however, does not mean that SUSY is wrong. It simply means that the masses of the supersymmetric partners are not the same as their SM partners (i.e. SUSY is broken somehow).

3. THE LHC AND CMS DETECTOR

3.1 The Large Hadron Collider

The Large Hadron Collider (LHC) [92] is, in many people's estimation, the largest and most complex machine ever built by humanity. The main accelerator at the European Organization for Nuclear Research (CERN), the LHC is located both in France and Switzerland due to its enormous size (Fig. 3.1). It was built between 1998 and 2008 and installed in the 26.7 km tunnel dug for its predecessor, the Large Electron-Positron Collider (LEP), which is located between 50 m and 170 m underground. It is the highest energy collider in the world, eclipsing the previous record holder, the Tevatron at Fermilab in Batavia, IL. The following section is a description of the LHC and CERN accelerator complex, a more detailed description can be found in [92] and [93].

The LHC provides beams for several experiments located along its beam line, though we will only concern ourselves with the four highest profile experiments (Fig. 3.1):

- The CMS (Compact Muon Solenoid) [12] and ATLAS (A Toroidal LHC ApparatuS) [94] experiments are both general purpose detectors. Their goals include precision measurements to test the Standard Model and searches for new physics, including the Higgs boson.

- LHCb (Large Hadron Collider beauty) [95] was designed to do precision measurements of CP-violation and the physics of B-mesons.

- ALICE (A Large Ion Collider Experiment) [96] studies heavy ion collisions.

The LHC was designed to collide two beams of protons (pp), heavy ions (PbPb), or a combination of the two (pPb) at specific interaction points around the beam line. For the purposes of this thesis we will only cover proton-proton collisions from this point forward. The protons come from a single bottle of hydrogen gas, which is then disassociated and stripped of electrons to form a proton beam. Interestingly, only 1 ng of hydrogen is required per day in order to form the LHC beams. The protons next travel through the Linac2 machine where they are bunched by radio frequency

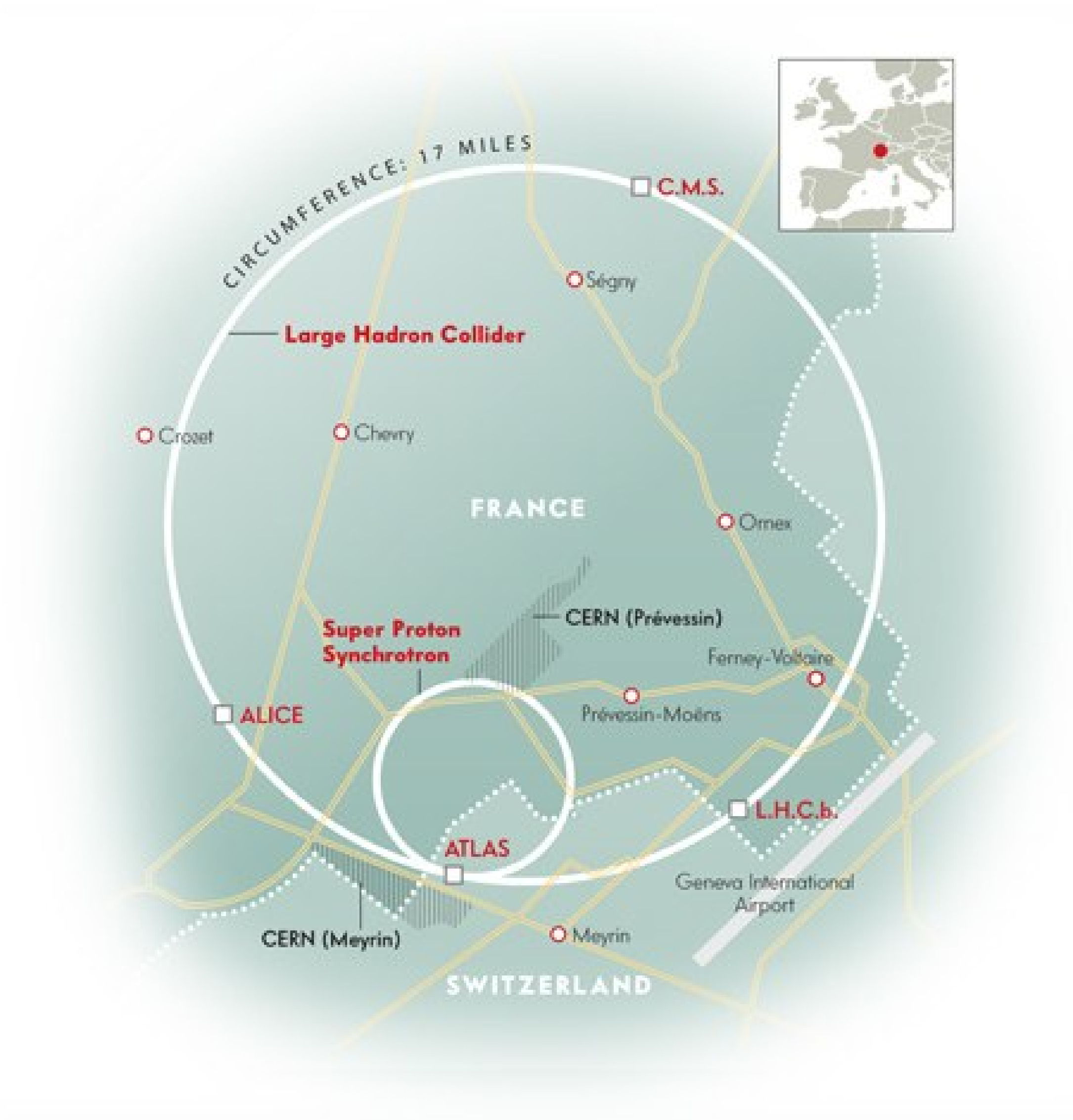


Figure 3.1: Overhead view of CERN and its main experiments, CMS; ATLAS; LHCb; and ALICE, as well as two of the larger accelerators, the LHC and SPS. The schematic is overlaid on a map of Switzerland and France. Reprinted from [4].

(RF) electromagnetic fields and are accelerated to 50 MeV. This chain continues through the Proton Synchroton Booster (PSB), the Proton Synchrotron (PS), and the Super Proton Synchrotron (SPS) where the protons are accelerated to 1.4 GeV, 26 GeV, and 450 GeV respectively (Fig. 3.2).

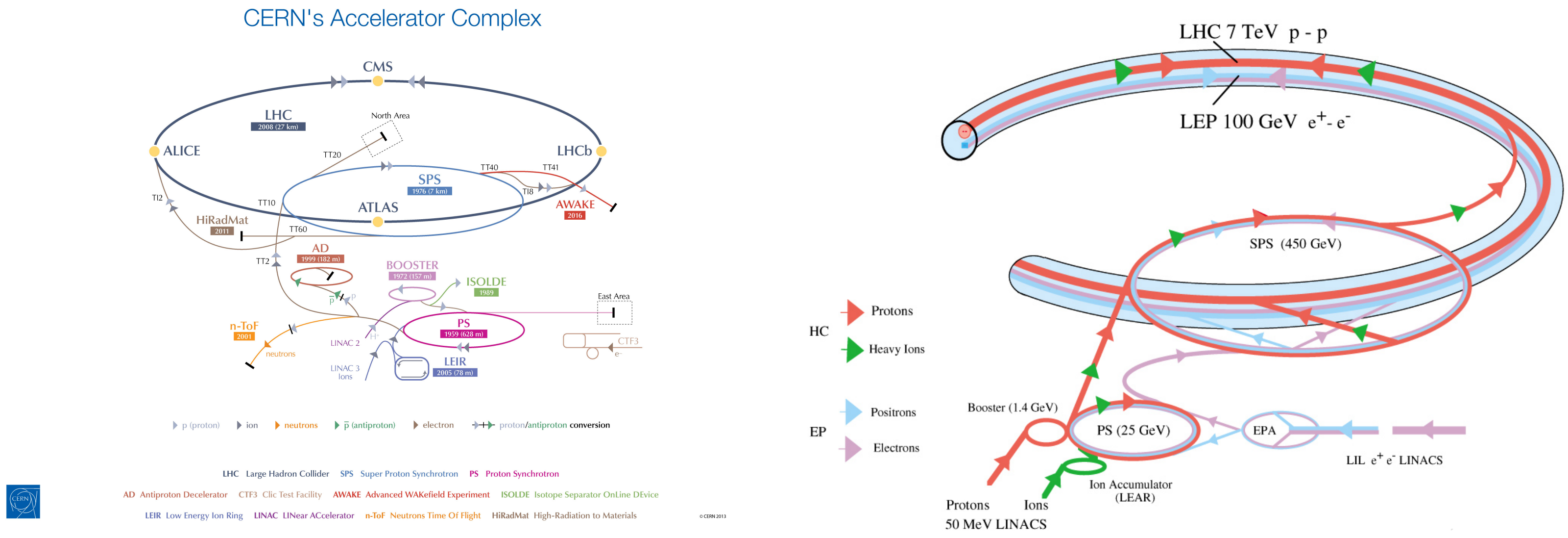


Figure 3.2: Left: A schematic of the CERN accelerator complex [5]. Right: A diagram of the LHC injection chain. Also included is a diagram of the heavy ion and LEP injection chains. Reprinted from [6].

After being accelerated in the SPS, the proton bunches are injected into the two LHC beam pipes, which were designed to accelerate the two proton beams to 7 TeV (Fig 3.3). Size limitations in the tunnel dictated that the the beam lines be formed by twin bore magnets. Each magnet is formed by a single mechanical structure and cryostat while containing two coils and two beam channels. The coils are made out of superconducting NbTi Rutherford cables cooled to 1.9 K by 120 t of superfluid helium. This forms the 8.33 T magnets necessary for bending the 7 TeV protons (Fig. 3.4). The LHC contains 1232 superconducting dipole magnets for bending the protons and 392 superconducting quadrupole magnets for focusing the beams. The beam line also contains sextapole, octopole, and decapole magnets, which are also used to focus and correct the direction of the beams. The original LHC design calls for a bunch spacing of 25 ns, 10^{11} protons per bunch, and 2808 bunches per beam.

The original plan was to start the LHC accelerator complex in September 2008. However, due to a catastrophic incident damaging the machine, the startup was delayed until November 23, 2009; even then colliding beams only had a center-of-mass energy of 900 GeV. From March 30, 2010 through the end of 2011 the LHC operated with a center-of-mass energy of 7 TeV. Then in 2012 the energy was again increased to 8 TeV (4 TeV per beam), which is the energy of the beams during the data-taking period focused on by this thesis. It is important to note, though, that the machine has continued to operate after the 2012 data taking period and increased the center-of-mass energy to 13 TeV starting in 2015 (there was a planned shutdown from 2013 through early 2015).

In addition to the center-of-mass energy, collider physicists are interested in the rate at which a specific physics process occurs. This in turn is related to the cross sections, the probability that two particles will collide and react a certain way, and the luminosity. The rate of events is given by equation 3.1, where $\mathcal{L}$ is the collision luminosity and σ is the cross section for a given physical process.

$$dN/dt = \mathcal{L}{\cdot}\sigma \tag{3.1}$$

The luminosity as it is described here is often called the “instantaneous luminosity” as this

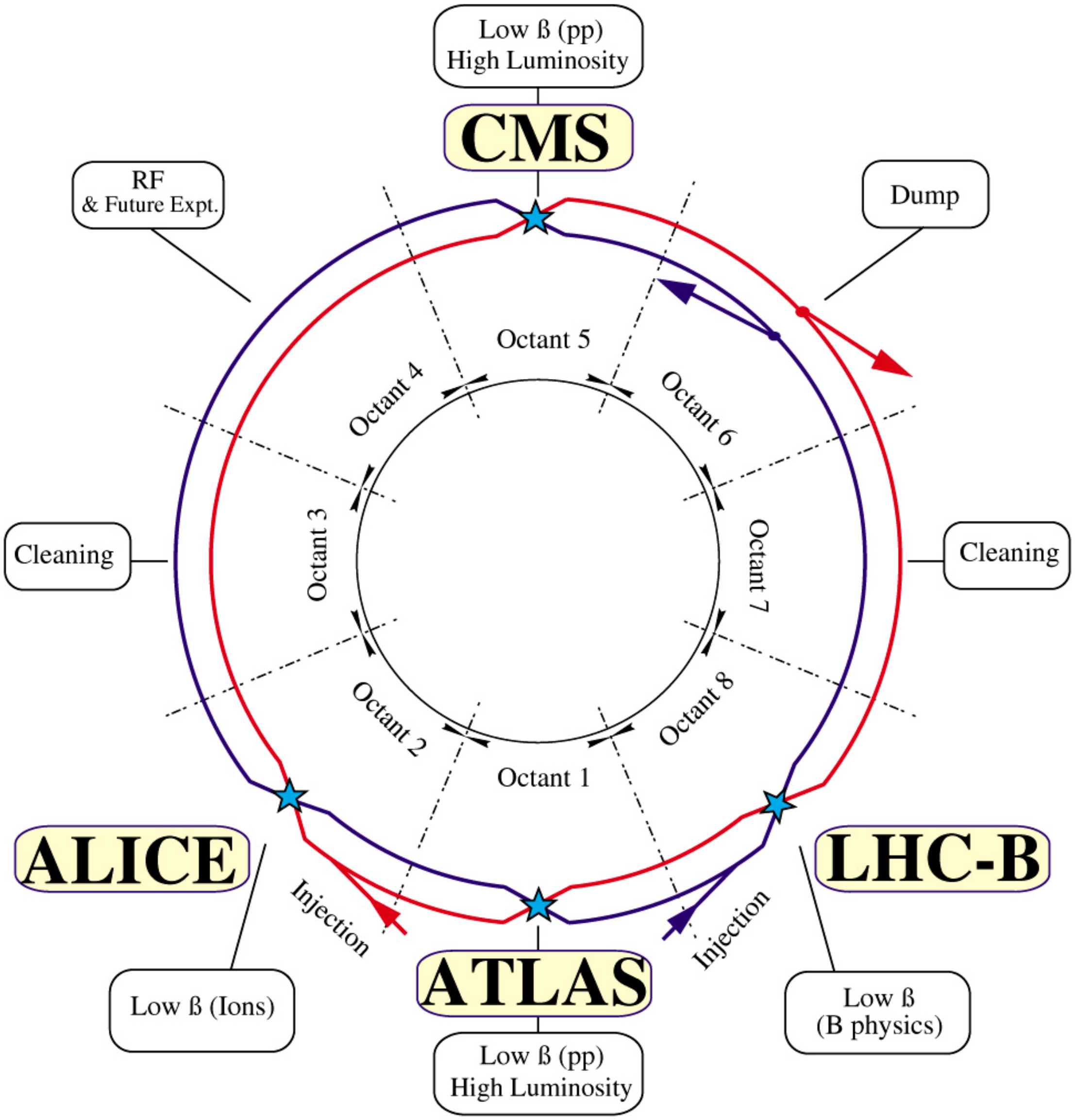


Figure 3.3: A diagram of the LHC beams along with the four major experiments. Reprinted from [7].

value can change from moment to moment. The "integrated luminosity" is then a measure of the total amount of data collected. The instantaneous luminosity itself depends upon the parameters of the LHC beams and the optical properties of the focusing system at the interaction point. This information is summed up in equation 3.2 [97]:

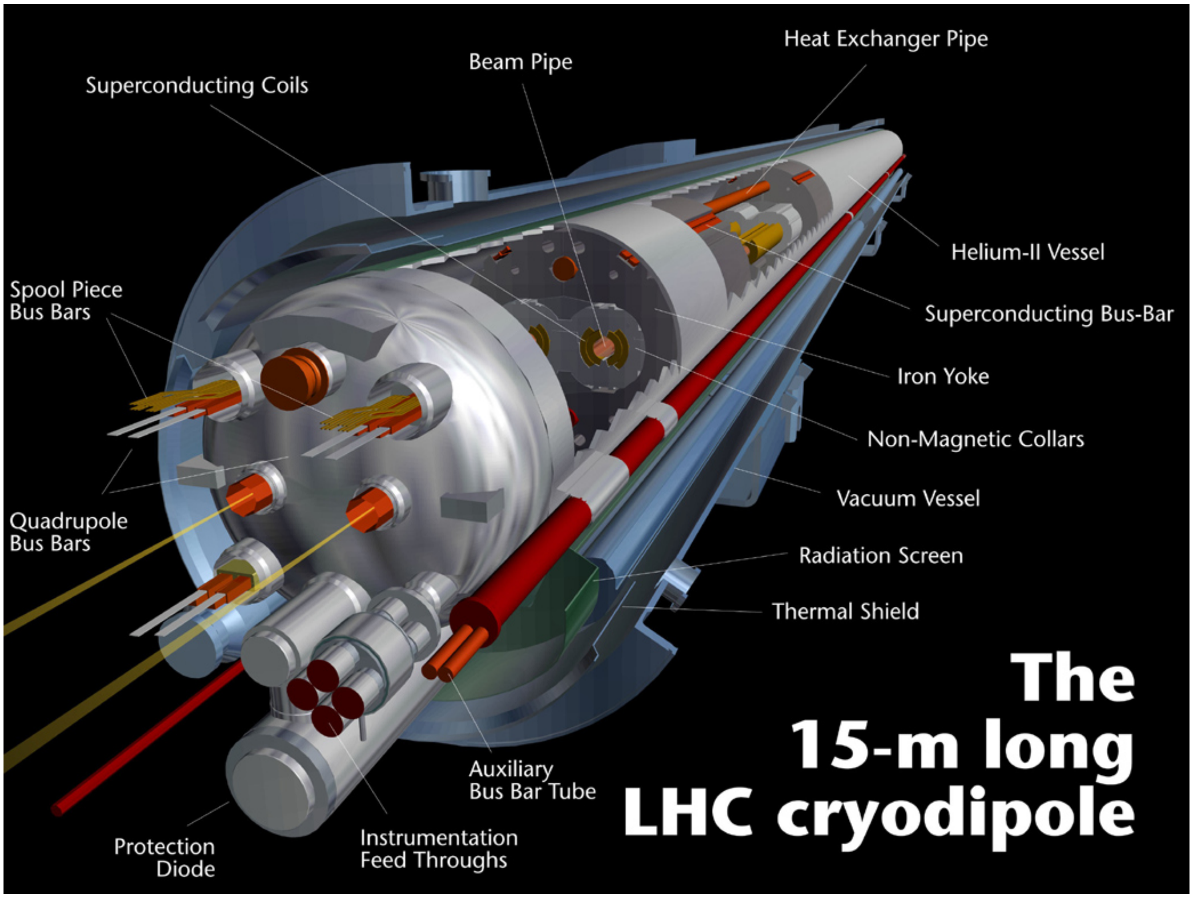


Figure 3.4: A diagram of an LHC dipole magnet and cryostat. Reprinted from [8].

$$\mathcal{L} = \frac{N^2 n_b f \gamma}{4\pi \epsilon_n \beta^*} F \tag{3.2}$$

where:

- N: protons per bunch
- n_b: bunches in the LHC ring
- f: frequency of bunch revolutions around the ring
- γ: relativistic factor for the protons
- ϵ_n: normalized emittance of the proton beams

- β^*: beta function at the interaction point
- F: geometrical reduction factor due to the crossing angle of the beams

The maximum design luminosity of the LHC is $1 \times 10^{34}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$. During the 2010 and 2011 run periods (7 TeV center-of-mass energy) the instantaneous luminosity increased from $1 \times 10^{32}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$ to $5 \times 10^{33}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$. During the 2012 data-taking period, the peak instantaneous luminosity was $7.67 \times 10^{33}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$ with a bunch spacing of 50 ns, a maximum number of bunches of 1380, and $\sim 2.2 \times 10^{14}$ protons per beam ($\sim 1.6 \times 10^{11}$ protons per bunch). The LHC delivered 23.30 fb^{-1} of integrated luminosity to the CMS detector of which 21.79 fb^{-1} was recorded. As of the end of 2017, the LHC is still running at 13 TeV (6.5 TeV per beam) with a peak luminosity of $2.04 \times 10^{34}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$, 1868 bunches, and 1.25×10^{11} protons per bunch [98, 9]. Figures 3.5 and 3.6 show the total integrated luminosity delivered by the LHC and recorded by the CMS experiment for the various data-taking periods [9].

3.2 The CMS Detector

The CMS experiment is one of two general purpose detectors at the LHC tasked with a wide variety of physics analyses. The goals of the physics program range from precision Standard Model measurements to the search for physics beyond the Standard Model and even includes a hugely successful heavy ion program. The detector itself is located 100 m underground near Cessy, France on the opposite side of the LHC from the main CERN site in Meyrin (see fig. 3.1). It was largely built on the surface and then lowered into the collision cavern in 15 pieces, which then had to be assembled. The detector has a cylindrical design which is 22 m in length, 15 m in diameter, and weights 14000 tonnes. The shape and positioning of the detector around the interaction point (IP) gives the experiment nearly 4π coverage of the proton-proton collisions. In total, there are $\sim 10^8$ data channels checked in each bunch crossing owing to the high granularity of the CMS sub-detectors. The layout of the detector can be seen in fig. 3.7. The following sections will describe each of the sub-detectors and its properties [12].

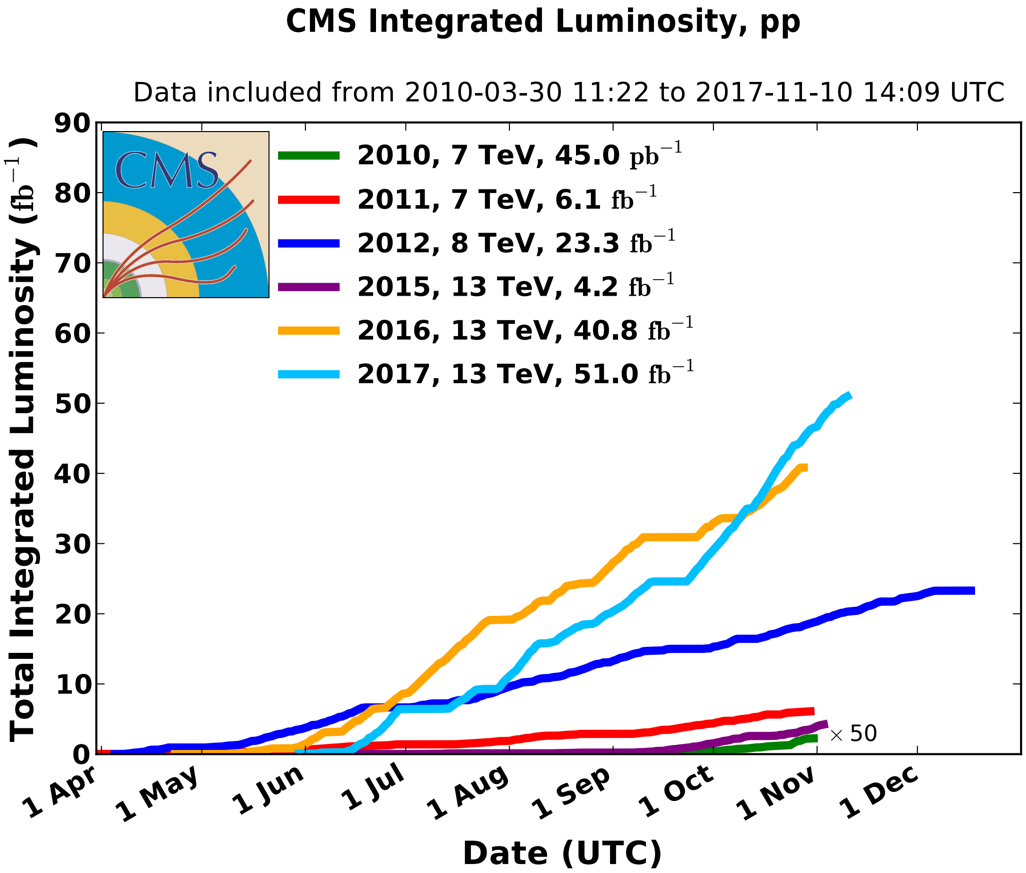


Figure 3.5: Total integrated luminosity versus time delivered to the CMS experiment for the 2010, 2011, 2012, 2015, 2016, and 2017 p-p data-taking periods. Reprinted from [9].

3.2.1 Coordinate System

The IP is at the center of the detector and is the origin of the right-handed coordinate system used to describe the detector and the physics being measured (location and direction). The z-axis is defined along the LHC beam line. Instead of using the polar angle, θ, which would go from 0° along the positive z-axis to 90° pointing straight up from the interaction point, collider physicists use the quantity pseudorapidity defined as $\eta = -ln\left[\tan\left(\theta/2\right)\right]$. The benefits of using the pseudorapidity are that differences in this coordinate, $\Delta\eta$, are invariant under boosts in the z-direction and particle production is roughly uniform in η. The x- and y-axes form the plane perpendicular to the z-axis, where positive x points to the center of the LHC ring and positive

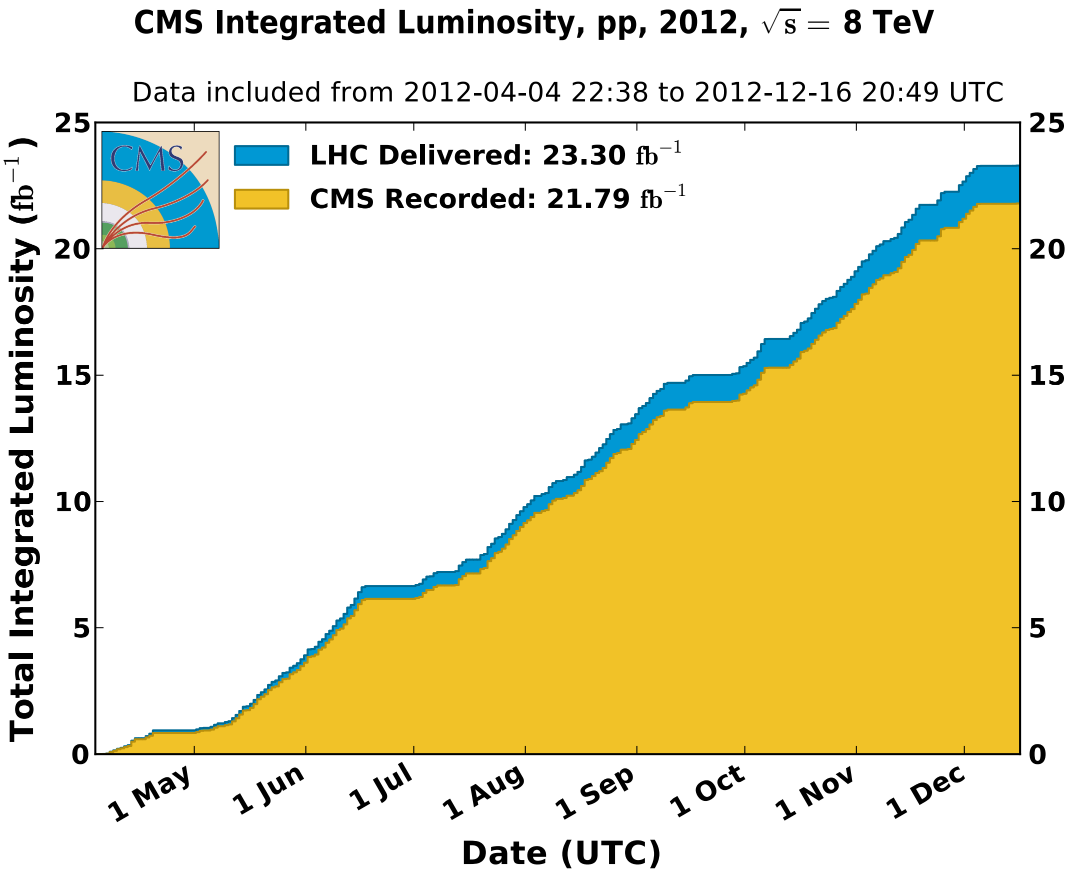


Figure 3.6: Total integrated, offline luminosity versus day in 2012. The blue graph shows the delivered luminosity while the orange graph shows the luminosity recorded by the CMS experiment. This graph shows only the luminosity collected for p-p collisions during stable beams. Reprinted from [9].

y points upward. The azimuthal angle, φ, and radial coordinate, r, are also defined in this same plane. It is sometimes more useful to use φ and r due to the bending of the particles in the magnetic field. Lastly, this paper will often refer to the quantity p_{T}, which is the magnitude of the component of the momentum vector in the transverse plane. A schematic of the coordinate system described above is shown in fig. 3.8.

3.2.2 Tracker and Pixel Detector

The CMS all-silicon tracker is the closest sub-detector to the LHC beam pipe. Its purpose is twofold; to determine the charged-particle direction at its production vertex and to measure the

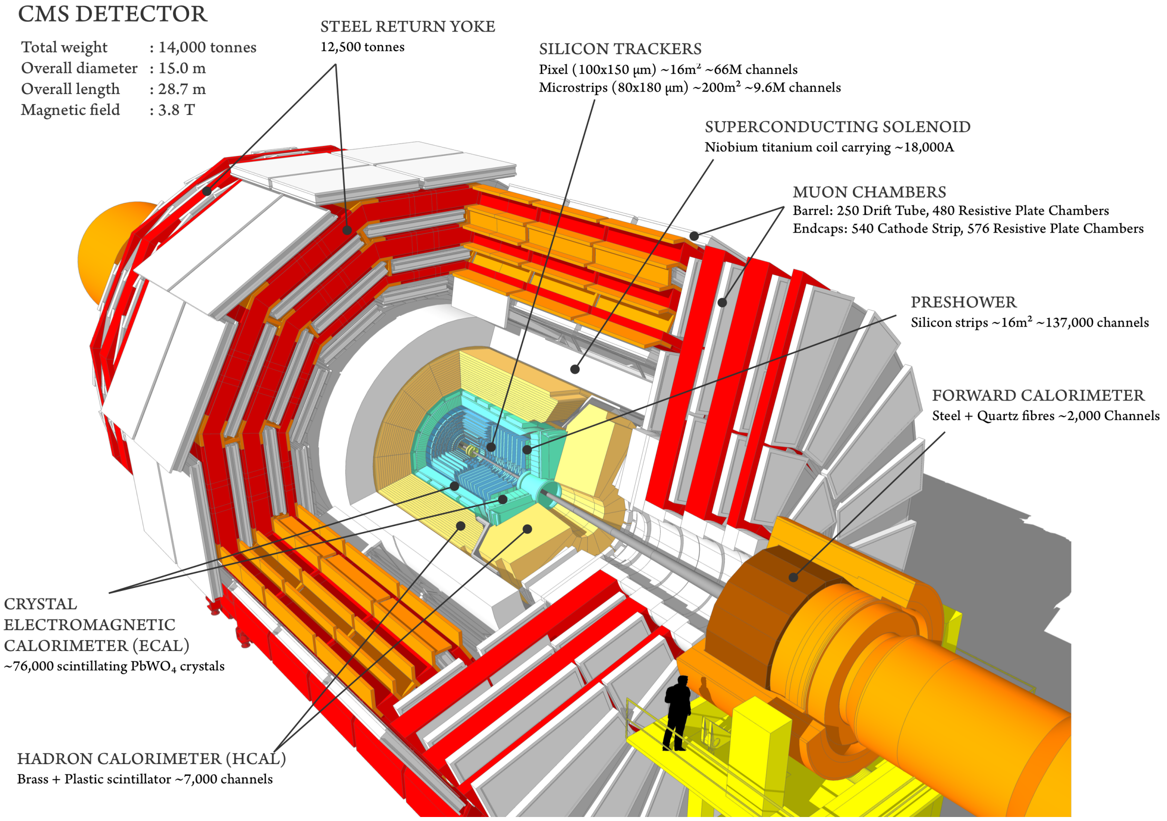

CMS DETECTOR
Total weight : 14,000 tonnes
Overall diameter : 15.0 m
Overall length : 28.7 m
Magnetic field : 3.8 T
STEEL RETURN YOKE
12,500 tonnes
SILICON TRACKERS
Pixel (100x150 μm) ~16m^2 ~66M channels
Microstrips (80x180 μm) ~200m^2 ~9.6M channels
SUPERCONDUCTING SOLENOID
Niobium titanium coil carrying ~18,000A
MUON CHAMBERS
Barrel: 250 Drift Tube, 480 Resistive Plate Chambers
Endcaps: 540 Cathode Strip, 576 Resistive Plate Chambers
PRESHOWER
Silicon strips ~16m^2 ~137,000 channels
FORWARD CALORIMETER
Steel + Quartz fibres ~2,000 Channels
CRYSTAL ELECTROMAGNETIC CALORIMETER (ECAL)
~76,000 scintillating PbWO$_4$ crystals
HADRON CALORIMETER (HCAL)
Brass + Plastic scintillator ~7,000 channels

Figure 3.7: A view of the CMS detector with major sub-detectors labeled and notable facts. A human silhouette is included for scale. Reprinted from [10].

momentum of charged particles. In the later case, the tracker is far superior to the calorimeter systems for p_{T} up to several hundred GeV. The sub-detector is 5.8 m long and 2.5 m in diameter, covering a pseudorapidity range of $|\eta| < 2.5$. It is, by necessity, highly granular, to keep the occupancy low, and relatively radiation hard. The tracker is exposed to extreme doses of radiation ranging from 0.18 to 84 Mrad after 500 fb^{-1} of data. The radiation tolerance was a key factor in determining the materials and design of the sensors and on-board electronics of the tracker. To keep the radiation damage as low as possible, among other benefits, the tracker is kept at $-10\,^{\circ}$C. For non-isolated particles of $1 < p_{\mathrm{T}} < 10\,$GeV and $|\eta| < 1.4$, the track resolutions are typically 1.5% in p_{T} and 25–90 (45–150) μm in the transverse (longitudinal) impact parameter. On the other hand, isolated particles of $p_{\mathrm{T}} = 100\,$GeV emitted at $|\eta| < 1.4$ have track resolutions of 2.8% in p_{T} and 10 (30) μm in the transverse (longitudinal) impact parameter [99]. At higher η the reduced

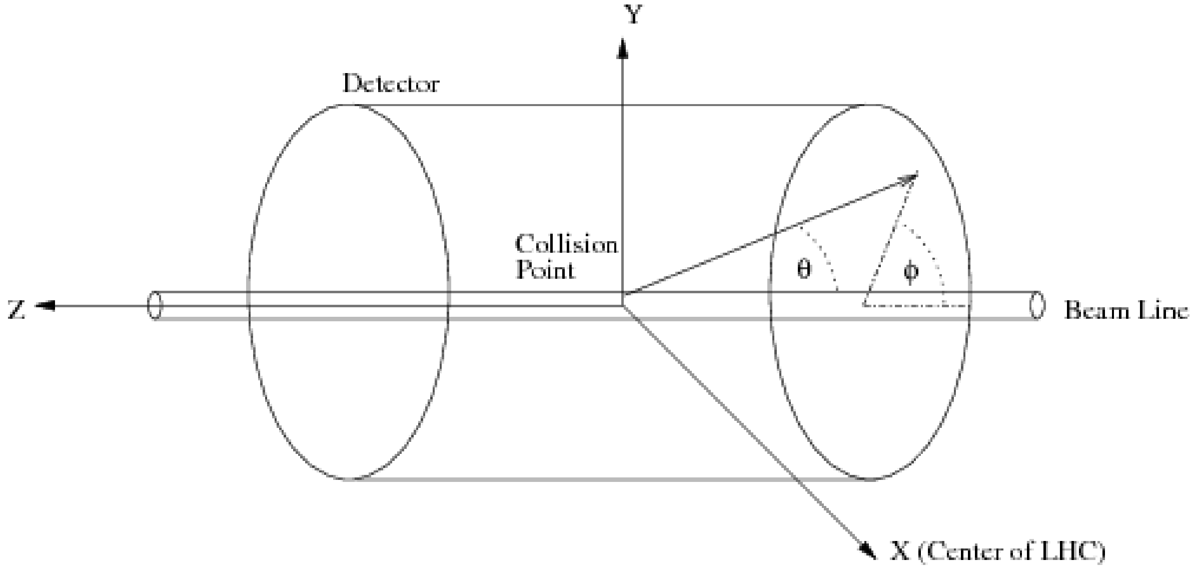


Figure 3.8: Schematic of the CMS coordinate system. Reprinted from [11].

transverse depth of the tracker degrades the resolution (particles traverse fewer layers). Fig. 3.9 shows the layout of the tracker and its subsystems. The tracker is formed by two major subsystems, the pixel detector and the silicon strip tracker.

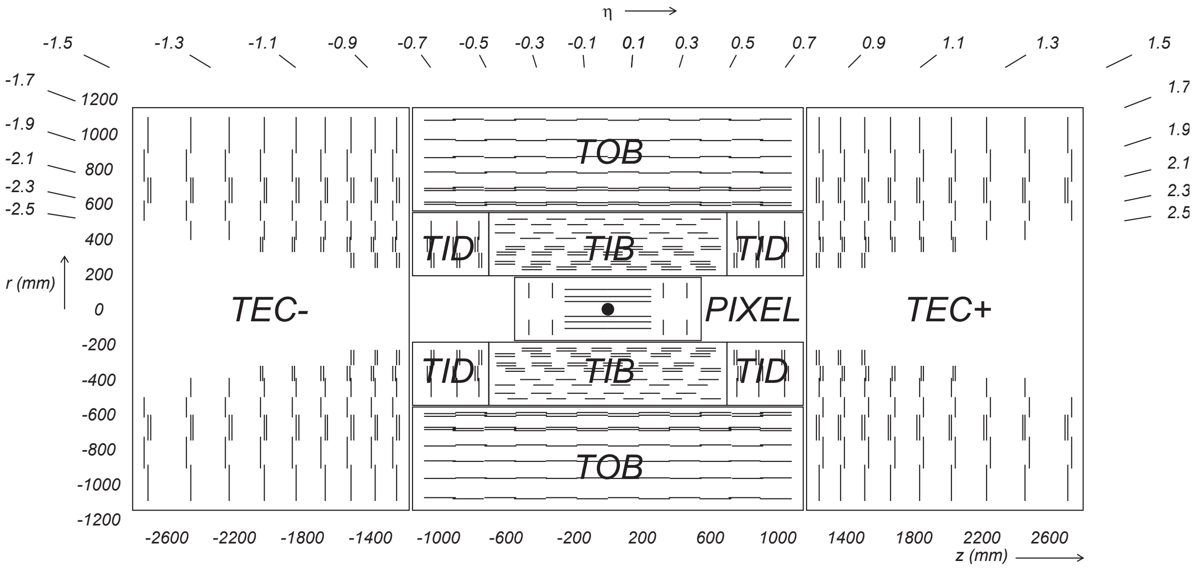


Figure 3.9: Layout of the CMS tracker with subsystems labeled.

The pixel detector is made up of three barrel layers, called the BPIX, and two endcap layers called the FPIX. The BPIX contains 48 million pixels and the FPIX contains another 18 million pixels. In total it consists of 1440 hybrid silicon detector modules, each with a dimension of $100 \times 150\,\mu\text{m}^2$. The small pixel size enables track resolutions of $10\,\mu$m in the transverse plane and $20\,\mu$m in the z-direction. The pixel detector is what gives CMS its excellent secondary vertex tagging ability in addition to producing seed tracks for the strip tracker and the high level trigger.

Just as the pixel detector was made up of the BPIX and FPIX subsystems, the silicon strip detector is made up of four subsystems. The Tracker Inner Barrel (TIB) has four layers of $320\,\mu$m strips. At each end of the TIB is a three-layer Tracker Inner Disks (TID), which contains strips of the same thickness. The Tracker Outer Barrel (TOB) is the six layer system which surrounds the TIB/TID. The first four layers of the TOB use $500\,\mu$m thick strips, and the last two layers use $122\,\mu$m thick strips. The Tracker EndCaps (TEC) are on either side of the previous setup and contains nine disks with up to seven layers of strips. These strips are $320\,\mu$m thick in the inner four rings and $500\,\mu$m thick in the outer three rings. In total, the strip detector contains 9.3 million silicon strips (15 148 modules).

The 2012 LHC run was an excellent year for the tracker. The BPIX maintained 97.7% of its channels operational while the FPIX had 92.8% of its channels operational. The reconstruction efficiencies were also quite high, 99.5%, for each later of the pixel detector (>99.2% for the first layer). The strip detector maintained 97.5% of its channels active and had a reconstruction efficiency greater than 99% for each layer[100].

3.2.3 Electromagnetic Calorimeter

The electromagnetic calorimeter (ECAL) is a homogeneous detector consisting entirely of 75 848 lead tungstate crystals ($PbWO_4$). The detector is divided up into two sections which provide coverage in pseudorapidity $|\eta| < 1.479$ in a barrel region (EB) and $1.479 < |\eta| < 3.0$ in two endcap regions (EE). There are also preshower detectors (PS) in each of the endcaps, in front of the EE, which cover a pseudorapidity range of $1.653 < |\eta| < 2.6$. Fig. 3.10 shows the structure of the detector with the key η values labeled.

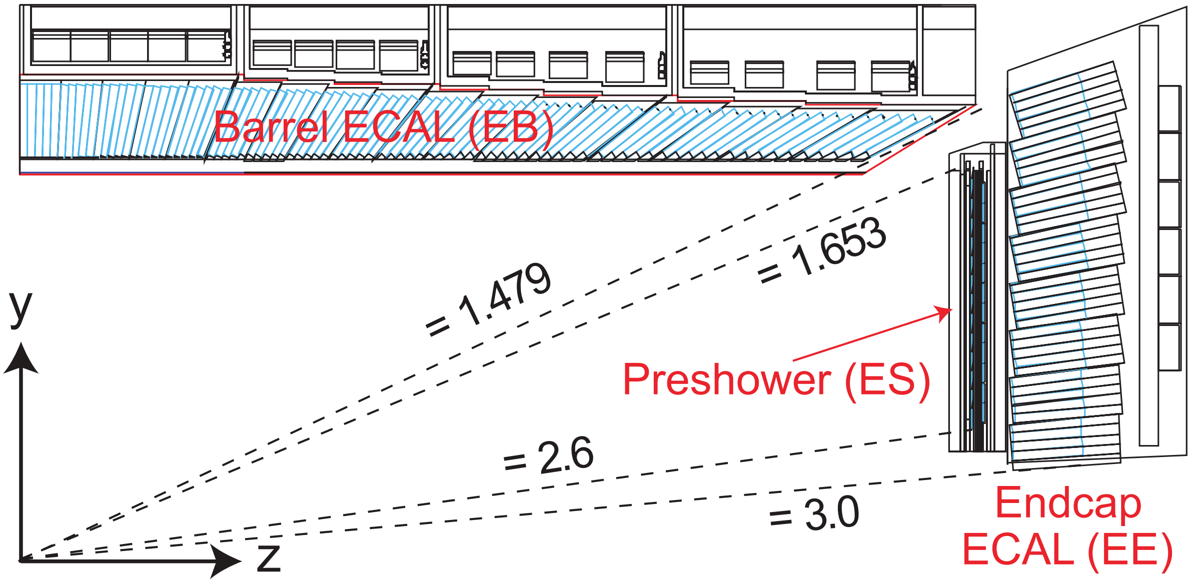


Figure 3.10: A schematic of the CMS ECAL detect with labeled subsystems and key η ranges marked.

The barrel region of the ECAL consists of 61 200 crystals with a tapered shape arranged in a projective geometry. Each crystal is about 0.0174×0.0174 in $\eta - \varphi$, which corresponds to 22×22 mm^2 at the front face and 26×26 mm^2 at the back face. Each crystal has a depth of 230 mm, which for $PbWO_4$ corresponds to 25.8 radiation lengths (X_0). The scintillation light produced in the crystals is read out by avalanche photodiodes (APDs), which produce approximately 4.5 photoelectrons per MeV at 18 °C. The dark current of the APDs is sensitive to radiation exposure. During the 2012 run, the dark current ranged from 0.13 to 1.3 μA on average, which corresponds to an average noise of 47 to 57 MeV [101].

The EE contains 14 648 $PbWO_4$ crystals arranged in a non-projective $x - y$ geometry (see fig. 3.10). The crystal dimensions are 28.62×28.62 mm^2 at the front face and 30×30 mm^2 at the back face with a depth of 220 mm or 24.7 X_0. Instead of using APDs link in the EB, the EE uses vacuum phototriodes (VPTs) to read out the scintillation light. Again holding the photodetectors at 18 °C, the phototriodes produce 4.5 photoelectrons per MeV. The average noise in the VPTs for 2012 was 180–200 MeV, but it could reach 600 MeV at high η due to the higher radiation doses in the more forward regions [101].

The ES is located in front of each of the EE detectors. It consists of two planes of silicon strip sensors interleaved with a total of $3X_0$ of lead absorber (2 X_0 for the first layer and 1 X_0 for the second layer). The silicon strips are 320 μm thick and can collect 3.6 fC of charge from a minimum ionizing particle (MIP).

One of the main goals of the CMS experiment was to discover the Higgs boson. Because of its low irreducible standard model background, the $H \to \gamma\gamma$ channel was considered the "golden channel". Due to this, a significant amount of money and time was spent on the design and the materials for the ECAL. $PbWO_4$ is a great choice for an ECAL because its properties, listed in table 3.1, lead to a precision energy measurement for EM objects (by this I mean a fine small resolution).

Property	Value
Peak emission wavelength	425 nm
High density	8.28 g/cm^3
Short radiation length	0.89 cm
Short Molière radius	2.2 cm
Fast decay time	6 ns

Table 3.1: $PbWO_4$ properties and their measured values

The energy resolution, σ, of deposits in the ECAL vary as a function of energy (E) (in units of GeV). This is typically modeled using an NSC function as in equation 3.3:

$$\left(\frac{\sigma}{E}\right)^2 = \left(\frac{N}{E}\right)^2 + \left(\frac{S}{\sqrt{E}}\right)^2 + C^2 \tag{3.3}$$

where N is the noise term, S is the stochastic term, and C is the constant term. Typical values for these terms come from test beam studies and are listed in table 3.2 [101]. In the barrel section of the ECAL, an energy resolution of about 1% is achieved for unconverted or late-converting photons in the tens of GeV energy range. The remaining barrel photons have a resolution of about 1.3% up to a pseudorapidity of $|\eta| = 1$, rising to about 2.5% at $|\eta| = 1.4$. In the endcaps, the resolution of unconverted or late-converting photons is about 2.5%, while the remaining endcap photons have a resolution between 3% and 4% [102].

Term	Typical Value
N	12%
S	2.8%
C	0.30%

Table 3.2: Typical values for the noise, stochastic, and constant terms of the ECAL energy resolution function. These values are obtained from test beam studies.

3.2.4 Hadron Calorimeter

The CMS hadron calorimeter (HCAL) is, as its name suggests, designed to measure the energy of hadrons. This is especially important for neutral hadrons which leave no tracks and, to a large extent, do not register in the ECAL. The HCAL is a sampling calorimeter, meaning that is contains both an active, energy measurement material as well as a material which induces the hadrons to shower. The HCAL is made up of four subsystem: HCAL barrel (HB), HCAL endcap (HE), HCAL outer (HO), and HCAL forward (HF). The HB, HE, and HO subsystems all use the same technology, while the HF uses a different technology. Fig. 3.11 shows the structure and position of the HCAL subsystems. When both the ECAL and HCAL work together, the CMS calorimeters can measure a charged pion with a resolution of $\sigma/E \approx 100\%/\sqrt{E[GeV]} \oplus 5\%$, where E is the jet energy.

The HB occupies the region $|\eta| < 1.3$ and contains alternating layers of brass and scintillator. The number of nuclear interaction lengths (λ_0) ranges from 5.82 at $\eta = 0$ to 10.6 at $\eta = 1.3$. Additionally the EB, which is directly in front of the HB, has $1.1\lambda_0$ and can measure a portion of early developing hadronic showers, though not as accurately. The properties of the brass used can be found in table 3.3 while the layer thicknesses and materials can be found in table 3.4. Most of the plastic scintillating layers are 3.7 mm thick, but layer 16 is 9 mm thick so that it can sample more from late developing showers. There is also an additional 9 mm thick layer 0 before

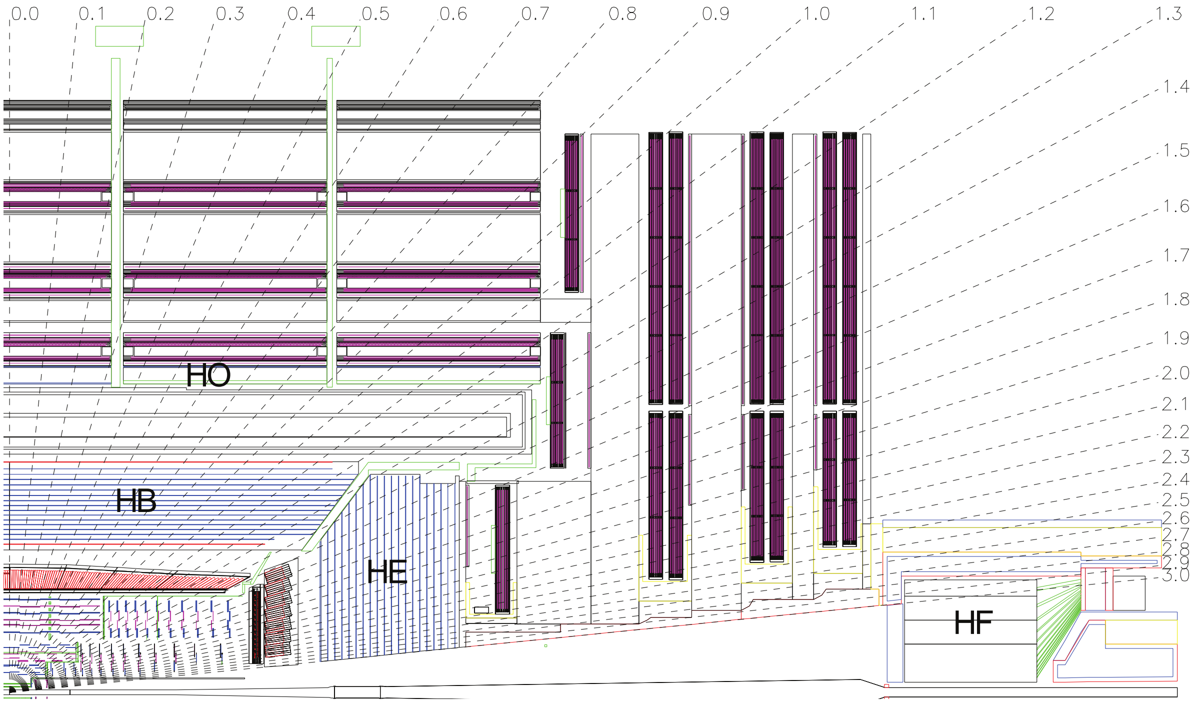


Figure 3.11: A schematic of the CMS HCAL detector with its major subsystems labeled: HB, HE, HO, and HF.

the first absorbing layer to catch the showers which are initiated in the dead material between the EB and HB. The scintillating tiles are arranged in a projective geometry (pointing close to the nominal interaction point) with the tiles occupying 0.087 × 0.087 in $\eta - \varphi$. For $|\eta| < 1.479$, the HCAL cells map on to 5×5 arrays of ECAL crystals to form calorimeter towers. Within each tower, the energy deposits in ECAL and HCAL cells are summed to define the calorimeter tower energies, subsequently used to provide the energies and directions of hadronic jets. The scintillator is separated into 16 η section and 36 φ sections with almost 70000 tiles used. The light is collected by wavelength shifting (WLS) fibers that encircle the tiles. Fibers from several layers are read out by one hybrid photodiode (HPD), which are used for their large dynamic range and low sensitivity to magnetic fields.

In the central region of the detector there are too few λ_0 to fully contain a hadronic shower. For this reason the HO system was added as a scintillating tile extension to the HB. The HO consists of five rings, each with a width of 2.536 m in the z-direction. The most central ring, Ring 0, has

Property	Value
Materials	Brass (70% Copper and 30% Zinc) or Steel
Density	8.53 g/cm^3 %
Radiation Length	1.49 cm
Nuclear Interaction Length	16.42 cm

Table 3.3: Properties of the brass absorber used for the CMS HB.

Layer number(s)	Material	Thickness (mm)
1	Steel	40
2-9	Brass	50.5
10-15	Brass	56.5
16	Steel	75

Table 3.4: Absorbing layer thicknesses and materials for the CMS HB

two scintillating layers, one inside the solenoid and one outside the solenoid. The other rings have only one layer outside of the solenoid, which acts as a 19.5 cm iron absorber layer. This addition to the HB brings the total depth of the CMS calorimeter systems to $11.8\lambda_0$.

The HE, a 17 layer sampling calorimeter, covers the $1.3 < |\eta| < 3.0$ region. It consists of 79 mm brass absorbing layers and uses the same scintillating material as is used in the HB, but contains only 20916 tiles. Within $|\eta| < 1.6$ the granularity of these tiles is the same as for the HB, but at higher η the approximate granularity becomes 0.174×0.174 in $\eta - \varphi$. Like the HB, the HE also has a layer 0. However, unlike the HB, the scintillating layers in the HE are grouped into "depths" before the light reaches the HPDs. Fig. 3.12 shows a schematic of the CMS HCAL system where the different colors corresponds to the various depths. This depth segmentation allows for a more precise recalibration of the HE, which receives a higher radiation dose than the HB. When combined with the EE, this section of the detector corresponds to a length of $10\lambda_0$.

The HF uses steel as an absorber and embedded quartz fibers as the sensitive material. The reason for the change in technology is that the HF needs to be able to withstand at least 100 Mrad/year. The two halves of the HF are located 11.2 m from the interaction region, one on each end, and together they provide coverage in the range $3.0 < |\eta| < 5.2$. Unlike the other hadronic calorimeter

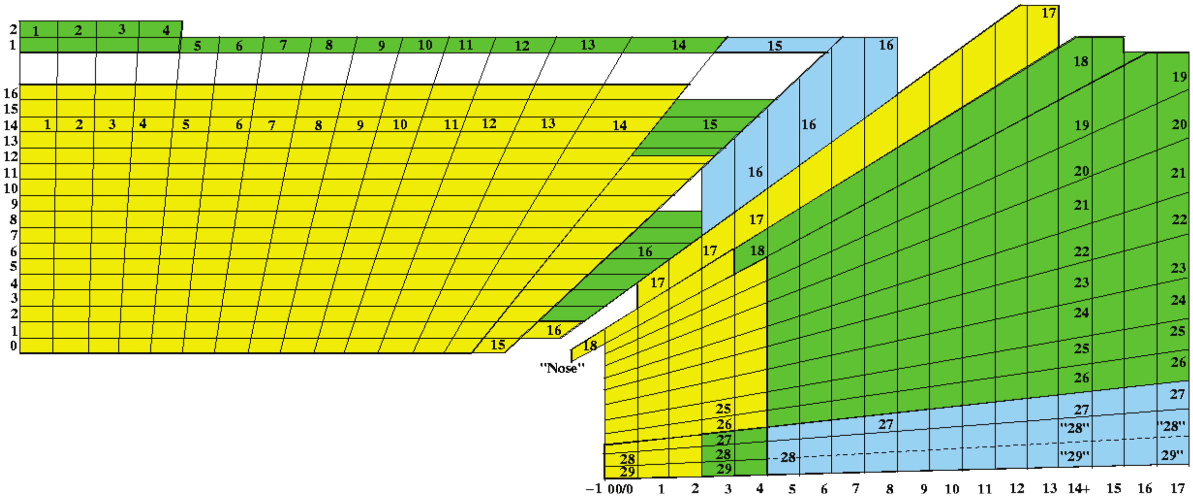


Figure 3.12: A schematic of the HB and HE depth segmentation.

systems, the HF does not have a piece of the ECAL in front of it. Each HF calorimeter consist of 432 readout towers, containing almost 1000 km of 800 μm diameter long and short quartz fibers running parallel to the beam with a granularity of 0.175 × 0.175 in $\eta - \varphi$. The long fibers run the entire depth of the HF calorimeter (165 cm, or approximately 10 interaction length), while the short fibers start at a depth of 22 cm from the front of the detector. By reading out the two sets of fibers separately, it is possible to distinguish EM showers generated by electrons and photons, which deposit a large fraction of their energy in the long-fiber calorimeter segment, from those generated by hadrons, which produce on average nearly equal signals in both calorimeter segments. The fibers make use of Cherenkov light read out by photomultiplier tubes (PMTs), which receive approximately 1 photoelectron for every 4 GeV of deposited energy.

3.2.5 Solenoid

One of the namesake features of the CMS apparatus is a superconducting solenoid of 6 m internal diameter, providing a magnetic field of 3.8 T. The solenoid thus surrounds both the barrel and endcap parts of the silicon pixel and strip tracker, the ECAL, and the HCAL. The high magnetic field allows CMS to have a relatively small size while also having sufficiently high bending of the high energy charged particles to measure their momenta in the tracker.

The magnet itself is made up of a 4-layer winding of reinforced NbTi superconductor cooled to 4.5 K. Like the rest of CMS, this system needed to be modular and is constructed of 5 rings of equal length. The cold mass of the magnet is 220 tonnes and it stores 2.35 GJ when the current is fully on. Fig. 3.13 shows an artist's rendering of the solenoid.

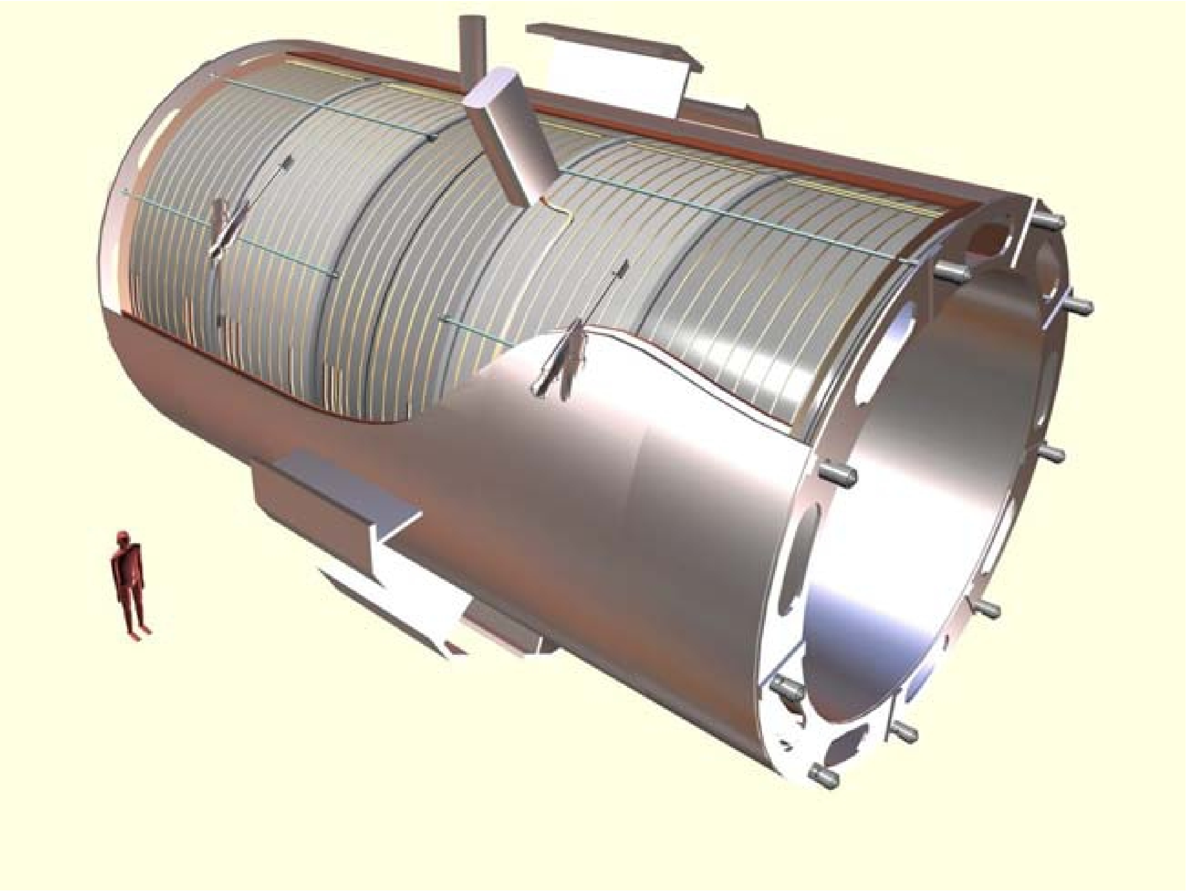


Figure 3.13: An artists rendering of the CMS solenoid. The five superconducting rings can be seen inside the cryostat and support structure. A human figure is shown for comparison.

3.2.6 Muon System

Muons are measured in gas-ionization detectors embedded in the steel flux-return yoke outside the solenoid in the pseudorapidity range $|\eta| < 2.4$, with detection planes made using three technologies: drift tubes (DTs), cathode strip chambers (CSCs), and resistive plate chambers (RPCs).

The barrel region of the detector contains DTs and RPCs, while the endcap region contains CSCs and RPCs. The layout of the muon system can be seen in fig. 3.14. The iron yoke not only returns the flux from the solenoid, but also shields the muon chambers from stray hadrons. The entire muon detection system has nearly 1 million electronic channels and weights in excess of 10000 tons. The muon system on its own has a resolution of 15–40% depending on $|\eta|$. Matching muons to tracks measured in the silicon tracker results in a relative transverse momentum resolution for muons with $20 < p_{\rm T} < 100$ GeV of 1.3–2.0% in the barrel and better than 6% in the endcaps. The p_T resolution in the barrel is better than 10% for muons with $p_{\rm T}$ up to 1 TeV [103].

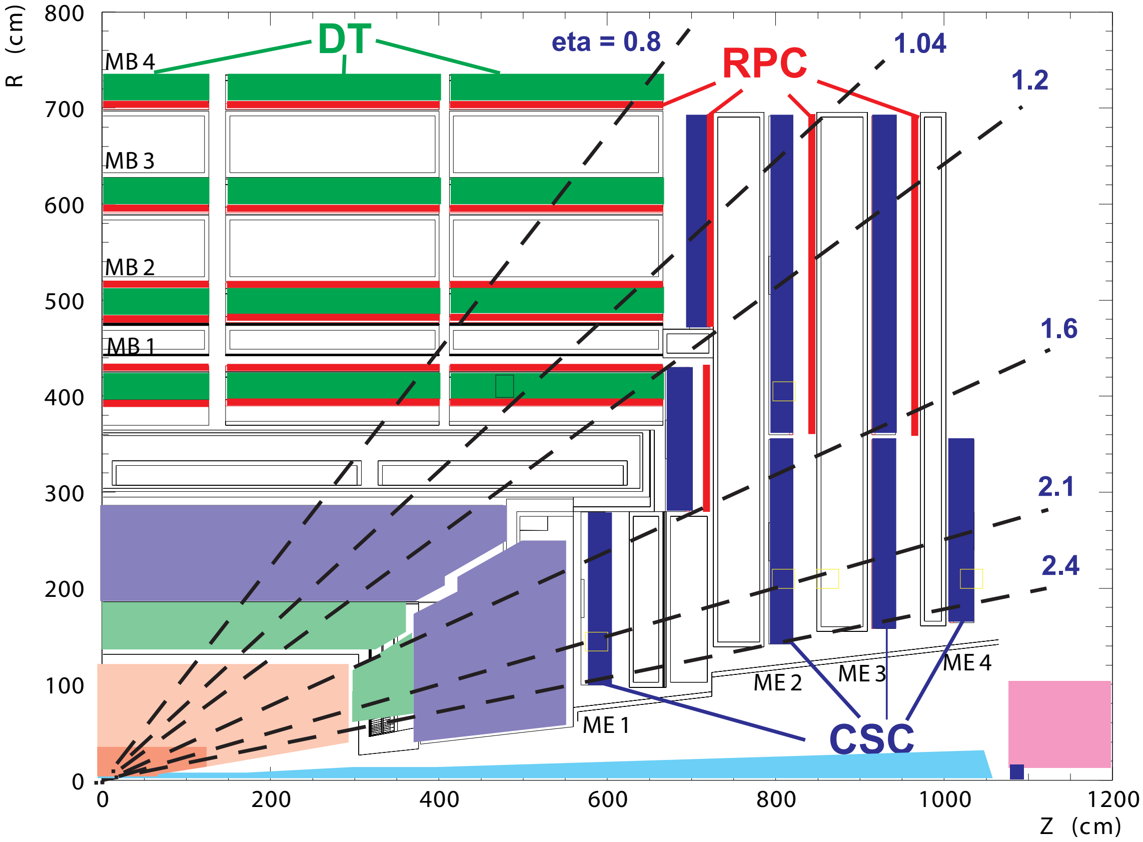


Figure 3.14: Layout of the muon system with the three different detector technologies labeled.

The DTs are divided into four stations named MB1 through MB4 (Muon Barrel), starting radially from the center of the detector outward. The first three stations contain 12 chambers divided into three groups of four. Two of the groups measure the $r - \varphi$ coordinates of the muon while the third group measures the z coordinate. However, MB4 does not have a group of chambers which measured the z coordinate. The four stations contain 250 DTs in total with a collective 172000 sensitive wires, covering an η range of $|\eta| < 1.2$. The chambers themselves contain a gas mixture of 85% Ar and 15% CO_2 and have gold-plated, stainless steel anode wires with a diameter of 50 μm. Within $|\eta| < 0.8$, the MB stations can reconstruct a high-p_{T} muon track with an efficiency greater than 95%. The global $r - \varphi$ resolution is 100 μm. Fig. 3.15 gives a transverse view of the DTs in one of the five wheels of CMS.

The CSCs are separated into four stations as well, names ME1 through ME4 (Muon Endcap), and cover $0.9 < |\eta| < 2.4$. ME1 has three groups of 72 CSC, ME2 and ME3 each have one group of 36 CSCs and one group of 72 CSCs, and ME4 has one group of 36 CSCs. Thus each endcap contains 468 CSCs total. Within a CSC the cathode strips are arranged radially in order to measure the $r - \varphi$ coordinate of the muon. The anode wires are then arranged perpendicular to the strips in order to measure the η coordinate. The cathode strips themselves are made of a fiberglass and epoxy material called FR4, which is coated with 36 μm of copper. The anode wires are gold-plated tungsten with a diameter of 50 μm (the first group of ME1 uses 30 μm wires). There are approximately 220000 cathode strip readout channels and 180000 anode wire readout channels in total. Each CSC contains a gas mixture of 40% Ar, 50% CO_2, and 10% CF_4.

The RPCs are meant to aid in triggering on muons. They cover out to $|\eta| < 1.6$ and can provide information to the trigger system much faster than the DTs or CSCs. The time resolution for the RPCs is less than 3 ns, whereas the DTs and CSCs have a maximum drift time of 400 ns and 60 ns, respectively. With such a small time resolution, the RPCs can precisely identify the bunch crossing time of a muon candidate. MB1 and MB2 have one internal and one external group of RPCs, relative to the DTs. MB3 and MB4 each have two internal groups of RPCs. This amounts to 480 RPCs for the barrel. The endcap has 3 stations of RPCs, 144 chambers in total, arranged in concentric circle on the iron return yoke. The RPCs are a type of parallel plate detector with a gas mixture of 96.2% $C_2H_2F_4$, 3.5% C_4H_{10}, and 0.3% SF_6.

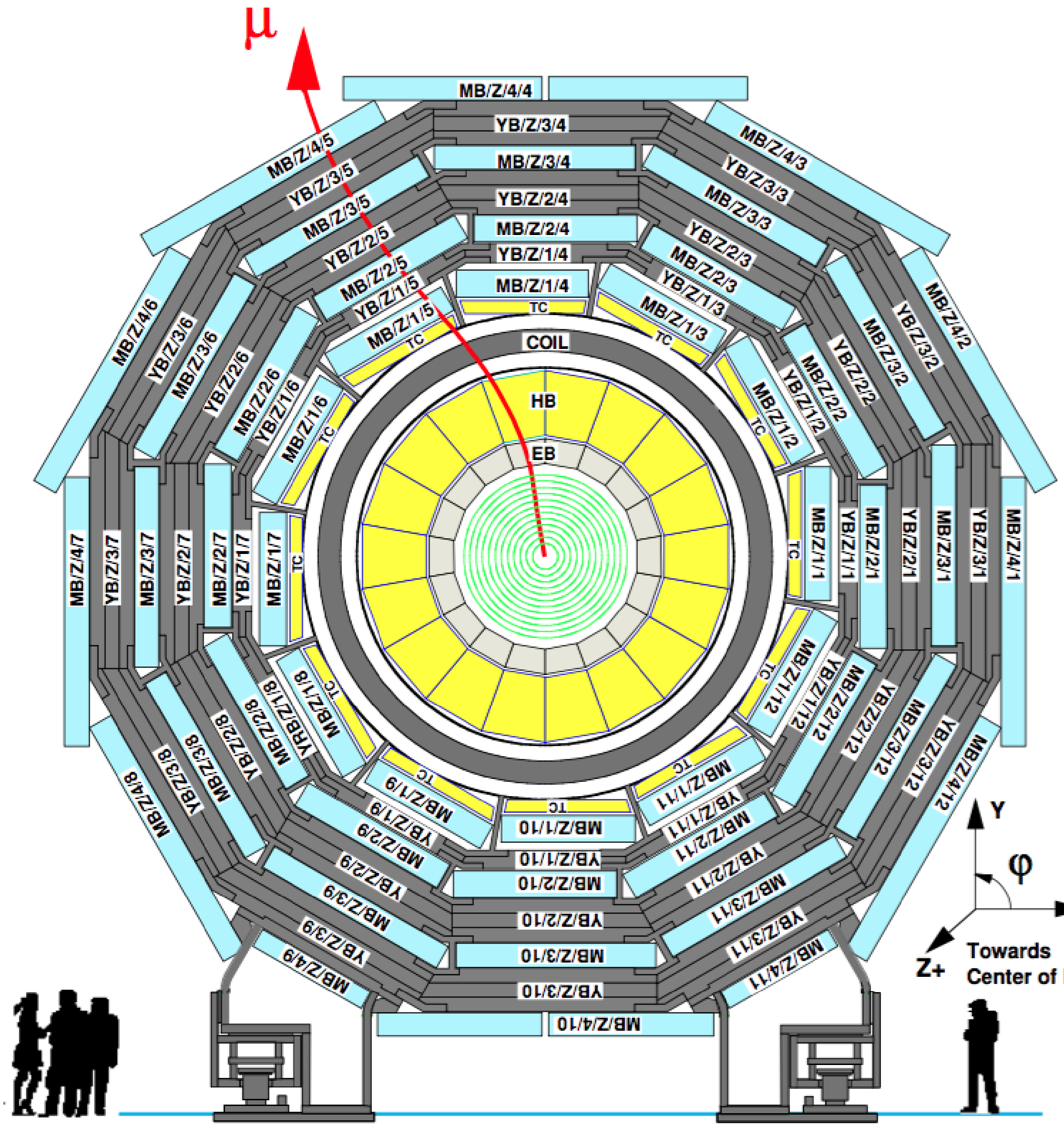


Figure 3.15: Transverse view of one of the five wheels of the CMS detector. The DTs and their layout can be clearly seen. Reprinted from [12].

3.2.7 Trigger

In order to provide as many collisions as possible to the experiments, the LHC must operate at a high luminosity (see sec. 3.1). At the proposed LHC center-of-mass energies the p-p collision cross section is about 100 mb. This, combined with the luminosity, gives us a collision rate of approximately 1 MHz. At this rate it would be impossible for the experiment to store and process all of the raw information coming from the detector. A trigger system is implemented to reduce this rate and keep only the most interesting, and hopefully relevant, events. CMS has implemented a two-tiered trigger system [104]. The first level (L1), composed of custom hardware processors, uses information from the calorimeters and muon detectors to select events at a rate of around 100 kHz within a time interval of less than 4 μs. The second level, known as the high-level trigger (HLT), consists of a farm of processors running a version of the full event reconstruction software optimized for fast processing, and reduces the event rate to less than 1 kHz before data storage.

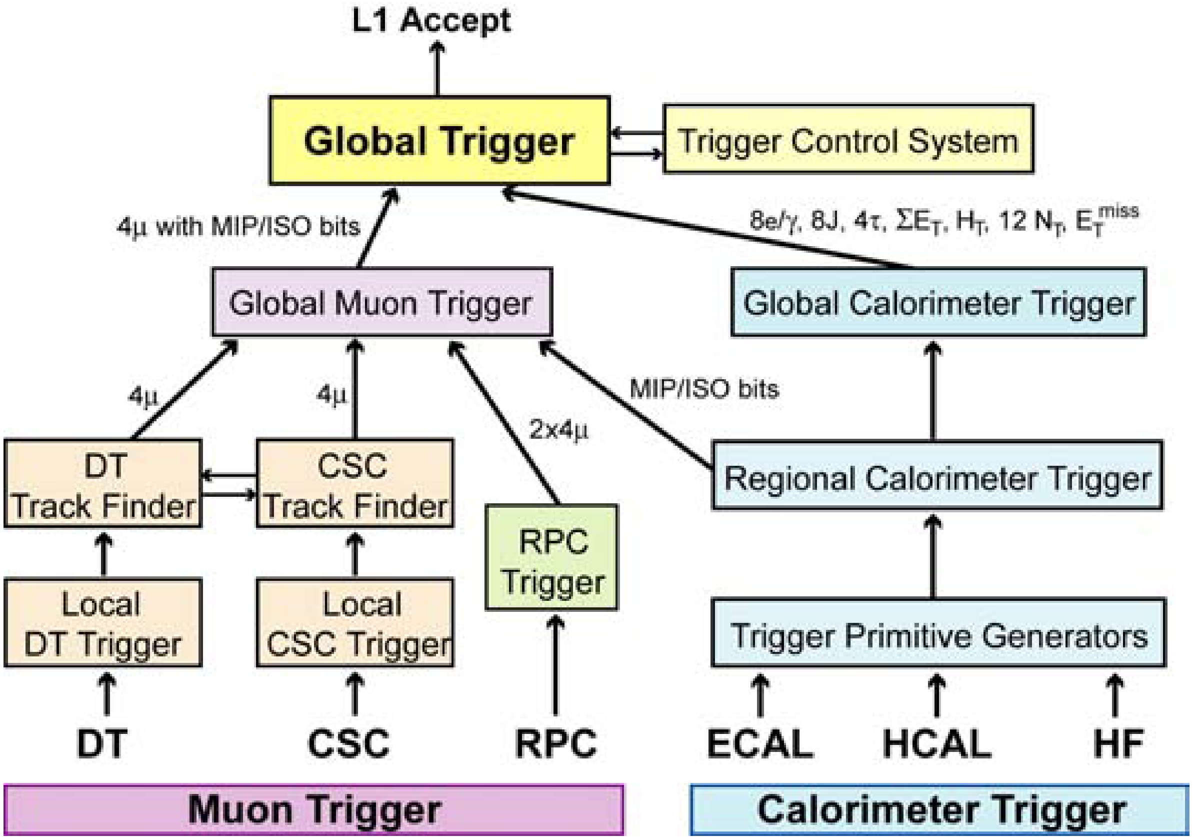


Figure 3.16: The architecture of the L1 trigger system.

The L1 trigger is is composed of custom built, programmable electronics including field programmable gate arrays (FPGAs), memory lookup tables (LUTs) and application specific integrated circuits (ASICs). The components of this trigger system are arranged so that there can be local, regional, and global decision making (see fig. 3.16). Most of the sub-detectors send information to this trigger system, but due to the algorithmic complexity of track finding, the process would take too long if the tracker was included in the decision making process. A new "track trigger" system is in the process of being developed, which would allow tracking information to be included in the L1 trigger decision making process.

The calorimeter side of the L1 trigger system starts with the Trigger Primitive Generators (TPG), which are constructed from energy deposits in the ECAL, HCAL, and HF. These are then combined in the Regional Calorimeter Trigger (RCT), which groups the calorimeter towers into regions. A region is defined as four towers for the barrel and endcap and one tower for the HF. The regions are used to find photon and electron candidates, measure transverse energy sums (ΣE_T), and determine tau-jet vetoes. The RCT also sends information to the Global Muon Trigger (GMT) about energy deposits to help determine if a muon candidate is isolated. The information is then sent to the Global Calorimeter Trigger (GCT), which determines the jet candidates, providing up to four jets and four tau-jets from the central HCAL and four jets from the HF. The GCT also calculates the E_T, $\not\!\!E_\mathrm{T}$, and H_T, which is calculated as ΣE_T for all jets above a certain threshold.

Each of the muon sub-detector's technologies (DT, CSC, and RPC) has a local trigger system. The Regional Muon Trigger (RMT) takes the local trigger information from the DT and CSC and makes tracks using the DT and CSC Track Finders (DTTF and CSCTF). In contrast, the RPCs are a form of dedicated trigger due to their small time resolution. The Global Muon Trigger (GMT) combines the information from the RMT and RPCs to produce up to four muon candidates in each of the barrel and endcap regions. The GMT also contains information about the p_T, charge, η, φ, quality, MIP, and isolation of each of the muon candidates.

Finally, the Global Trigger (GT) combines the GCT and GMT information to decide whether or not to store the event; a decision which is called a Level-1 Accept (L1A). The GT also makes use of information about the sub-detector readouts and DAQ systems from the Trigger Control System (TCS). The L1A is returned to the sub-detectors by the Timing, Trigger, and Control (TTC) system. This entire process takes $3.2\,\mu$s, an equivalent of $\mathcal{O}(100)$ bunch crossings, which means that the data must be pipelined in order to synchronize the steps in the trigger system. Meanwhile, the high resolution data used for offline analysis is stored in memory. In 2012, the L1 Trigger rate was as high as 100 kHz with a dead time of only 3% [105].

After the L1A decision, the High Level Trigger (HLT), a farm of more than 13000 central processing units (CPUs), further analyzes the events. The HLT system uses a form of the full offline reconstruction algorithms described in section 4, but also includes several optimizations to make the process faster. This is needed because, in contrast to offline processing, the HLT is limited by the number of events that can be stored in the pipeline. These optimizations include making the fasted algorithm run first, skipping a trigger path after the first failing quality filter, and considering smaller regions of the detector based on the L1 candidates. The menu of triggers to be run changes as the LHC and Monte Carlo (MC) simulation conditions change, even while CMS is operational. In 2012, the HLT had an output rate of 100 kHz and took 200 ms per event, $\mathcal{O}(100)$ times faster than the offline reconstruction [106]. Events that pass the HLT are then sorted into primary datasets (PDs) according to the passed triggers with as little overlap as possible.

3.2.8 Luminosity Measurement

Besides measuring the kinematics of each of the particles traversing the detector, CMS must also measure the instantaneous luminosity delivered by the LHC. Both the pixel detector (section 3.2.2) and the HF (section 3.2.4) are able to measure the luminosity to varying degrees of accuracy.

The pixel detector has a very small granularity, which means that any given pixel is activated by at most one track per bunch crossing. We can then create cluster by grouping nearby activated pixels, with the typical cluster containing an average of 5 pixels. A minimum bias event typically creates 200 clusters [107]. Even for events with 100 pileup (PU) interactions, a number significantly higher than was reached in 2012, the total pixel detector occupancy could be as low as 0.1%. This means that the number of pixel hits should scale linearly with the number of interactions per bunch crossing, which is shown in equation 3.4 [108].

$$\mathcal{L} = \frac{\nu \langle n \rangle}{\sigma_{vis}} \tag{3.4}$$

Here the luminosity, $\mathcal{L}$, is proportional to average number of pixel clusters, $\langle n \rangle$. The other parameters are the LHC revolution frequency, $\nu = 11246\,\text{Hz}$ and the visible cross section, σ_{vis}, as calibrated by a Van der Meer scan [109]. In 2012 this technique was used to measure the total integrated luminosity with a systematic uncertainty of 2.6%.

Another method to measure the luminosity makes use of the HF, but due to some sever limitations in its accuracy, this measurement is only used as a cross-check for the pixel counting method. What makes the HF suitable for this type of measurement is that it can safely be run during unstable beams [108]. The average transverse energy per tower can be directly related to the luminosity or the average fraction of empty towers can be related to the mean number of interactions per crossing, which is more of in indirect measurement. The benefit of using the HF is that it can make an online determination of the luminosity within 1 s to an accuracy of 1%. One downside is that even in 2012 the levels of pileup made the luminosity relationship non-linear. Additionally, the calibration of this measurement can change due to drifts in the gains of the HF PMTs [110].

4. EVENT RECONSTRUCTION

The CMS detector is designed to identify the various particle species which travel through it after a proton-proton collision. As discussed in section 3, the sub-detector technologies were chosen so that particles could be identified by where they deposit their energy as well as how their trajectories change in a magnetic field. Fig. 4.1 shows how various types of particles interact within the CMS sub-detectors. All of the charged particles (i.e. electrons, muons, and charged hadrons) will deposit some energy in the tracker, while neutral particles (i.e. photons and neutral hadrons) will not. Electrons and photons will deposit all of their energy inside of the ECAL while hadrons, both charged and neutral, will deposit most of their energy in the HCAL. Muons are the only visible particle which will be able to travel to the muon chambers. Neutrinos will pass through all layers of the detector unseen and their presence must be inferred by missing transverse energy ($E_{\mathrm{T}}^{\mathrm{miss}}$ or $\not\!\!E_{\mathrm{T}}$); the idea being that if the sum of the transverse momentum is not conserved, then that missing momentum must correspond to at least one unseen particle.

The process of translating abstract detector objects to physical particles takes several steps within the CMS software framework (CMSSW). The first of this process is local reconstruction, where the various subsystems of each sub-detector create what are called reconstructed hits, or RecHits for short. RecHits in the tracker contain information about the position of energy clusters (groups of contiguous strips or pixels which contain a signal) as well as energy deposition information which aids in particle identification. The muon RecHits ostensibly contain information about the position of the signal. However, the RecHits from the DTs and CSCs can be combined to form three-dimensional track segments, which also provide directional information. The ECAL and HCAL RecHits contain information about the energy deposited, the position of those deposits, and the time at which they occurred.

The next step is to process this information in a global manner, where the subsystems within each sub-detector are combined. Pattern recognition algorithms are run on the tracker RecHits to reconstruct the path that the particles take through the sub-detector (a.k.a tracks). The ECAL

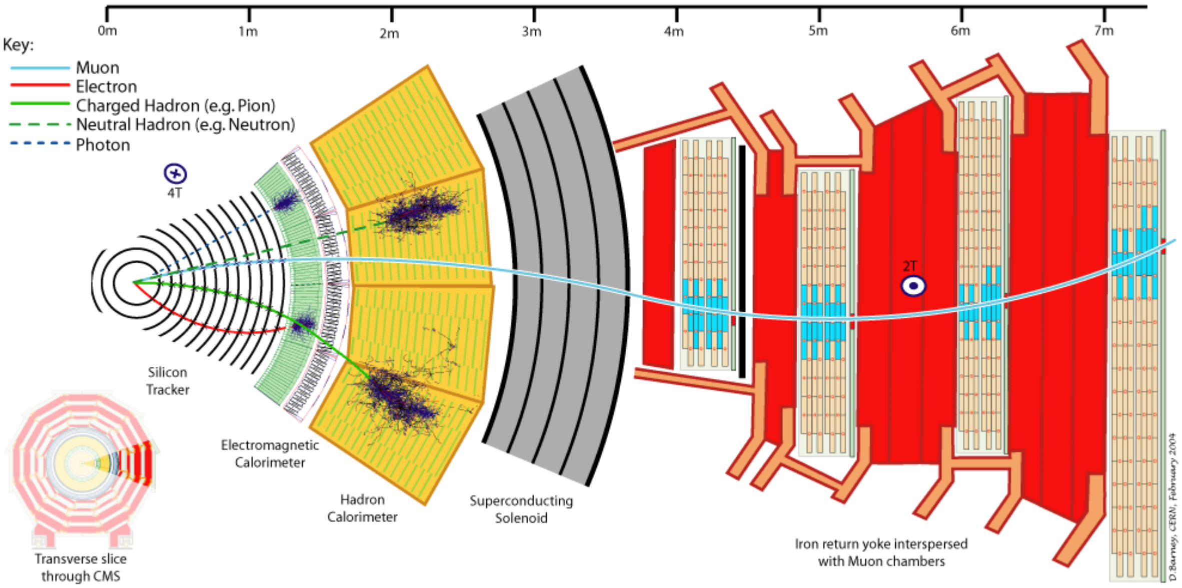

0m
1m
2m
3m
4m
5m
6m
7m
Key:
Muon
Electron
Charged Hadron (e.g. Pion)
Neutral Hadron (e.g. Neutron)
Photon
4T
Silicon Tracker
Electromagnetic Calorimeter
Hadron Calorimeter
Superconducting Solenoid
2T
Iron return yoke interspersed with Muon chambers
Transverse slice through CMS
D.Barney, CERN, February 2004

Figure 4.1: Cross-sectional view of the CMS detector with all of the sub-detectors labeled.The colored lines correspond to different particle species, which interact with different pieces of the detector and may or may not be bent by the magnetic field. Reprinted from [13].

and HCAL RecHits within a tower are summed to form "CaloTowers" which have a projective $\eta - \varphi$ geometry. The muon system creates "standalone" muons by associating RecHits and track segments with compatible radial trajectories. This process takes into account the bending a muon undergoes before reaching and within the muon system due to the magnetic field.

At this point, all of the reconstruction information is combined to form particles that can be used for physics analysis. The process of reconstructing and classifying every stable particle is called Particle Flow (PF) and will be discussed further in section 4.2. This analysis focuses on electrons, muons, jets, b-jets, and $\not\!\!E_{\mathrm{T}}$, the reconstruction of which will be described in the following sections. Additional information about the reconstruction process beyond the scope of this thesis can be found in [111].

4.1 Tracks and Vertices

While CMS analyses cover a wide range of final states, a majority of them will include jets in some fashion, including this one. It's important that the particle flow algorithm identify and measure each particle inside a jet in order to improve the jet energy response and resolution. Section 4.5 will cover the reconstruction and properties of jets in more detail, but it is important to note that two thirds of the constituents inside of a jet are charged particles. This motivates the need for excellent tracking capabilities. Tracks are created from the RecHits using the Combinatorial Track Finder (CTF) algorithm, which is an iterative process [99]. This process seeks to find the appropriate balance between high reconstruction efficiency and low fake rate (see fig. 4.2) [14].

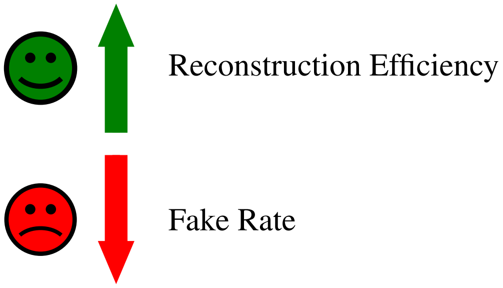


Figure 4.2: A diagram showing the goals of the iterative tracking process.

The track finding procedure begins by finding track seeds using only a few hits and very tight criteria. A track is built by extrapolating from the trajectory of the seed and adding new hits that match this trajectory, keeping in mind that charged particles will bend in the presence of the magnetic field. The tight requirements on this first step lead to a moderate tracking efficiency and a vanishingly small fake rate. After a track is found, all of the hits are used in a fit to determine the track parameters (i.e. p_{T}, χ^2, etc.), which are then used to judge the quality of the track. If a track doesn't meet certain quality requirements on the p_{T}, the transverse impact parameter d_0, and the longitudinal impact parameter d_z, it isn't kept. Additionally, a trajectory cleaning step to remove duplicate tracks is applied to each iteration and to the final track collection. A duplicate track can form either from different seeds or from the same seed which forms two very similar tracks. If a pair of any two tracks share more than 19% of hits as determined by equation 4.1, where N_1^{hits} and N_2^{hits} are the number of hits used in forming the tracks and 19% is an empirically determined value, then the track with the fewest number of hits or the largest χ^2 is removed. The hits which are unambiguously assigned to the tracks are removed from consideration in the next iteration and their tracks saved for later use.

$$f_{shared} = \frac{M^{hits}_{shared}}{min\left(N^{hits}_1, N^{hits}_2\right)} \tag{4.1}$$

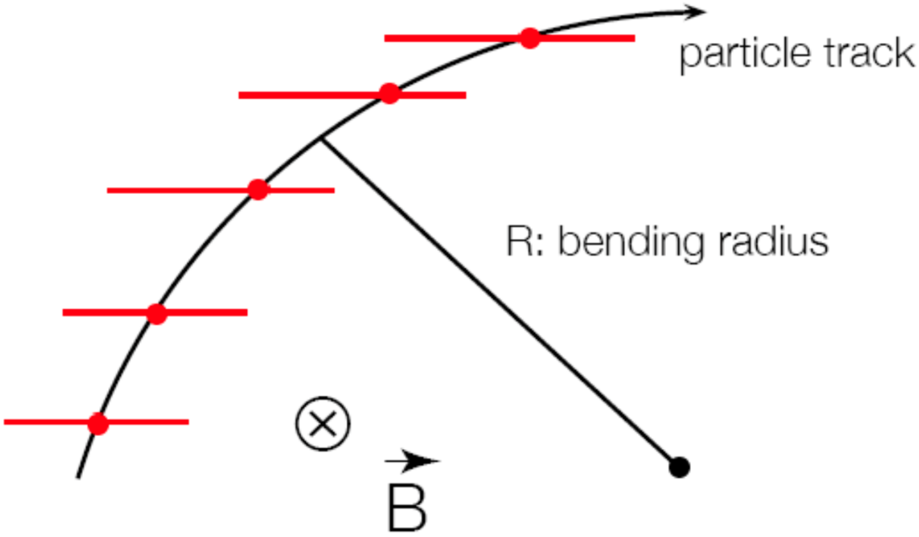


Figure 4.3: Schematic view of a particle track with hits labeled.

In each subsequent iteration the track seeding criteria is loosened and the same procedure occurs. The looser seeding requirements boosts the tracking efficiency, while the removal of the hits from the previous iteration keeps the fake rate low due to the reduced combinatorics. The specific seeding criteria for each iteration can also be found in table 4.1. After three iterations, 90% of charged hadron tracks within jets are reconstructed and 99.5% of muons in the tracker acceptance are found. Subsequent iterations loosen the constraints on the origin vertex, which allows for the reconstructions of tracks associated with a secondary vertex (i.e. $\gamma \rightarrow e^+e^-$ conversions, long-lived particles, nuclear interactions in the tracker material). Tracks meeting this set of criteria can be reconstructed with as little as three hits, a p_{T} as low as 150 MeV, and a vertex more than 50 cm away from the beam axis. Nevertheless, the fake rate is still kept on the order of 1% [14].

step	seed type	seed sub-detectors	p_T [GeV/c]	d_0 [cm]	$\lvert z_0 \rvert$
0	triplet	pixel	>0.6	<0.02	$<4.0\sigma$
1	triplet	pixel	>0.2	<0.02	$<4.0\sigma$
2	pair	pixel	>0.6	<0.015	<0.09 cm
3	triplet	pixel	>0.3	<1.5	$<2.5\sigma$
4	triplet	pixel/TIB/TID/TEC	>0.5–0.6	<1.5	<10.0 cm
5	pair	TIB/TID/TEC	>0.6	<2.0	<10.0 cm
6	pair	TOB/TEC	>0.6	<2.0	<30.0 cm

Table 4.1: The seed criteria used in each iteration during the 2012 run. The seed types, pair and triplet, indicate if two or three RecHits are used, respectively. The σ in the z_0 criteria indicated the length of the beam spot in the z-direction as determined by a Gaussian fit [29].

The tracks found will have a helical shape of with a given radius of curvature as in fig. 4.3. The softest, low p_T particle trajectories can form small rings, while the higher p_T particles will be bent less. The momentum of each track can be extracted from the radius of curvature (R), given by a circular fit to the track, the magnetic field strength, as well as η and φ of the track at the interaction point[1]. The following system of equations can be used to determine the particles 3-momenta at the interaction point:

$$\begin{aligned} p_x &= p_T \cos\varphi \\ p_y &= p_T \sin\varphi \\ p_z &= p_T \sinh\eta \\ p_T &= 0.3 \cdot B \cdot R \end{aligned} \tag{4.2}$$

After the collection of high purity tracks is created, CMS uses these to reconstruct the location of the vertices where proton-proton interactions occurred [99]. The vertex finding algorithm is agnostic to whether or not the vertices come from the main hard scatter vertex of interest or any of the pileup vertices from additional proton-proton interactions. However, there is a need to select prompt tracks occurring near the interaction point instead of tracks from secondary vertices. CMS

[1] The η of the track is determined as if the interaction point was at the center of the detector.

requires that the significance of the transverse impact parameter $d_0 < 5$, the number of pixel hits be ≥ 2, the number of pixel and strip hits be ≥ 5, and the track $\chi^2 < 20$. Once there is a collection of prompt tracks they are clustered together in z at their closest approach to the beam spot. A balance must be struck between vertex finding efficiency and the splitting of good vertices. To do this, a deterministic annealing (DA) algorithm is employed and is useful in cases where one wants to find the approximate global minimum of a problem with many degrees of freedom; specifically where an approximate global minimum is preferred over a more accurate local minimum. More information about DA can be found in [112], but simply put the process is similar to what happens when one heats a system and then slowly cools it to minimize the "free energy," which in this analogy is the χ^2 of the vertices. In this case there is a system of z_i^T with uncertainty σ_i^z and an unknown number of vertices z_k^V. There is a probability $0 \leq p_{ik} \leq 1$ for any track i to be assigned to vertex k and in the beginning, the algorithm assumes that every possible assignment is equally likely. The free energy to be minimized can be found in equation 4.3, where p_i is a constant weight for each tracks representing their consistency with originating from the beam spot and z_k^V are the vertices with weights ρ_k.

$$F = -T \sum_{i}^{\#tracks} p_i \log \sum_{k}^{\#vertices} \rho_k \exp\left[-\frac{1}{T}\frac{\left(z_i^T - z_k^V\right)^2}{\sigma_i^{z2}}\right] \tag{4.3}$$

The number of vertices can be arbitrarily large, but any extra vertices used in the method will overlap with the effective vertices already found at distinct positions. The probability that a given track corresponds to a specific vertex is given by equation 4.4.

$$p_{ik} = \frac{\rho_k \exp\left[-\frac{1}{T}\frac{\left(z_i^T - z_k^V\right)^2}{\sigma_i^{z2}}\right]}{\sum_{k'} \rho_{k'} \exp\left[-\frac{1}{T}\frac{\left(z_i^T - z_{k'}^V\right)^2}{\sigma_i^{z2}}\right]} \tag{4.4}$$

At high temperature all tracks belong to a single vertex and all p_{ik} are equal. As $T \to 0$ each track becomes compatible with exactly one vertex. The number of vertices grows each time the temperature falls below the critical temperature of a given vertex, T_c^k, given by equation 4.5, where that vertex is replaced by two nearby vertices. As this happens the tracks are reassigned according to their probabilities before the temperature is lowered again. The starting temperature of the whole process is chosen to be above the first critical temperature where $\rho_1 = p_{i1} = 1$. The temperature is lowered by a cooling factor of 0.6 down to $T_{min} = 4$, which balances the need to resolve all true vertices with the risk of splitting a true vertex.

$$T_c^k = 2\sum_i \frac{p_i p_{ik}}{\sigma_i^{z2}} \left(\frac{z_i^T - z_k^V}{\sigma_i^z} \right) / \sum_i \frac{p_i p_{ik}}{\sigma_i^{z2}} \tag{4.5}$$

By the time the T_{min} condition is reached it is still possible for a track to be assigned to multiple vertices. Thus, for the final track assignment, the temperature is cooled to $T = 1$, without more splitting of the vertices. For a track to be assigned to a given vertex it must have a minimum probability of 0.5 and have passed the outlier mitigation criteria.

After all of the candidate vertices are found using the DA method, the candidates with at least two tracks assigned to them are passed through the adaptive vertex fitter (AVF) to compute all of the vertex parameters. Key among those parameters are the spacial coordinates and the number of degrees of freedom given by equation 4.6, where w_i is a weight, between 0 and 1, given to each track depending on the likelihood that the track actually belongs to that vertex. Additional quality requirements for a good track are $n_{dof} > 4$ (at least four associated tracks), $|z| < 24\,\text{mm}$, and $|\rho| < 2\,\text{mm}$, where ρ is the transverse position of the vertex [113]. If the track $\chi^2/N_{dof} < 20$, then the track is matched to that vertex and only that vertex [16].

$$n_{dof} = -3 + 2\sum_{i=1}^{\#tracks} w_i \tag{4.6}$$

As mentioned before, a single vertex is classified as the “primary” vertex, with all other proton-proton collisions being classified as secondary, pileup vertices. The leading vertex is the one with the greatest sum of the squares of the associated tracks’ transverse momenta ($\sum |p_{\mathrm{T}}^{\mathrm{track}}|^2$).

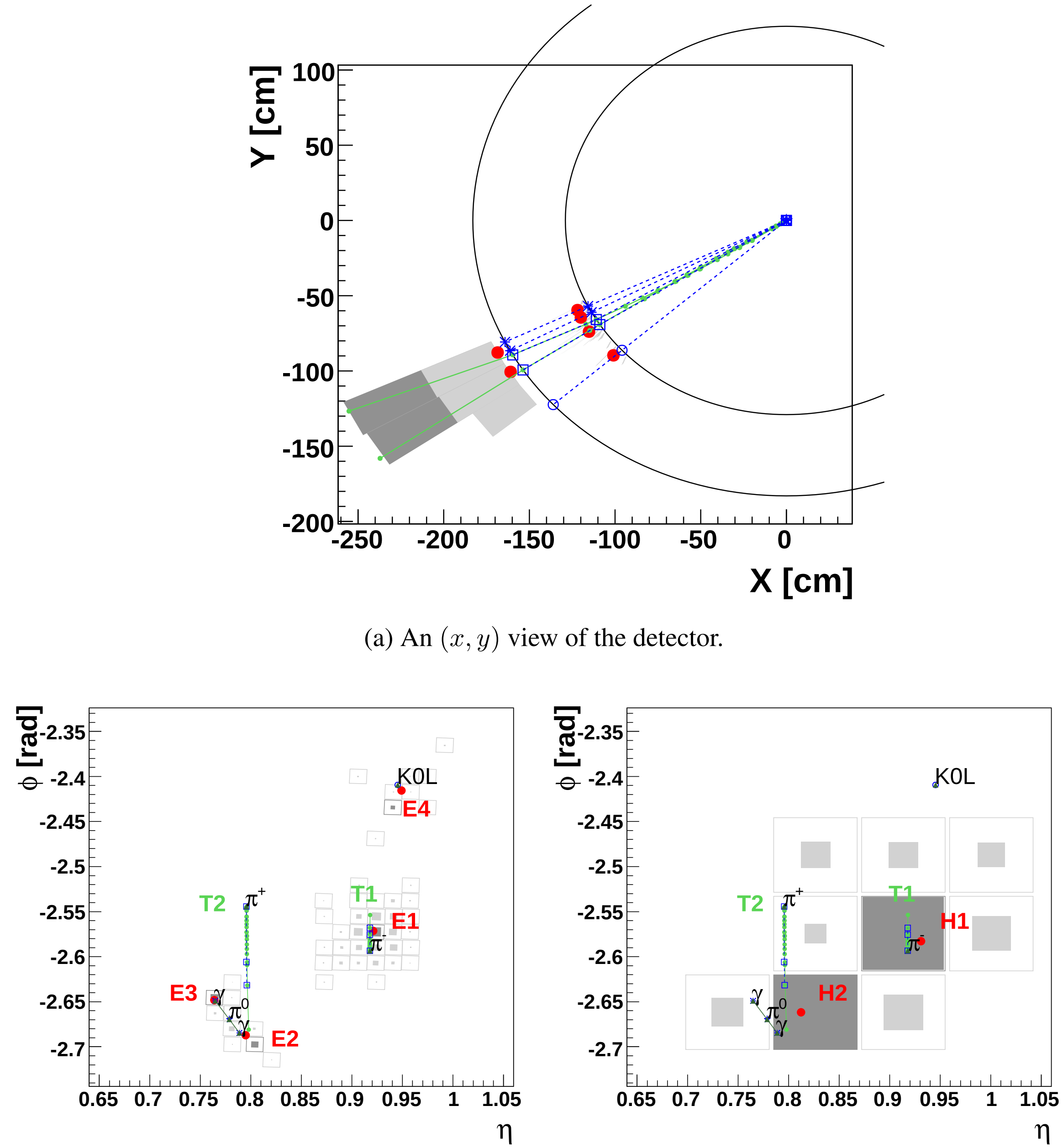
The methodology above is a simplification of the actual track and vertex finding algorithms, but is sufficiently detailed for the purposes of this document. The subsequent sections will discuss how the RecHits, tracks, and vertices are used to reconstruct particles.

4.2 Particle Flow

The CMS experiment has decided to use a holistic approach to reconstructing the event produced by a proton-proton collision. The particle flow (PF) event reconstruction algorithm uses information from all of the sub-detectors in order to identify as accurately as possible each individual particle in the event as described in the first part of this section [14, 114] and to reconstruct their direction and energy. Other quantities that can be determined from a particle level reconstruction algorithm are the charged lepton isolation and the likelihood that a jet was initiated by a B hadron. The CMS detector is ideally suited for a particle flow approach because of its extremely granular sub-detectors and high magnetic field. This approach has been validated in [115, 116, 117, 118], where an improvement over simpler techniques was shown each time. The output of the PF algorithm is a list of particles known as “PF candidates,” which are used to build the higher level objects that physicists analyze, such as jets, taus, and $E_{\mathrm{T}}^{\mathrm{miss}}$. Because the CMS detector is so granular, the occupancy and event complexity play almost no role in the PF algorithm efficiency. With the current algorithm, charged-particle tracks out to $|\eta| <$ 2.6 can be reconstructed even with a p_{T} as low as 150 MeV, all while maintaining a high reconstruction efficiency and low fake rate. The algorithm can even identify the difference between photons and charged-particles in high multiplicity environments like jets. This is largely due to the tracking information, which more accurately determines the p_{T} than the calorimeter system for for charged particles up to several hundred GeV. Additionally, the tracker can measure the direction of a charged particle before its trajectory can be changed in the magnetic field. Fig. 4.1 shows, in graphical terms, how the reconstruction algorithm can classify a particle based on the sub-detectors with which it interacts.

The inputs to the PF algorithm come from the local reconstruction products, RecHits, as described at the start of section 4. More specifically the RecHits are turned into either tracks or energy clusters, which are then used by the algorithm. The tracks may come from the tracker, as described in section 4.1, or from the muon system. The clusters are created by the calorimeter RecHits and are treated slightly different than the CaloTowers previously discussed. A local energy maxima above a threshold value, also known as a "cluster seed," is chosen as the beginning of a calorimeter cluster. From there "topological clusters" are grown by adding neighboring hits above a two standard deviation threshold energy set by the subsystem to remove photo-detector noise in the ECAL (i.e. 80 MeV in the barrel and up to 300 MeV in the endcaps) or HCAL (i.e. 800 MeV). Some clusters are removed if its characteristics match those of an expected noise source, but otherwise a topological cluster will will create as many "particle-flow clusters" as there are seeds. Energy is shared among the cells in the cluster according to the cell-cluster distance.

Once all of the tracks and clusters have been found, the two collections are associated using a "linking algorithm" to create "blocks." First, the track is extrapolated from its last hit in the tracker subsystem to the two layers of the PS, the ECAL at a depth corresponding to the expected maximum of a typical electron shower, and to the HCAL at a depth of one interaction length. A link is made if this extrapolated position is within the cluster boundaries, which can be enlarged by one cell size in each direction to account for non-instrumented areas, multiple scattering of low-momentum charged particles, and the uncertainty in the position of the shower maximum. This linking algorithm is based on minimizing the $\eta - \varphi$ distance ($\Delta R = \sqrt{\Delta\eta - \Delta\varphi}$) between the track and cluster. As an additional complication, tangents are drawn from the intersection between the track and the tracker layers to the ECAL. If one of these tangents falls within an ECAL cluster then the cluster is marked as a potential Bremsstrahlung photon. Linking between calorimeter systems (i.e. HCAL and ECAL or ECAL and PS) is done similarly, but the cluster position in the more granular system must be within the envelope of the less granular system. A track and muon track are linked when an acceptable χ^2 is returned by a global fit between two tracks. If there are multiple track matches for a single muon track, then the match with the minimum χ^2 is chosen to form a "global muon." Fig. 4.4 shows a graphical representation of what the linking algorithm sees and how the links between tracks and clusters are made.



(a) An (x, y) view of the detector.

(b) An (η, φ) view of the ECAL.

(c) An (η, φ) view of the HCAL.

Figure 4.4: These three figures show a representation of how the PF algorithm sees a hadronic jet. (a) An (x, y) view of the detector with elements from the tracker, ECAL, and HCAL shown. The ECAL and HCAL surfaces shown in (b) and (c) are represented by the concentric circles centered around the interaction point in (a). (b) shows the energy clusters from the K^0_L, π^-, and the two photons from the π^0 decay. While the π^+ doesn't deposit any energy in the ECAL, it does show up as a cluster in the HCAL along with the π^- (c). The tracks from these charged particles show up as vertical lines in the (η, φ) plane, but as curved lines in the (x, y) plane. The cluster positions are represented by dots, the simulated particles by dashed lines, and the position at which the particles impact the calorimeter surfaces by the open marker. Reprinted from [14].

The blocks are classified as a specific type of particle based on which sub-detectors were linked and then removed from the list of unclassified blocks to prevent double counting. To begin with, if the momentum of the combined charged-particle and muon tracks is equal to the momentum of the charged-particle track alone, then the particle is classified as a PF muon. The minimum ionization energy expected to be deposited by a muon is subtracted from the remaining clusters. The other charged-particle tracks are checked to see if they match the properties of an electron, which is to say that electrons tend to radiate energy via bremsstrahlung, which causes the curvature of the tracks to increase as they move away from the interaction point. A Gaussian Sum Filter (GSF) is used to match these tracks with ECAL clusters and a successful match is classified as a PF electron. More information about the GSF and its improvements over the standard CMS tracks finding algorithms can be found at [119].

Tracks which aren't matched to muons or classified as electrons are matched to clusters, if possible, and form PF charged hadrons. In this case the total cluster energy must be similar to, but smaller than, the total track momentum. Only the closest cluster may be linked to any given track, but a given cluster may have multiple track links due to the large granularity of the calorimeters. The energy of charged hadrons is determined from a combination of the track momentum and the corresponding ECAL and HCAL energy, corrected for zero-suppression effects and for the response function of the calorimeters to hadronic showers. Any excess energy remaining after removing the track energy from the clusters is assumed to come from neutral particles. If this excess energy is in the ECAL then the neutral particle is classified as a PF photon and its energy is directly obtained from the ECAL measurement, corrected for zero-suppression effects. After the removal of the PF photons, the remaining excesses are classified as neutral hadrons and their energy is obtained from the corresponding corrected ECAL and HCAL energy. Clusters which are not matched to any tracks are used to make PF photons in the ECAL and neutral hadrons in the HCAL.

4.3 Electrons

Broadly speaking, the PF electron candidate identification process discussed in section 4.2 can be considered "tracker-driven" [120]. This method is ideal for low-p_{T} electrons and electrons in high multiplicity environments like jets. On the other hand, high p_{T} electrons need an "ECAL-driven" approach. In this case the ECAL clusters are grouped into "superclusters" for the purpose of trying to capture energy from two sources, photons produced due to bremsstrahlung and the spread of energy in φ due to the magnetic field [121]. These superclusters are then matched to track seeds and a GSF is used to reconstruct the track trajectory. The GSF is necessary to account for changes in direction due to bremsstrahlung [119]. After the ECAL-driven list is created it can be compared to the list of PF electron candidates to prevent double counting.

The electron four momentum is estimated by combining the energy measurement in the ECAL, the momentum measurement in the tracker at the main interaction vertex, and the energy sum of all bremsstrahlung photons attached to the track. The momentum resolution for electrons with $p_{\mathrm{T}} \approx 45\,\mathrm{GeV}$ from $\mathrm{Z} \rightarrow ee$ decays ranges from 1.7% for non-showering electrons in the barrel region to 4.5% for showering electrons in the endcaps. The di-electron mass resolution for $\mathrm{Z} \rightarrow ee$ decays when both electrons are in the ECAL barrel is 1.9%, and is 2.9% when both electrons are in the endcaps. [122].

Only electron selection has been discussed so far. However, once there is a complete list of electron candidates, quality cuts are imposed to identify genuine electrons [123, 124, 125]. There are two similar methods for evaluating these quality requirements. One is a purely cut based technique and the other merges these requirements, plus some additional variables, into a single MVA based training. This analysis used the MVA based training in order to extract as much performance from the selection cuts as possible. However, it is still informative to list the cut based requirements since they are all used inside of the MVA training. The η width of the supercluster, $\sigma_{i\eta i\eta}$, is taken from the covariance matrix of a weighted difference between the η positions of the crystals and the seed cluster. A modified η is used in this calculation to account for the crystal spacing and each crystals contribution is weighted by $\log\left(E_{crystal}/E_{sc}\right)$ [126]. Two additional variables are calculated as the differences between the positions of the supercluster, $(\eta_{sc}, \varphi_{sc})$, and the extrapolated track, $\left(\eta_{in}^{extrap}, \varphi_{in}^{extrap}\right)$, thus defined as $|\Delta\eta_{in}| = |\eta_{sc} - \eta_{in}^{extrap}|$ and $|\Delta\varphi_{in}| = |\varphi_{sc} - \varphi_{in}^{extrap}|$. The ratio of the leakage energy, H, in the HCAL tower behind the ECAL seed cluster is compared to the energy of that seed cluster in the variable H/E. The transverse and longitudinal impact parameters compared to the associated vertex, d_0^{vtx} and d_z^{vtx}, and a comparison of the electron energy and momentum, $|1/E - 1/p|$, are used. Both identification schemes also make use of the PF based isolation variable shown in equation 4.7. However, rather than using the base isolation value, the relative isolation $I_e^{PF}/p_{\mathrm{T}}^{\mathrm{e}}$ is used. The isolation variable is simply the sum of the p_{T} of the charged hadron (CH), neutral hadron (NH), and photon (γ) PF candidates within a cone of $\Delta R < 0.3$ around the electron candidate. The expected amount of energy due to pileup is then removed by multiplying the median energy density by the electron effective area, A_{eff}, but it is protected from becoming a negative value.

$$I_e^{PF} = \sum_{\Delta R < 0.3} p_{\mathrm{T}}^{(CH)} + max\left(\sum_{\Delta R < 0.3} p_{\mathrm{T}}^{(NH)} + \sum_{\Delta R < 0.3} p_{\mathrm{T}}^{(\gamma)} - \rho A_{eff}, 0\right) \tag{4.7}$$

A set of values for the identification requirements is called a working point (WP) and there are several WP based upon the desired identification efficiency and fake rate. This analysis makes use of the tight working point for the selected electron and the loose working point to veto on additional electrons. Table 4.2 lists the cut based identification requirements for the tight and loose WP. Similarly, table 4.3 lists the requirements for the MVA based identification. In addition to the identification requirements, selected electrons must have a $p_{\mathrm{T}} > 30\,\mathrm{GeV}$ and be in the barrel, $|\eta_{\mathrm{sc}}| < 1.4442$, or endcap, $1.566 < |\eta_{\mathrm{sc}}| < 2.5$. They must also pass a conversion veto to make sure the aren't produced by a converted photon. Loose electrons have the same η and conversion requirements, but are only required to have a $p_{\mathrm{T}} > 15\,\mathrm{GeV}$. The p_{T} requirements are selected to match the HLT requirements of the PD listed in section 5.1.1.

Cut Variable	Cut Value			
	Tight		Loose	
	Barrel	Endcap	Barrel	Endcap
$I_e^{PF}/p_{\mathrm{T}}^{(e)} <$	0.1	0.1	0.15	0.15
$\sigma_{i\eta i\eta} <$	0.01	0.03	0.01	0.03
$\|\Delta\varphi_{in}\| <$	0.03	0.02	0.8	0.7
$\|\Delta\eta_{in}\| <$	0.004	0.005	0.007	0.01
$H/E <$	0.12	0.1	0.15	0.07
$\|d_0^{vtx}\| <$	0.02	0.02	0.04	0.04
$\|d_z^{vtx}\| <$	0.1	0.1	0.2	0.2
$\|1/E - 1/p\| <$	0.05	0.05	-	-

Table 4.2: Cut based electron identification requirements for the tight and loose working points.

Supercluster Pseudorapidity	Cut Value			
	Tight		Loose	
	MVA	$I_e^{PF}/p_{\mathrm{T}}^{(e)}$	MVA	$I_e^{PF}/p_{\mathrm{T}}^{(e)}$
$\|\eta_{\mathrm{sc}}\| < 0.8$	>0.977	<0.093	>0.877	<0.426
$0.8 < \|\eta_{\mathrm{sc}}\| < 1.479$	>0.956	<0.095	>0.811	<0.481
$1.479 < \|\eta_{\mathrm{sc}}\| < 2.5$	>0.966	<0.171	>0.707	<0.390

Table 4.3: MVA based electron identification requirements for the tight and loose working points. The tight MVA requirements were trained using triggering electrons whereas the loose MVA requirements, usually used as a veto, were trained on non-triggering electrons.

4.4 Muons

In addition to using the PF algorithm to identify muon candidates, CMS uses two supplementary methods to identify high and low-momentum muon candidates [127]. The union of these collections will me used for the final muon reconstruction. To capture the low-momentum muons, charged particle tracks which have a p_{T} and p above a threshold are extrapolated out to the muon sub-detector. If the track position matched a track segment in the muon sub-detector, then the track is made into "tracker muon." The other method, able to capture the high-momentum muons, is to find a match in the tracker for the standalone muons made by the muon sub-detector, which is the reverse of the previous method. If a match is found, then the candidate is considered a "global muon" and a global fit of the two tracks is made to improve the momentum measurement and resolution. The global muons, tracker muons, and standalone muons are then combined into a single collection which avoids double counting.

Just like for the electrons, there are identification requirements which each muon must pass. This helps to remove cosmic ray muons, muons from heavy flavor decays, and leakage from hadronic showers which may enter the muon collection. Just like the electrons, the the distance between the primary vertex and the transverse and longitudinal impact parameters, d_0^{vtx} and d_z^{vtx}, are used. Additionally, there are requirements on the number of hits in the muon system, the number of stations used in the muon system, the number of pixel hits in the tracker, the overall number of tracker hits, and the reduced χ^2 of the global muon fit. The isolation, which can be seen in equation 4.8, is calculated using the PF candidates within a cone of $\Delta R < 0.4$ around the muon.

$$I_{\mu}^{PF} = \sum_{\Delta R<0.4} p_{\mathrm{T}}^{(CH)} + max\left(\sum_{\Delta R<0.4} p_{\mathrm{T}}^{(NH)} + \sum_{\Delta R<0.4} p_{\mathrm{T}}^{(\gamma)} - \Delta\beta \sum_{\Delta R<0.4} p_{\mathrm{T}}^{(PU)}, 0\right) \qquad (4.8)$$

The variable is very similar to the one used for electrons except that instead of an effective area pileup correction, the muons use a pileup correction based on the sum p_{T} of the charge particles which don't come from the same vertex as the muon candidate. The $\Delta\beta$ term is set to 0.5 and is the ratio of charged to neutral particles in pileup [115].

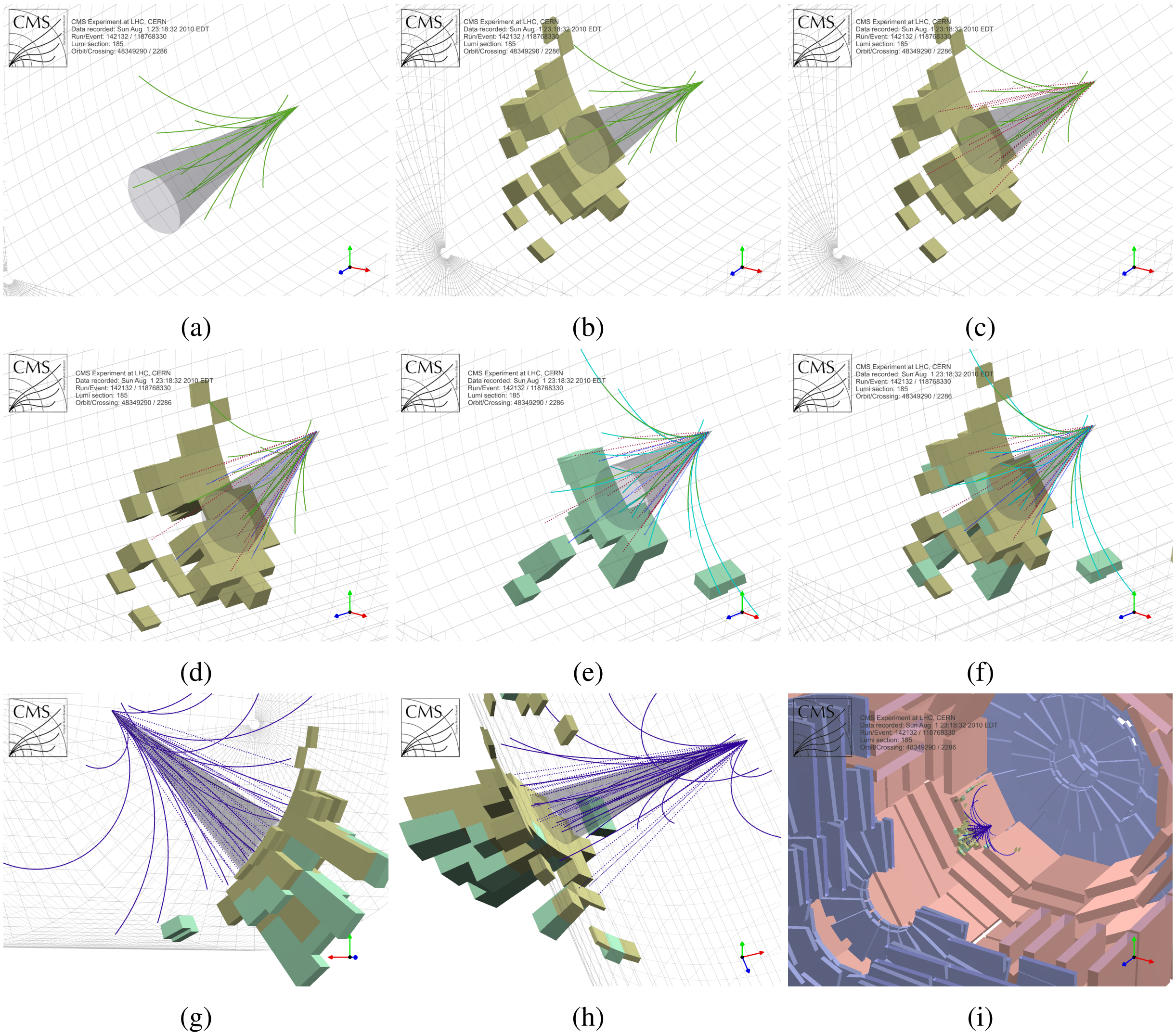
The identification requirements for muons also relies on two WP, a set of tight cuts to select for muons to use in the analysis and a set of loose cuts to veto on additional muons. One again, additional p_{T} and η requirements are imposed on the tight and loose muons to ensure that they match the requirements of the PD as stated in 5.1.1. The tight muons must have $p_{\mathrm{T}} > 25\,\mathrm{GeV}$ and be within $|\eta| < 2.1$ whereas the loose muons must have $p_{\mathrm{T}} > 10\,\mathrm{GeV}$ and be within $|\eta| < 2.5$. The cuts used to identify good, prompt muons are listed in table 4.4.

Cut Variable	Cut Value	
	Tight	Loose
Is PF muon	True	True
Muon category	Global muon	Global muon OR tracker muon
$I_{\mu}^{PF}/p_{\mathrm{T}}^{(\mu)} <$	0.12	0.2
$\lvert d_0^{vtx}\rvert <$	0.02	-
$\lvert d_z^{vtx}\rvert <$	0.5	-
Global track fit $\chi^2/n_{\mathrm{dof}} <$	10	-
Global track fit $n_{\text{muon segment}} >$	0	-
n_{hits} (pixel) $>$	0	-
n_{layers} (tracker) $>$	5	-
$n_{stations}$ (muon) $>$	1	-

Table 4.4: Cut based muon identification requirements for the tight and loose working points.

4.5 Jets

The protons that make up the LHC beams are bound states of quarks and gluons, which are particles that carry color charge. If, during a proton-proton collision a quark or gluon is freed, it must create other colored particles to combine with and form color singlet bound states, hadrons, in a process known as hadronization. This is because a colored state cannot exist alone due to QCD confinement, which only allows for free colorless states. The cascade of particle production will continue until there are no free color states and there is not enough energy in the gluon field to continue hadronizing. The hadronization products themselves may still decay into other particles, including colorless leptons and photons. The CMS detector will not see the initiating parton, but it will certainly measure this cascade of particles as a narrow cluster of tracks and energy, which are collectively referred to as a jet [128]. While this is the behavior of most light quarks and gluons, top quarks are so heavy that they decay into a W boson and a b quark without hadronizing first.


CMS
CMS Experiment at LHC, CERN
Data recorded: Sun Aug 1 23:18:32 2010 EDT
Run/Event: 142132 / 118768330
Lumi section: 185
Orbit/Crossing: 48349290 / 2286

(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Figure 4.5: Different views of the same 115 GeV PF jet are shown with varying amounts of information displayed. The panels are ordered sequentially from left to right and top to bottom where each subsequent panel includes additional information. The image is of a jet with its (a) tracks, (b) ECAL deposits, (c) photon candidates, (d) neutral hadrons. Panel (e) shows the jet with its charged hadrons, but replacing the ECAL deposits for the HCAL deposits. Panels (f)-(h) show various views of the same jet with all of its constituents, while panel (i) shows the jet as it would appear in the CMS detector. The distance between the primary vertex and the interior of the red muon chambers is 7.5 m and the calorimeter deposits are scaled to about 10 GeV/m. Reprinted from [15].

While the best way to cluster the cascade is still an open topic of discussion[2], this analysis clusters PF candidates using the anti-k_T algorithm [17] as defined in the FASTJET package [129][3]. The anti-k_T is a sequential recombination clustering algorithm which is both infrared and collinear safe. Infrared safety means that the jet clustering algorithm is insensitive to the emission of soft, wide angle particles. In other words, the jet is invariant under $\vec{p_i} \rightarrow \vec{p_j} + \vec{p_k}$, where the particle with momentum $\vec{p_i}$ is split into two particles, each carrying momentum $\vec{p_j}$ and $\vec{p_k}$ respectively. As an example, two jets should not be merged together just because one of them produced a 1 GeV particle between them. Collinear safety means that if there is a splitting which results in two parallel high-p_T particles, a single jet is produced and the jet properties will not be different from a jet where this splitting did not occur. When an algorithm obeys these two properties, they are referred to as being IRC safe. Simply put, the anti-k_T algorithm results in jets which have physical properties (i.e. p_T, mass, etc.) that are representative of the partons in the event.

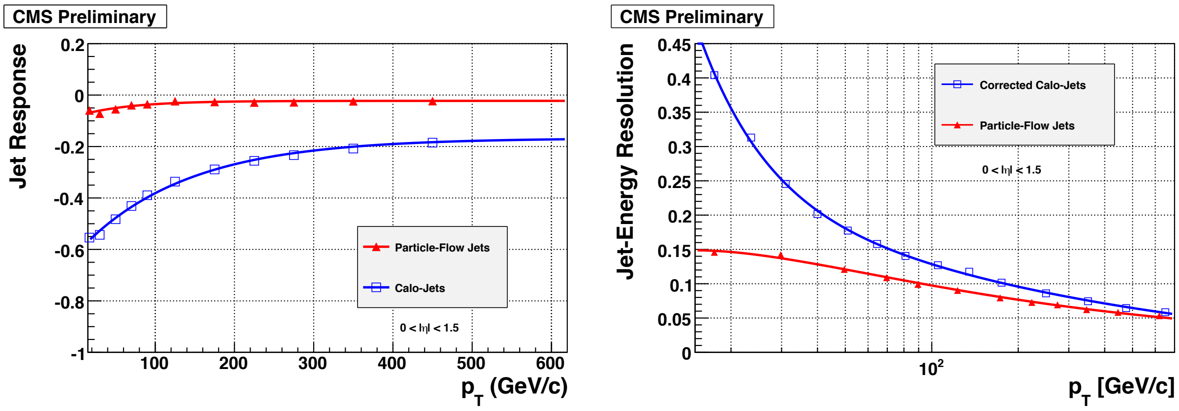
The use of PF candidates, with their built in tracking information, provides a huge benefit to the reconstruction and clustering of jets in CMS. About 65% of the energy within a jet is carried by the charged particles and thus a lot of information about a jet comes from the tracker.[4] An alternative to clustering PF candidates is to cluster the energy deposits in the calorimeter towers, but that provides both less spacial information as well as a lower response, where jet response is defined as $\langle \left(p_\mathrm{T}^\mathrm{reco} - p_\mathrm{T}^\mathrm{gen}\right) / p_\mathrm{T}^\mathrm{gen} \rangle$ and $p_\mathrm{T}^\mathrm{reco}$ ($p_\mathrm{T}^\mathrm{gen}$) is the reconstructed (generated) p_T. Fig. 4.6a shows a comparison of the PF based jet response (PF jets) versus calorimeter based jet responses (calo jets). The use of tracking information also improves the jet resolution, where typical values for a PF jet are 15% at 10 GeV, 8% at 100 GeV, and 4% at 1 TeV. This is compared to about 40%, 12%, and 5% when using calo jets [14]. A comparison of the resolution curves can be see in fig. 4.6b.

Before clustering the PF candidates, a pileup mitigation algorithm called charged hadron sub-

[2] Researchers are constantly asking themselves, "What is a jet?" The question is referring more to the idea of how to reconstruct a jet rather than the concept of a jet.

[3] In addition to providing fast, sequential clustering algorithms, the FASTJET package is able to calculate the jet area, which is a non-trivial quantity [130].

[4] 25% of the energy is carried by photons and the remaining 10% is carried by neutral hadrons.



(a)

(b)

Figure 4.6: Jet response (left) and resolution (right) as a function of p_{T} for jets made by clustering PF candidates and those clustering calorimeter towers. These figures were made using a MC sample with a center-of-mass energy of 10 TeV, requiring the jets' p_{T} to be less than 750 GeV, and that the jets are within $|\eta| < 1.5$. Reprinted from [14].

traction (CHS) is performed. As discussed in section 4.1, CMS can associate a track to a specific vertex. If these tracks are unambiguously associated to a pileup vertex, they are removed from the collection of PF candidates used to cluster jets and calculate the $\vec{\not\!E}_{\mathrm{T}}$. As shown in fig. 4.7, the CHS algorithm is able to remove about 50% of the pileup energy produced during the same bunch crossing as the primary vertex. Any remaining energy from charged hadrons is coming from tracks that are not associated with a high quality vertex or which simply have too large a χ^2/N_{dof}. Some of this is explained by the vertex reconstruction and identification inefficiency of about 30% [16].

Like most other clustering algorithms (i.e. k$_{\mathrm{T}}$, Cambridge/Aachen, SisCone, etc.), anti-k$_{\mathrm{T}}$ is iterative, wherein at each iteration two distance parameters are calculated. d_{ij} and d_{iB}, as defined in equation 4.9, are the distance between two entities (PF candidates or existing clusters) and the distance from any one entity and the beam, respectively. y is the rapidity and R is a radius parameter, which is 0.5 in this analysis. The anti-k$_{\mathrm{T}}$ algorithm is achieved when $p = -1$, whereas if $p = 1$ ($p = 0$) the k$_{\mathrm{T}}$ (Cambridge/Aachen) algorithm is used instead. If $d_{ij} < d_{iB}$, then entity i and j are combined vectorially. However, if $d_{ij} > d_{iB}$, then entity i is classified as a jet and is removed from further clustering. This process continues until all PF candidates have been

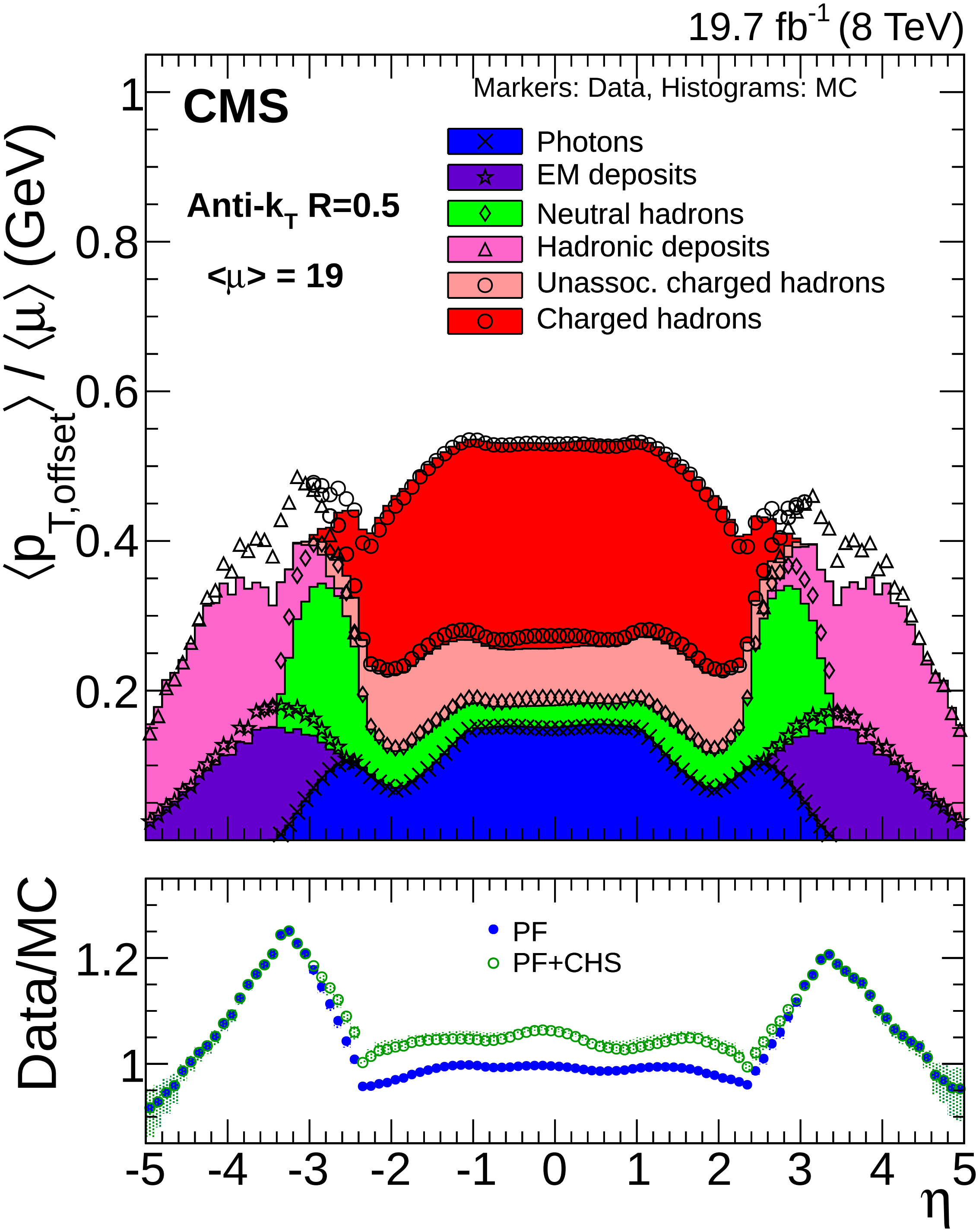


Figure 4.7: Pileup energy within a jet per additional proton-proton interaction (μ) separated by type PF type. The fraction labeled "charged hadrons" will be removed by the CHS algorithm. The ratio of the data to the simulation is shown in the lower panel. Reprinted from [16].

clustered [17] and the momentum of the jet is the vectorial sum of all of the PF candidate momenta. The result of this process can be seen in fig. 4.8.

$$d_{iB} = {p_T}_i^{2p} \tag{4.9a}$$

$$d_{ij} = min\left({p_T}_i^{2p}, {p_T}_j^{2p}\right) \frac{\left(y_i - y_j\right)^2 + \left(\varphi_i - \varphi_j\right)^2}{R^2} \tag{4.9b}$$

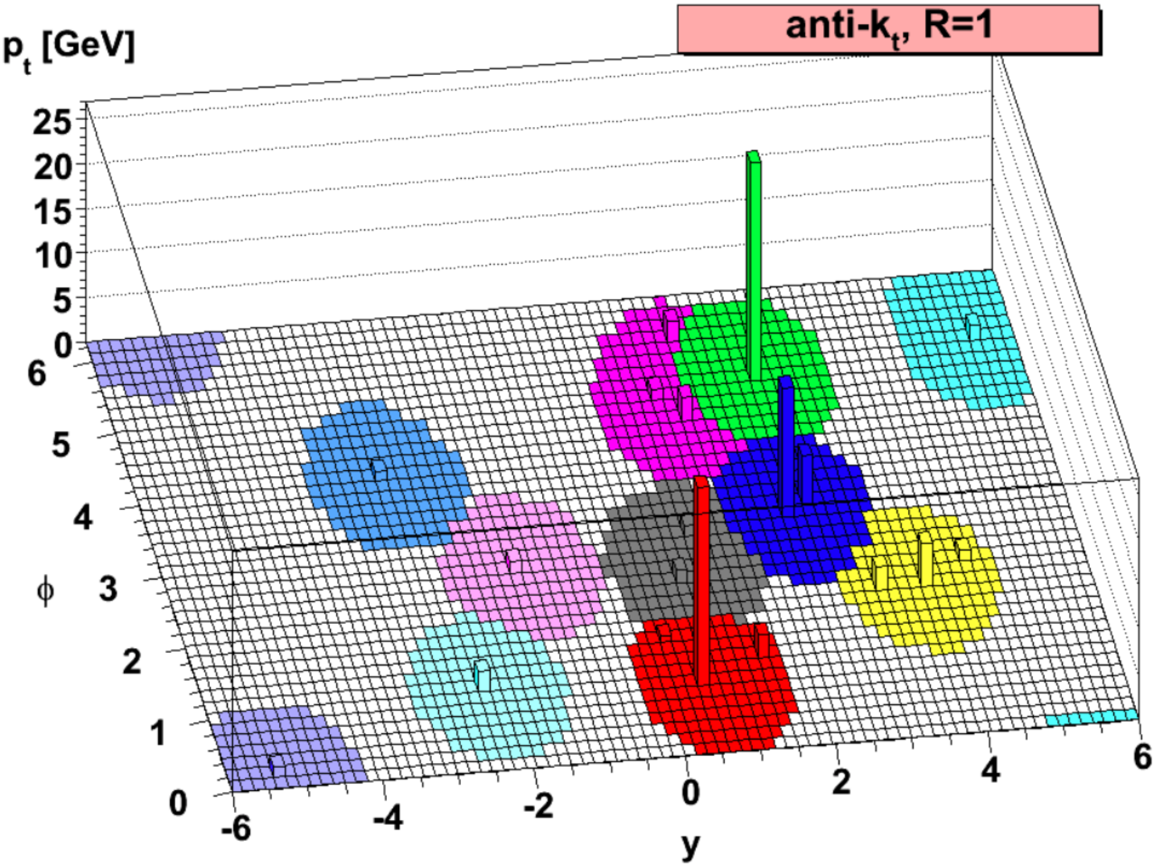


Figure 4.8: Jets clustered from generator level partons using the anti-k$_T$ algorithm. This produces roughly circular jets with stable areas that are insensitive to additional soft particles. Reprinted from [17].

After CHS and the clustering procedure, the momentum and energy of the jets still might not be the same as those from the initial parton, whether because of pileup or detector effects. To correct for this, CMS uses a factorized approach, wherein each level of correction targets a specific effect and each correction is applied in order. The goal is to make sure each jet has a relative response ($\mathcal{R}_{rel} = RelRsp = \frac{p_{\mathrm{T}}^{\mathrm{reco}}}{p_{\mathrm{T}}^{\mathrm{ref}}}$) of 1.0, where $p_{\mathrm{T}}^{\mathrm{reco}}$ is the reconstructed jet p_{T} and $p_{\mathrm{T}}^{\mathrm{ref}}$ is the true or reference p_{T} of the jet without all of the deleterious effects. This type of scaling is commonly referred to as a jet energy correction (JEC)[5]. The first level of correction, commonly referred to as the L1FastJet corrections, starts by removing any remaining pileup[6] or electronic noise energy that may have made it into the jet reconstruction. This multiplicative correction will only remove energy from within the jet and will take the form in equation 4.10, where ρ is the median energy density of the event, A is the jet area, and f is an estimate of the offset inside the jet per unit of jet area [131, 132].

$$p_{\mathrm{T}}^{\mathrm{L1Corrected}} = p_{\mathrm{T}}^{\mathrm{uncorrected}} \cdot \left(1 - A\frac{f\left(\eta, \rho, A\right)}{p_{\mathrm{T}}^{\mathrm{uncorrected}}}\right) \tag{4.10}$$

The L2Relative correction seeks to correct for the non-linearity in the jet response as a function of η while the L3Absolute correction does the same thing as a function of p_{T}. These are again multiplicative corrections that can either increase or decrease the energy of the jet. All three corrections are applied to both data and simulation. An additional level of correction, termed L2L3Residual, is applied only to data to correct for the difference in scale between the data and simulation.

A final level of modification to the reconstructed objects is an η dependent smearing factor applied to the jet 4-momenta coming from the MC samples. The distribution of jet energies within the MC simulation tends to be more sharply peaked and less broad than the same distribution in data. In other words, the MC has a smaller jet energy resolution (JER) than we can realistically measure using the CMS detector. The deterministic "smearing" method recommended by CMS seeks to make the jet energy resolution in MC match the jet energy resolution in data. The reconstructed jet

[5] The terms p_{T} and energy will be used interchangeably only when discussing the jet energy corrections. This is because the corrections will affect both the energy and p_{T} terms within the jet 4-momentum.

[6] CHS was able to remove pileup energy coming from charged hadrons, but not energy added to the jet from, for example, neutral hadrons or photons as seen in fig. 4.7.

p_{T} is scaled by a correction factor C_{JER} as determined in equation 4.11, where C_η is a correction factor derived as a function of η whose values can be found in table 4.5. The multiplicative JER correction factor is then used to modify the jet 4-momentum as in equation 4.12.

$$C_{JER} = max\left(0.0, \frac{p_T^{GEN}}{p_T^{RECO}} + C_\eta \cdot \left(1 - \frac{p_T^{GEN}}{p_T^{RECO}}\right)\right) \tag{4.11}$$

$$\mathbf{X}_{Jet}^{corrected} = C_{JER} \cdot \mathbf{X}_{Jet}^{RECO} \tag{4.12}$$

Although the reconstruction of the $\vec{\not\!\!E}_{\mathrm{T}}$ object will not be discussed until section 4.7, it is important to note that its value is intrinsically tied to that of the jets. Any modification to the jet energies must also be propagated to the $\vec{\not\!\!E}_{\mathrm{T}}$. The propagation of the corrections due to the JER scaling is shown in equations 4.13 and 4.14. The propagation of the JEC, on the other hand, will be discussed in section 4.7 [133].

$$\not\!\!E_{\mathrm{x}}^{\mathrm{corrected}} = (1 - C_{JER})\, Jet_x^{RECO} + \not\!\!E_{\mathrm{x}}^{\mathrm{RECO}} \tag{4.13}$$

$$\not\!\!E_{\mathrm{y}}^{\mathrm{corrected}} = (1 - C_{JER})\, Jet_y^{RECO} + \not\!\!E_{\mathrm{y}}^{\mathrm{RECO}} \tag{4.14}$$

$\lvert\eta\rvert$	Correction Factor C_η
< 0.5	$1.052^{+0.012}_{-0.012}$ (stat.)$^{+0.062}_{-0.061}$ (syst.)
$\geqslant 0.5$ & < 1.1	$1.057^{+0.012}_{-0.012}$ (stat.)$^{+0.056}_{-0.055}$ (syst.)
$\geqslant 1.1$ & < 1.7	$1.096^{+0.017}_{-0.017}$ (stat.)$^{+0.063}_{-0.062}$ (syst.)
$\geqslant 1.7$ & < 2.3	$1.134^{+0.035}_{-0.035}$ (stat.)$^{+0.087}_{-0.085}$ (syst.)
$\geqslant 2.3$ & < 5.0	$1.288^{+0.127}_{-0.127}$ (stat.)$^{+0.155}_{-0.153}$ (syst.)

Table 4.5: Jet energy resolution (JER) scale factors.

A set of quality cuts, collectively called PF jet identification, are applied to the resulting collection of jets to ensure that only real, hard scatter PF jets are used during the analysis [134]. Several working points are defined at varying levels of efficiency and purity, but this analysis makes use of the loose criteria shown in table 4.6 [135]. The variables used in these cuts include the fraction of neutral hadrons in the jet f_{NH}, the fraction of neutral EM particles f_γ, the fraction of charged hadrons f_{CH}, the fraction of charged EM particles f_{EM}, the number of constituents $n_{constituents}$, and the multiplicity of charged particles $n_{charged}$. All cuts on the jet energy fractions are made on the raw jets, before any energy correction are applied. In addition to the PF jet quality cuts, this analysis requires that all jets be within 2.4 $< |\eta| <$, the leading jet has a $p_{\mathrm{T}} > 30\,\mathrm{GeV}$, and all other jets have $p_{\mathrm{T}} > 25\,\mathrm{GeV}$. Additionally, all jets are required to be at least $\Delta R(\mathrm{jet}, \mathrm{lepton}) > 0.3$ away from any isolated, selected lepton.

Cut Variable	Cut Value
	Loose
$f_{CH} >$	0.0
$f_{NH} <$	0.99
$f_\gamma <$	0.99
$f_{EM} <$	0.99
$n_{charged} >$	0
$n_{constituents} >$	1

Table 4.6: Cut based PF jet identification requirements for the loose working point.

4.6 b-tagging

Bottom quarks are interesting because they are often associated with the decays of the top quark and the Higgs boson. The experimental signature of a hadronizing bottom quark will be a *b-jet*. This flavor of jet is identifiable because of the unique decay kinematics of b hadrons, including their long lifetime (1.5 ps $\Rightarrow c\tau \approx 450\,\mu$m) and high p_{T} decay products [2, 136]. Additionally, b hadrons have a relatively large mass ($\sim 5\,\mathrm{GeV}$), which means they have a higher track multiplicity than other quark jets, about 5 on average. The displaced tracks will form a secondary vertex with a large impact parameter which can be measured by the tracking sub-detector. CMS uses the Combined Secondary Vertex (CSV) algorithm to tag jets as either being initiated by a bottom quark or some other parton (u, d, s, c, and g) [137, 138].

In order to identify secondary vertices, the algorithm starts from a subset of well-reconstructed tracks. These tracks must have a p_{T} greater than 1 GeV, $\chi^2/N_{dof} < 5$, a transverse (longitudinal) impact parameter less than 0.2 cm (17 cm), and a ΔR to the jet axis less than 0.3. Each track is also required to have at least 8 hits in the tracker, of which 2 must be from the pixel detector. To reduce the effects of pileup the track's distance of closest approach to the jet axis (primary vertex) must be less than 700 μm (5 cm). Once the tracks are selected, the secondary vertices are reconstructed using the AVF described in section 4.1. At each iteration, if the track weight is greater than 0.5 the track is removed and the iterations continue until no more secondary vertices are found. In order to increase the purity of the secondary vertices they are required to share no more than 65% of their tracks with the primary vertex, to be more then 3σ away from the primary vertex in the $\eta - \varphi$ plane, and the ΔR between the vertex and the jet direction must be less than 0.5. A secondary vertex candidate is also rejected if its radial distance to the primary vertex is greater than 2.5 cm and its invariant mass is close to that of the K^0. The jets are then assigned as being associated to a real secondary vertex, a *pseudo-vertex*, or no vertex. A *pseudo-vertex* if created when the AVF fails to find a secondary vertex, but there are at least two tracks with $S_{ip} > 2$, where S_{ip} is the significance of the track's impact parameter defined as the value of the impact parameter divided by its uncertainty.

The following are used as inputs to the CSV tagger:

- The significance of the flight distance in the transverse plane between the secondary and primary vertices.
- The invariant mass of the secondary vertex (the mass of all of the tracks associated with that vertex).
- The number of tracks associated to the secondary vertex.
- The ratio of the energy carried by the tracks associated to the secondary vertex and all tracks in the jet.
- The $\Delta\eta$ between the jet axis and the tracks associated to the secondary vertex.

- The transverse impact parameter significance of the tracks which raises the invariant mass above 1.5 GeV, the charm threshold. The tracks are ordered by decreasing significance and combined one-by-one until the charm threshold is met.

- The number of tracks in the jet.

- The three-dimensional impact parameter significance of each track.

- The secondary vertex category (real, pseudo, or none).

All of the inputs are computed for jets with at least one associated real secondary vertex. The first input is not computed for jets with only a *pseudo-vertex* because it doesn't have a well-defined position. Only the last three inputs, which are track based, are computed when no secondary vertex is found for the jet.

The inputs to the algorithm are combined using a likelihood-based discriminator, where the likelihood is defined in equation 4.15.

$$\mathcal{L}^{b,c,q} = f^{b,c,q}\left(\alpha\right) \times \prod_{i} f_{\alpha}^{b,c,q}\left(x_i\right) \tag{4.15}$$

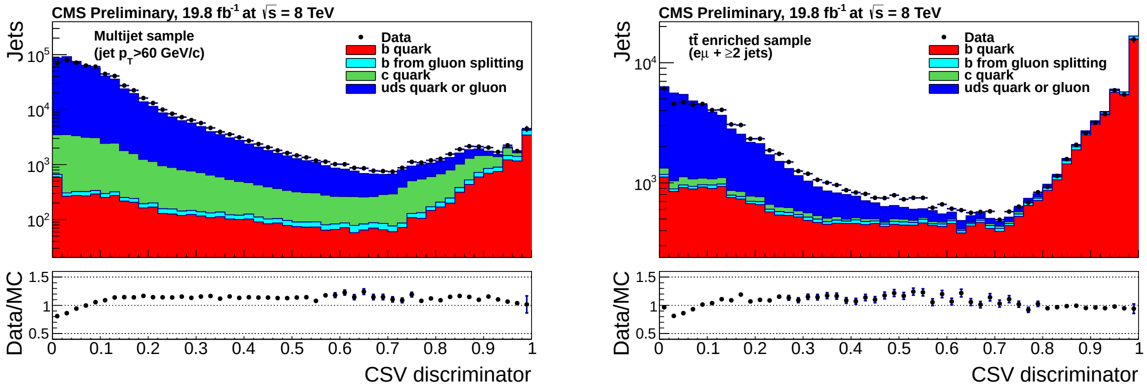
Here b,c, and $q = \{u, d, s, q\}$ are the flavor of the jet, α is the vertex category, $f^{b,c,q}\left(\alpha\right)$ is the probability density function (PDF) for the jet of a given flavor to have a vertex of category α, x_i is one of the inputs, and $f_{\alpha}^{b,c,q}\left(x_i\right)$ is the PDF for x_i given the jet flavor and vertex category. The discriminator is then defined in equation 4.16.

$$d_{CSV} = f_{BG}\left(c\right)\frac{\mathcal{L}^b}{\mathcal{L}^b + \mathcal{L}^c} + f_{BG}\left(q\right)\frac{\mathcal{L}^b}{\mathcal{L}^b + \mathcal{L}^q} \tag{4.16}$$

In this case, $f_{BG}\left(c\right) = 0.25$ and $f_{BG}\left(q\right) = 0.75$ are weights that approximate the expected background (BG) composition.

Working points for this discriminator are defined using probability to mis-identify a light quark or gluon jet as a b-jet [18]. The loose, medium, and tight working points have a 10%, 1%, and

0.1% mistag rate, respectively. This analysis uses the medium working point ($d_{CSV} > 0.679$), which has a tagging efficiency of $\geq 60\%$ as shown in fig. 4.10. For example, a b-jet with a p_{T} of 80 GeV has a tagging efficiency of 75%. Fig. 4.9 shows the CSV discriminator distribution in both a QCD dominated and $\mathrm{t\bar{t}}$ dominated sample. The MC simulation is separated by jet flavor to show the discrimination power of the CSV algorithm.



(a)

(b)

Figure 4.9: The d_{CSV} distribution in a (left) QCD dominated sample and (right) a $\mathrm{t\bar{t}}$ dominated sample. Reprinted from [18].

4.7 Missing Transverse Energy

While CMS is designed to detect as many particles as possible, some particles may be outside of the detector acceptance, may be mis-measured, or may simple not interact with the detection elements. Examples of this are a particle which is beyond an η of 5.0 or a neutrino, which will make it through the detector without ever interacting or a BSM particle which do not interact with the detector. Furthermore, there may be additional particles in the event due to pileup which can can lead to fake $\not\!\vec{E}_{\mathrm{T}}$ due to calorimeter thresholds and response nonlinearities. Because the proton beams have near zero momentum in the x and y directions, only traveling in the z direction, any imbalance in the momentum in the transverse plane indicates additional, missing, or mis-measured particles. This imbalance is called missing transverse momentum and is the negative vector sum

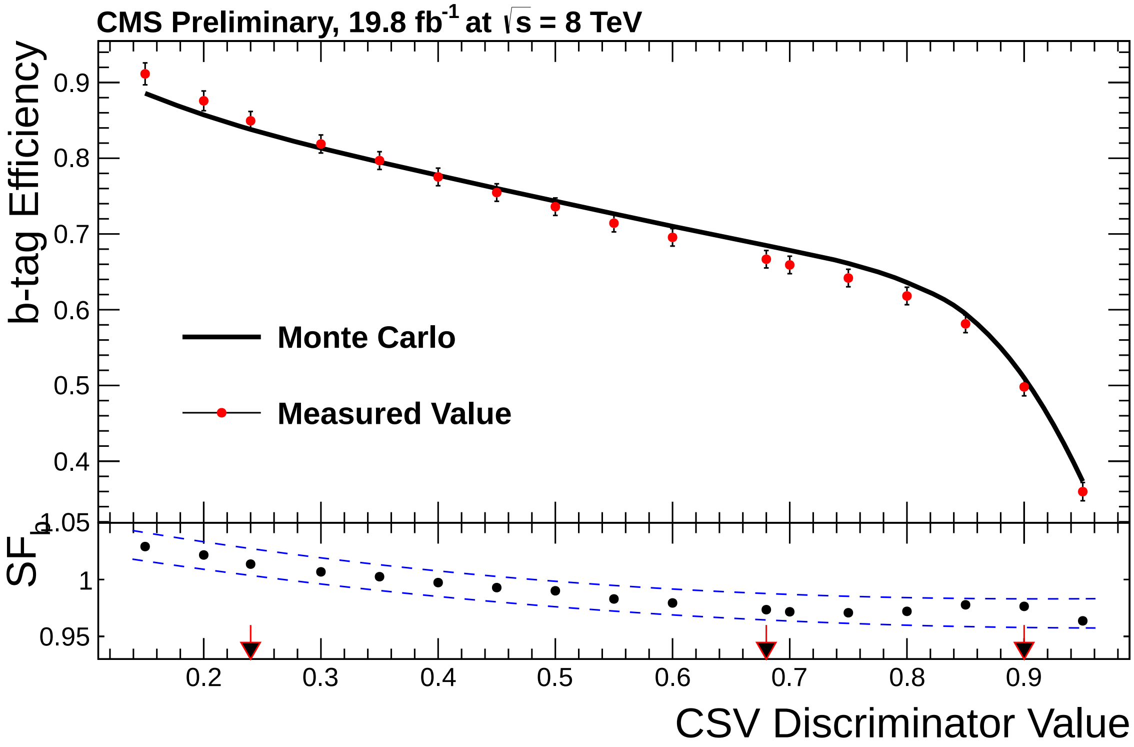


Figure 4.10: The b-tagging efficiency as a function of the CSV discriminator (d_{CSV}) value for both data and MC. The lower panel shows the ratio of the data and MC efficiencies and the arrows along the x-axis show the loose, medium, and tight working point values. Reprinted from [18].

of $\vec{p}_{\mathrm{T}}$ for all PF candidates in the event as seen in equation 4.17. The magnitude of this quantity is known as missing transverse energy and is represented as $\not\!\!E_{\mathrm{T}}$ [139]. A schematic of these two quantities is shown in fig. 4.11.

$$\vec{\not\!\!E}_{\mathrm{T}}^{\mathrm{uncorr.}} = -\sum_{i} \vec{p}_{T}^{\,i} \tag{4.17}$$

Because the $\vec{\not\!\!E}_{\mathrm{T}}$ is affected by every *visible* particle in the event, meaning particles which interact using the electromagnetic or strong forces, it is particularly sensitive to minimum energy thresholds in the calorimeters, inefficiencies and p_{T} thresholds in the tracker, and the non-linear, non-compensating response of the ECAL and HCAL. While electrons and muons have a very good resolution and are typically measured correctly, composite objects like jets have a non-negligible affect on the $\vec{\not\!\!E}_{\mathrm{T}}$ and the bias due to these effects can be reduced by correcting the jets and prop-

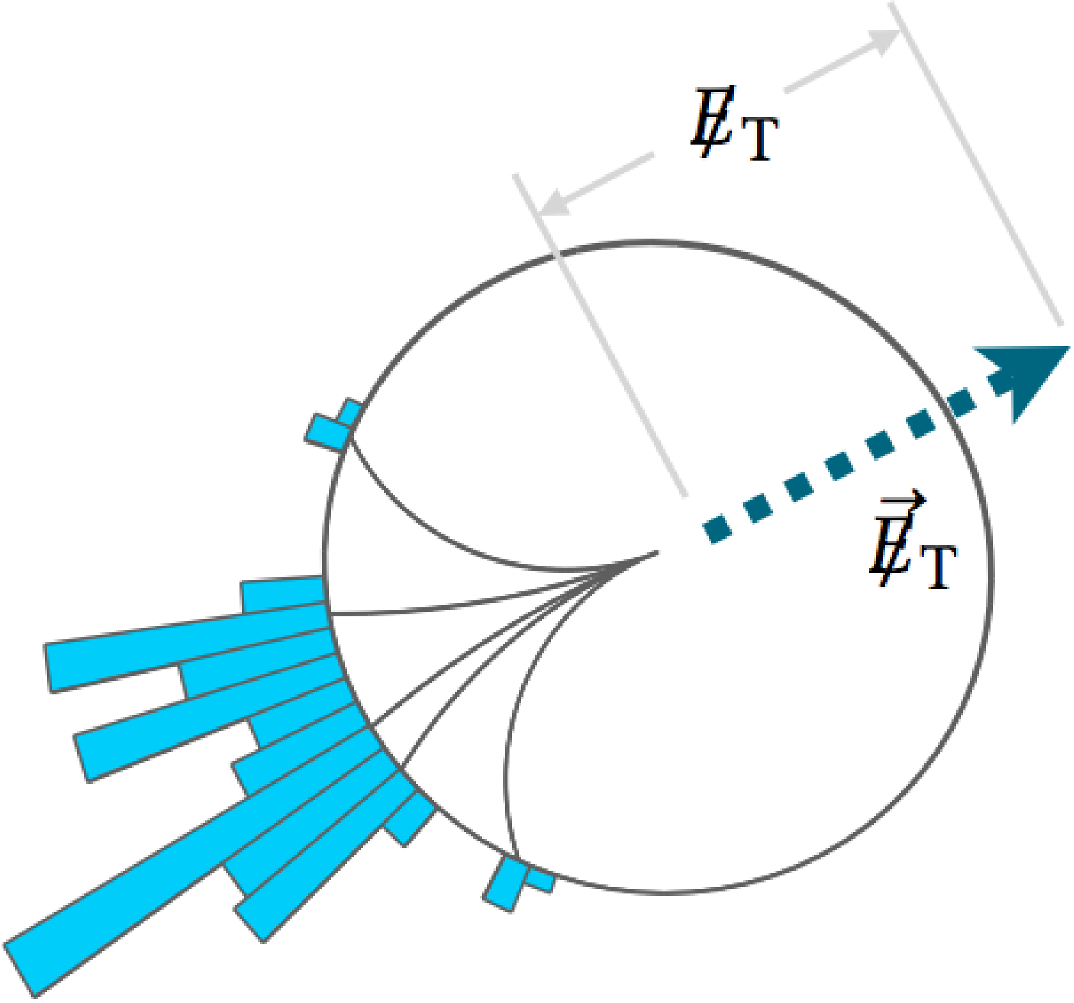


Figure 4.11: A schematic of the $\vec{\not\!\!E}_{\mathrm{T}}$ and $\not\!\!E_{\mathrm{T}}$ quantities Reprinted from [19].

agating those corrections to the $\vec{\not\!\!E}_{\mathrm{T}}$. Unfortunately the corrections discussed in section 4.5 are applied to the composite object and not the the individual constituents[7]. The jet energy corrections are therefore propagated to the $\vec{\not\!\!E}_{\mathrm{T}}$ with the requirements that $f_{EM} < 0.9$ and $p_{\mathrm{T}} > 10\,$GeV so as to exclude electrons which may sometimes produce a non-genuine jet. This type of correction is called Type-1 corrected $\vec{\not\!\!E}_{\mathrm{T}}$ and is what is used in this analysis as a proxy for the undetected

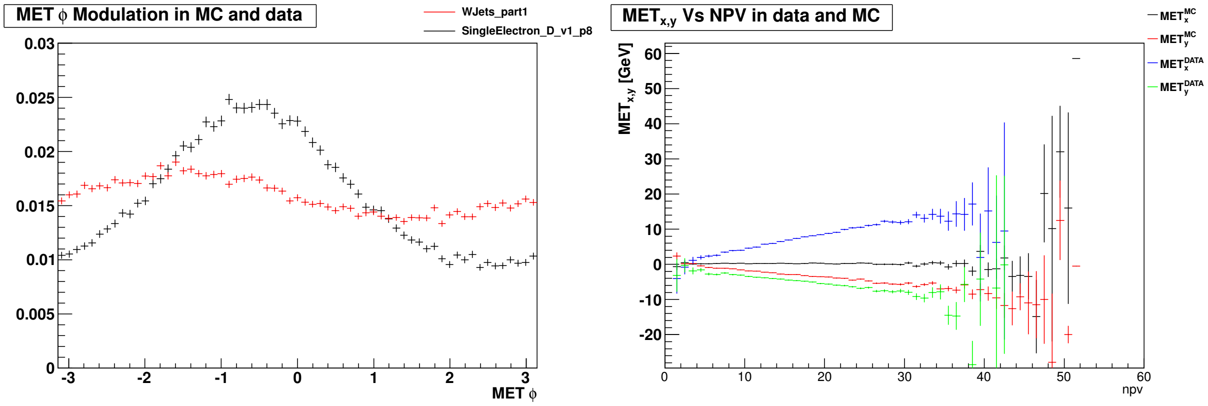
[7] A method which is being actively worked on.

neutrino coming from the decay of one of the W boson.

$$
\begin{aligned}
\vec{\not{E}}_{\mathrm{T}}^{\text{uncorr.}} &= -\sum_{i\in\text{jets}} \vec{p}_{\mathrm{T,\,i}} - \sum_{i\notin\text{jets}} \vec{p}_{\mathrm{T,\,i}} && \text{(4.18a)}\\
&= -\sum_{\text{jet}} \vec{p}^{\,\text{uncorr.}}_{\mathrm{T,\,jet}} - \sum_{i\notin\text{jets}} \vec{p}_{\mathrm{T,\,i}} && \text{(4.18b)}\\
&= -\sum_{\substack{\text{jet}\\ \vec{p}^{\,\text{L123}}_{\mathrm{T,\,jet}}>10\,\text{GeV}}} \vec{p}^{\,\text{uncorr.}}_{\mathrm{T,\,jet}} - \sum_{\substack{\text{jet}\\ \vec{p}^{\,\text{L123}}_{\mathrm{T,\,jet}}<10\,\text{GeV}}} \vec{p}^{\,\text{uncorr.}}_{\mathrm{T,\,jet}} - \sum_{i\notin\text{jets}} \vec{p}_{\mathrm{T,\,i}} && \text{(4.18c)}\\
&= -\sum_{\substack{\text{jet}\\ \vec{p}^{\,\text{L123}}_{\mathrm{T,\,jet}}>10\,\text{GeV}}} \vec{p}^{\,\text{L1}}_{\mathrm{T,\,jet}} - \sum_{\substack{\text{jet}\\ \vec{p}^{\,\text{L123}}_{\mathrm{T,\,jet}}>10\,\text{GeV}}} \left(\vec{p}^{\,\text{uncorr.}}_{\mathrm{T,\,jet}} - \vec{p}^{\,\text{L1}}_{\mathrm{T,\,jet}}\right) && \\
&\quad - \sum_{\substack{\text{jet}\\ \vec{p}^{\,\text{L123}}_{\mathrm{T,\,jet}}<10\,\text{GeV}}} \vec{p}^{\,\text{uncorr.}}_{\mathrm{T,\,jet}} - \sum_{i\notin\text{jets}} \vec{p}_{\mathrm{T,\,i}} && \text{(4.18d)}\\
\vec{\not{E}}_{\mathrm{T}}^{\text{Type}-1} &= -\sum_{\substack{\text{jet}\\ \vec{p}^{\,\text{L123}}_{\mathrm{T,\,jet}}>10\,\text{GeV}}} \vec{p}^{\,\text{L123}}_{\mathrm{T,\,jet}} - \sum_{\substack{\text{jet}\\ \vec{p}^{\,\text{L123}}_{\mathrm{T,\,jet}}>10\,\text{GeV}}} \left(\vec{p}^{\,\text{uncorr.}}_{\mathrm{T,\,jet}} - \vec{p}^{\,\text{L1}}_{\mathrm{T,\,jet}}\right) && \\
&\quad - \sum_{\substack{\text{jet}\\ \vec{p}^{\,\text{L123}}_{\mathrm{T,\,jet}}<10\,\text{GeV}}} \vec{p}^{\,\text{uncorr.}}_{\mathrm{T,\,jet}} - \sum_{i\notin\text{jets}} \vec{p}_{\mathrm{T,\,i}} && \text{(4.18e)}
\end{aligned}
$$

Equation 4.17 shows the simplified and uncorrected model of $\vec{\not{E}}_{\mathrm{T}}$ which can be broken into two categories, those particles which are contained in jets and those which are not. This is shown in equation 4.18a, but can be simplified further to equation 4.18b by noting that the first term is simply the sum of p_{T} for the uncorrected jets. In equation 4.18c the jets are further broken into two classes based on their corrected p_{T} and in equation 4.18d the jets are broken into the pileup corrected jets term and a term for the pileup itself, where $\left(\vec{p}^{\,\text{uncorr.}}_{\mathrm{T,\,jet}} - \vec{p}^{\,\text{L1}}_{\mathrm{T,\,jet}}\right)$ is the additional energy due to pileup (offset). Now that the $\vec{\not{E}}_{\mathrm{T}}$ is fully broken down it can be corrected by replacing $\vec{p}^{\,\text{L1}}_{\mathrm{T,\,jet}}$ in the first term with $\vec{p}^{\,\text{L123}}_{\mathrm{T,\,jet}}$ to give equation 4.18e. This is a correction on the clustered energy in the event above a given threshold [19].

An additional modification to the $\vec{\not{E}}_{\mathrm{T}}$ to remove a modulation in the φ component is also used. $\vec{\not{E}}_{\mathrm{T}}$ should be independent of φ because the proton-proton collisions are rotationally symmetric around the beam axis, so any asymmetry must be due to an error in the simulation or reconstruction. However, we observed a sinusoidal modulation of period 2π in the $\not\!\!E_{\mathrm{T}\varphi}$ for both data and simulation after reconstruction, as can be seen in fig. 4.12a. This effect can be caused by an anisotropic detector responses, inactive calorimeter cells, detector misalignment (for even one of the sub-detectors), or a displacement of the beam spot. All of these will cause the same effect.



(a)

(b)

Figure 4.12: (a) Distribution of the $\not\!\!E_{\mathrm{T}\varphi}$ for data (black) and simulation (red). Only the $\mathrm{W}+\mathrm{jets}$ simulation is shown here, although all of the simulations suffer from the same modulation. (b) Distributions of $\not\!\!E_{\mathrm{x,\,y}}$ as a function of the number of primary vertices. The black and red markers represent the x and y distributions for simulation, respectively, while the blue and green markers are for data.

While the exact cause of the modulation might be unknown, we do know that the amplitude of the modulation increases linearly with the number of proton-proton interactions as each additional particle in the event will increase the φ asymmetry.[8] The dependence on the number of primary vertices can be seen in fig. 4.12b, which shows the x and y components of the $\not\!\!E_{\mathrm{T}}$ 4-vector as a function of the number of primary vertices. Without first correcting this modulation, any cut on the p_{T} of the $\not\!\!E_{\mathrm{T}}$ would preferentially select events on a specific side of the detector. Luckily, the

[8] Assuming the additional particles are created isotropically.

amplitude of the modulation can be reduced by using the transformation in equation 4.19.

$$\vec{p}_T^{\,i} \rightarrow \vec{p}_T^{\,i} - \vec{c} \tag{4.19}$$

Thus the $\vec{\not\!\!E}_{\mathrm{T}}$ becomes:

$$\begin{aligned} \vec{\not\!\!E}_{\mathrm{T}}^{\mathrm{xy}} &= -\sum_{i\in \mathrm{all}} (\vec{p}_{\mathrm{T,\,i}} - \vec{c}) \\ &= -\sum_{i\in \mathrm{all}} \vec{p}_{\mathrm{T,\,i}} + \sum_{i\in \mathrm{all}} \vec{c} \\ &= \vec{\not\!\!E}_{\mathrm{T}}^{\mathrm{raw}} + n\vec{c} \\ &= \vec{\not\!\!E}_{\mathrm{T}}^{\mathrm{raw}} + \vec{C}_{\mathrm{T}}^{\mathrm{xy}} \end{aligned} \tag{4.20}$$

However, instead of applying this correction based on the number of particles, the correction is parameterized based on the number of vertices as a proxy for the number of particles. The correction as a function of the number of vertices is:

$$\vec{C}_{\mathrm{T}}^{\mathrm{xy}} = \vec{c}_A + n_{vtx}\vec{c}_B \tag{4.21}$$

where $\vec{c}_A$ and $\vec{c}_B$ are constant vectors and n_{vtx} is the number of reconstructed primary vertices.

Practically this corrections is accomplished by fitting the distributions in fig. 4.12b with a first order polynomial to obtain $\vec{c}_A$ and $\vec{c}_B$. Those coefficients can be found in table 4.7. Then we use equations 4.22 and 4.23, which differ from equation 4.20 only due to the sign of the coefficients.

$$\not\!\!E_{\mathrm{x}}^{\mathrm{corrected}} = \not\!\!E_{\mathrm{x}}^{\mathrm{RECO}} - ([0]_{\mathrm{x}} + [1]_{\mathrm{x}} \cdot \mathrm{N_{PV}}) \tag{4.22}$$

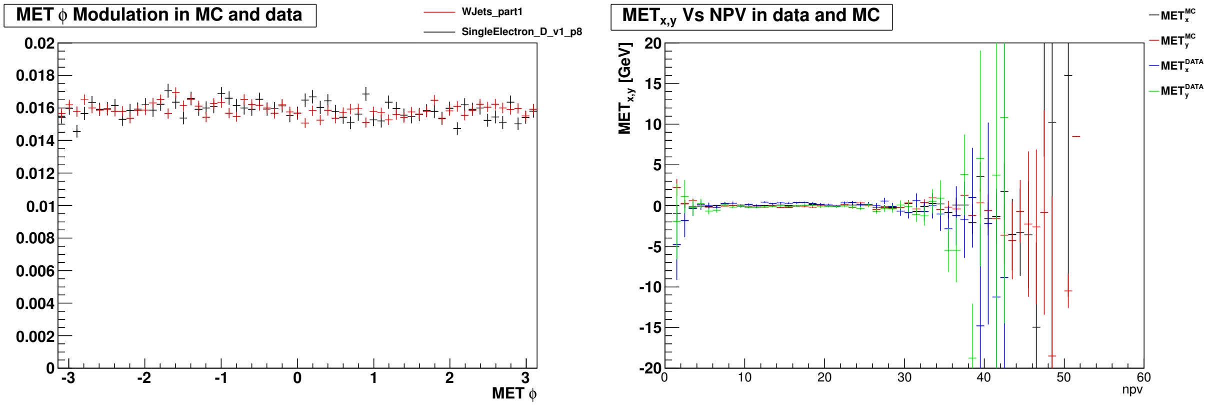
$$\not\!\!E_{\mathrm{y}}^{\mathrm{corrected}} = \not\!\!E_{\mathrm{y}}^{\mathrm{RECO}} - ([0]_{\mathrm{y}} + [1]_{\mathrm{y}} \cdot \mathrm{N_{PV}}) \tag{4.23}$$

The results of this correction are shown in fig. 4.13, where both the modulation in φ and the slope as a function of $\mathrm{N_{PV}}$ are gone. From here we can place a cut on the p_{T} of the $\vec{\not\!\!E}_{\mathrm{T}}$ object without biasing our selection.

In this analysis the resulting $\not\!\!E_{\mathrm{T}}$ is required to have at least 25 GeV in order to reduce the QCD

Coordinate	Parameter 0	Parameter 1
	Data	
x	$2.0105E - 01$	$4.2663E - 01$
y	$-9.1350E - 01$	$-2.3120E - 01$
	MC	
x	$2.9059E - 01$	$-3.5293E - 03$
y	$3.0183E - 01$	$-1.9974E - 01$

Table 4.7: The fit parameters for the $\vec{\not\!\!E}_{\mathrm{T}}^{\varphi}$ corrections.



(a)

(b)

Figure 4.13: (a) Distribution of the $\not\!\!E_{\mathrm{T}\varphi}$ for data (black) and simulation (red) with the correction for the modulation applied. Only the W+jets simulation is shown here, although all of the simulations suffer from the same modulation. (b) Distributions of $\not\!\!E_{\mathrm{x,\,y}}$ as a function of the number of primary vertices after the φ modulation correction has been applied. The black and red markers represent the x and y distributions for simulation, respectively, while the blue and green markers are for data.

events making it through the selection process. A pileup correction to the $\vec{\not\!\!E}_{\mathrm{T}}$ was also available, but was not implemented in this analysis. It is nevertheless discussed in appendix B.1. In addition to propagating the JEC to the $\vec{\not\!\!E}_{\mathrm{T}}$, CMS also filters events and or $\vec{\not\!\!E}_{\mathrm{T}}$ contributions which might introduce noise from the calorimeters or beam halo [140]. These filters are discussed further in appendix B.2.

4.8 Event Generation

In the search for new physics, a signal will generally appear as a small deviation from the SM prediction. In order to disentangle the SM background from a rare signal, the SM and new physics predictions must be extremely accurate. These predictions are ensembles of simulated events made by Monte Carlo (MC) event generators which are broken up by physics process and final state and then recombined during the analysis [20, 141, 142]. These generators are able to simulate a full event (bunch crossing) at the parton level, which is nicely illustrated in fig. 4.14. The image shows a $t\bar{t}h$ final state including final state gluons (QCD) and hadronization. While the entire event from hard scatter production to hadronization cannot be described using perturbation theory, the hard process can be calculated using fixed order perturbation theory and matrix elements (ME). The parton showers, red lines in fig. 4.14, then connect the hard process with the hadronization scale. Phenomenological models are used to simulate the hadronization into stable particles and the underlying event (UE), which is the usually softer interactions by the constituents of the protons which did not take part in the hard scatter process. Photon and gluon emission from the initial protons and final state partons, respectively called initial state radiation (ISR) and final state radiation (FSR), must also be simulated.

Because protons are not elementary particles, it is important to discuss these interactions in terms of the partons inside the proton. Hadrons, like the proton, are made up of valence quarks, sea quarks, and gluons[9]. In essence, the simulation of a proton-proton hard scatter interaction is the calculation of a cross section for an N-particle final state, seen in equation 4.24, where a (b) is a parton carrying a fraction of the momentum x_a (x_b) for hadron A (B).

$$\sigma_N^{AB}(s) = \int dx_a dx_b f_a\left(x_a, \mu^2\right) f_b\left(x_b, \mu^2\right) \hat{\sigma}_N^{ab}\left(\hat{s}, \mu^2\right) \tag{4.24}$$

The parton distribution function (PDF) of the form $f_{a,b}\left(x_{a,b}, \mu^2\right)$ gives the probability density of finding such a parton with momentum fraction $x_{a,b}$ renormalized to scale μ^2. PDFs cannot be obtained using perturbative nor lattice QCD calculations. Instead they are measured within the resolution of the existing experiments. CMS makes use of the Martin-Stirling-Thorne-Watt (MSTW) [21] and Coordinated Theoretical-Experimental Project on QCD (CTEQ) PDFs. Fig. 4.15 shows the NLO MSTW PDFs calculated for two different momentum scales. Other terms include the center-

[9] More precisely, protons are a bound state of two up quarks and a down quark.

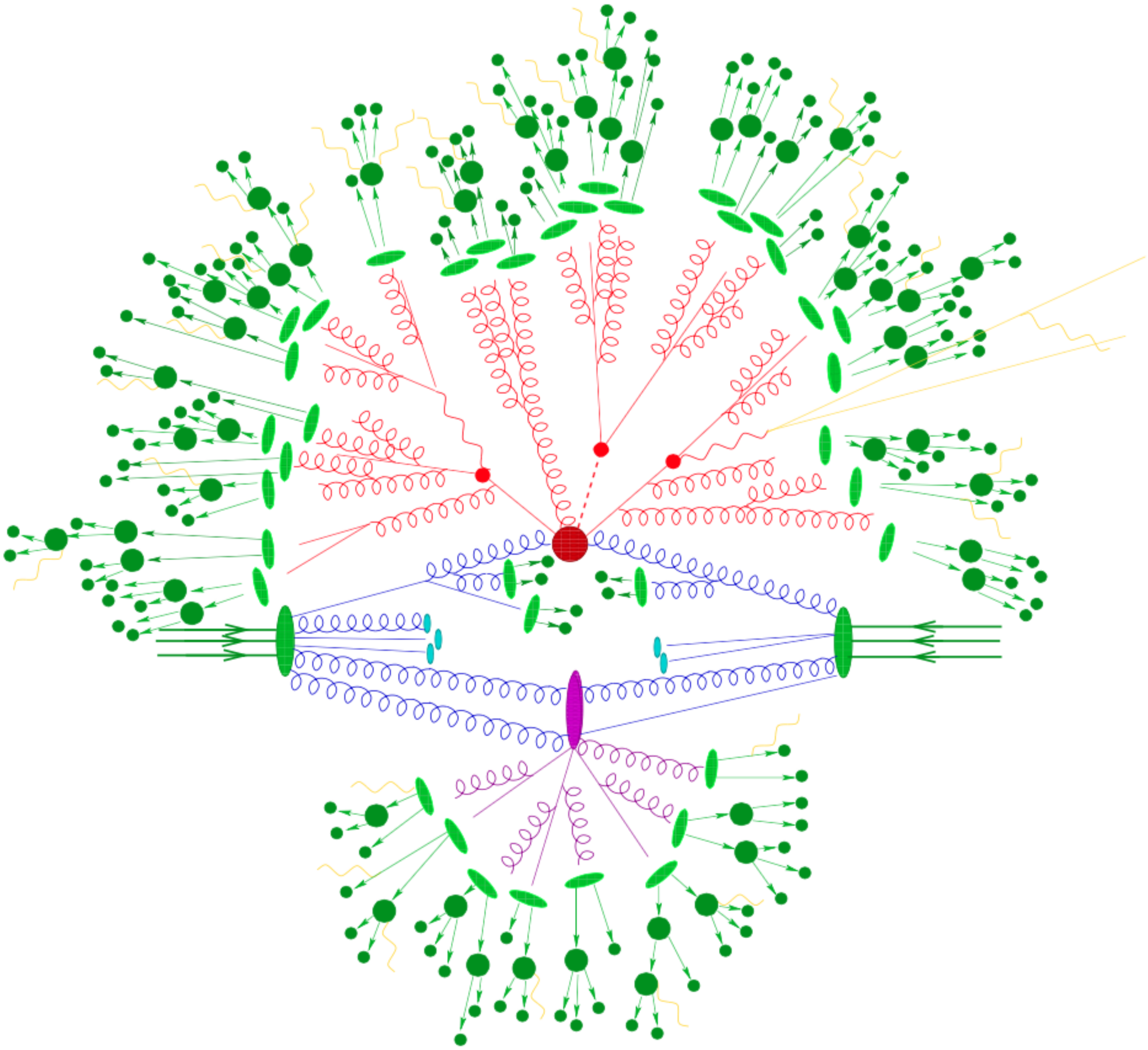


Figure 4.14: A graphical representation of a $t\bar{t}h$ event as seen by a MC event generator. The hard scatter interaction is represented by the red circle being produced by the two gluons coming of the incoming protons. The three small red dots represent the top quarks and the Higgs boson which then decay to additional hard QCD radiation. The underlying event is represented by the purple shapes and lines while the light green shapes are the final-state partons, which then hadronize and decay into the dark green circles. The yellow lines show the photon radiation which can occur at any state in the event generation process. Reprinted from [20].

of-mass energy of the interaction $\sqrt{\hat{s}} = \sqrt{x_a x_b s}$ where $\sqrt{s}$ is the center-of-mass energy of the proton-proton system and $\hat{\sigma}_{ab\rightarrow X}\left(\hat{s}, \mu^2\right)$, which is the cross section for having a given set of initial state partons. The full form of the partonic cross section is given by

$$\begin{aligned}\hat{\sigma}_N^{ab} = \int_{cuts} d\hat{\sigma}_N^{ab} = & \frac{(2\pi)^4 S}{4\sqrt{(p_1{\cdot}p_2)^2 - m_1^2 m_2^2}} \times \\ & \int_{cuts} \left[\prod_{i=1}^{N} \frac{d^3 q_i}{(2\pi)^3 2E_i}\right] \delta^4 \left(p_1 + p_2 - \sum_i^N q_i\right) |\mathcal{M}_{p_1 p_2 \to \{\vec{q}\}}^{ab}|^2\end{aligned} \tag{4.25}$$

where p_i are the four-momenta of the incoming partons, q_i and E_i are the outgoing particle four-momenta and energies, S is the product of $1/j!$ for j identical particles in the final state, and $\mathcal{M}_{p_1 p_2 \to \{\vec{q}\}}^{ab}$ is the ME associated to the kinematic configuration $p_1 p_2 \to \{\vec{q}\}$ with initial partons a and b [20, 143]. In order to evaluate the parton level ME the event generator must either have the ME hard coded or it must be able to compute all of the Feynman diagrams associated with a given process. A good example of this type of calculation can be found in Table 1.1 of [20]. While the number of diagrams for a $2 \to 2$ or $2 \to 3$ process is limited and can be built and computed automatically, the problem becomes much more difficult for next to leading order (NLO) computations as the number of diagrams grows factorially [144]. The growth of the number of diagrams can be seen in fig. 4.16. In many cases the LO MEs and PDFs are used to generate events and a K-factor is used to scale the events to their NLO or NNLO predictions. Besides computing the MEs, the the multi-dimensional phase space integration is quite complicated and requires the use of Monte-Carlo integration techniques [145].

The parton shower takes the partons created by the hard process and UE and perturbatively evolves them down to the hadronization scale, at which point they form colorless hadrons. The partons are initially produced at a scale t' and the parton shower determines the scale $t < t'$ at which the parton should branch into two daughter particles, selecting the kinematics and flavors of those new particles. This process continues recursively and only ceases once the hadronization scale is reached, $\mathcal{O}\left(\text{GeV}\right)$, where α_s becomes large and perturbative methods are no longer applicable. Generators make use of any number of phenomenological models, including the cluster hadronization model [146, 147] and the Lund string model [148, 149], to turn this list of colored partons into colorless hadrons. No matter the model used, the hadrons which result from these

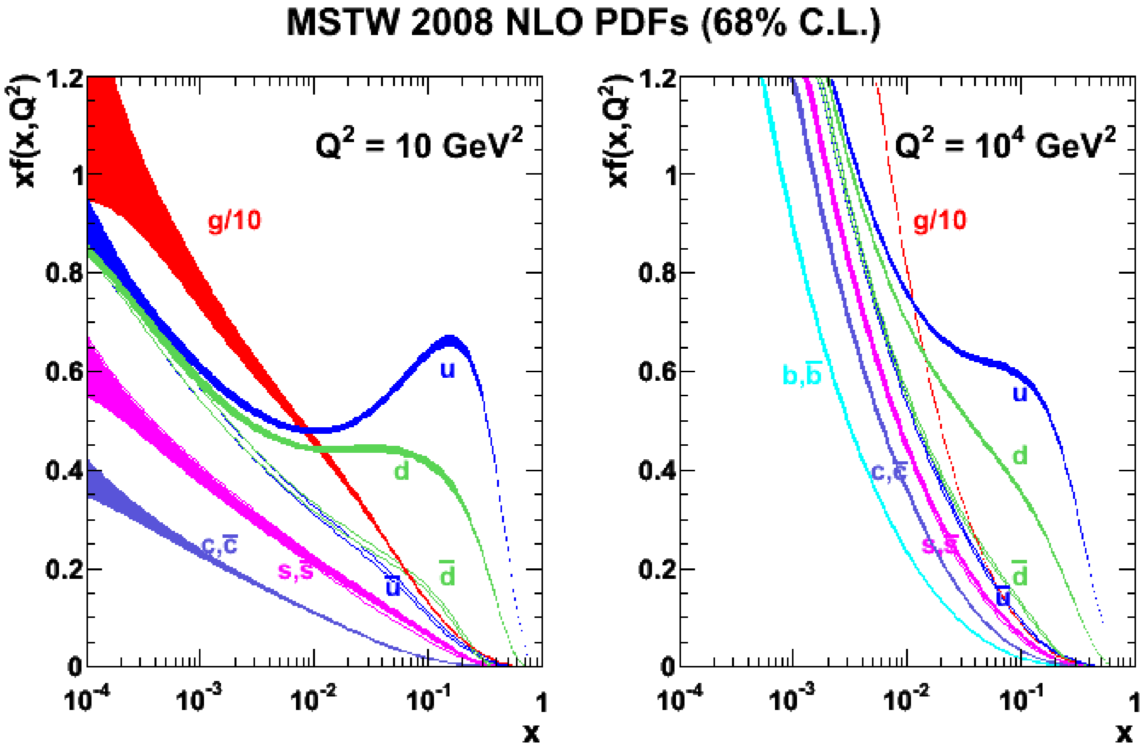
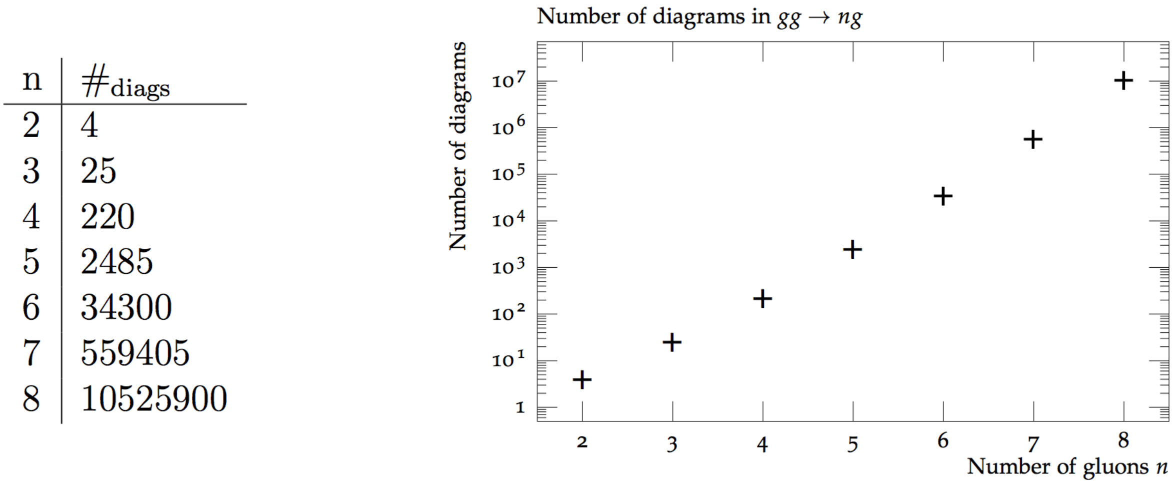


Figure 4.15: The MSTW PDFs calculated to NLO as a function of the momentum fraction for two different interaction momentum scales Q^2. In the case of synchrotron collisions Q^2 is the square of the total four-momentum of the proton-proton interaction. The right plot shows the momentum scale more commonly found at the LHC. Reprinted from [21].



n	#diags
2	4
3	25
4	220
5	2485
6	34300
7	559405
8	10525900

Figure 4.16: The number of diagrams which must be calculated to fully calculate the $gg \to ng$ amplitude. Reprinted from [20].

models are often unstable and will be forced to decay into stable hadrons, which are defined to have a mean lifetime above a given threshold as defined by the experiment.

Various generators are used in this analysis, each with their own benefits and drawbacks. PYTHIA [150] is a general purpose event generator capable of handling many $2 \to 1,\ 2,\ 3$ processes. It is capable of handling all of the needed generation steps including generating the hard scatter process, parton showering to the leading log (LL) level, hadronization, and the UE simulation. PYTHIA makes use of the Lund string model for hadronization and describes the UE as additional, but not quite independent perturbative $2 \to 2$ scatterings. Another generator used is MADGRAPH [151], which more accurately simulates hard parton emission (i.e. ISR and FSR), but must be interfaced with PYTHIA for showering soft and collinear radiation. The POWHEG [152, 153] generator uses NLO matrix elements and PDFs and then matches this with a modified shower simulation. Both MADGRAPH and POWHEG are interfaced with PYTHIA for hadronization. For more accurate tau lepton decays CMS often uses the TAUOLA [154] software package.

4.9 Detector Simulation

Event generation simulates the particle kinematics for a given event, but doesn't examine how the particles will interact with the detector and it's constituent materials or how the readout electronics will behave. To simulate the response of the CMS detector, the generators are interfaced with a sophisticated detector simulation based on the GEANT4 [155, 156] software package, which takes into account the exact detector geometry as well as all materials used. The alignment, calibration, and other conditions which may change over time are periodically checked and are stored in a database. These conditions are used both for offline simulation and reconstruction as well as for online activities. A snapshot of the conditions at some point in time is called a global tag. For reference, this analysis uses the GR_R_53_V10 and START53_V7A global tags for data and simulation, respectively [157]. The final state particles from the event generator are sent to the detector simulation, which tracks the particles as they move through the detector depositing energy into what are called simulated hits (SimHits). While the models of electromagnetic interactions are extremely precise, the hadronic interactions have a greater uncertainty associated with them. The simulation goes through the data acquisition process, event simulating the responses of the photodetectors and readout electronics. The resulting information is then analyzed by the same reconstruction process that the real data goes through and is stored using the ROOT [158] software library.

5. HIGGS ANALYSIS

This thesis presents a search for the SM Higgs boson decaying to the $\mathrm{l}\nu\mathrm{jj}$ final state making use of data collected by the CMS detector at the LHC. To study the efficacy of various object and event selection criteria we make use of signal and background MC simulations. While the signal samples are fully MC based, some of the background samples use data-driven techniques, which will be discussed later in this chapter. The matrix element probabilities for an event final state being created by a specific diagram are computed. Several multivariate techniques are studied and used to distinguish between signal-like and background-like events. We use the discriminator outputs from these multivariate classifiers to set limits on the SM $\mathrm{H} \to \mathrm{WW}$ cross section.

5.1 Data and Monte Carlo Samples

5.1.1 Data

As mentioned previously, this analysis makes use of the full 2012 CMS dataset of 8 TeV data. Fig. 5.1 shows the cumulative delivered, recorded, and validated luminosity versus time. Only fully validated data, where both the LHC and CMS are completely operational, are use used for CMS analyses [9]. Table 5.1 shows the data samples used for this analysis, which corresponds to $\sim$19.2 fb^{-1}. The datasets are split by the two HLT paths used, one which selects for a single high p_{T} electron and one for a single high p_{T} muon. These two separate PDs correspond to the HLT_Ele27_WP80_v* and HLT_IsoMu24_eta2p1_v* trigger paths, respectively. The HLT_Ele27_WP80_v* path requires a reconstructed electron with $p_{\mathrm{T}} > 27$ GeV along with several other criteria grouped into a working point with 80% efficiency of selecting true electrons. The HLT_IsoMu24_eta2p1_v* criteria requires an isolated, reconstructed muon with $p_{\mathrm{T}} > 24$ GeV within $|\eta| < 2.1$. The luminosities listed in the table are associated with a 2.6% uncertainty as specified in [107] and were collected using the HF luminosity measurements [108].

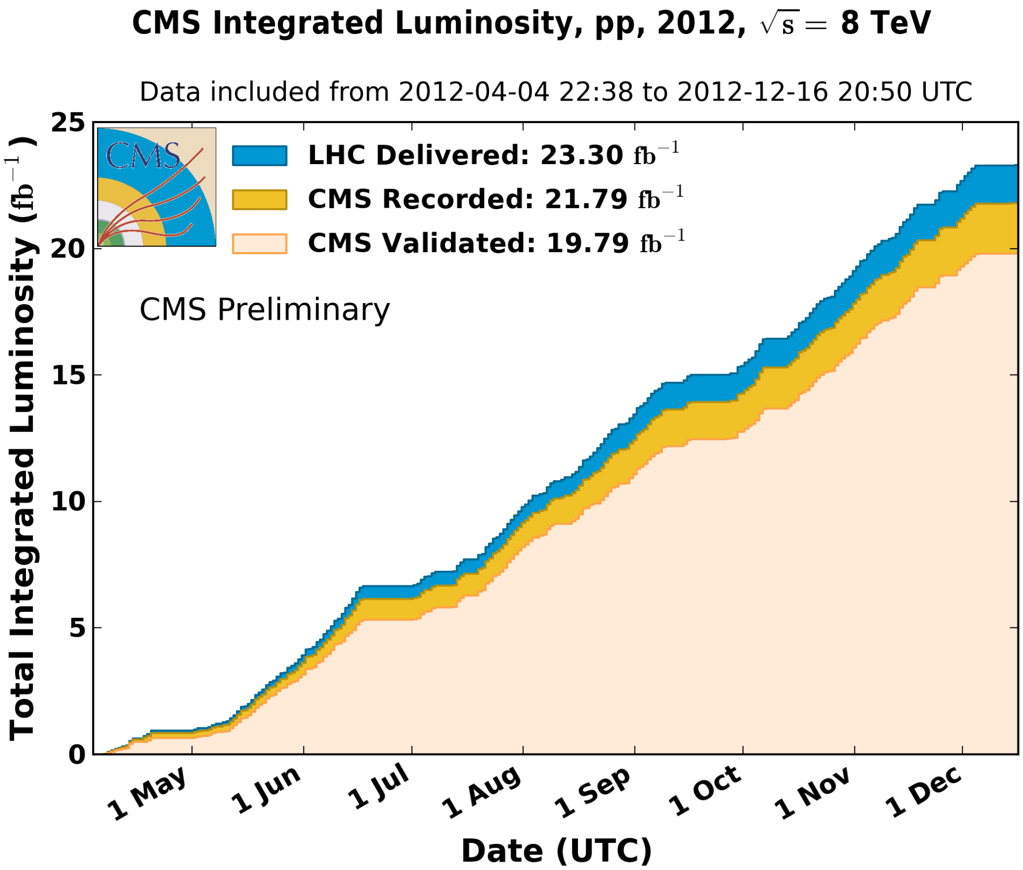


Figure 5.1: Cumulative day-by-day integrated luminosity in 2012 delivered by the LHC (blue), recorded by CMS (dark orange), and validated for physics use (light orange). Reprinted from [22].

5.1.2 Monte Carlo

This analysis makes use of MC simulation to study the background processes which have similar final states to that of the $\mathrm{H} \to \mathrm{WW} \to \mathrm{l}\nu\mathrm{jj}$ signal. Both the kinematic distributions and the final yields are extracted from these samples. The MC simulation is used for all backgrounds except for the multijet process, where a data-driven approach is used instead. The process of developing this sample is described in detail in the section 5.1.3. The signal sample kinematics and yields are also taken from MC. Tables 5.2 and 5.3 list all of the MC sample for the Higgs signals and SM background processes, respectively. The SM background and volunteer signal samples are centrally produced by the CMS collaboration. The ggH samples were produced specifically for

Dataset	Run Range	Integrated Luminosity
/SingleMu/Run2012A-13Jul2012-v1/AOD	190645-196531	0.809 fb^{-1}
/SingleMu/Run2012A-recover-06Aug2012-v1/AOD	190782-190949	0.082 fb^{-1}
/SingleMu/Run2012B-13Jul2012-v1/AOD	193834-196531	4.383 fb^{-1}
/SingleMu/Run2012C-24Aug2012-v1/AOD	198022-198523	0.489 fb^{-1}
/SingleMu/Run2012C-PromptReco-v2/AOD	194631-203002	6.285 fb^{-1}
/SingleMu/Run2012D-PromptReco-v1/AOD	194480-208686	7.231 fb^{-1}
Total SingleMu	**190645–208686**	**19.279 fb^{-1}**
/SingleElectron/Run2012A-13Jul2012-v1/AOD	190645-196531	0.809 fb^{-1}
/SingleElectron/Run2012A-recover-06Aug2012-v1/AOD	190782-190949	0.082 fb^{-1}
/SingleElectron/Run2012B-13Jul2012-v1/AOD	193834-196531	4.336 fb^{-1}
/SingleElectron/Run2012C-24Aug2012-v1/AOD	198022-198523	0.489 fb^{-1}
/SingleElectron/Run2012C-PromptReco-v2/AOD	194631-203002	6.194 fb^{-1}
/SingleElectron/Run2012D-PromptReco-v1/AOD	194480-208686	7.238 fb^{-1}
Total SingleElectron	**190645–208686**	**19.148 fb^{-1}**

Table 5.1: The datasets analyzed for this analysis.

this analysis. All of the samples, regardless of who produced them, are stored in a database called the Data Aggregation System (DAS) and organized by the “Dataset Name” field. The backgrounds were modeled by MC samples generated with MADGRAPH [151] and PYTHIA6 [150]. The signal MC samples were also generated by PYTHIA6. Tables 5.3 and 5.2 list all of the MC for the Higgs signal and SM background processes, respectively.

The $\mathrm{t\bar{t}}$, W + jets, and Z + jets SM background samples are generated using MADGRAPH v5.1.3.30 [151]. The $\mathrm{t\bar{t}}$ sample is inclusive, meaning that it includes all decay modes of the W boson coming from the top decay. The $\mathrm{W}+$jets and $\mathrm{Z}+$jets samples are also inclusive, but in this case it means that in addition to the leptonic decay of the boson there are any number of final state jets. The single top quark samples are modeled using the POWHEG 1.0 r138 [159, 160, 161] generator. The diboson processes use the PYTHIA v6.4.24 generator [150]. The cross sections for the $\mathrm{t\bar{t}}$ and single top quark processes are calculated at next-to-next-to-leading logarithmic (NNLL) accuracy [162] while the inclusive $\mathrm{W}+$jets and $\mathrm{Z}+$jets processes are calculated at next-to-next-to-leading order (NNLO) accuracy [163]. The diboson cross sections are calculated at next-to-leading order (NLO) accuracy [164].

The $\mathrm{H} \rightarrow \mathrm{WW}$ signal samples are generated with PYTHIA v6.4.24 [150], where one W is

Signal Processes			
Production & Decay Modes	Dataset Name	Cross Section [pb]	BR
ggH; $M_{\mathrm{H}} = 125\,\mathrm{GeV}$, $\mathrm{H} \rightarrow \mathrm{WW} \rightarrow \mathrm{l}\nu\mathrm{jj}$	/LQ-ggh125_BIG_SIM_ggH125_part1/aperloff-LQ-ggh125_AODSIM_Summer12_START53_V7E-768a14b04b0ac2af0d20e6783fbdb759/USER	19.27	0.0947
	/LQ-ggh125_BIG_GEN_part2/aperloff-LQ-ggh125_BIG_RECO_part2-33e909ff21293ad9fa8564de2959fe54/USER	19.27	0.0947
	/LQ-ggh125_BIG_GEN_part3/aperloff-LQ-ggh125_BIG_RECO_part3-33e909ff21293ad9fa8564de2959fe54/USER	19.27	0.0947
	/LQ-ggh125_Part6_SIM/goodell-LQ-qqh125_RECO_Part6-33e909ff21293ad9fa8564de2959fe54/USER	19.27	0.0947
	/LQ-ggh125_Part7_SIM/goodell-LQ-qqh125_RECO_Part7-33e909ff21293ad9fa8564de2959fe54/USER	19.27	0.0947
	/LQ-ggh125_Part8_GENSIM/goodell-LQ-ggh125_Part8_RECO-33e909ff21293ad9fa8564de2959fe54/USER	19.27	0.0947
qqH; $M_{\mathrm{H}} = 125\,\mathrm{GeV}$, $\mathrm{H} \rightarrow \mathrm{WW} \rightarrow \mathrm{l}\nu\mathrm{jj}$	/LQ-vbf125_GENSIM/ajkumar-LQ-qqh125_AODSIM_Summer12_START53_V7A-c8f8ed334db8a7d6f56c62266b1dfa5b/USER	1.578	0.0947
WH, ZH, ttH; $M_{\mathrm{H}} = 125\,\mathrm{GeV}$, $\mathrm{H} \rightarrow \mathrm{WW}$, inclusive	/WH_ZH_TTH_HToWW_M-125_8TeV-pythia6	1.249	0.215
Non signal Higgs Production			
WH, ZH, ttH; $M_{\mathrm{H}} = 125\,\mathrm{GeV}$, $\mathrm{H} \rightarrow \mathrm{ZZ}$, inclusive	/WH_ZH_TTH_HToZZ_M-125_8TeV-pythia6	1.249	0.0264
WH; $M_{\mathrm{H}} = 125\,\mathrm{GeV}$, $\mathrm{H} \rightarrow \mathrm{b\bar{b}}$, $\mathrm{W} \rightarrow \mathrm{l}\nu$	/WH_WToLNu_HToBB_M-125_8TeV-powheg-herwigpp	0.7046	0.1879
ttH; $M_{\mathrm{H}} = 125\,\mathrm{GeV}$, $\mathrm{H} \rightarrow \mathrm{b\bar{b}}$	/TTH_HToBB_M-125_8TeV-pythia6	0.1293	0.577

Table 5.2: List of signal datasets and cross sections. All of the centrally produced sample names are followed by /Summer12_DR53X-PU_S10_START53_V7A-v1/AODSIM.

	Background Processes	
Process	Dataset Name	Cross Section [pb]
$\mathrm{W}+\mathrm{jets}$	/WJetsToLNu_TuneZ2Star_8TeV-madgraph-tarball	37509
$\mathrm{t\bar{t}}$	/TTJets_MassiveBinDECAY_TuneZ2star_8TeV-madgraph-tauola	225.197
$\mathrm{Z}+\mathrm{jets}$	/DYJetsToLL_M-50_TuneZ2Star_8TeV-madgraph-tarball	3387.6
WW	/WW_TuneZ2star_8TeV_pythia6_tauola	54.838
WZ	/WZ_TuneZ2star_8TeV_pythia6_tauola	33.21
ZZ	/ZZ_TuneZ2star_8TeV_pythia6_tauola	17.654
$\mathrm{t} \rightarrow \mathrm{b}\ell\nu$ (s-channel)	/T_s-channel_TuneZ2star_8TeV-powheg-tauola	3.79
$\mathrm{t} \rightarrow \mathrm{b}\ell\nu$ (t-channel)	/T_t-channel_TuneZ2star_8TeV-powheg-tauola	56.4
$\mathrm{t} \rightarrow X$ (tW-channel)	/T_tW-channel-DR_TuneZ2star_8TeV-powheg-tauola	11.1
$\bar{\mathrm{t}} \rightarrow \mathrm{b}\ell\nu$ (s-channel)	/Tbar_s-channel_TuneZ2star_8TeV-powheg-tauola	1.76
$\bar{\mathrm{t}} \rightarrow \mathrm{b}\ell\nu$ (t-channel)	/Tbar_t-channel_TuneZ2star_8TeV-powheg-tauola	30.7
$\bar{\mathrm{t}} \rightarrow X$ (tW-channel)	/Tbar_tW-channel-DR_TuneZ2star_8TeV-powheg-tauola	11.1
QCD (e-channel)	See table 5.1 for a list of SingleElectron datasets	N/A
QCD (μ-channel)	See table 5.1 for a list of SingleMu datasets	N/A

Table 5.3: List of background MC datasets and cross sections used in the analysis. Every dataset name is followed by /Summer12_DR53X-PU_S10_START53_V7A-v1/AODSIM. In addition to v1, this analysis also uses v2 of the $\mathrm{W}+\mathrm{jets}$ sample.

required to decay leptonically while the other is required to decay hadronically. The cross sections for the Higgs production are calculated at NNLL QCD and NLO EW accuracies. The calculations for gluon-gluon fusion and VBF production cross sections use the complex-pole-scheme (CPS) while the associated production cross section are calculated with the zero-width-approximation (ZWA) [165]. These samples were privately produced because the centrally produces samples did not include enough events and had large statistical fluctuations.

5.1.3 Multijet-QCD Background

It is well known that the QCD process is difficult to model to the desired level of accuracy. Additionally, the event selection in this analysis requires two isolated jets and an isolated lepton, which vastly reduces the number of QCD MC events that pass the selection criteria. Although the probability to mis-reconstruct a jet as a lepton is fairly low, the production cross section for the multijet process is extremely high and thus cannot be ignored. When using the MC samples we are left with a statistically limited sample that is almost useless for describing this background.

Rather than relying on MC for the QCD background sample, a data-driven sample was created by using the same trigger requirements as the data, but removing the isolation requirement for the lepton and inverting the lepton particle flow isolation cut during selection. The main idea of the method is to utilize differences in lepton identification properties that separate prompt, isolated leptons from W and Z decays, also known as “real leptons,” from non-prompt, non-isolated leptons, also known as “fake leptons.” The normal signal selection requires an isolated lepton, without other particles around it, to limit this sort of “fake lepton,” but this is exactly the type of property we want to select for when forming a QCD sample from data. This process provides a completely orthogonal sample of QCD events from data that won’t, and shouldn’t be used for signal extraction. Since we make use of the entire 2012 dataset[1], we end up with statistically rich samples containing lots of mis-identified leptons.

A complete description of the event selection will be discussed in section 5.2, but here I will just talk about the isolation requirements. The loosest lepton PF isolation requirement used to

[1] The QCD events are scaled slightly to account for failed jobs (missing luminosity) during processing.

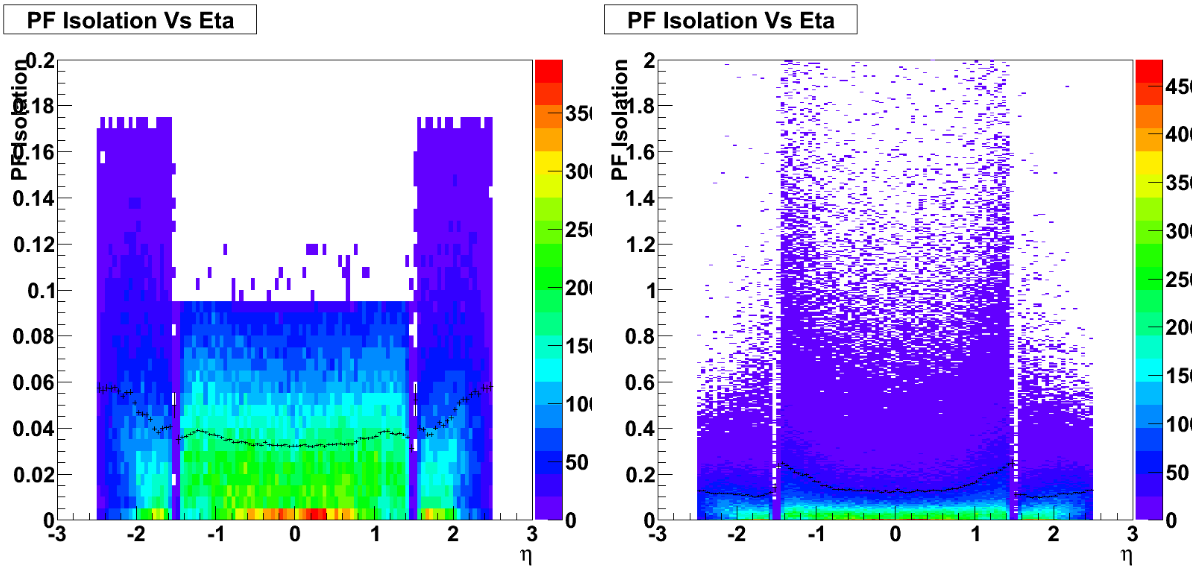
determine the signal region is Isolation$_{\text{PF}} < 0.2$, which is used to veto on “loose” or questionable leptons. The assumption is that any lepton with Isolation$_{\text{PF}} > 0.2$ is a mis-reconstructed lepton coming from QCD. For electrons we must also turn off the MVA-based identification requirements as they are stringent enough that they won’t allow for any fake leptons to pass our selection. As mentioned before, the electrons must still pass the “HLT_Ele27_WP80_v*” electron trigger used for the data containing our signal. On the other hand, the muon trigger is changed to be “HLT_Mu24_eta2p1_v*” to remove the isolation requirement that was included in the trigger used to select for the signal.

In order to gain greater separation from the signal selection to ensure as little non-QCD contamination as possible, we actually use a minimum isolation requirement of Isolation$_{\text{PF}} > 0.3$. We also put an upper limit on the PF isolation value to keep the sample from having a bias towards high nPV values. For electrons the upper limit was 0.7 and for muons it was 2.0. The 1σ systematic uncertainty bands for electrons (muons) were selected to be $0.2 <$ Isolation$_{\text{PF}} < 0.3$ on the low side and >0.7 (2.0) for the high side Fig. 5.2 shows the pf isolation values contained in the electron multijet and data samples as a function of η.

5.2 Event & Object Selection

As described in section 4, CMS provides to every analysis a list of reconstructed objects (i.e. jets, electrons, etc.) which may be used. However, these reconstruction algorithms are intentionally generic so that the objects they return are applicable to a wide array of physics analyses. Specific groups within CMS called physics object groups (POGs) are responsible for developing object quality criteria which must be implemented by each analysis to prevent fake or poorly reconstructed objects. This section will discuss the object selection criteria used to identify vertices, electrons, muons, jets, and $\vec{\not{E}}_{\text{T}}$, which all meet or exceed the object requirements as set by the relevant POGs. Only events which contain objects of the right quality and multiplicities will be used in the analysis.

Like most analyses, this one selects for a single good quality primary vertex, although the presence of additional vertices (pileup) does not disqualify the event. The primary vertex must



(a)

(b)

Figure 5.2: The PF isolation for the electron channel as a function of η (left) with and (right) without the lepton isolation and electron MVA-based identification requirements.

pass certain additional quality criteria. There must be at least four degrees of freedom used to find the vertex, the absolute value of the z-coordinate of the vertex must be less than 24 cm, the absolute value of the ρ-coordinate (cylindrical coordinate system) must be less than 2.0 cm, and the vertex must not be identified as a fake vertex. These criteria are summarized in table 5.4.

Cut	Value
N_{DOF}	$\geqslant$4
$\lvert z \rvert$	$\leqslant$24 cm
$\lvert \rho \rvert$	$\leqslant$2.0 cm

Table 5.4: The primary vertex selection requirements for this analysis.

As mentioned before, this analysis selects for the presence of one lepton, either an electron or muon, at least two jets, and some amount of $\vec{\not{E}}_{\mathrm{T}}$. In practical terms this means that we select for one tight electron (muon) as defined in section 4.3 (4.4) and veto the event if there are any additional tight or loose electrons and muons (muons and electrons). Some additional cuts beyond those of the identification requirements are imposed to cut out some of the background events while maximizing the number of signal events we could use for the multivariate analysis techniques. The additional p_{T} and η requirements as specified in the same sections are also applied. For the tight electrons this meant raising the p_{T} requirement from 27 GeV to 30 GeV, which avoids using events right on the trigger turn on threshold while only removing ∼5% of signal events, as seen in fig. 5.3. Because muon reconstruction and identification in CMS is very good, we only raised the p_{T} requirement to 25 GeV from 24 GeV.

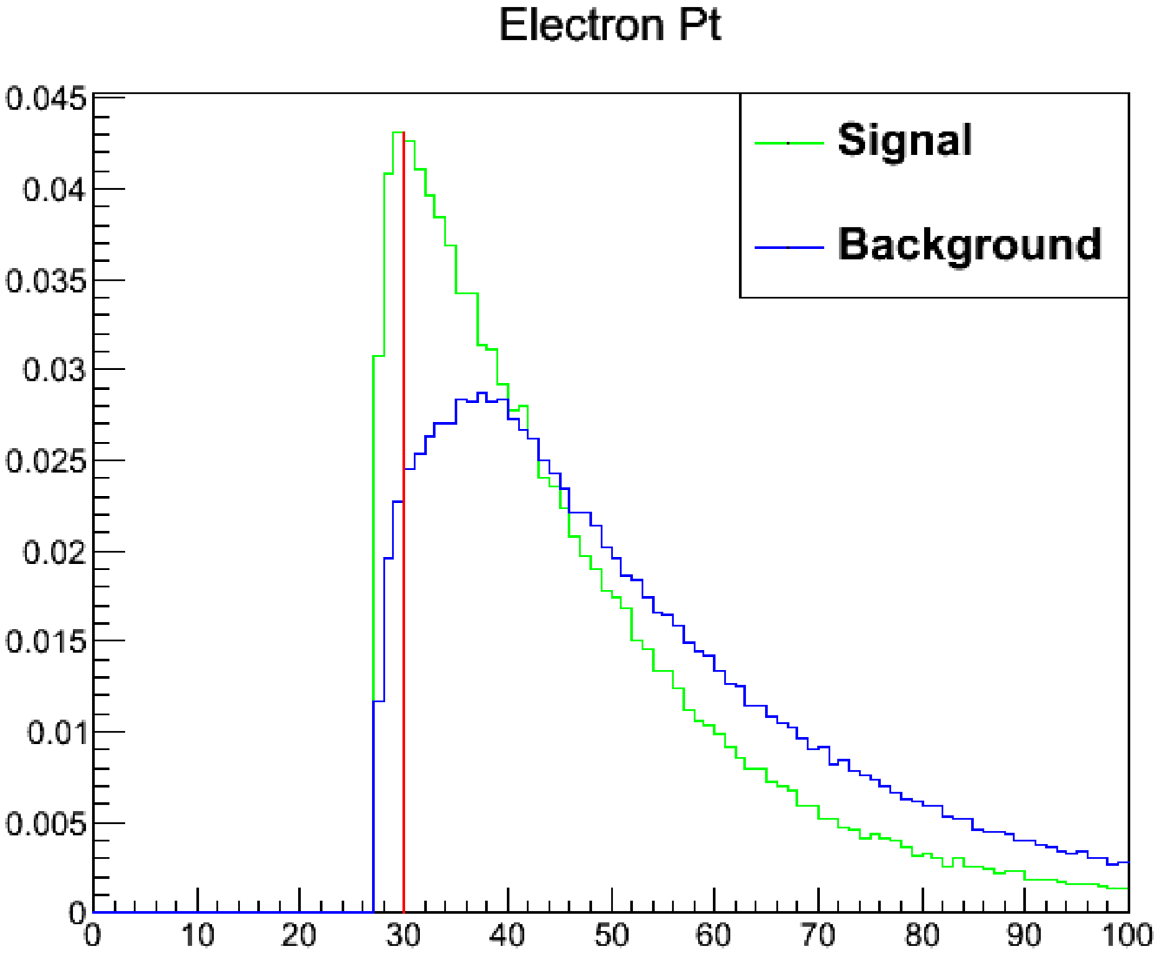


Figure 5.3: Histograms of the electron p_{T} distribution where the gluon-gluon fusion signal is in green and the $\mathrm{W}+$jets background is in blue. The histograms are normalized to unit area. The red line show the cut on electron p_{T} where 5% of the signal is lost.

Beyond the lepton requirements, this analysis selected for any number of jets as long as they pass the selection criteria found in section 4.5. As the hadronic W decay will have at least two jets, that is the minimum number of jets needed to make it into the signal region, but we do not veto on additional jets which might come from ISR or FSR. The requirement of the leading jet having a $p_{\mathrm{T}} > 30$ GeV was implemented to reduce the impact of the multijet background while minimally impacting the signal. Besides the logical splitting of events based on lepton flavor, we also split events into three categories based on the number of jets in the event; exactly two jets, exactly three jets, and four or more jets. As stated in section 4.7 we also require at least 25 GeV of $\vec{\not{E}}_{\mathrm{T}}$.

Given that our signal has only one hadronic W boson, we don't expect the $\mathrm{W} \to \mathrm{b\overline{b}}$ branching fraction to contribute much to our signal. However, we also want to remove as many $\mathrm{t\bar{t}}$ or single top events as possible, which are commonly associated with $\mathrm{b}$quarks. Thus we decided to veto events with b-tagged jets in order to reduce our backgrounds as much as possible. An additional reason to do this is to keep the orthogonality between this analysis and another CMS analysis which was looking at the VH production channel where $\mathrm{H} \to \mathrm{b\overline{b}}$. That analysis uses the same final state as this one, but requires two b-tagged jets [166]. To prevent overlap, we only ever considered events with one or fewer b-tagged jets and then we separate the events into two categories based on the number of b-tags. The zero b-tag events are used for signal extraction while the one b-tag events, which have a much larger impact from $\mathrm{t\bar{t}}$ and a higher $\mathrm{H} \to \mathrm{b\overline{b}}$ signal yield, are used for validation purposes and to check the volunteer signal contribution.

5.3 MC Corrections

Although a significant amount of work and time goes into making sure the MC simulation properly models the data, there can still exist discrepancies between the observed data and simulation Often this occurs because the exact data taking conditions are not known in advance, like the pileup conditions that will exist. Another reason the MC might not exactly mimic the data is that even state of the are generators are limited in their precision; much of the physics of hadronization is still unknown and hard physics processes can often only be computed up to NLO precision. Data, on the other hand, contains all hadronization effects and all orders of precision. I have already discussed some object specific corrections like the jet energy corrections, jet energy resolution, and $\vec{\not{E}}_{\mathrm{T}}$corrections in sections 4.5 and 4.7. For other discrepancies it is often necessary to reweight the full event rather than a specific object.

Broadly speaking these corrections can be separated into two categories: those which are common to all CMS analyses and those which are specific to this analysis. The first category includes the b-tagging CSV discriminant weights and top quark p_{T} spectrum weights for the $\mathrm{t\bar{t}}$ simulation while the second category includes the weights for our multijet sample. These event weights are applied after selecting for the events as they do not change the object kinematics.

5.3.1 Pileup Reweighting

Pileup is an important quantity as it can affect the reconstruction efficiency and even the observed kinematics of all the objects used in this analysis. Up to this point it has been described as additional proton-proton interactions within an event, besides the interaction that produced the physics objects we are interested in studying. There are several other properties of pileup which are worth noting. I have so far either referred to pileup in a general sense or as relating to additional objects (tracks or energy) which might be found in the same bunch crossing as the event under study. In reality there are two different categories of pileup. There is indeed the pileup which comes from additional proton-proton interactions within the same bunch crossing, known as “in-time” pileup. There is also energy from pileup added to objects because it was left in the sub-detectors from bunch crossings before or after the current one. This is known as “out-of-time” pileup and comes about because the integration window of the sub-detectors can be larger than 25 ns. An additional property is somewhat obvious in that the true number of proton-proton interaction within an event, μ, is related to the instantaneous luminosity, which can vary within any given data taking period and even within a luminosity section (LS). As a benchmark, the average number of proton-proton interactions per bunch crossing in 2012 was 21 [9].

The MC samples used in CMS are usually generated before the data is taken and are thus created with an assumption of what the pileup conditions will look like in data. A broad distribution of μ values, the number of min-bias pileup events overlaid on the hard scatter event, is generally

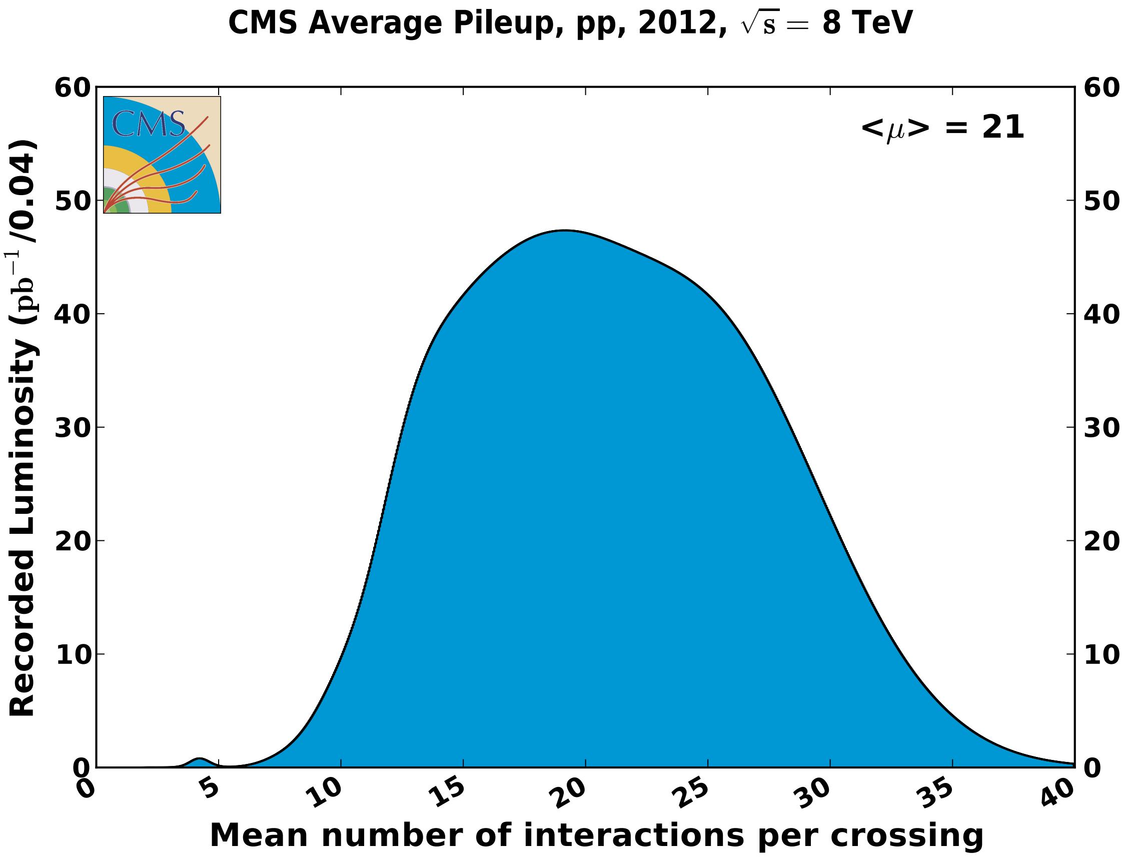


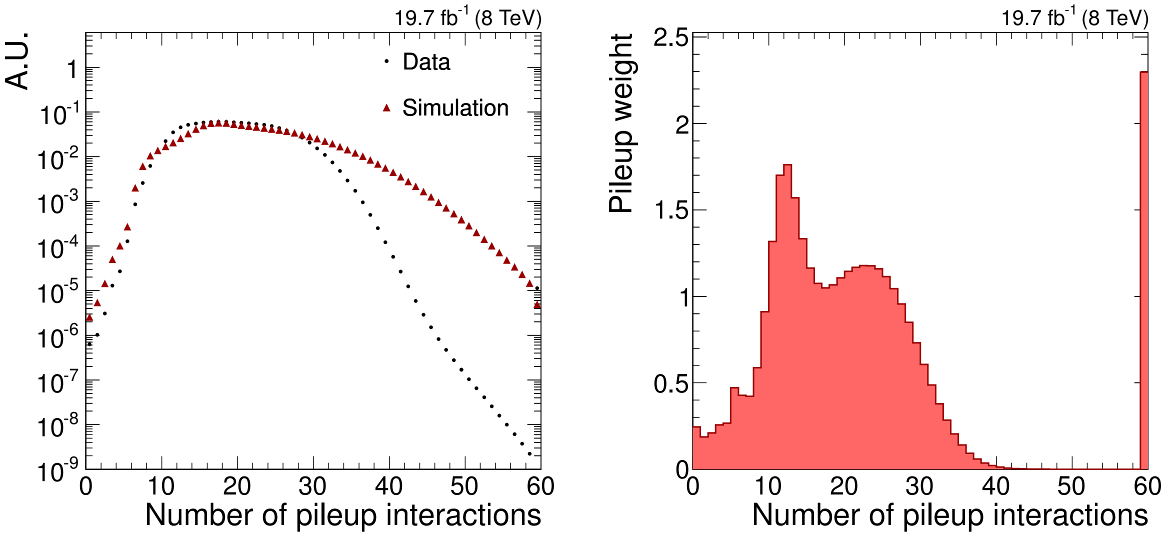
Figure 5.4: The mean number of interactions per bunch crossing in 2012. The min-bias cross section used for the calculation is 80 mb.

chosen so as to cover all pileup conditions which might be experienced over the course of a data taking period. Somewhat unsurprisingly the anticipated μ distribution rarely matches the one one observed in the data and thus the MC must be reweighted such that the μ distributions match [167]. To generate a histogram for the average number of interactions per bunch crossing coming from data we make use of the approved pileupCalc tool provided by CMS. This tool takes as input the total inelastic cross section $\sigma_{\text{inelastic}} = 69.3\,\text{mb}$[2], a file in JSON format with every run number and luminosity section matched to a given average instantaneous luminosity and integrated luminosity for that given LS, and another JSON formatted file with the run numbers and LS used in the given analysis[3]. All of the MC samples used contain the same μ distribution scenario denoted by the

[2] This is the CMS approved best fit value, not the theoretical value.

[3] This analysis uses the full 2012 "golden" JSON file called

"S10" notation in the dataset name. The per event weights as a function of μ are created by dividing the normalized distribution from data by the normalized MC based distribution. The weights are then applied to each MC event by looking up the weight for the mean number of pileup interactions used to generate that specific event [168]. The distributions of pileup interactions in MC and data a well as the corresponding pileup weights can be seen in fig. 5.5. Unfortunately, because the weights are not at unity, the statistical precision of the MC samples is reduced. Fig. 5.6 shows the data to MC comparison of the N_{PV} distribution before and after the pileup reweighting scheme has been applied.



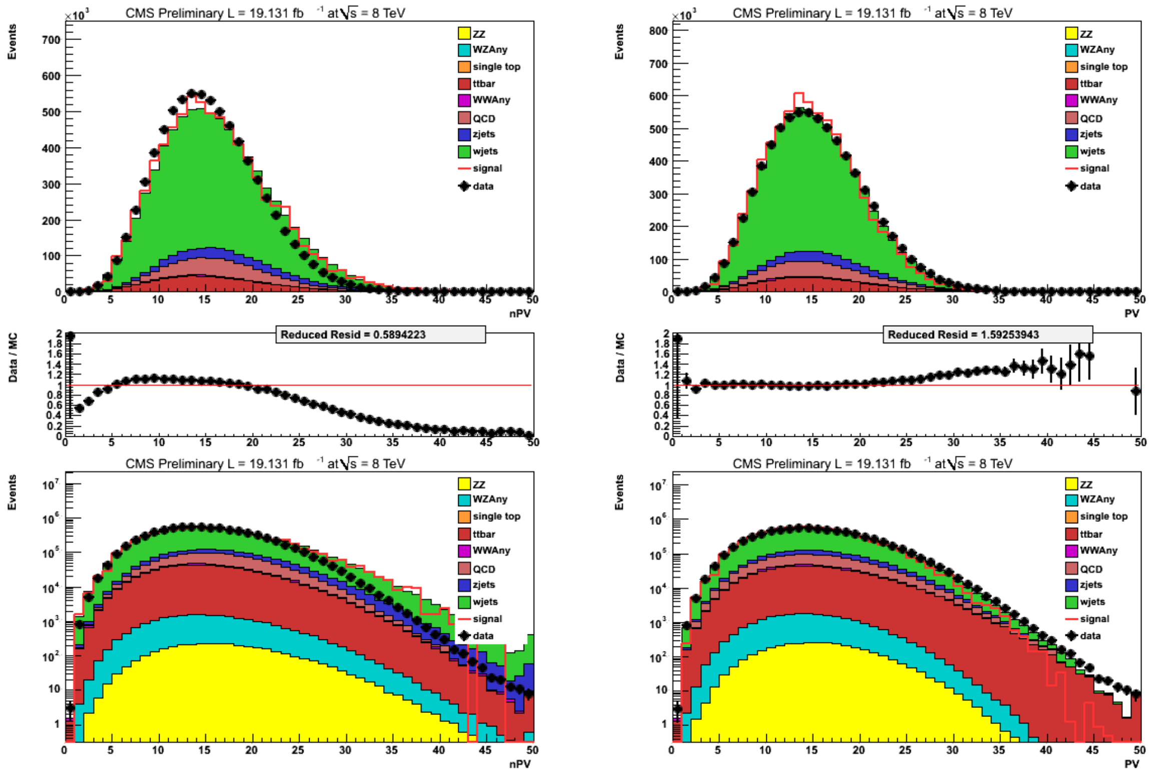
(a)

(b)

Figure 5.5: (a) Distributions of the number of pileup interactions in data and in simulation. (b) The derived pileup weights as a function of the number of interactions.

While this methodology is sufficient for the simulated backgrounds, it does not work for the data-driven multijet background. As can be seen from figs. 5.7a and 5.7c, the distributions for the number of primary vertices between data and the QCD samples do not match, indicating some bias due to the selection. Since the QCD sample does not contain the truth level number of pileup

Cert_190456-208686_8TeV_PromptReco_Collisions12_JSON.txt.



(a)

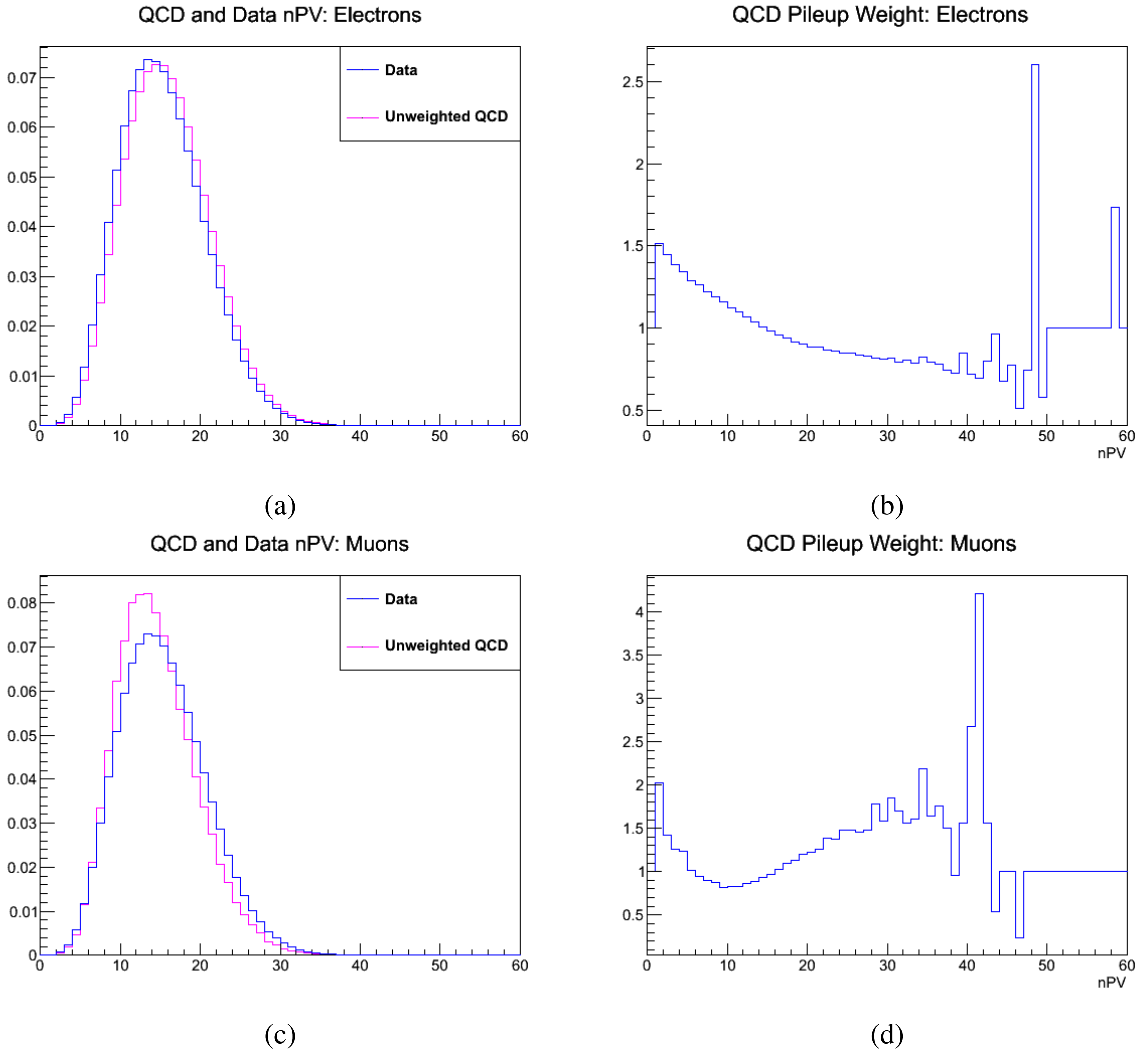
(b)

Figure 5.6: Comparison of the number of primary vertices (N_{PV}) in data and in MC (a) before the the pileup weights are applied and (b) after the weights are applied. These distributions correspond to the 19 fb^{-1} collected during the 2012 data taking period and include both the electron and muon categories.

interactions, this is data after all, it would be improper to look up pileup weights using the same weight distribution as for simulation. Instead, a new set of weights is derived using the number of primary vertices for data in the signal region and anti-isolated region, assuming that the vertex finding efficiency is the same in both regions and only the selection of the lepton changes. These weights can be seen in figs. 5.7b and 5.7d and are applied in the same manner as before.

5.3.2 CSV Reweighting

Section 4.6 introduced the criterion for tagging a jet as being produced by a bquark and the use of the Combined Secondary Vertex (CSV) discriminant. The derivation of this discriminant is described in [137, 138]. This analysis relies heavily on the identification of bjets to veto the $t\bar{t}$



(a)

(b)

(c)

(d)

Figure 5.7: Distribution of the number of primary vertices for data and QCD (a,c) and the associated weights (b,c). Figures (a) and (b) show the electron channel while figures (c) and (d) show the muon channel.

background, so it is absolutely crucial that it behave the same in both data and MC and accurately describe the rate of observing a bjet. [169] notes that the tagging efficiency in data is not the same as that in MC, so a correction to the CSV discriminant must be made. The corrections described there both correct the rate of observing a jet in MC with a CSV value above a given threshold as well as the general shape of the CSV distribution. If at the end of the procedure the shape of the data and MC distributions agree, then they will also properly assess the rate of events passing a given CSV threshold.

The method is based on calculating a scale factor for both heavy and light flavor quarks which is parameterized by the CSV value, jet p_{T}, and, in the case of light flavor quarks, jet η. We first retrieve the truth level jet flavor in order to determine the correct category: bjet, cjet, or light flavor (anything else). The cjets are given a flat scale factor of 1, meaning that there is no need to correct the CSV value for this flavor. The bjet scale factors are divided into five p_{T} bins of $p_{\mathrm{T}} < 40\,\mathrm{GeV}$, $40\,\mathrm{GeV} < p_{\mathrm{T}} < 60\,\mathrm{GeV}$, $60\,\mathrm{GeV} < p_{\mathrm{T}} < 100\,\mathrm{GeV}$, $100\,\mathrm{GeV} < p_{\mathrm{T}} < 160\,\mathrm{GeV}$, and $p_{\mathrm{T}} > 160\,\mathrm{GeV}$. The light flavor scale factors are divided into only three p_{T} bins of $p_{\mathrm{T}} < 40\,\mathrm{GeV}$, $40\,\mathrm{GeV} < p_{\mathrm{T}} < 60\,\mathrm{GeV}$, and $p_{\mathrm{T}} > 60\,\mathrm{GeV}$, but are also divided into three eta bins of $|\eta| < 0.8$, $0.8 \leqslant |\eta| < 1.6$, and $1.6 \leqslant |\eta| < 2.4$. An individual scale factor is retrieved for each jet, which is then combined as in equation 5.1 in order to create an event weight.

$$SF_{\mathrm{total}} = \prod_{i}^{N_{\mathrm{jets}}} SF_{\mathrm{jet}_i} = SF_{\mathrm{jet}_1} \cdot SF_{\mathrm{jet}_2} \cdot \ldots \tag{5.1}$$

The CSV value for each jet is unchanged, but the event is weighted by SF_{total}.

5.3.3 $\mathrm{t\bar{t}}$ Reweighting

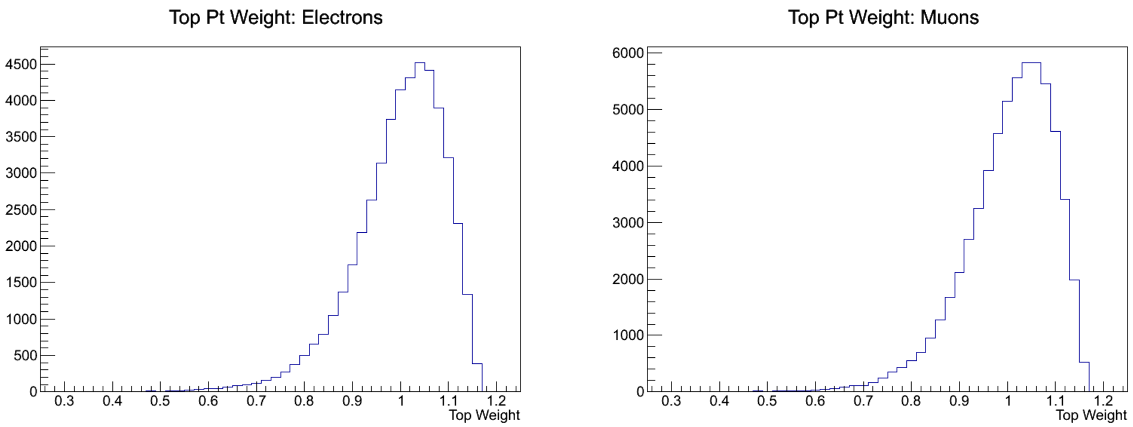
Differential top-quark-pair analyses have shown that the shape of the p_{T} spectrum for top quarks is softer in data than predicted by simulation [170, 171]. Although it has been shown that NNLO predictions show reasonable agreement [172], this analysis must correct for the discrepancy in the $\mathrm{t\bar{t}}$ simulation. Events are reweighted based on the p_{T} of the generator level $\mathrm{t}$and $\mathrm{\bar{t}}$in only the $\mathrm{t\bar{t}}$ simulation. The weight w_{TopPt} is calculated as:

$$w_{\mathrm{TopPt}} = \sqrt{SF_{\mathrm{t}} \cdot SF_{\mathrm{\bar{t}}}} \tag{5.2}$$

$$SF\left(p_{\mathrm{T}}^{\mathrm{gen}}\right) = \exp\left(a + b p_{\mathrm{T}}^{\mathrm{gen}}\right) \tag{5.3}$$

with $a = 0.156$ and $b = -0.00141$. Fig. 5.8 shows the distribution of weights for electron and muon events separately. The bulk of the weights are centered around 1, indicating that no correction is necessary, with a long low side tail, indicating that a good fraction of events require the top

p_{T} to be scaled down. Some events do require that the top p_{T} be increased.



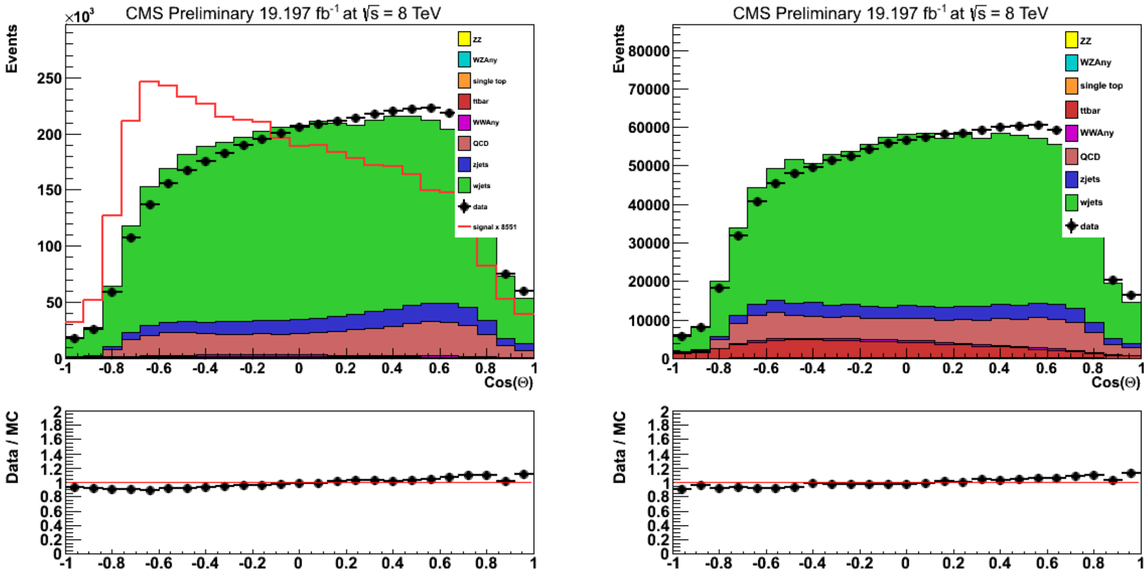
(a) (b)

Figure 5.8: Top p_{T} weight distributions for (a) electrons events and (b) muon events.

5.3.4 cos(theta$_l$) Reweighting

A linear trend in the data to MC comparison of the $\cos{(\theta_1)}$variable was discovered, indicating a mis-modeling problem in the simulation. $\cos{(\theta_1)}$is one of the angular variables involved in the WW system and is the cosine of the angle between the daughter lepton and the WW decay plane, which corresponds to $\cos{(\theta_2)}$ in fig. 5.22. Fig. 5.9a shows this trend in the two jet bin, though the trend is the same in the other jet bins.

We correct for the trend in the $\mathrm{W+}$jets MC as this is the biggest background and correcting it will improve the overall agreement. We create the corrections in the one b-tag control region shown in fig. 5.9b so as to not bias our backgrounds in the signal region. It's clear from fig. 5.9 that the trend in the one b-tag region is the same as the trend in the signal region. Although the regions are similar, the $\mathrm{t\bar{t}}$ MC plays a much larger role in the control region because it contains two real $\mathrm{b}$jets. Therefore we subtract the expected $\mathrm{t\bar{t}}$ yield from the data before creating the weights. The new weights shown in fig. 5.10a are combined multiplicatively with the pileup and CSV weights for the $\mathrm{W+}$jets sample. The corrected distribution is shown in fig. 5.10b where it is clear that the



(a) (b)

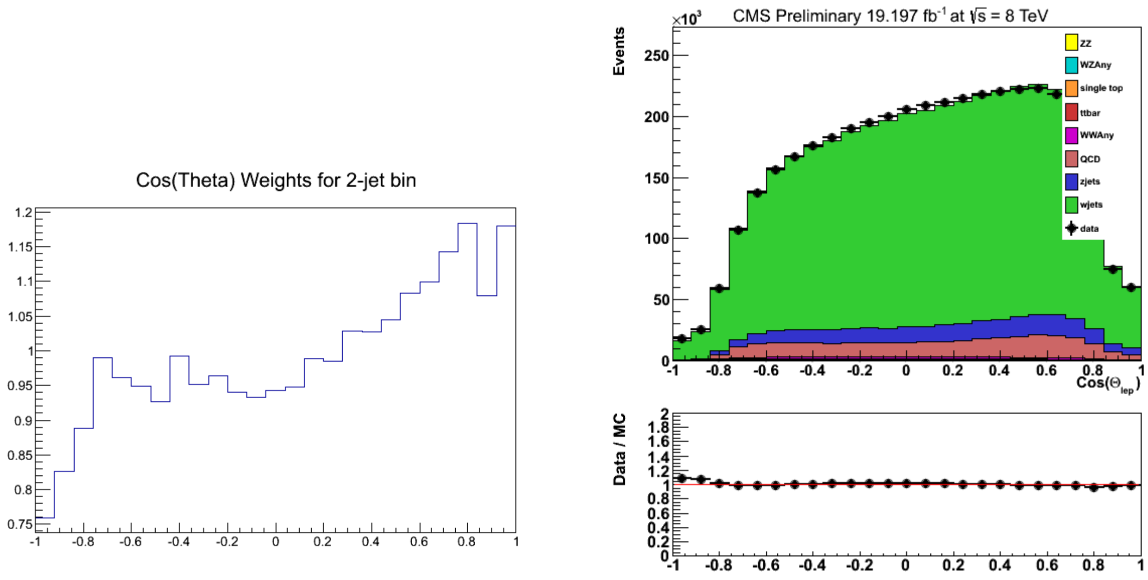
Figure 5.9: Distribution of $\cos{(\theta_1)}$for data and MC in the two jet bin for (a) the signal region and (b) the one b-tag region. The top of each figure shows the data and MC expectations while the bottom shows their ratio with a clear linear trend.

trend has been removed.

5.3.5 QCD Reweighting

As stated in section 5.1.3, the QCD sample is obtained by selecting on anti-isolated leptons, as opposed to the isolated signal selection. Although these regions are similar kinematically, the ratio of the number of events in the signal region to the number of events in the anti-isolated region changes significantly as a function of η. This effect was first noticed in MC, which was used to check the anti-isolation procedure despite its limited statistics in the low $\hat{p}_{\mathrm{T}}$-binned samples. Fig. 5.11 shows the suspect ratio as a function of η in the different QCD $\hat{p}_{\mathrm{T}}$ bins. The effect seems to be particularly large in the endcap regions ($|\eta| > 1.3$). A weighting procedure is necessary to make sure that the expected yield as a function of η for the data-driven QCD sample is correct when used in the signal region.

To derive the weights we use the one jet control region separated into 13 (12) bins of lepton



(a) (b)

Figure 5.10: (a) Weights created in the one b-tag control region used to correct the $\cos\left(\theta_{\mathrm{l}}\right)$mis-modeling. (b) $\cos\left(\theta_{\mathrm{l}}\right)$distribution in the signal region after applying the weights.

$|eta|$ for the electron (muon) channel. We want to find the scale factor S_{QCD} such that:

$$N^{\mathrm{QCD}}_{\mathrm{anti-isolated}}\left(\eta\right) S_{\mathrm{QCD}}\left(\eta\right) = N^{\mathrm{QCD}}_{\text{signal region}}\left(\eta\right), \tag{5.4}$$

where $N^{\mathrm{QCD}}_{\mathrm{anti-isolated}}$ and $N^{\mathrm{QCD}}_{\text{signal region}}$ represent the number of events in the anti-isolated and signal regions, respectively, given the same luminosity in both. In order to determine the scale factor needed to to modify the QCD contribution in each bin, we perform a fit to the data using the $\not{E}_{\mathrm{T}}$ distribution. The QCD and $\mathrm{W}+$jets contributions are allowed to float while the contributions from all of the other backgrounds are fixed to their SM expectations. The $\not{E}_{\mathrm{T}}$ distributions post-fitting as well as the χ^2/NDF for all of the fits are shown in fig. 5.12. The fit returns both the scale factor S_{QCD} as well as a scale factor for the $\mathrm{W}+$jets, $S_{\mathrm{W+jets}}$, which are shown in fig. 5.13. The shape of the weights follows very closely the shape of the ratio in MC from fig. 5.11, which is a very good indication that we are indeed correcting for the intended effect. The same procedure is performed for the muon events, yielding the weights shown in fig. 5.14. Note that the absolute

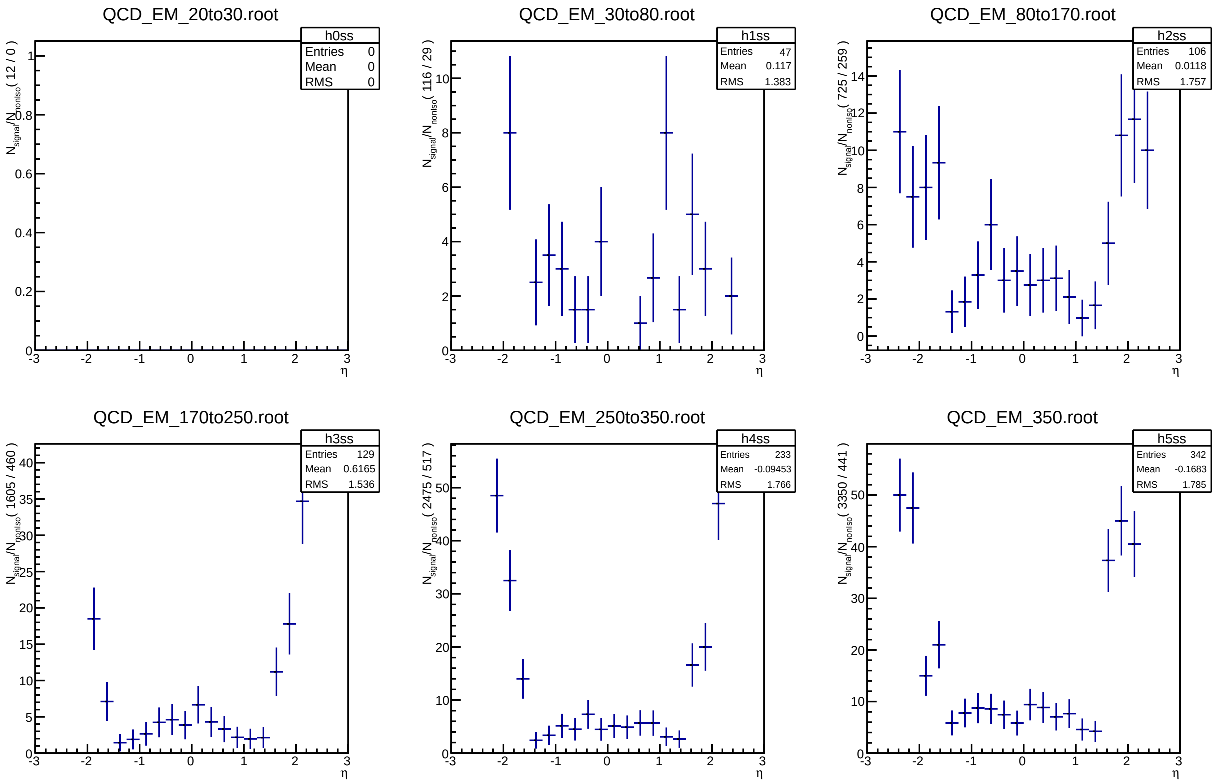


Figure 5.11: The ratio of the number of events in the signal region to the number of events in the anti-isolated region for six QCD $\hat{p}_{\mathrm{T}}$ bins. The total number of events in the two regions is shown in parentheses along the y-axis. The first plot is empty due to the low number of MC events which pass the selection criteria.

value of the scale factors is not what matters, only their relative values, as the sample will undergo an additional normalization in order to obtain the correct yield.

As an additional cross check, the same procedure was done to the ⩾2 jets bin to see if the distribution of weights was similar to that of the control region. From figs. 5.15 and 5.16 we see that the procedure, done on the signal region, does indeed return similar scale factors to those found in the one jet control region. This gives us high confidence that the scale factors from the one jet bin will correct the shape of the QCD distributions in the signal region.

The right plot of fig. 5.16 shows that the $\mathrm{W}+$jets normalization also needs to be measured as a fit to the ratio of measured events and expected event yields. This ratio has a value of 0.953 ± 0.008, which is not consistent with the 2.56% error on the theoretical cross section. To find the correct

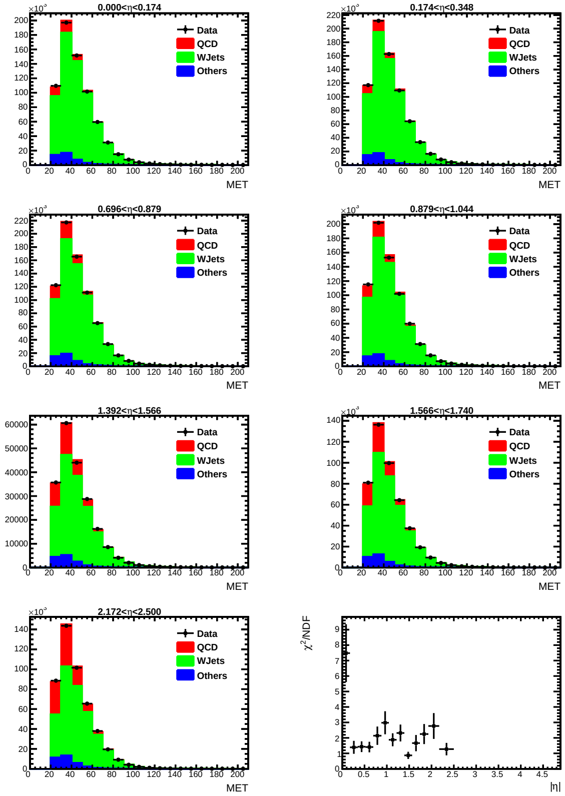


Figure 5.12: The $\not\!\!E_{\mathrm{T}}$ distributions used to derive the QCD weights in the 13 different bins of lepton $|\eta|$ after the fitting the QCD (red) and $\mathrm{W}+\mathrm{jets}$ (green) contributions to the data (black markers). The contribution from the other SM processes (blue) is held fixed to their SM expectation. The last pad in the plot show the χ^2/NDF of the fits.

$\mathrm{W}+\mathrm{jets}$ and QCD normalizations a two component fit to the $\not\!\!E_{\mathrm{T}}$ distribution of the data is used, allowing only the $\mathrm{W}+\mathrm{jets}$ and QCD fractions to float. The expected yields of the other SM backgrounds are held constant during the fit. A Gaussian constraint is imposed on the $\mathrm{W}+\mathrm{jets}$ scale factor because its theoretical cross section uncertainty is known. The derived scale factors are shown in table 5.5 and the $\not\!\!E_{\mathrm{T}}$ distribution for data and MC after the reweighting can be seen in fig. 5.17.

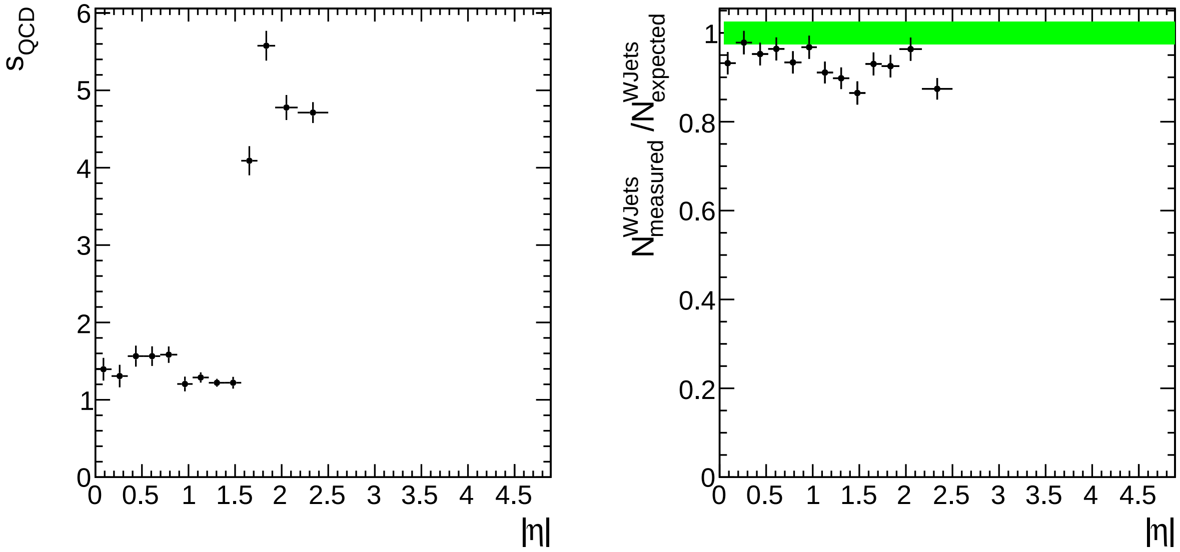


Figure 5.13: S_{QCD} (left) and $S_{\text{W+jets}}$ (right) scale factors as a function of lepton $|\eta|$ derived in the electron channel for the one jet bin. The green band indicates the uncertainty on the $\text{W} + \text{jets}$ expectation due to the theoretical uncertainty in the SM cross section.

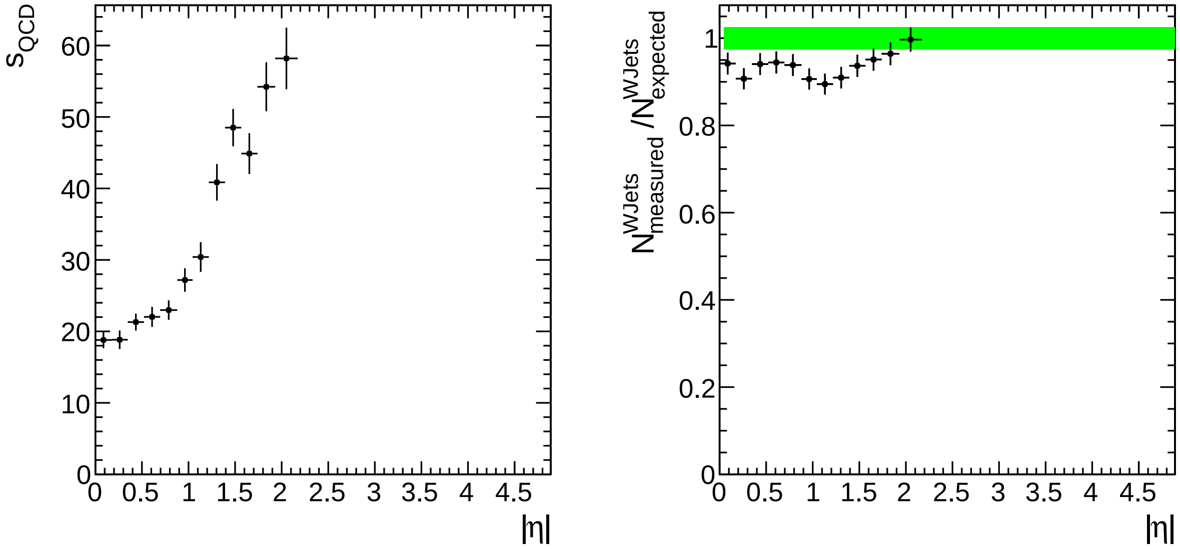


Figure 5.14: S_{QCD} (left) and $S_{\text{W+jets}}$ (right) scale factors as a function of lepton $|\eta|$ derived in the muon channel for the one jet bin. The green band indicates the uncertainty on the $\text{W} + \text{jets}$ expectation due to the theoretical uncertainty in the SM cross section.

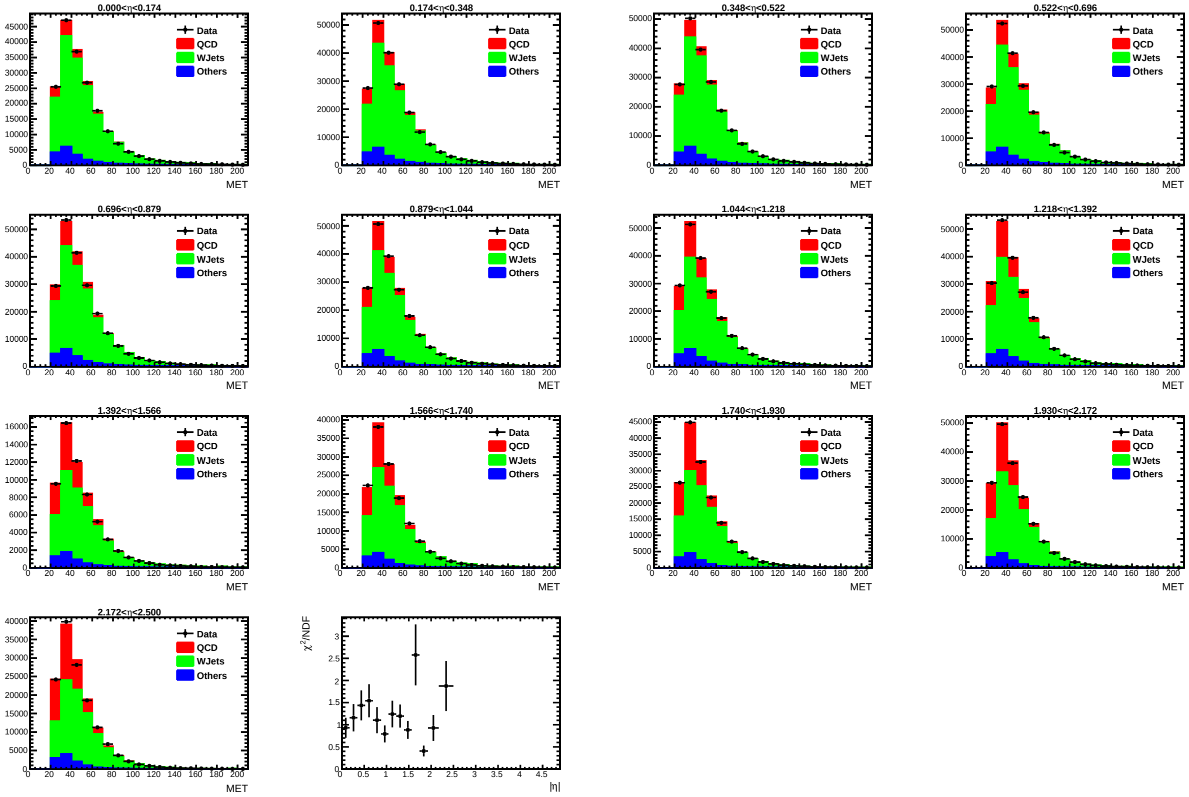


Figure 5.15: The $\not\!\!E_{\mathrm{T}}$ distributions in the $\geqslant$2 jet bin used to derive the QCD weights in the 13 different bins of lepton $|\eta|$ after the fitting the QCD (red) and W + jets (green) contributions to the data (black markers). The contribution from the other SM processes (blue) is held fixed to their SM expectation. The last pad in the plot show the χ^2/NDF of the fits.

Lepton Category	W + jets SF	QCD SF
Electron	1.04515 ± 0.00509474	0.248858 ± 0.0131115
Muon	0.969517 ± 0.00442517	0.145418 ± 0.00669525

Table 5.5: W + jets and QCD scale factors as derived from a two component fit to the $\not\!\!E_{\mathrm{T}}$ distribution.

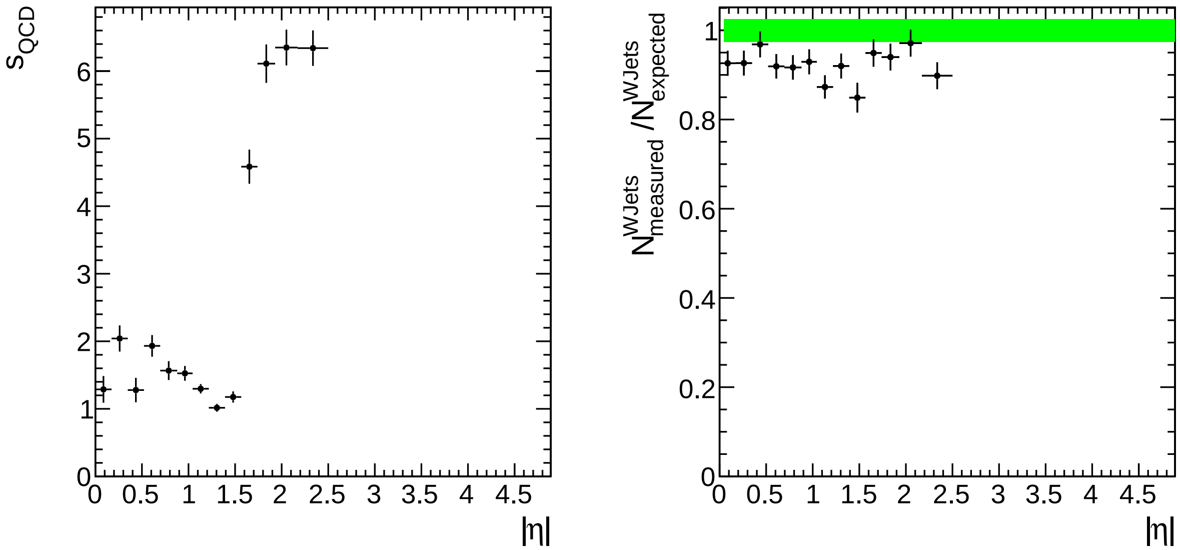


Figure 5.16: S_{QCD} (left) and $S_{\mathrm{W+jets}}$ (right) scale factors as a function of lepton $|\eta|$ derived in the electron channel for the $\geqslant$2 jet bin. The green band indicates the uncertainty on the W + jets expectation due to the theoretical uncertainty in the SM cross section.

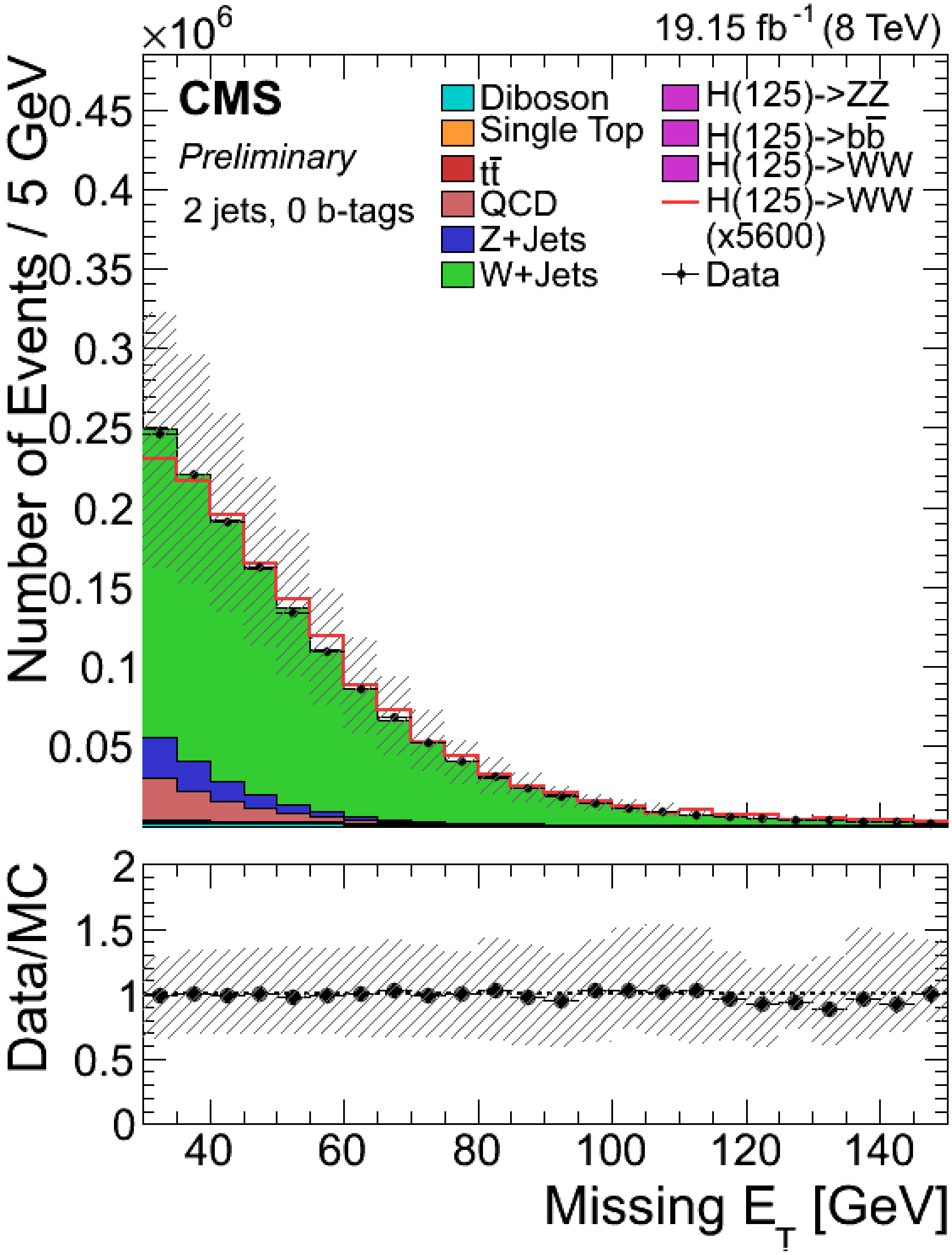


Figure 5.17: The $\not\!\!E_{\mathrm{T}}$ distribution for the two jets, electron channel showing good agreement between data and MC after the QCD reweighting.

5.4 Data-to-MC Comparisons & Yields

After applying all of the object and event selections, object corrections, and event weights we can now look at the expected yields for the simulated signals and backgrounds. Table 5.6 shows the event yields for our signal selection separated by jet bin, but combining the electron and muon categories. Table 5.7, on the other hand, shows the percentage yields where the numbers from table 5.6 have been normalized to the sum of the events in background and signal sections. In both tables, Higgs events where the Higgs boson does not decay to two W bosons are referred to as 'volunteer signal'. This is in contrast to true $\mathrm{H} \to \mathrm{WW}$ events, which we sometimes refer to as 'true signal'. Both of these categories are normalized to the $\mathrm{H} \to \mathrm{WW}$ yields in order to be able to compare the volunteer signal contamination to the true signal.

It is clear from these tables that the dominant background for all jet bins is $\mathrm{W} + \mathrm{jets}$. Its expected yield is by far much larger than all of the other backgrounds. From table 5.7 one can also see that the sum of the volunteer signal is at most 7% of the $\mathrm{H} \to \mathrm{WW}$ signal, which means the b-tag cut is keeping the non-$\mathrm{H} \to \mathrm{WW}$ contamination to a minimum. If the b-tag cut was not used the $\mathrm{t\bar{t}}$ background would become much more significant, even becoming the dominant background in the $\geqslant$4 jet bin. Additionally, the volunteer signal would become as high as 87% of the $\mathrm{H} \to \mathrm{WW}$ signal, which means that there would be a lot of overlap between this analysis and other CMS analyses. Comparison plots for several kinematic variables can be found in appendix C.

5.5 Multivariate Analysis

One of the problems of past analyses, such as cut-and-count experiments, is that they ignore the additional information that comes from using the many correlated bins of a shape analysis. By doing a cut-and-count experiment across many bins an analysis is able to gain in discrimination power. That being said, it would be wasteful and suboptimal to use a single discriminating kinematic distribution, which means the discrimination power of the unused variables is missed. This analysis uses the output of a boosted decision tree (BDT) classifier as the template used for limit setting, choosing to combine the discrimination power of several kinematic variables. This

Process	**2 Jets**	**3 Jets**	**$\geqslant$4 Jets**
Diboson	46495.97 ± 78.55	15049.18 ± 44.70	4150.48 ± 23.47
$\mathrm{W}+$jets	3446003.06 ± 6434.30	756463.35 ± 3008.63	189815.29 ± 1515.40
$\mathrm{Z}+$jets	270460.62 ± 822.24	69061.73 ± 415.90	19829.24 ± 222.71
$\mathrm{t\bar{t}}$	22452.06 ± 142.85	27902.44 ± 160.86	31218.33 ± 170.54
Single t	16587.13 ± 84.31	7193.89 ± 59.29	3068.60 ± 40.25
Multijet	275465.33 ± 952.52	74168.89 ± 504.39	22109.53 ± 282.94
Total Background	4077464.17 ± 6558.75	949839.48 ± 3083.93	270191.47 ± 1567.59
ggH, $\mathrm{H} \rightarrow \mathrm{WW}$ M_{H} =125 GeV	552.09 ± 1.92	211.15 ± 1.19	79.51 ± 0.73
qqH, $\mathrm{H} \rightarrow \mathrm{WW}$ M_{H} =125 GeV	106.60 ± 0.56	52.66 ± 0.39	17.51 ± 0.23
WH_ZH_TTH, $\mathrm{H} \rightarrow \mathrm{WW}$ M_{H} =125 GeV	136.20 ± 2.22	84.35 ± 1.75	42.32 ± 1.22
Total $\mathrm{H} \rightarrow \mathrm{WW}$	794.89 ± 2.99	348.16 ± 2.15	139.34 ± 1.44
WH_ZH_TTH, $\mathrm{H} \rightarrow \mathrm{ZZ}$ M_{H} =125 GeV	10.30 ± 0.17	5.30 ± 0.12	2.35 ± 0.08
WH, $\mathrm{H} \rightarrow \mathrm{b\bar{b}}$ M_{H} =125 GeV	45.34 ± 0.40	14.22 ± 0.23	3.86 ± 0.12
ttH, $\mathrm{H} \rightarrow \mathrm{b\bar{b}}$ M_{H} =125 GeV	0.59 ± 0.03	1.33 ± 0.05	3.77 ± 0.09
Total Volunteer Signal	56.23 ± 0.44	20.85 ± 0.26	9.98 ± 0.17
Signal $_{\mathrm{H} \rightarrow \mathrm{WW}}$/Bkg	0.000195	0.000367	0.000516
Signal $_{\mathrm{H} \rightarrow \mathrm{WW}}/\sqrt{\text{Bkg}}$	0.394	0.357	0.268
Data	4057594	953513	272713

Table 5.6: Expected yields for both the electron and muon categories when normalized to the SM cross sections and collected luminosity. The table is broken up into three sections; the top section contains all of the background processes, the middle section shows the $\mathrm{H} \rightarrow \mathrm{WW}$ contributions, and the bottom section shows the other Higgs processes that could mimic our final state, but do not originate from a $\mathrm{H} \rightarrow \mathrm{WW}$ process. This table contains the yields for the zero b-tag category. Only statistical uncertainties are shown.

Process	**2 Jets**	**3 Jets**	**$\geqslant$4 Jets**
Diboson	0.011	0.016	0.015
W + jets	0.845	0.796	0.703
Z + jets	0.066	0.073	0.073
$\mathrm{t\bar{t}}$	0.006	0.029	0.116
Single t	0.004	0.008	0.011
Multijet	0.068	0.078	0.082
Total Background	1.000	1.000	1.000
ggH, H → WW M_{H} =125 GeV	0.695	0.606	0.571
qqH, H → WW M_{H} =125 GeV	0.134	0.151	0.126
WH_ZH_TTH, H → WW M_{H} =125 GeV	0.171	0.242	0.304
Total H → WW	1.000	1.000	1.000
WH_ZH_TTH, H → ZZ M_{H} =125 GeV	0.013	0.015	0.017
WH, H → $\mathrm{b\bar{b}}$ M_{H} =125 GeV	0.057	0.041	0.028
ttH, H → $\mathrm{b\bar{b}}$ M_{H} =125 GeV	0.001	0.004	0.027
Total Volunteer/Total H → WW	0.071	0.060	0.072

Table 5.7: Expected percent yields for both the electron and muon categories separated by jet bin. The background samples are normalized by the total background, while the H → WW and volunteer signal samples are normalized by the H → WW total. The Dominant background in all jet bins, W + jets, is highlighted in green. This table contains the percent yields for the zero b-tag category.

type of multivariate analysis (MVA) is useful in quantifying the separation of the signal samples (H → WW) from the background samples.

5.5.1 Boosted Decision Tree

Multivariate techniques are used to model the dependence of one or more target variables on a set of input variables. Boosted decision trees are a more robust alternative to artificial neural networks and were first introduced to the high energy physics (HEP) community by the MiniBooNE collaboration [173]. This machine learning (ML) technique has since been used countless times throughout the HEP community. This analysis makes use of the BDT algorithm implemented in the ROOT TMVA package [174]. The key ingredient here is the boosting technique, which helps to mitigate the problem of “overtraining,” which is common to ML algorithms, and increases the overall performance of the algorithm [175]. The issue with overtraining is that the output of the ML algorithm becomes overly dependent on the multivariate inputs. In other words this means that a small change in the input variable $x \rightarrow x + \delta x$ can cause a large change in the output of the algorithm $f\left(x+\delta x\right) - f\left(x\right) \gg \epsilon$. The ML algorithm may be picking up on minute changes in the simulation or statistical fluctuations, both of which are not true features of the target classification. While these jumps may seem to indicate a higher amount of discrimination power in the training sample, they are not indicative of the underlying physics being modeled and must be suppressed. The BDT algorithm train many weak decision trees, which are then combined using the namesake "boosting algorithm." This algorithm "boosts" the events that are misclassified in the previous tree so that each successive generation of tree contains fewer misclassified events. Some of the benefits of boosting are that weak or less discriminating input variables will have a reduced impact and that many input variables can be included to improve the overall classification performance. This section will describe the general process of training of a BDT classifier while the subsequent sections will explain how the BDTs were trained for this analysis.

A decision tree is a binary tree structure made up of nodes which are meant to provide higher purity samples of signal and background at each subsequent layer of the tree. A set of input variables is chosen by the analyzer before the start of the training sequence. The higher purity is achieved by placing a cut on the single input variable which will achieve the best separation (highest purity) at any given node. This can be thought of as each node creating a boundary $t\left(x\right)$ in multi-dimensional space and estimating the likelihood ratio $\frac{\mathcal{L}(t|S)}{\mathcal{L}(t|B)}$ in a small portion of that space. The input variables should be chosen for their discrimination power, which can be quantified at each stage of the tree as $S/\left(S+B\right)$. The signal purity P, on the other hand, is defined at every node as the number of signal events divided by the total number of events in the sample, both signal and background. For purity P, a cut value can be chosen to minimize the Gini Index $Gini = G_{\text{left}} + G_{\text{right}}$, where $G = P\left(1-P\right)$ and G_{side} is calculated one both sides of the cut. This cut will then define the population of signal and background for two nodes in the next layer. A perfect cut which completely separates signal from background will achieve $Gini = G_{\text{left}} = G_{\text{right}} = 0$ while any impurity will mean $Gini \neq 0$. The Gini Index will reach a maximum when the samples are fully mixed. For training purposed, the starting node will have the same mixture of signal and background as the training sample, while each successive cut level will reduce the impurity as shown in figs. 5.18 and 5.19. It can be seen from both figures that a single variable may be used to define a cut at more than one node in the tree, as in the jet2dRLep variable in fig. 5.19. It is also possible that a variable will not be used at all. The granularity of the cuts tested by the algorithm is a user specified parameter, which must be wisely chosen to allow for flexibility in the cut space, but not so granular as to adversely increase the computing time. The stopping point of the algorithm can be based on the minimum number of training events remaining in each node, the maximum number of layers from the root node, a requirement on the purity, or a combination of two or more of those criteria. At this point the multi-dimensional space is split into many regions, which are classified as either signal or background depending upon the purity level of the final node. A purity > 0.5 is classified as signal and a purity < 0.5 is classified as background [174].

Training many independent decision trees without boosting will not prevent overtraining as each tree would have a different misclassification rate. The boosting algorithm solves this by combining many decision trees (a "forest" of trees) to minimize the ensemble misclassification rate. This analysis makes use of the adaptive boost (AdaBoost) procedure, which weights higher in subsequent trees events which are mis-classified in the current tree [176]. The event weights are initialized to 1, but change after the first tree. Nevertheless, the weights in each tree are always normalized such that the sum of the weights remains constant. The events in each new tree are weighted by multiplying the previous event weights by a boost weight α common to the tree. α is defined as:

$$\alpha = \frac{1 - err}{err}, \tag{5.5}$$

where err is the mis-classification rate of the previous tree. The weighted sum of the tree outputs is given by:

$$y_{\text{Boost}}\left(\mathbf{x}\right) = \frac{1}{M}\sum_{m=0}^{M} \ln\left(\alpha_m\right) h_m\left(\mathbf{x}\right), \tag{5.6}$$

where there are m trees, **x** are the input variables, and $h\left(\mathbf{x}\right) \in \{-1, 1\}$ is the single event classifier

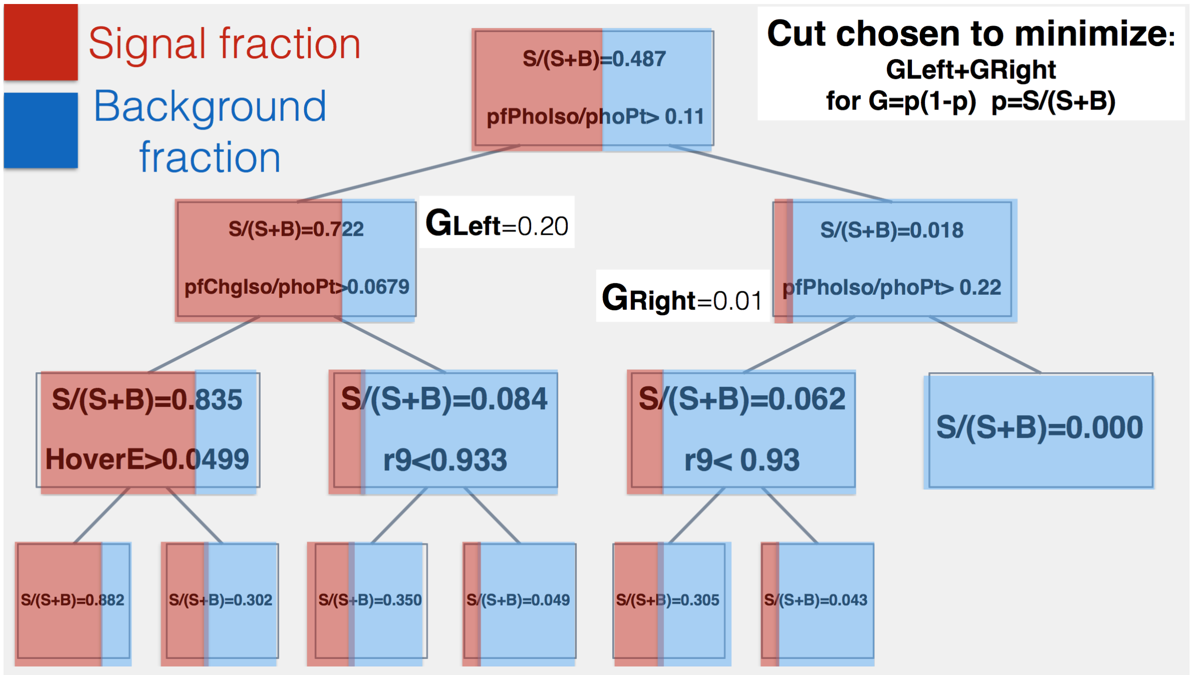


Figure 5.18: Example BDT classifier tree showing the cut optimization procedure to separate signal and background events. The colors within each node represent the purity p. The root note contains equal amounts of signal and background, but after the first layer the right-most node contains almost pure background while the left-most node contains 70% signal. The base of the tree provides a node with more than 80% signal purity.

indicating if the event is signal, $h(\mathbf{x}) = 1$, or background, $h(\mathbf{x}) = -1$. The resulting discriminant on the event at the end of the training, $y_{\text{Boost}}(\mathbf{x})$, is a number in the range $[1, -1]$, where 1 is most signal-like and -1 is most background-like.

The AdaBoost procedure is ideal for use with shallow trees with two or three levels each, leaving a relatively large population of events in each of the final nodes. These are also known as weak classifiers and provide little discrimination power on their own. The benefit to using these is that they are much less prone to overtraining, but they can be grouped together, through the boosting procedure, to provide good discrimination power. Had the trees been allowed to reach a state where a single event was left in a node, this would imply that there was a cut sequence that would lead to perfect signal versus background classification, a practical impossibility. It is therefore important that the analyzer keep this in mind when specifying the stopping hyper-

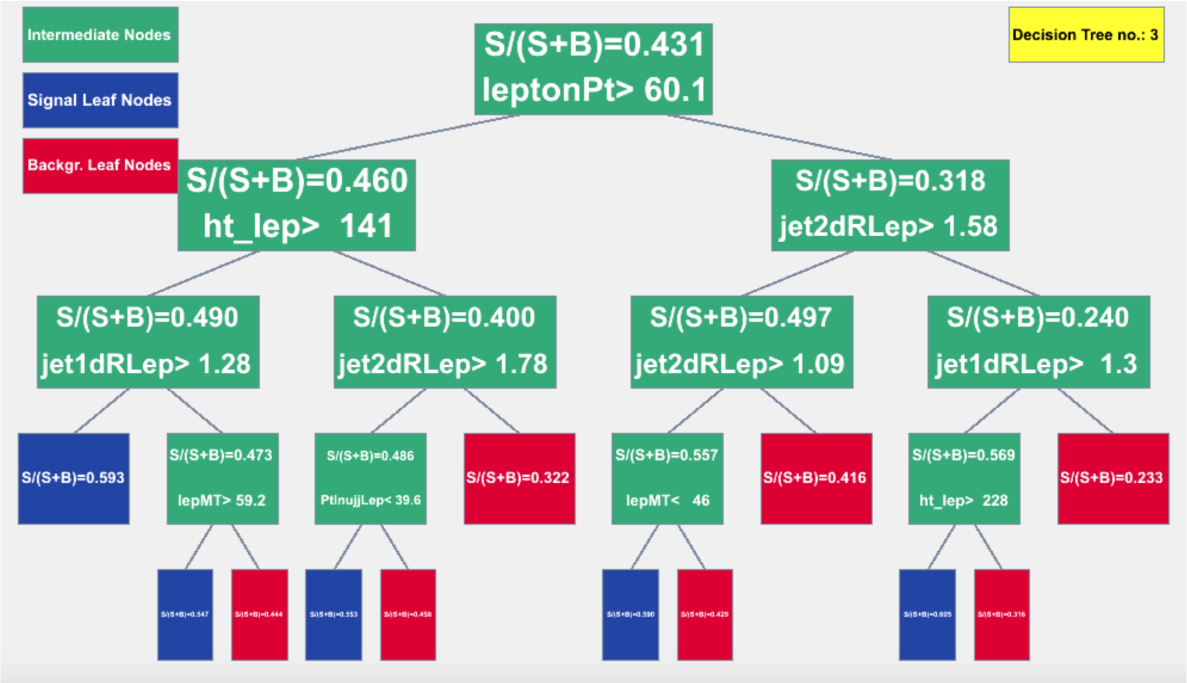


Figure 5.19: An example decision tree used by this analysis. This tree will be combined with a forest of other trees using the boosting algorithm. The bottom nodes are defined as being more signal or background like based on the majority population left in the node.

parameters. One of the other hyper-parameters specific to the AdaBoost procedure is the boost weight exponent, where $\alpha \rightarrow \alpha^{\beta}$. By changing β one can slow down the learning rate, allowing for a larger number of boost iterations. The list of tunable hyper-parameters is as follows:

- NTrees: The number of trees in the forest.
- nEventsMin: The minimum number of events allowed in a node after the splitting.
- MaxDepth: The maximum number of levels in the tree aside from the root node.
- BoostType: The boosting method to use. This analysis used the adaptive boost (AdaBoost) method, but other options are available.
- AdaBoostBeta: The exponent of the AdaBoost weight. This analysis used $\beta = 0.5$.
- SeparationType: While this analysis used the Gini Index there are other choices for measuring the separation of signal and background.

- nCuts: The number of steps available for a single variable when determining the cut value. Increasing this number leads to finer granularity, the benefit of which was not seen by this analysis. We chose to use a step size of 20.

- PruneMethod: It is possible to prune away some branches to increase performance. This was unnecessary for this analysis as it used a boost procedure which limited the size of the tree.

- NodePurityLimit: This parameter determines at which purity ($P >$ NodePurityLimit) the final node is considered a signal node. This analysis used a value of 0.5.

As an additional way to check for overtraining, one can reserve a set of events to use as a testing sample to check the efficacy of the classifier response. The amount of signal and background to split off is tunable, but this analysis used half of the events for training and the other half for testing. When comparing the training and testing distributions the Kolomogrov-Smirnoff test[4] is used to determine their compatibility. For this analysis a separate BDT is trained for each jet category and is individually optimized based on the chosen input variables and the hyper-parameters of the training algorithm. Section 5.5.2 will discuss the selection of the potential kinematic variables while sections 5.5.3 and 5.5.4 will discuss the optimization of the inputs and parameters, respectively, for the individual trainings.

5.5.2 Kinematic Variable Selection

While it may be tempting to use the 4-vectors of the final state objects as inputs to the BDT, shallow networks, like the ones used here, are not very good at learning the intricacies necessary to discriminate physics processes based on simple inputs. Conversely, networks can be subject to sever overtraining if too many high level variables are used as input. Instead, the user must choose a select set of input distributions to use, preferably ones that already has some separation between the signal(s) and background(s). It is also a good idea to provide the BDT only the dominant signal and background to train on, so as to develop a classifier with the maximal amount

[4] The value returned by this test is the probability that the two distributions originated from the same probability distribution.

of separation power. Given table 5.6, we used the normalized $\mathrm{W+jets}$ background and $\mathrm{H} \rightarrow \mathrm{WW}$ signal MC as input samples. A list of variables with possible separation powers was then created. Each variable's separation power was quantified using the two figures of merit (FOM) listed in equation 5.7 and 5.8, where i denotes the bin number in the distribution.

$$FOM1 = \sum_{i=1}^{nBins} (\text{signal} - \text{background})^2 \tag{5.7}$$

$$FOM2 = \sum_{i=1}^{nBins} \frac{(\text{signal} - \text{background})^2}{(\text{signal} + \text{background})^2} \tag{5.8}$$

Fig. 5.20 shows several of these distributions with their associated figures of merit.

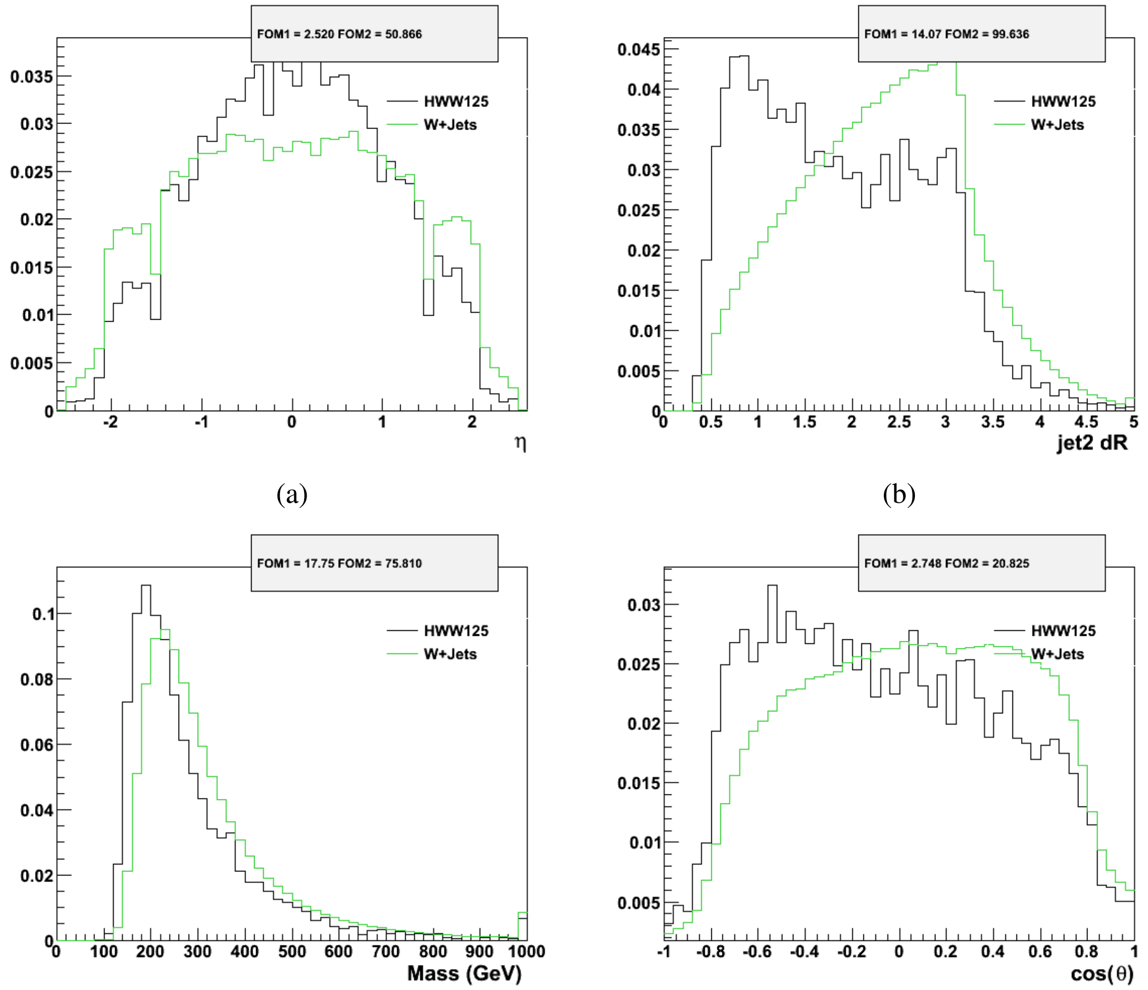
An additional method for determining useful variables is to calculate the cumulative distribution function (CDF) for each of the variables being tested. The CDF histograms are built bin-by-bin from the nominal distributions of each variable. The contents of any given bin in the CDF are equal to the sum of that bin and all of the previous bins in the nominal distribution as shown in equation 5.9.

$$C_i^{\text{CDF}} = \sum_{\text{bin}=0}^{i} C_i^{\text{nominal}} \tag{5.9}$$

Fig. 5.21 shows the PDF for the lepton p_{T} variable and the corresponding CDF. We are looking for variables which maximize the difference between the signal and background curves. To this end we also calculate FOM1 and FOM2 for the CDF distributions.

The variables were then ranked based on these four FOM values, separately for each jet bin, and only the top 20 variables in each jet bin were chosen to move on. The final ranking was achieved by averaging the rankings of the four methods, the purpose of which was to remove any undue method bias. Section 5.5.3 will discuss the specific variables chosen for each jet bin. However, a list of all variables considered can be found in table 5.8. The lepton and jet 4-vectors are denoted with l and j, respectively, with the jets sorted in order of descending p_{T}. Some of the variable definitions are listed below:

- ${p_{\mathrm{T}}}_x$: The p_{T} of object x in the event.



(a)

(b)

(c)

(d)

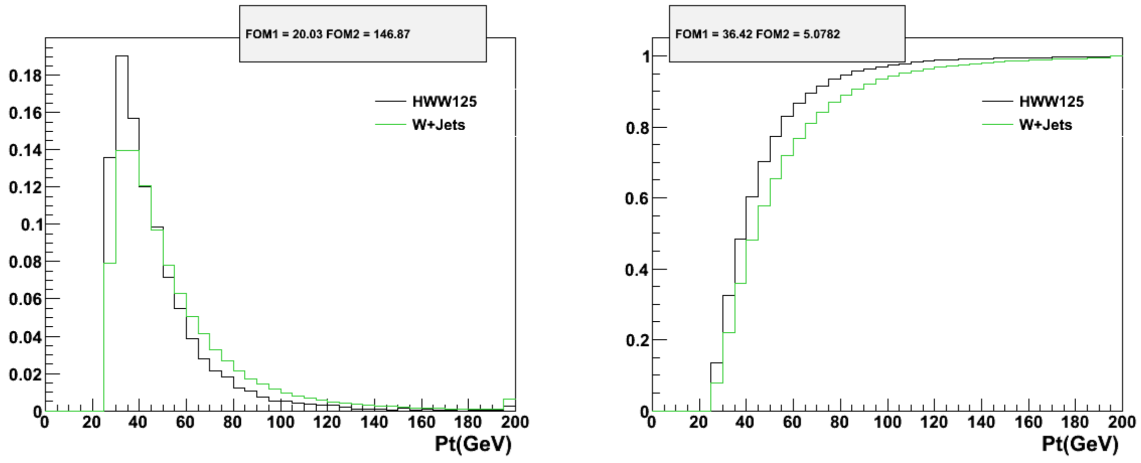
Figure 5.20: Example distributions used to examine possible input variables to the BDT trainings. The ggH (black) and W + jets (green) samples are both unit normalized. The two FOMs calculated are shown, but since these are normalized distributions the resulting numbers would be quite small. Thus the FOM have been multiplied by 10^5 for ease of reading. The four distributions are (a) lepton η, (b) $\Delta R\,(\mathrm{l}, \text{jet2})$, (c) $M_{\mathrm{l}\nu\mathrm{jj}}$, and (d) $\cos(\theta_{\mathrm{l}})$.

- η_x: The η of object x in the event.

- φ_x: The φ of object x in the event.

- M_{t}: The transverse component of the mass of the leptonically decaying W boson.

- $\Delta R\,(\mathrm{l}, \mathrm{j}_1)$: The distance in R between the lepton and the highest p_{T} jet ($\Delta R = \sqrt{\Delta\Phi^2 + \Delta\eta^2}$).



(a)

(b)

Figure 5.21: (a) Nominal and (b) CDF distributions for the lepton p_{T} variable. The signal is shown in black and the background in green.

- HT: The scalar sum of the lepton p_{T} and the E_{T} of all jets in the event
- $M_{\mathrm{l}\nu\mathrm{jj}}$: The 4-body mass defined as the mass of the vector sum of the lepton, $\vec{\not{E}}_{\mathrm{T}}$, and the two highest p_{T} jets in the event.
- $p_{\mathrm{T}\,\mathrm{l}\nu\mathrm{jj}}$: The p_{T} of the 4-body system created by summing the 4-vectors of the lepton, $\vec{\not{E}}_{\mathrm{T}}$, and the two highest p_{T} jets in the event.
- $\Delta R\left(\mathrm{l},\mathrm{jj}\right)$: The ΔR, as defined above, between the lepton and the di-jet system formed by the two highest p_{T} jets in the event.
- $\Delta\varphi\left(\vec{\not{E}}_{\mathrm{T}},\mathrm{j}\right)$: The $\Delta\varphi$ between the leading jet and the $\vec{\not{E}}_{\mathrm{T}}$.
- $\Delta\varphi\left(\mathrm{j},\mathrm{j}\right)$: The $\Delta\varphi$ between the two highest p_{T} jets in the event.
- $\Delta\varphi_{\mathrm{min}}\left(\mathrm{l},\mathrm{j}\right)$: The smallest $\Delta\varphi$ between the lepton and any of the jets in the event.
- $\Delta\eta\left(\mathrm{j},\mathrm{j}\right)$: The η between the two highest p_{T} jets in the event.
- $\mathrm{CSV}_{disc.}\left(\mathrm{j}_i\right)$: The CSV discriminant value for jet i.

$\cos\left(\Delta\Phi_{\mathrm{WH}}\right)$	$\cos\left(\Delta\Phi_{\mathrm{WW}}\right)$
$\cos\left(\theta_{\mathrm{j}}\right)$	$\cos\left(\theta_{\mathrm{l}}\right)$
$\cos\left(\theta_{\mathrm{WH}}\right)$	$\Delta\eta\left(\mathrm{j},\mathrm{j}\right)$
$\Delta\varphi\left(\mathrm{j},\mathrm{j}\right)$	$\Delta\varphi\left(\vec{\not\!\!E}_{\mathrm{T}},\mathrm{j}\right)$
$\Delta\varphi\left(\vec{\not\!\!E}_{\mathrm{T}},\mathrm{l}\right)$	$\Delta R\left(\mathrm{l},\mathrm{jj}\right)$
$\eta\left(\mathrm{j},\mathrm{j}\right)$	HT
$\mathrm{CSV}_{disc.}\left(\mathrm{j}_1\right)$	$\mathrm{CSV}_{disc.}\left(\mathrm{j}_2\right)$
$\Delta R\left(\mathrm{l},\mathrm{j}_1\right)$	$\Delta R\left(\mathrm{l},\mathrm{j}_2\right)$
$\Delta R\left(\mathrm{l},\mathrm{j}_3\right)$	$\Delta R\left(\mathrm{l},\mathrm{j}_4\right)$
η_{j_1}	η_{j_2}
φ_{j_1}	φ_{j_2}
$p_{\mathrm{T}}\left(\mathrm{j}_1\right)$	$p_{\mathrm{T}}\left(\mathrm{j}_2\right)$
$\mathrm{Charge}_{\mathrm{l}}$	η_{l}
$\mathrm{Charge}_{\mathrm{l}} \times \eta_{\mathrm{l}}$	p_{Tl}
$\not\!\!E_{\mathrm{T}}$	$\varphi_{\not E_{\mathrm{T}}}$
$\Delta\varphi_{\mathrm{min}}\left(\mathrm{l},\mathrm{j}\right)$	$\Delta\varphi_{\mathrm{min}}\left(\vec{\not\!\!E}_{\mathrm{T}},\mathrm{j}\right)$
M_{jj}	$M_{\mathrm{l}\nu\mathrm{jj}}$
M_{t}	$\mathrm{nBTag}_{\mathrm{CSV_m}}$
$\mathrm{n_j}$	$\mathrm{n}_{\mathrm{j_{low}}}$
$\mathrm{n_{PV}}$	$p_{\mathrm{Tl}\nu\mathrm{jj}}$
$\sum E_{\mathrm{Tj}}$	p_{Tjj}

Table 5.8: A list of all of the kinematic variables considered for inclusion in the BDT training. The variables are listed in no particular order and thus placement within the table is unimportant.

In additions to the definitions listed above, the are a whole host of angular variables which in part specify the kinematics of the Higgs and W boson decays. When defining these variables, much of which was done in [177], it will help to refer to the diagram in fig. 5.22. To start with, a kinematic fit is used to calculate the longitudinal momentum of the neutrino, p_z, and to also constrain the invariant mass of the leptonic W, $M_{\mathrm{l}\nu}$. Because the angle definitions are agnostic as to the type of particle decaying, the initiating particle will be referred to as particle X. The angles are as follows:

- θ^* is the polar angle between the collision axis z and the X decay axis z' as defined in the rest frame of particle X.
- Φ_1 is the azimuthal angle between the zz' plane and the decay plane of the hadronic W.
- Φ is the angle between the decay planes of the WW system in the rest frame of particle X.
- θ_1 is the angle between the z' axis the highest p_{T} jet, defined from 0 to π.
- θ_2 is the angle between the z' axis and the lepton.

Rather than using the bare angles, we have chosen to use the cosine of the angles and have thus named them as:

- $\Phi \rightarrow \cos\left(\Delta\Phi_{\mathrm{WW}}\right)$
- $\Phi_1 \rightarrow \cos\left(\Delta\Phi_{\mathrm{WH}}\right)$
- $\theta_1 \rightarrow \cos\left(\theta_{\mathrm{j}}\right)$
- $\theta_2 \rightarrow \cos\left(\theta_{\mathrm{l}}\right)$
- $\theta^* \rightarrow \cos\left(\theta_{\mathrm{WH}}\right)$

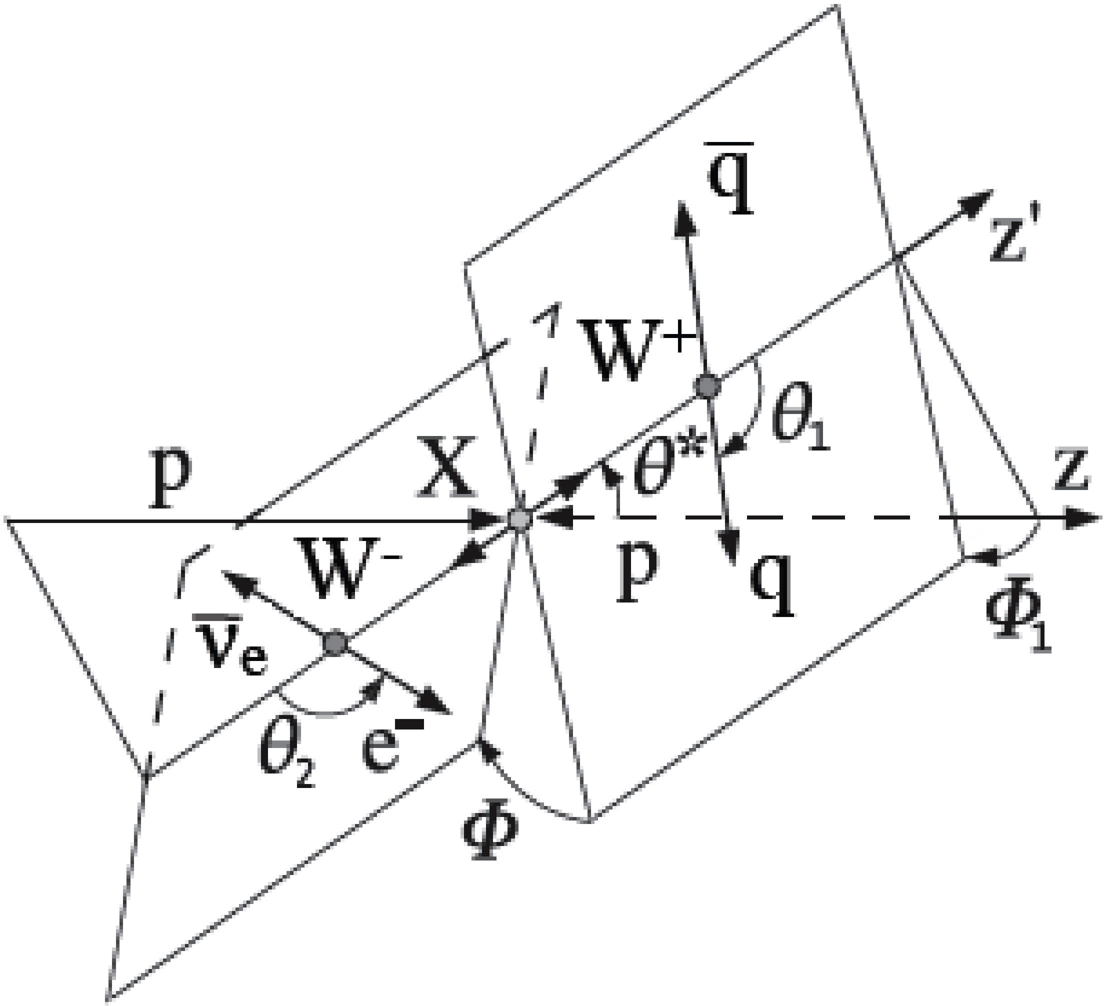


Figure 5.22: Planes and angular variables in the $\mathrm{H} \rightarrow \mathrm{WW} \rightarrow \mathrm{l}\nu\mathrm{qq}$ decay process [23].

5.5.3 BDT Input Optimization

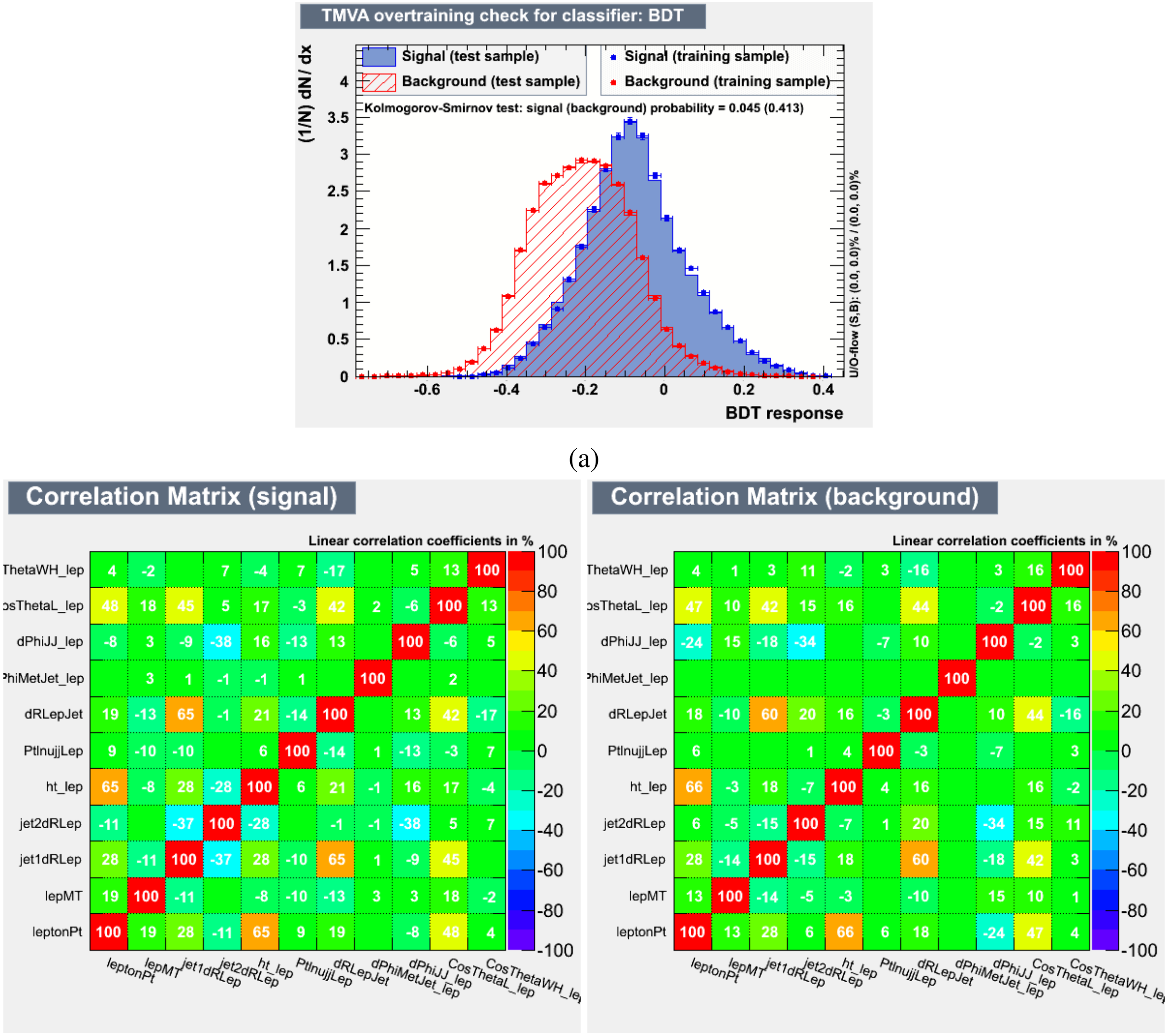
Remember that only the $\mathrm{W} + \text{jets}$ sample is being used to represent the background while a combination of the three $\mathrm{H} \rightarrow \mathrm{WW}$ samples is being used to represent the signal. When training a BDT the absolute normalization of the signal samples is not what is important. Instead, they must be normalized to their expected fractions relative to each other. Thus, the two other samples are normalized to the $\mathrm{ggH} \rightarrow \mathrm{WW}$ sample as shown in table 5.9. After setting up the samples the BDTs in each jet bin had to be optimized. The procedure in section 5.5.2 was used to select the individual variables with the most discrimination power. This section will describe how that list of input variables was optimized for the BDT trainings in each jet bin.

Process	**2 Jets**	**3 Jets**	**$\geqslant$4Jets**
ggH; $M_{\mathrm{H}} = 125\,\mathrm{GeV}$, $\mathrm{H} \to \mathrm{WW} \to \mathrm{l}\nu\mathrm{jj}$	1.0	1.0	1.0
qqH; $M_{\mathrm{H}} = 125\,\mathrm{GeV}$, $\mathrm{H} \to \mathrm{WW} \to \mathrm{l}\nu\mathrm{jj}$	0.195	0.248	0.239
WH, ZH, ttH; $M_{\mathrm{H}} = 125\,\mathrm{GeV}$, $\mathrm{H} \to \mathrm{WW}$	0.256	0.416	0.608

Table 5.9: The scale factors used to normalize the input signal samples for the BDT trainings.

To begin with, a BDT was trained using the best 20 variables specified in the previous section. After that, the BDT was checked for redundant variables by looking at the input variable correlation plots and for overtraining by using the Kolmogorov-Smirnov values between the training and test samples. If the Kolmogorov-Smirnov score was too low, this setup was rejected as it would indicate that the training and test samples didn't come from same underlying PDF, which they must. The variables used to train the BDT were also ranked by TMVA in order of their importance, which is measured by how much a given variable was used to discriminate signal from background. On the next iteration the two lowest performing variables were removed and the BDT was retrained. This process continued until only three input variables remained. Fig. 5.23 contains an example response curve examining overtraining as well as the correlation plots for signal and background. Based on the minimal correlation shown, none of the 11 variables are redundant in this training.

To quantitatively compare the various trainings we made use of their respective receiver operating characteristic (ROC) curves, an example of which is shown in fig. 5.24a. These curves measure the performance of a binary classification system by testing the signal efficiency and background rejection assuming a cut is placed on the classifier output. There are several ways of using the ROC curve to test the overall performance of any one BDT training. While a common method is to use the area under the ROC curve (greater area means better performance), we chose to use a different measure. Since the point (1,1) represents perfect signal acceptance and background rejection, that is the ideal point. A training with a ROC curve whose distance to that point is minimized will



(a)

Correlation Matrix (signal)

	leptonPt	lepMT	jet1dRLep	jet2dRLep	ht_lep	PtlnujjLep	dRLepJet	dPhiMetJet_lep	dPhiJJ_lep	CosThetaL_lep	CosThetaWH_le
ThetaWH_lep	4	-2		7	-4	7	-17		5	13	100
osThetaL_lep	48	18	45	5	17	-3	42	2	-6	100	13
dPhiJJ_lep	-8	3	-9	-38	16	-13	13		100	-6	5
PhiMetJet_lep		3	1	-1	-1	1		100		2	
dRLepJet	19	-13	65	-1	21	-14	100		13	42	-17
PtlnujjLep	9	-10	-10		6	100	-14	1	-13	-3	7
ht_lep	65	-8	28	-28	100	6	21	-1	16	17	-4
jet2dRLep	-11		-37	100	-28		-1	-1	-38	5	7
jet1dRLep	28	-11	100	-37	28	-10	65	1	-9	45	
lepMT	19	100	-11		-8	-10	-13	3	3	18	-2
leptonPt	100	19	28	-11	65	9	19		-8	48	4

Correlation Matrix (background)

	leptonPt	lepMT	jet1dRLep	jet2dRLep	ht_lep	PtlnujjLep	dRLepJet	dPhiMetJet_lep	dPhiJJ_lep	CosThetaL_lep	CosThetaWH_le
ThetaWH_lep	4	1	3	11	-2	3	-16		3	16	100
osThetaL_lep	47	10	42	15	16		44		-2	100	16
dPhiJJ_lep	-24	15	-18	-34		-7	10		100	-2	3
PhiMetJet_lep								100			
dRLepJet	18	-10	60	20	16	-3	100		10	44	-16
PtlnujjLep	6			1	4	100	-3		-7		3
ht_lep	66	-3	18	-7	100	4	16			16	-2
jet2dRLep	6	-5	-15	100	-7	1	20		-34	15	11
jet1dRLep	28	-14	100	-15	18		60		-18	42	3
lepMT	13	100	-14	-5	-3		-10		15	10	1
leptonPt	100	13	28	6	66	6	18		-24	47	4

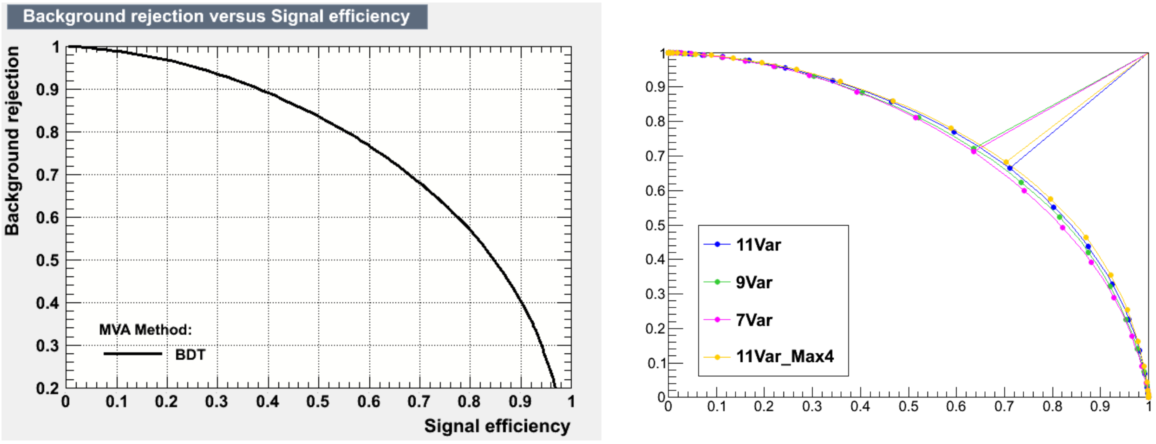
(b)

(c)

Figure 5.23: Example validation plots after the BDT training. (a) The response distributions for signal and background for both the training (markers) and test samples (filled histograms). The Kolmogorov-Smirnov value is used to decide how much overtraining has occurred. The correlation matrices of the input variables for the (b) signal and (c) background samples.

perform better than any other training. Thus we chose to use this distance as our FOM between the various trainings.

The ROC curves from the multiple trainings were compared as seen in fig. 5.24b. Although reducing the number of variables can help to prevent overtraining, there comes a point when this process begins to negatively impact the performance of the BDT. By comparing the ROC curves



(a)

(b)

Figure 5.24: Example ROC curved produced after the BDT training. (a) A standard ROC curve produced by TMVA. (b) ROC curves from multiple trainings with their associated distances of closest approach calculated. The training using 11 input variables showed the most discrimination power.

we were able to identify the trainings with the best performance. The variables used in these trainings are identified in table 5.10. The validation plots for the input variables can be found in appendix D.1.

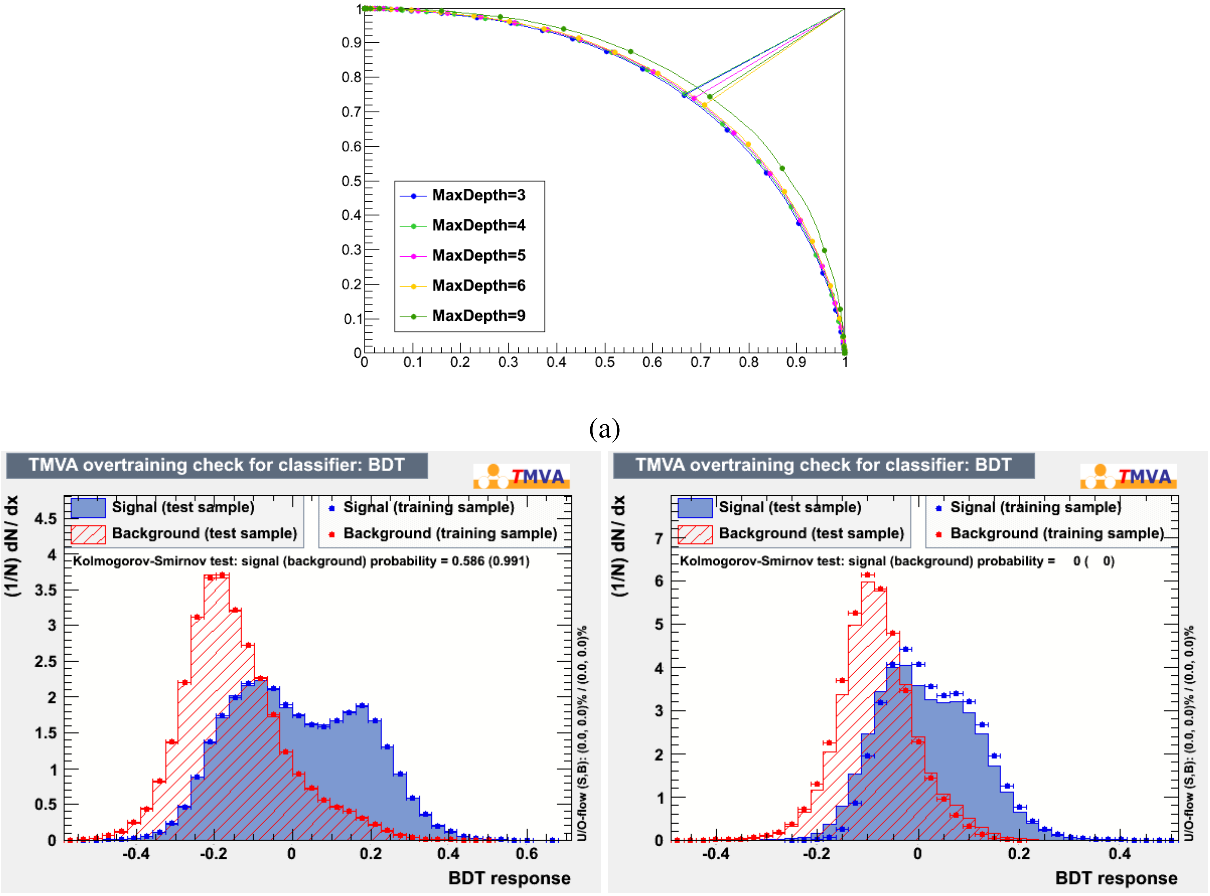
5.5.4 BDT Parameter Optimization

Besides the number of input variables, there are several hyper-parameters for the BDT trainings which must also be optimized to extract the maximum amount of performance and reduce the amount of overtraining. These hyper-parameters include the maximum number of trees to using in the training (nTrees), the value of β used in the boosting procedure (adaBoostBeta), the maximum depth allowed for each tree (MaxDepth), the minimum number of events allowed to remain in a node after splitting (nEventsMin), and the fraction of signal versus background events used during training. The trainings optimized for the input variables were used as a baseline for these next trainings. Each hyper-parameter was varied individually to see its effect on the performance of the BDT.

The ROC curves and overtraining plots for the tests on the MaxDepth parameter are shown in fig. 5.25. Although increasing the depth of the trees results in improved performance, it also

Variable	**2 Jets**	**3 Jets**	**$\geqslant$4Jets**
$p_{\mathrm{T}\,\mathrm{l}}$	★	✓	
Charge$_{\mathrm{l}} \times \eta_{\mathrm{l}}$		✓	✓
M_{t}	✓		
$p_{\mathrm{T}\,\mathrm{l}\nu\mathrm{jj}}$	✓		
$M_{\mathrm{l}\nu\mathrm{jj}}$		✓	✓
HT	✓	★	★
$\Delta R\left(\mathrm{l}, \mathrm{j}_1\right)$	✓		
$\Delta R\left(\mathrm{l}, \mathrm{j}_2\right)$	✓	✓	✓
$\Delta R\left(\mathrm{l}, \mathrm{j}_3\right)$		✓	✓
$\Delta R\left(\mathrm{l}, \mathrm{jj}\right)$	✓	✓	
$\Delta\varphi_{\text{min}}\left(\mathrm{l}, \mathrm{j}\right)$		✓	
$\Delta\eta\left(\mathrm{j}, \mathrm{j}\right)$		✓	
$\Delta\varphi\left(\vec{\not{E}}_{\mathrm{T}}, \mathrm{j}\right)$	✓	✓	✓
$\Delta\varphi\left(\vec{\not{E}}_{\mathrm{T}}, \mathrm{l}\right)$			✓
$\Delta\varphi\left(\mathrm{j}, \mathrm{j}\right)$	✓		
$\cos\left(\theta_{\mathrm{l}}\right)$	✓	✓	
$\cos\left(\theta_{\mathrm{WH}}\right)$	✓	✓	
$\cos\left(\theta_{\mathrm{j}}\right)$		✓	

Table 5.10: A list of the input variables chosen for each BDT training. The variables are optimized separately in each jet bin. The check marks denote the chosen variables for each jet bin while the stars denote the the best performing variable.



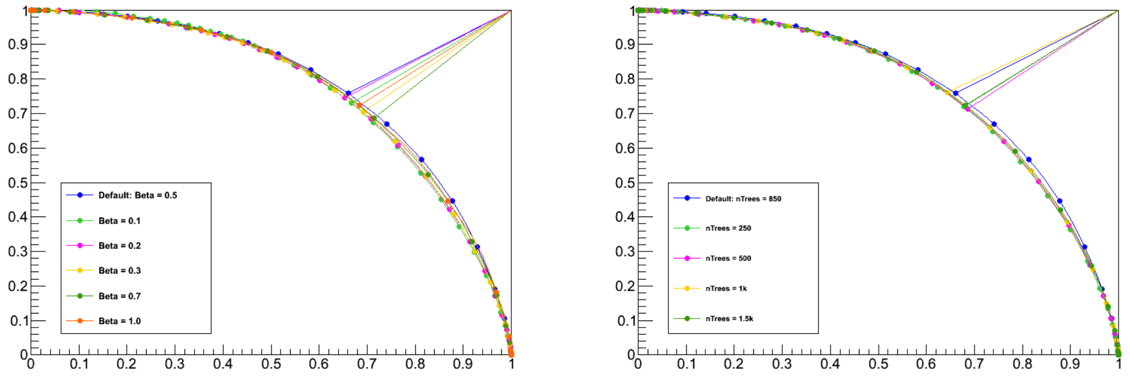
(a)

(b)

(c)

Figure 5.25: (a) The ROC curve used to test the performance of five different values of the MaxDepth parameter. The overtraining plots showing the BDT response for MaxDepth values of (b) 3 and (c) 9. The Kolmogorov-Smirnov scores for a MaxDepth of 3 are far superior to those for a MaxDepth of 9.

significantly increases the overtraining. Thus we chose the largest MaxDepth value that didn't result in overtraining. The ROC comparisons for the parameters adaBoostBeta and nTrees is are shown in fig. 5.26. In these cases, the default values turned out to be the best performing. The final hyper-parameter values used for our trainings can be found in table 5.11. We also found that using the maximum possible amount of signal and background events was best, meaning we fed the trainings all the events that we had. The events were then split evenly between the test and training samples. The resultant BDT classifier distributions are shown in appendix D.2.



(a)

(b)

Figure 5.26: The ROC curves used to test the performance of different values of (a) the adaptive boost factor β and (b) the number of trees used in the trainings. The best performing value is represented by the dark blue curve.

Hyper-Parameter	**2 Jets**	**3 Jets**	**$\geqslant$4Jets**
MaxDepth	4	3	3
nTrees	850	850	850
adaBootBeta	0.5	0.5	0.5
nEventsMin	100	100	100

Table 5.11: Hyper-parameters used for the BDT trainings.

5.6 Matrix Element Analysis

Table 5.6 clearly shows that the total signal, in every channel, is at least an order of magnitude smaller than even the statistical uncertainty of the background. A simple cut and count experiment will not lead to any significant results. The previous $H \to WW \to l\nu jj$ analyses have performed a fit to sensitive distributions like the 4-body mass, the mass of the system made out of the two jets, lepton, and $\not{E}_{\rm T}$, which is sensitive to the Higgs mass peak. However, this approach only includes a small amount of available information, leaving out additional sensitive kinematic distributions. It is also felt that a BDT analysis using only kinematic variable would be sub-optimal because shallow classifiers are not robust against non-linear correlations and are only as good as the input variables chosen. While the BDT classifiers described in the last section show a good amount of signal to background discrimination, there is another method which can also be used to separate signal from background. Instead, this analysis uses a matrix element method (MEM), which starts from the differential cross section calculation from quantum field theory to classify how likely and event is to come from a given process [178, 24].

The output of the MEM will be a set of differential cross sections, correct up to a normalization factor. The original application of this technique was in [178], where the outputs were referred to as probabilities. Because the only purpose of these outputs is to construct a discriminant between signal and background, it is inconsequential whether we call them probabilities or likelihoods. Therefore, I will refer to the outputs as probabilities in keeping with tradition.

5.6.1 Differential Cross Section

The probability $P\left(x;\alpha\right) = P_{evt}$ of a signal is proportional to the differential production cross section, where α is the parameter we wish to measure, like the mass of the Higgs boson, and x is a set of physical variables. This is true if the detector resolution is sufficiently small and the beam energies are well known, as it is in the case of CMS. For the scattering of two particles the differential cross section can be written as [2]:

$$d\sigma = \frac{\left(2\pi\right)^4 \left|\mathcal{M}\right|^2}{4\sqrt{\left(q_1 \cdot q_2\right)^2 - m_{q_1}^2 m_{q_2}^2}} d\Phi_n\left(q_1 + q_2; p_1, ..., p_n\right) \tag{5.10}$$

where $|\mathcal{M}|$ is the Lorentz invariant matrix element (ME) [143]; q_1, q_2 and m_1, m_2 are the 4-momenta and masses of the incident particles; p_i are the 4-momenta of the n final state particles; and $d\Phi_n$ is the n-body phase space. The phase space term is written as:

$$d\Phi_n\left(q_1 + q_2; p_1, ..., p_n\right) = \delta^4\left(q_1 + q_2 - \sum_{i=1}^{n} p_i\right) \prod_{i=1}^{n} \frac{d^3 p_i}{\left(2\pi\right)^3 2E_i} \tag{5.11}$$

If CMS could measure all of the final state particles with 100% accuracy, no detector effects or uncertainties, and all of the information about the initial state particles was known - including the energy, momentum, and particle type - we could analytically solve this equation and normalize it to the total cross section to define an event probability $P_{evt} \sim \frac{d\sigma}{\sigma}$. Using the differential cross sections for each of the processes being tested we could create a perfect discriminant for each event. Unfortunately, this is not the case and there are several unknowns which must be accounted for:

1. Some particles involved in the ME are either not measured at all or not fully measured. The initial state partons are held within protons, making their exact energies unknown. The neutrino in the final state is not fully measured by CMS. We use the $\not\!\!E_{\mathrm{T}}$ as a proxy for the neutrino, but we can't measure the p_z component of its momentum vector.

2. The partons in the final state are only measured after showering and hadronizing to form jets. While every effort is taken to measure the jet energies with great accuracy, this is no substitute for the parton level energies.

3. The energy resolution of the CMS sub-detectors cannot be ignored, especially for jets.

4. For practical reasons the ME cannot be exactly calculated. The more precise a probability one wants to calculate, the more diagrams one must include in the ME calculation. This increases the computational complexity of the problem significantly. That is why this analysis used mainly tree-level diagrams, with some sub-leading diagrams for our biggest background, $\mathrm{W}+$jets.

Despite our best efforts, each of these effects leads to some loss in sensitivity.

5.6.2 Parton Distribution Functions and Phase Space

if final state fully known (momenta and energies), then can calculate the initial state momenta and energies from conservation of energy and momentum, assuming the transverse momentum of the initial state particles is zero (a fairly good assumption). Without knowing the full final state, a set of PDFs will determine the likelihood of a given initial state configuration. The PDF scale varies with the process dependent momentum transfer Q^2. For $\mathrm{W}+$jets, for example, $\mathrm{Q}^2 =$

$M_{\mathrm{W}}^2 + \left(\sum_{\mathrm{jets}} p_{\mathrm{T}}\right)^2$ while for Drell-Yan scattering $\mathrm{Q}^2 = \hat{s} = |q_1 + q_2|^2$, where q_1 and q_2 are 4-vectors of the initial state quarks. Because Q^2 comes from perturbative calculations and cannot be measured, its value is not well defined [24].

Taking this into account, the differential cross section calculation becomes:

$$d\sigma = \frac{(2\pi)^4 |\mathcal{M}|^2}{4\sqrt{(q_1 \cdot q_2)^2 - m_{q_1}^2 m_{q_2}^2}} f(x_1) f(x_2) \, d\Phi_n (q_1 + q_2; p_1, ..., p_n), \tag{5.12}$$

where $f(x_i)$ are the PDFs for the incoming protons and $x_i = E_{q_i}/E_{beam}$ is the fraction of the proton momentum carried by the incident parton i.

The differential cross section can be further simplified by using $\sqrt{(q_1 \cdot q_2)^2 - m_{q_1}^2 m_{q_2}^2} \simeq 2E_{q_1}E_{q_2}$. In other words by considering the input partons to be massless and ignoring any small transverse momentum they may have we arrive at:

$$d\sigma = 2\pi^4 |\mathcal{M}|^2 \frac{f(x_1)}{|E_{q_1}|} \frac{f(x_2)}{|E_{q_2}|} d\Phi_n (q_1 + q_2; p_1, ..., p_n). \tag{5.13}$$

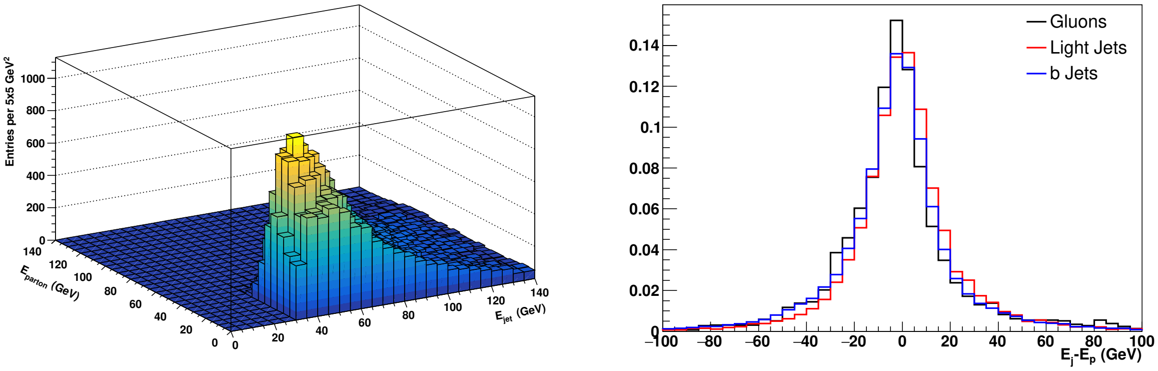
5.6.3 Transfer Functions

While leptons in CMS can be measured with a high degree of accuracy, the lepton resolution is relatively small, the jet energies are not the same as the energies of their initiating partons. Even worse are the neutrinos which pass all the way through the CMS detector without being measured. The solution is to use a transfer function, which maps between the energies and momenta of the final state partons to those of the measured objects. Adding a transfer function W into the differential cross section calculation we find:

$$d\sigma = 2\pi^4 |\mathcal{M}|^2 \frac{f(x_1)}{|E_{q_1}|} \frac{f(x_2)}{|E_{q_2}|} W_l^3 (p_l, p_{l_{\mathrm{meas}}}) \, W_\nu^3 (p_\nu, p_{\nu_{\mathrm{meas}}}) \prod_{i=1}^{n_{\mathrm{jets}}} W_i^3 (p_i, p_{i_{\mathrm{meas}}}) \, d\Phi_n (q_1 + q_2; p_1, ..., p_n) \tag{5.14}$$

Here W^3 refers to the three transfer functions necessary to map the energy, polar angle, and azimuthal angle of the parton to the observed quantity.

Luckily we can start making some simplifications right off the bat. The lepton quantities and jet angles are assumed to be measured well enough that the transfer function approaches a Dirac delta function. Even if this isn't exactly true it would only reduce the sensitivity of the analysis and would not affect the final result. Thus those terms will disappear from the equation. Unfortunately, the same assumptions cannot be made about the jet energies. The jet energy transfer functions are modeled as a ten parameter double Gaussian fit to the difference in parton and jet energies from a MC sample. The underlying distribution of energies can be seen in fig. 5.27a. While matching parton energies and jet energies might sound a lot like the L5Flavor corrections in CMS, the JEC only correct the jet energies back to the most probable value. By using a transfer function we can integrate across all possible jet energies to extract more information. Although the transfer functions will vary across η we had limited statistics available in out MC sample used to derive these and thus were forced to use a single bin of $|\eta| < 2.4$. Three different sets of TF were derived for b quark jets, light quark jets, and gluon jets as seen in fig. 5.27b. All three types of jets will have different kinematics and thus will produce different transfer functions.



(a) (b)

Figure 5.27: (a) The distribution of parton energy versus jet energy in Monte Carlo events for light flavored jets. (b) Distributions of the difference in parton energy and jet energy for different kinds of jets. This shows that a separate transfer function is necessary for each flavor of jet.

As stated before, the neutrino momentum is not measured; nor can the z component of the neutrino momentum cannot be calculated. This stems from the fact that the longitudinal momentum of the initial state partons is not know, only the momentum of the protons. We don't know how the momentum is split between the various partons that make up the proton. During the computation of the differential cross section we integrate over the unknown quantities, which includes the neutrino's longitudinal momentum. The momentum is allowed to vary from 0 GeV to 4 TeV, the beam energy, which is motivated by the conservation of energy and momentum. At this point, assuming a choice for neutrino p_z and jet energies, the x and y components of the neutrino momentum as well as the z component of the momenta for the initial partons can be derived from conservation of energy and momentum.

After accounting for all the simplifications, the PDFs, and the transfer functions the differential cross section becomes:

$$d\sigma = \int dp_{z_\nu} 2\pi^4 |\mathcal{M}|^2 \frac{f\left(x_1\right)}{|E_{q_1}|} \frac{f\left(x_2\right)}{|E_{q_2}|} \prod_{i=1}^{n_{\text{jets}}} \frac{dE_i W\left(E_i, E_{i_{\text{meas}}}\right)}{E_i} \frac{\delta^4\left(q_1 + q_2 - p_l - p_\nu - \sum_{i=1}^{n_{\text{jets}}} p_i\right)}{E_l E_\nu} \tag{5.15}$$

As you can see all of the parton level quantities have been replaces by their measured counterparts, which allows us to perform the calculations with the measurements taken by CMS. This equation can then be normalized to the total cross section to form an event probability:

$$P(x;\alpha) = \frac{1}{\sigma} \int 2\pi^4 |\mathcal{M}|^2 \frac{f\left(x_1\right)}{|E_{q_1}|} \frac{f\left(x_2\right)}{|E_{q_2}|} W\left(y, x\right) d\Phi_4 dE_{q_1} dE_{q_2} \tag{5.16}$$

where $f\left(x_i\right)$ are the PDFs, $x_i = E_{q_i}/E_{beam}$ is the fraction of the proton momentum carried by the incident parton i, and $W\left(y, x\right)$ is the transfer function mapping measured jet energies x to the parton energies y. For simplicity the equation has been returned to a more compact form. At this point we can use numerical integration to calculate the the probability densities of interest.

5.6.4 Matrix Elements

As I alluded to before, no analytic form for a scattering process matrix element to all orders exists. On top of an already computationally difficult problem, the loop corrections in higher-order calculations become too costly, which is why we chose to use mostly leading order diagrams (except for $\mathrm{W} + \mathrm{jets}$). The matrix elements were generated by MADGRAPH in FORTRAN and then converted by C++ to speed up computation.[5] MADGRAPH makes use of a library called HELAS [179] to do the leading-order matrix element calculations. Each matrix elements can have contributions from multiple subprocesses (i.e. $\mathrm{pp} \to \mathrm{WW} \to \mathrm{l}\nu\mathrm{jj}$ includes diagrams from $\mathrm{u}\overline{\mathrm{u}} \to e^{+}\nu_{\mathrm{e}}\overline{\mathrm{u}}\mathrm{d}$, $\mathrm{u}\overline{\mathrm{u}} \to e^{-}\overline{\nu}_{\mathrm{e}}\mathrm{u}\overline{\mathrm{d}}$, $\mathrm{d}\overline{\mathrm{d}} \to e^{-}\overline{\nu}_{\mathrm{e}}\mathrm{u}\overline{\mathrm{d}}$, etc.). Additionally, each subprocess can be generated from a number of diagrams as seen in fig. 5.28 for the $\mathrm{gd} \to e^{-}\overline{\nu}_{\mathrm{e}}\mathrm{ug}$ process.

Matrix elements were calculated for all of the major signals and background in this analysis. There were 15 matrix elements which were eventually used: WW, WZ, WZbb, WLg, WLg (second order), Wgg, WLL, WLb, Wbb, ZLight, Single Top t-channel, Single Top s-channel, QCD, ggH ($M_{\mathrm{H}} = 125\,\mathrm{GeV}$), and WH ($M_{\mathrm{H}} = 125\,\mathrm{GeV}$). While some matrix element diagrams may be left out, the purpose of calculating the probabilities is to discriminate a signal event from a background event. The loss of a diagram will simply reduce the sensitivity of the classifier, not change the answer. The some of the Feynman diagrams used for calculating the matrix elements can be found in figs. 5.29 and 5.30.

Diagrams with more than two jets present a problem for the matrix element calculations because they doesn't have the same final state as the signal. As was stated in section 2.6, for a $\mathrm{t}\overline{\mathrm{t}}$ event to pass as a two jet event it must mean that some of the jets are missed. This can happen in two different ways; either both W bosons decay leptonically and one lepton is not detected or one of the W bosons decays hadronically and two of the four jets are missed. If we really confine the matrix element to two jets, then we can use the diagram with one leptonically decaying W boson where the other W is simply not observed (i.e. it decays outside our acceptance). In this case there are three additional unknown momentum components coming from the W boson which must be integrated over. If a third jet were allowed, then the the typical semileptonic W boson decay is used and one of the light quarks is assumed to be missed. This also adds three additional integrations to the calculation. Although it would have been nice to include a $\mathrm{t}\overline{\mathrm{t}}$ matrix element probability for

[5] Only the leading diagrams were converted. The C++ code was then run alongside the FORTRAN code to compare the outputs.

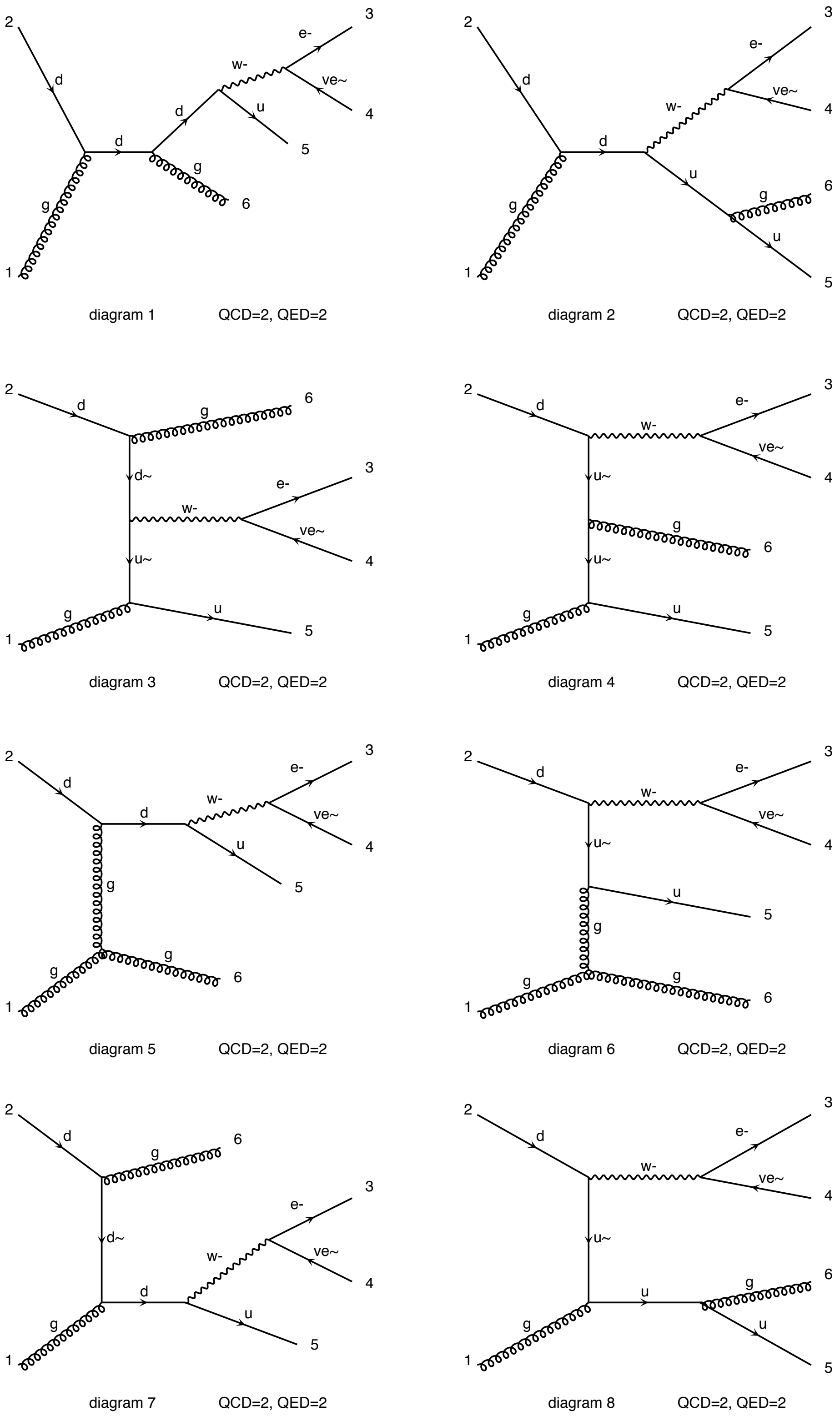


Figure 5.28: Feynman diagrams from the $\mathrm{gd} \to e^{-}\overline{\nu}_{\mathrm{e}}\mathrm{ug}$ process.

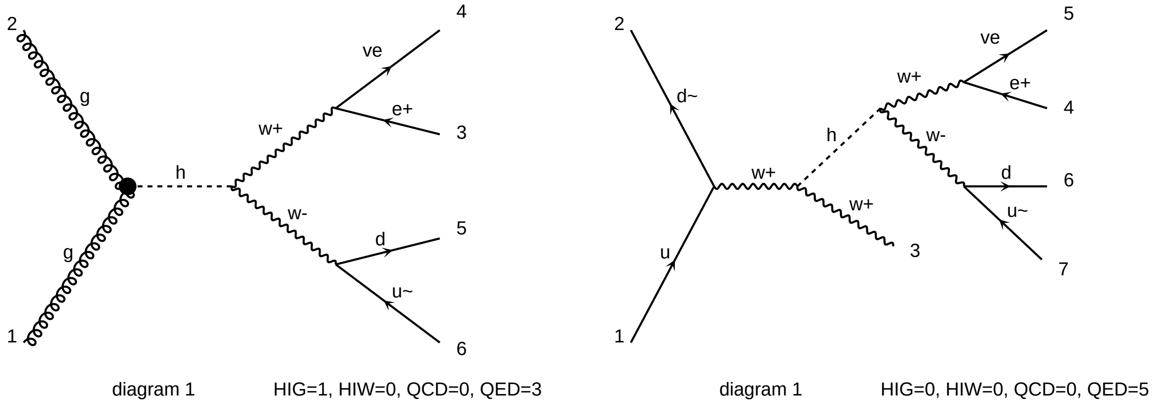


Figure 5.29: Feynman diagrams used to calculate the matrix element probabilities for the ggH and WH signals for two-jet events.

discrimination purposes, the additional integrals proved too costly to compute. Each $\mathrm{t\bar{t}}$ probability took over two minutes to compute, even using accelerated numerical integration packages. Therefore we did not compute the $\mathrm{t\bar{t}}$ probability and we rely on the signal probabilities being relatively low for $\mathrm{t\bar{t}}$ events.

5.6.5 Combinatorial Considerations

An ambiguity arises when there are multiple jets in the final state of the diagram. Therefore, we take the sum of the differential cross section for all combinations of matched partons and jets. We can reduce the number of combinations and increase the sensitivity of the computation when there is a b-tagged jet in the event and a bottom quark in the diagram. In general, when there is an ambiguity all of the various parton-jet combinations are used. However, in the case of the $\mathrm{t\bar{t}}$ diagram there are enough combinations to make this methodology computationally impractical. Therefore, only the two combinations where b-tagged jets are assigned to the two bottom quarks are used.

5.6.6 Numerical Integration

Obviously the differential cross section must be calculated many times for every value of the differential variables in both the data and MC samples. The integration is performed over the neu-

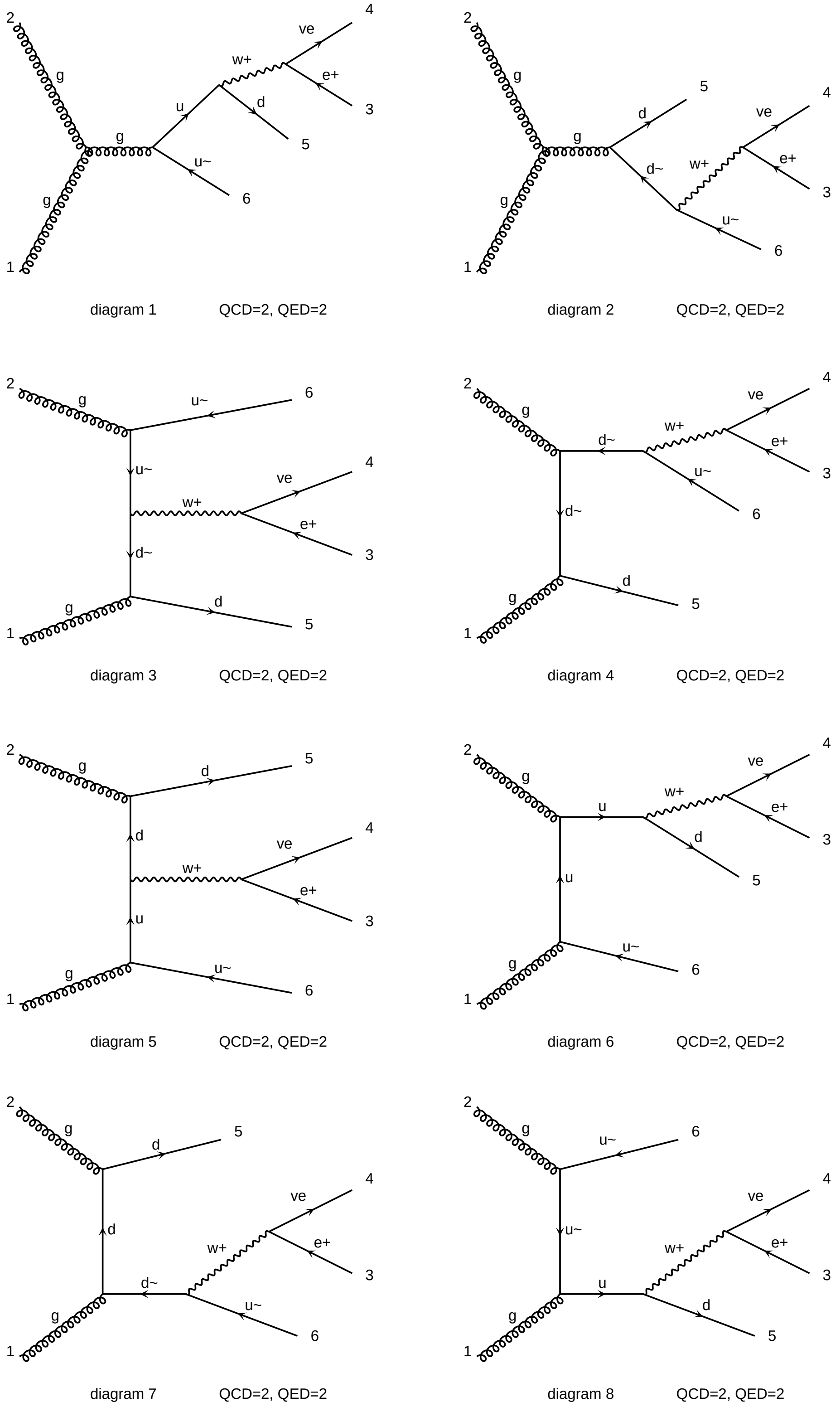


Figure 5.30: A sampling of Feynman diagrams used to calculate the W + jj matrix element probabilities for two-jet events.

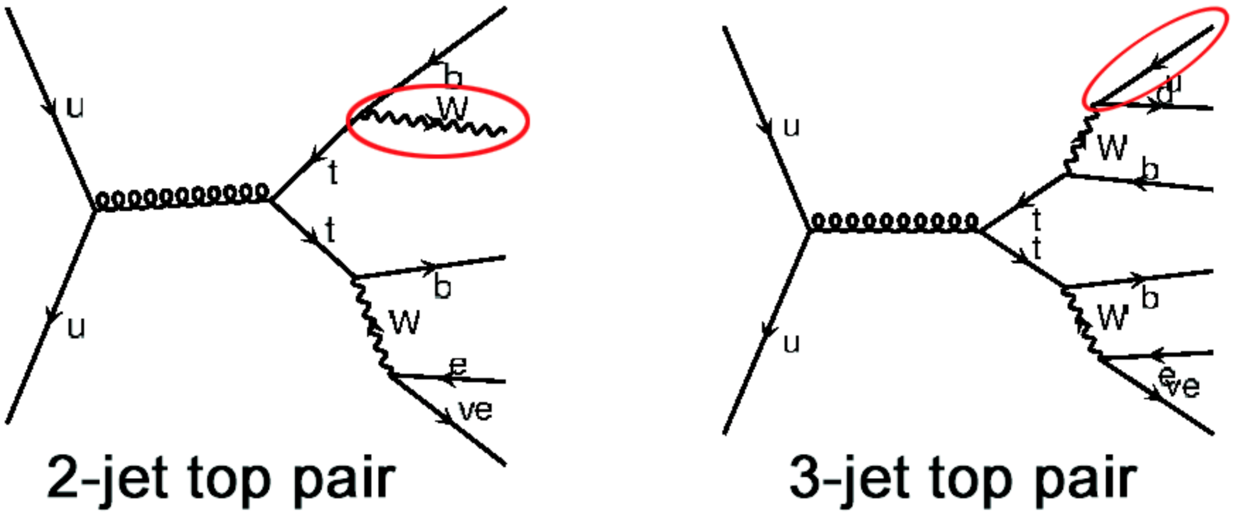


Figure 5.31: Feynman diagrams used to calculate top pair probabilities for two- and three-jet events. The circled particles are assumed to be unobserved and an integral is taken over their momenta. Figure and caption from [24].

trino longitudinal momentum, the jet energies, and in some cases over the momenta of missing particles, as in the case of $t\bar{t}$. The result is an integral with dimensionality of anywhere between three and seven dimensions; six in the case of top pair production with a two-jet final state. These types of equations can't be solved analytically, so we must instead use numerical integration techniques.

For the simpler cases involving three integrals and without any missing particles integration is performed using the adaptive quadrature [180] method based on the CERNLIB [181] RADMUL [182] routine, adapted for ROOT [183], then adapted again for the CDF single top analysis [24]. The algorithm iteratively divides the n-dimensional region to be integrated into equal-sized regions. At each iteration the uncertainty in each region is estimated and the region with the largest uncertainty is divided in half. The iterations continue until all of the regions have an error less than a user specified amount. In this analysis we used 1% for all of the matrix elements except for ZLight, where a 5% uncertainty is allowed. After the stopping condition is met, the integral in each region is estimated and returned. The benefit of using this method is that it is stable and its answers are reproducible; its calculations are deterministic and do not reply on pseudo random number generators (PRNG). Table 5.12 lists the computation times for each matrix element averaged over 1000 events computed in both the $\mathrm{W} +$ jets and ggH samples.

Diagram	$\mathrm{W} +$ jets Sample [s]	ggH $M_{\mathrm{H}} = 125\,\mathrm{GeV}$ Sample [s]
ggH	2.9	3.2
WH	4.5	3.8
QCD	0.4	0.5
Single Top s-channel	4.2	4.3
Single Top t-channel	2.9	3.3
Wbb	1.9	1.3
WLL	7.4	4.8
WLb	2.9	2.3
WLg (LO and NLO)	3.5	2.7
WW	1.5	1.1
WZ	3.9	2.9
WZbb	2.5	1.9
ZLight	39.1	26.3
Total	77.7	58.5
Total ($\mathrm{ggH} \times 35$, $\mathrm{WH} \times 14$)	233.7	217.9

Table 5.12: The computations times for each probability averaged over 1000 events and computed in both the $\mathrm{W} +$ jets and ggH samples. The Wgg diagram was not included in this test. The integration for each of these probabilities was performed using the ROOT integrator.

Although the ROOT integrator is deterministic and stable, good qualities in a numerical integrator, it starts to become prohibitively slow for higher dimensional integrals. Ref. [24] performed a test on the $\mathrm{t\bar{t}}$ computation and the ROOT integrator did not converge for a single integration even after running for an entire day. Instead, we used the DIVONNE Monte-Carlo integration algorithm from the CUBA library [184], which is based on CERNLIB's DIVON4 [185] function. The algorithm first uses stratified sampling, a method by which a population is subdivided into homogeneous, mutually exclusive[6], and collectively exhaustive[7] subpopulations. Sampling the individual subpopulations improves the representativeness of the estimate and reduces the variance and sampling error. In practical terms, the DIVONNE algorithm uses this type of sampling by

[6] Each element is assigned to one region.

[7] No element from the larger population is excluded.

partitioning the integration region into sub regions. Each subregion is required to have an equal value of the spread $\vec{s}$, defined as:

$$\vec{s}\,(r) = \frac{1}{2}V\,(r)\left(\max_{\mathbf{X}\in r}\vec{f}\,(\mathbf{X}) - \min_{\mathbf{X}\in r}\vec{f}\,(\mathbf{X})\right), \tag{5.17}$$

where $\mathrm{V}\,(r)$ is the multi-dimensional volume of region r and $\vec{f}\,(\mathbf{X})$. is the value of the function in the subregion. The Koksma-Hlawka inequality [186, 187, 188] shows that the variance is bounded by $\vec{s}$. Therefore the borders of each subregion are adjusted to reduce the spread and thus reach the user requested variance. Once the subregions are set, the integral is estimated summing the values of randomly selected points within each subregion. Once this first stage of integration is complete, the algorithm uses these results to estimate the number of samples necessary to reach the desired accuracy. Once the second of the two samples is chosen for a particular subregion, a χ^2 test is used to check if the samples averages are consistent within errors. If a subregion fails this test, then it is either subdivided again or more sampling points are used, depending upon the settings. In a test of 1000 events performed by [24], the DIVONNE algorithm returned results compatible with the RADMUL algorithm and was also stable to within 0.001%.

Diagram	$\mathrm{W+jets}$ Sample [s]	ggH $M_{\mathrm{H}} = 125\,\mathrm{GeV}$ Sample [s]
Single Top tW-channel	100.9	68.0
$\mathrm{t\bar{t}}$	134.7	133.6
Total	235.6	201.6

Table 5.13: The computation times for the unused single top and $\mathrm{t\bar{t}}$ diagrams. Including these would have doubled the overall computation time. The integration for each of these probabilities was performed using the DIVONNE integrator.

Even with the DIVONNE algorithm, the computation times for the $\mathrm{t\bar{t}}$ and single top tW channel ME probabilities were prohibitively large (see table 5.13) and were thus dropped from the list of computations. Besides the $M_{\mathrm{H}} = 125\,\mathrm{GeV}$ ggH and WH diagrams, 34 additional ggH and 13 additional WH probabilities were calculated corresponding to different Higgs mass hypotheses. In the end, however, these probabilities were not used. The total computation time for a single event was around four minutes, give or take some time for computing cluster overhead. This computation is by far the most time consuming aspect of this analysis, especially with tens of millions of events to process. The total computation time ended up costing $\sim$12 million CPU hours and spanned over 1.5 years, requiring the work of several analyzers and the entire Worldwide LHC Computing Grid (WLCG).

5.6.7 Standalone Matrix Element Based BDT

The fifteen probabilities $P(x;\alpha)$, corresponding to the leading order diagrams of the major background and signal processes, were computed for each event in both data and MC. Now that all of the leading order kinematics are encoded in these 15 numbers, they must be combined in order to discriminate signal from background. A BDT was used rather than combining the ME into a likelihood as in the Matrix Element Likelihood Analysis (MELA) used by $\mathrm{H} \to \mathrm{ZZ} \to 4\mathrm{l}$ or the event probability discriminants (EPD) used by the single top analysis done by CDF [24]. Three new BDTs were trained using the same settings as the BDTs with kinematic variables used as inputs. However, this time the inputs consisted of the 15 matrix element probabilities. The output discriminant distributions can be found in appendix D.2. Unfortunately, these BDTs (MEBDT), on their own, did not out perform the kinematic variable based BDTs (KinBDT). This might be due to the fact that we only used leading order diagrams (not even all of the diagrams), it might have to due with the combinatorics of the jets and partons, or it could have to do with sub-optimal transfer functions. However, by comparing the KinBDT to the MEBDT, we found that they had complimentary information.

5.6.8 Combined BDT

In order to combine the complimentary information from the kinematic variables and the MEs, with the purpose of discriminating a Higgs event from a background event, we combined the two sets of variables. An initial BDT was computed which combined the information from 15 of the computed MEs, as noted above. This gives a less discriminating shallow network the ability to create a better performing network because the inputs are already non-linear variables. The output of this BDT, along with previously selected kinematic variables, is then used as the input to a new BDT in order to combine all of this complimentary information. The combined BDT (KinMEBDT) has more discrimination power than either the MEs or the kinematic variables alone. Images of the output discriminant can be found in appendix D.2. Additionally, the ROC curves used to compare the various BDTs can be found in appendix D.4. Table 5.14 shows the FOM used to compare the BDTs.

The distribution of the KinMEBDT discriminant was chosen as the template for our limit setting procedure. However, before actually processing the limits, which will be discussed in section 6, the best way to bin the templates needed to be determined. There are two conditions which needed to be satisfied for for every bin:

1. There cannot be a bind which contains an observed count, but no background estimation. This would lead to an artificially high significance and an artificially low upper confidence level.

2. The sum of the background templates must have a statistical uncertainty of $\leqslant 10\%$ in each bin. This limits the effects of statistical fluctuations in any one bin.

To accomplish this optimization, we first started with a finely binned KinMEBDT distribution, bounded between -1 $\leqslant$ KinMEBDT $\leqslant$ 1. Starting with the lowest bin, we checked that each bin passed the two aforementioned conditions. If a bin failed either condition, then that bin and the next highest bin were merged. If both conditions were met, then the next highest bin was checked. The process continued until reaching the last bin, which could be merged into the previous bin if necessary. This resulted in leaving the maximum number of variable width bins reasonably allowable, a desirable property which leads to having the greatest discrimination power possible. Additionally, there were a different number of bins for each of the BDT trainings; 37, 29, and 20 bins for the two-jet, three-jet, and four of more jet categories, respectively.

BDT Input Variables	2 Jets	3 Jets	⩾4 Jets
Kin	0.7569 (0.4418)	0.7970 (0.3948)	0.7759 (0.4150)
ME	0.6568 (0.5497)	0.6698 (0.5321)	0.6598 (0.5462)
KinME	0.7581 (0.4402)	0.7983 (0.3926)	0.7973 (0.4051)

Table 5.14: The figures of merit (FOM) used to evaluate the various BDT trainings for the three jet bins and the three sets of input variables. The values outside of the parentheses are the areas under the curve (AUC) while the values in the parentheses are the shortest distances on the curve to the point (1,1).

5.7 Systematic Uncertainties

The input to the statistical analysis is a set of BDT discriminant histograms and their associated systematic uncertainties. Given that this is a shape analysis, it is important to consider systematic uncertainties that may change the expected yields (rate changes), the shape of the discriminating variable, or both. We consider many sources of uncertainty on both the background estimation and the signal normalization. Table 5.15 summarizes all of the systematic uncertainties considered for this analysis, with one systematic per line. The largest uncertainty comes from the W + jets normalization stemming from the QCD and W + jets rate estimation. Each source of systematic uncertainty will be described in more detail in the sections below.

5.7.1 LHC Luminosity

A flat rate uncertainty of 2.6% is applied to all of the simulated samples to account for the uncertainty on the LHC luminosity and thus the simulation normalizations [108].

5.7.2 Sample Cross Sections

The uncertainties on the theoretical cross sections used for the normalizations of the background simulations are taken from [189]. Likewise, the signal cross sections, branching ratios, and uncertainties are taken from CERN Yellow Report 3 [190]. The uncertainties on the background sample cross sections ranged from 3-5.7% while the signal cross section uncertainties range from 10-11% (PDF & QCDScale). The theoretical cross section uncertainties on the signal are broken into two components, the uncertainty on the QCD renormalization and factorization scales and the

Source	Type	Rate Uncertainty [%]	Notes
QCD Scale (ggH)	lnN	7-8	Scale uncertainty for NLO ggH prediction
QCD Scale (qqH)	lnN	0.2	Scale uncertainty for NLO qqH prediction
QCD Scale (ZH)	lnN	1	Scale uncertainty for NLO ZH prediction
QCD Scale (WH)	lnN	3.1	Scale uncertainty for NLO WH prediction
QCD Scale (ttH)	lnN	4-9	Scale uncertainty for NLO ttH prediction
PDF (gg)	lnN	6-7	PDF uncertainty for gg initiated processes $(\mathrm{ggH}, \mathrm{ttH})$
PDF $(\mathrm{q\overline{q}})$	lnN	2.6-2.8	PDF uncertainty for $\mathrm{q\overline{q}}$ initiated processes $(\mathrm{qqH}, \mathrm{WH}, \mathrm{ZH})$
QCD Scale $(\mathrm{t\bar{t}})$	lnN	5.7	Scale uncertainty for NLO $\mathrm{t\bar{t}}$ prediction
QCD Scale $(\mathrm{Z}+\text{jets})$	lnN	3.4	Scale uncertainty for NLO $\mathrm{Z}+\text{jets}$ prediction
QCD Scale (Single t)	lnN	5	Scale uncertainty for NLO single top prediction
QCD Scale (VV)	lnN	3	Scale uncertainty for NLO diboson prediction
$\mathrm{W}+\text{jets}$ Normalization	lnN	0.4-0.5	Scale uncertainty for $\mathrm{W}+\text{jets}$ prediction
QCD	lnN	10	Scale uncertainty for data-driven QCD prediction
$\mathrm{t\bar{t}}$	lnN	3	Scale uncertainty for $\mathrm{t\bar{t}}$ prediction
Luminosity 8 TeV	lnN	2.6	Signal and all backgrounds
Lepton Efficiency	lnN	2	Signal and all backgrounds
$\not\!\!E_{\mathrm{T}}$	lnN	0.2	Signal and all backgrounds
Jet Energy Scale	shape	0-20	Signal and all backgrounds
Pileup Weight	shape	0-8	Signal and all backgrounds
CSV Weight	shape	0-17	Signal and all backgrounds
Top p_{T} Weight	shape	0.5-2	$\mathrm{t\bar{t}}$ only
ME Matching	shape	-	$\mathrm{W}+\text{jets}$ only
Q^2 Scale	shape	-	$\mathrm{W}+\text{jets}$ only
$\cos{(\theta_{\mathrm{l}})}$Weight	shape	-	$\mathrm{W}+\text{jets}$ only
QCD Multijet η Weight	shape	6-30, 0.5-1	QCD and $\mathrm{W}+\text{jets}$ only

Table 5.15: Summary of the systematic uncertainties used in this analysis.

uncertainty on the PDFs. Table 5.16 shows a summary of the uncertainties used. An additional uncertainty of $\sim$0.5% is assigned to the W + jets backgrounds due to the uncertainty from the fit when determining the QCD sample normalization.

5.7.3 MET Uncertainty

With respect to $\not\!\!E_{\mathrm{T}}$, this analysis follow along the same line as the high mass lνjj group. Although we lowered the cut to be $\not\!\!E_{\mathrm{T}} \geqslant 25\,\mathrm{GeV}$, the uncertainty on the $\not\!\!E_{\mathrm{T}}$ should be similar. Thus we applied the same conservative estimate of a 0.2% uncertainty.

5.7.4 Lepton Selection and Trigger Efficiency

This analysis makes use of the single lepton triggers and requires a tight electron or muon in the event. Consequently we must account for any mis-modeling of the lepton identification or trigger efficiencies. A flat 1% uncertainty on the trigger efficiency is applied per [33]. A flat 2% uncertainty is applied for the lepton selection.

5.7.5 Pileup Weights

The necessity of the pileup weights were discussed in section 5.3.1. The number of pileup interactions in a single bunch crossing is given by:

$$N_i = \frac{\mathcal{L} \cdot \sigma_{\text{minimum bias}}}{v_{\text{orbit}}}, \tag{5.18}$$

where $\mathcal{L}$ is the instantaneous luminosity, $\sigma_{\text{minimum bias}}$ is the total minimum bias cross section for an event at the LHC, and v_{orbit} is the LHC orbit frequency (11246 Hz). In this calculation and the calculation of the pileup weights the minimum bias cross sections is used, but it's true value is not known.

In order to asses the effect of a systematic uncertainty due to choice of $\sigma_{\text{minimum bias}} = 69.3\,\mathrm{mb}$, a $\pm 7\%$ variation was used and the pileup weights were recalculated. Once that was done, the BDT templates were created again. As it turns out, the shape changes were negligible, but the rate changes due to this shift can be seen in table 5.17.

Process	PDF		QCD Scale					QCD Scale			
	gg	$q\bar{q}$	ggH	qqH	WH	ZH	ttH	$t\bar{t}$	V	VV	Single t
Single t											5%
Z + jets									3.4%		
Diboson										3%	
$t\bar{t}$								5.7%			
ggH	7-7.5%		7-8%								
qqH		2.6-2.8%		0.2%							
WH, ZH, ttH					1%	3.1%	3.8-9%				

Table 5.16: Uncertainties on the theoretical cross sections of the simulated signals and backgrounds.

Process	2 Jets	3 Jets	⩾4 Jets
Diboson	2-5%	3-6%	3.5-7%
$\mathrm{W}+\mathrm{jets}$	3%	4%	4%
$\mathrm{Z}+\mathrm{jets}$	7-8%	7-8%	7-8%
$\mathrm{t\bar{t}}$	2%	2%	2%
Single t	1-3%	2-8%	2-9%
Multijet	0-2%	0-3%	0-4%
ggH; $M_{\mathrm{H}} = 125\,\mathrm{GeV}$, $\mathrm{H} \rightarrow \mathrm{WW}$	2-3%	3%	3.5%
qqH; $M_{\mathrm{H}} = 125\,\mathrm{GeV}$, $\mathrm{H} \rightarrow \mathrm{WW}$	0.5-3%	1-3.5%	2.5-4%
WH, ZH, ttH; $M_{\mathrm{H}} = 125\,\mathrm{GeV}$, $\mathrm{H} \rightarrow \mathrm{WW}$	0-3%	1-3%	2-3.5%
WH, ZH, ttH; $M_{\mathrm{H}} = 125\,\mathrm{GeV}$, $\mathrm{H} \rightarrow \mathrm{ZZ}$	0.5-3%	2-4%	2-4%
WH; $M_{\mathrm{H}} = 125\,\mathrm{GeV}$, $\mathrm{H} \rightarrow \mathrm{b\bar{b}}$, $\mathrm{W} \rightarrow \mathrm{l}\nu$	0.5-3%	2-4%	3.5-4.5%
ttH; $M_{\mathrm{H}} = 125\,\mathrm{GeV}$, $\mathrm{H} \rightarrow \mathrm{b\bar{b}}$	1.5-4.5%	0-2.5%	2-4%

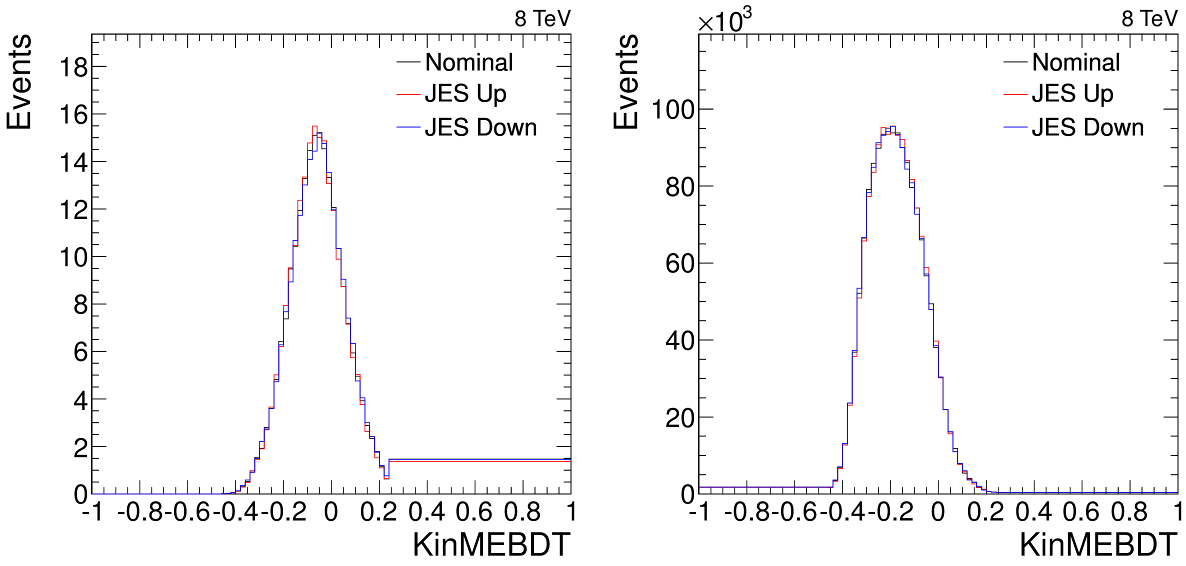
Table 5.17: Change in the expected yields due to the pileup weight uncertainties.

5.7.6 Jet Energy Scale (JES)

The jet energy corrections used to correct the jet energy scale back to the particle level were discussed in section 4.5. The uncertainty on this correction originates from several uncorrelated sources, but for simplicity we use the total combined uncertainty. For M uncorrelated sourced the total uncertainty $S\left(p_{\mathrm{T}}, \eta\right)$ is given by:

$$S\left(p_{\mathrm{T}}, \eta\right) = \sqrt{\sum_{i}^{M} s_{i}^{2}\left(p_{\mathrm{T}}, \eta\right)}, \tag{5.19}$$

where $s_i\left(p_{\mathrm{T}}, \eta\right)$ is the uncertainty for a single source i. The JES uncertainty varies as a function of p_{T} and η and is <4% in all regions of phase space [16]. To evaluate the effect this uncertainty has on the BDT discriminant we create the same distribution, but with the jet energies shifted by $\pm 1\sigma$ using the procedures given in [191, 192]. This is done before placing a cut on the p_{T} of the jets so as to allow for migration of events between jet bins. Some jets that once failed the p_{T} cut may not pass and some jets might then fail the p_{T} cut. Fig. 5.32 shows the the type of variations expected for the signal (ggH) and background ($\mathrm{W}+\mathrm{jets}$) samples. Additionally, table 5.18 lists the size of the yield uncertainty within each jet bin due to the JES.



(a) (b)

Figure 5.32: Combined kinematic and ME BDT discriminant distributions in the 2 jet, electron bin for the (a) ggH and (b) $\mathrm{W}+$jets samples. The black line shows the nominal yield while the red and blue lines show the change in shape if the JES is scales up and down by 1σ, respectively. The yields for the shifted samples are normalized to that of the nominal yield.

5.7.7 CSV Weights

Recommendation for how to treat the systematic uncertainties on the CSV weights were given by [169], which also details their derivation. In this analysis, however, the CSV weights were found to be very small and any change in them would have a negligible impact. It was decided to use a much simpler, yet conservative approach by. We overestimated the error by using weight^2 as the $+1\sigma$ variation and the unweighted distributions as the -1σ variation. The changes to the rate due to this methodology can be seen in table 5.19.

5.7.8 Top p_{T}

As discussed in section 5.3.3, the top-quark-pair cross section analyses found that the p_{T} spectrum of top quarks in data is softer than those in simulation. Thus we needed to reweight the top quark p_{T} spectrum in the $\mathrm{t\bar{t}}$ sample. In order to fully cover any uncertainty on the weights a 100% uncertainty is assumed. This means the one standard deviation up and down variations on the

Process	2 Jets	3 Jets	⩾4 Jets
Diboson	1-2%	2%	2%
$\mathrm{Z}+$jets	0-5.5%	<1%	<1%
$\mathrm{t\bar{t}}$	8-19%	4-7%	2-4%
Single t	2-0%	<1%	<1%
ggH; $M_{\mathrm{H}} = 125\,\mathrm{GeV}$, $\mathrm{H} \to \mathrm{WW}$	0-5%	0-2%	0-3%
qqH; $M_{\mathrm{H}} = 125\,\mathrm{GeV}$, $\mathrm{H} \to \mathrm{WW}$	<1%	4%	7%
WH, ZH, ttH; $M_{\mathrm{H}} = 125\,\mathrm{GeV}$, $\mathrm{H} \to \mathrm{WW}$	2-3%	0-5%	5-8%
WH, ZH, ttH; $M_{\mathrm{H}} = 125\,\mathrm{GeV}$, $\mathrm{H} \to \mathrm{ZZ}$	1.5%	0-6%	4-5%
WH; $M_{\mathrm{H}} = 125\,\mathrm{GeV}$, $\mathrm{H} \to \mathrm{b\bar{b}}$, $\mathrm{W} \to \mathrm{l}\nu$	8-9%	1-10%	2-13%
ttH; $M_{\mathrm{H}} = 125\,\mathrm{GeV}$, $\mathrm{H} \to \mathrm{b\bar{b}}$	4-17%	11-24%	18-21%

Table 5.18: Change in the expected yields due to the JES uncertainties.

Process	2 Jets	3 Jets	⩾4 Jets
Diboson	0.5-2%	1-3.5%	1-5%
$\mathrm{W}+$jets	0-3%	0-5.5%	0-8.5%
$\mathrm{Z}+$jets	2-5%	0-5.5%	2-5%
$\mathrm{t\bar{t}}$	5-11%	6-14%	6-17%
Single t	4-9%	4-12%	5-16%
ggH; $M_{\mathrm{H}} = 125\,\mathrm{GeV}$, $\mathrm{H} \to \mathrm{WW}$	1-3%	1-5%	1-7%
qqH; $M_{\mathrm{H}} = 125\,\mathrm{GeV}$, $\mathrm{H} \to \mathrm{WW}$	0-2%	1.5-2.5%	2-4%
WH, ZH, ttH; $M_{\mathrm{H}} = 125\,\mathrm{GeV}$, $\mathrm{H} \to \mathrm{WW}$	<1%	<1%	<1%
WH, ZH, ttH; $M_{\mathrm{H}} = 125\,\mathrm{GeV}$, $\mathrm{H} \to \mathrm{ZZ}$	<1%	<1%	<1%
WH; $M_{\mathrm{H}} = 125\,\mathrm{GeV}$, $\mathrm{H} \to \mathrm{b\bar{b}}$, $\mathrm{W} \to \mathrm{l}\nu$	<1%	<1%	<1%
ttH; $M_{\mathrm{H}} = 125\,\mathrm{GeV}$, $\mathrm{H} \to \mathrm{b\bar{b}}$	<1%	<1%	<1%

Table 5.19: Change in the expected yields due to the CSV weight uncertainties.

weights are taken to be:

$$+1\sigma : \; w_{\text{up}} = w_{\text{TopPt}} \cdot w_{\text{TopPt}}, \tag{5.20}$$

$$-1\sigma : \; w_{\text{down}} = 1. \tag{5.21}$$

This was the recommendation as provided by the TOP PAG [171] and results in an uncertainty of 0.5-2.1% on the $\mathrm{t\bar{t}}$ yield.

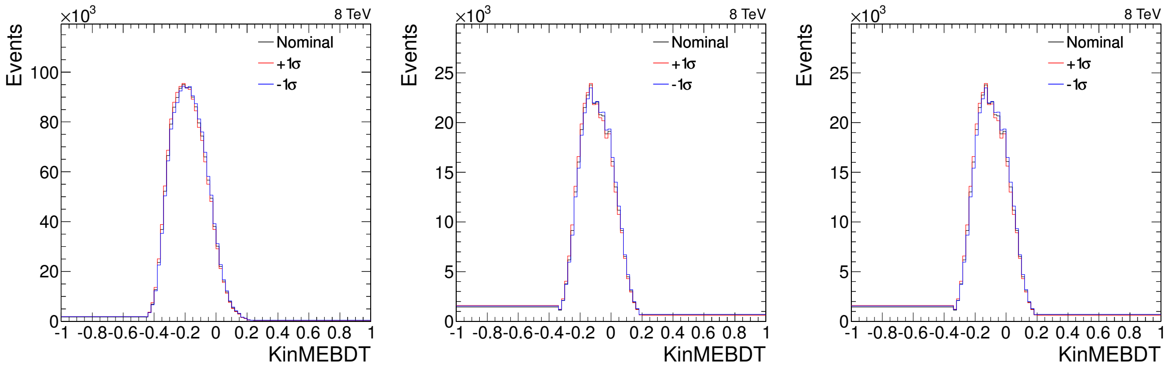
5.7.9 $\cos(\theta_l)$ Weight Uncertainty

Once again we assumed a 100% uncertainty on the $\cos(\theta_l)$weights. The one standard deviation up and down variations on the weights are taken to be:

$$+1\sigma : \; w_{\text{up}} = w_{\cos(\theta_l)} \cdot w_{\cos(\theta_l)}, \tag{5.22}$$

$$-1\sigma : \; w_{\text{down}} = 1. \tag{5.23}$$

These weights are then used as an uncertainty for the $\mathrm{W+jets}$ sample. As this is not a cut on the events and no change in selection has been made, this does not correspond to a change in the rate, only the $\mathrm{W+jets}$ shape.



(a) (b) (c)

Figure 5.33: Changes to the shape of the BDT discriminant for the $\mathrm{W+jets}$ sample due to variations on the $\cos(\theta_l)$weights for the (a) 2 jets bin, (b) 3 jet bin, and (c) ⩾4 jet bin.

5.7.10 $\mathrm{W}+$ jets Shape Uncertainties

In order to take into account variations on the Q^2 scale and matrix element parton matching new samples are generated, since these uncertainties cannot be applied after the generation stage. The samples used are listed in table 5.20. Since $\mathrm{W}+$ jets is our dominant background in all jet and lepton bins, it was deemed sufficient to apply the Q^2 and matching uncertainties only for this sample; generating new samples and/or processing existing large samples for all of the signals and backgrounds would be time consuming and would result in little to no change in the results.

The centrally produced $\mathrm{W}+$ jets events were generated using MADGRAPH, a matrix element level generator, which was then interfaced to PYTHIA to model the parton shower with its soft and collinear radiation. Because MADGRAPH generates tree-level diagrams a variation of the factorization and renormalization scales has a significant impact on the simulation. In this case the scales were varied by a factor of two.

Once the four samples listed in the table were processed, they went through the same selection and weighting procedure as the nominal $\mathrm{W}+$ jets sample. The new template histograms include only shape changes as the rate uncertainty for the nominal $\mathrm{W}+$ jets same is included in a different source.

Sample	Dataset Name	Cross Section
ME Matching Up	/WJetsToLNu_matchingup_8TeV-madgraph-tauola	37509 pb
ME Matching Down	/WJetsToLNu_matchingdown_8TeV-madgraph-tauola	37509 pb
Q^2 Scale Up	/WJetsToLNu_scaleup_8TeV-madgraph-tauola	37509 pb
Q^2 Scale Down	/WJetsToLNu_scaledown_8TeV-madgraph-tauola	37509 pb

Table 5.20: Samples used for $\mathrm{W}+$ jets systematic shape uncertainties. Each dataset name is appended with /Summer12_DR53X-PU_S10_START53_V7A-v1/AODSIM.

5.7.11 QCD η Weights Uncertainty

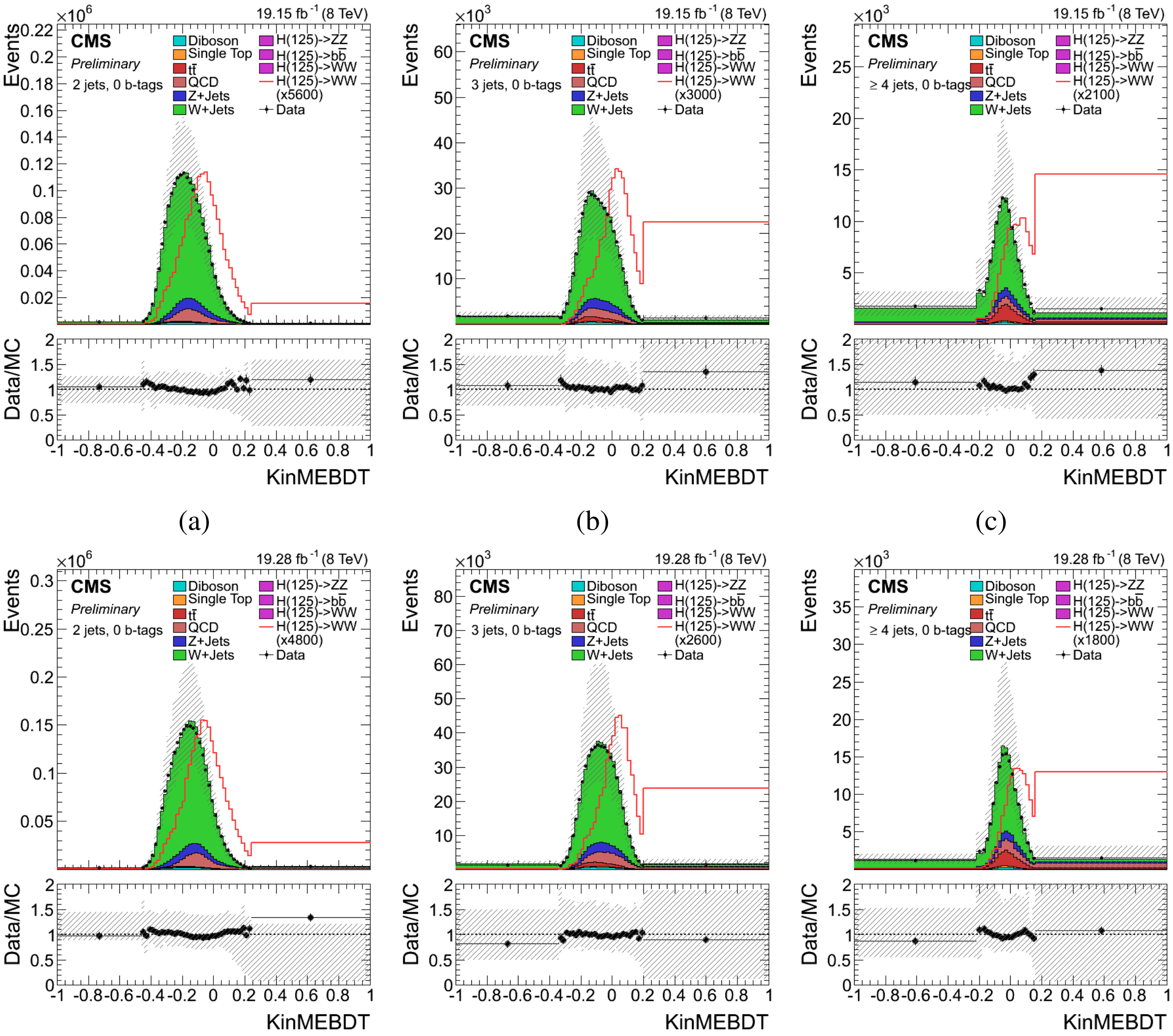
The uncertainty on the weights as a function of η for the data-driven QCD sample have to do with the choice of selection criteria, which was first discussed in section 5.1.3. The motivation for the chosen isolation windows was more practical than due to some deeper, underlying physics. Therefore the uncertainties for the weights are generated by varying the isolation criteria and creating alternate QCD samples with a modified set of events. One side of the isolation region was relaxed at a time to generate four new samples, two each for the electron and muon channels. These samples were then used to generate four new sets of weights, just as done in section 5.3.5. The resulting samples lead to a small variation in the QCD template shapes, but also lead to an uncertainty on the QCD yield of 6-30% and on the $\mathrm{W}+\mathrm{jets}$ yield of 0.1-0.5%.

6. RESULTS

The KinMEBDT distributions showing the background predictions and the observed data are shown in fig. 6.1. The Higgs signal hypothesis with $M_{\mathrm{H}} = 125\,\mathrm{GeV}$ is enlarged and shown using a red line to indicate where in the distribution the signal would lie. There is good agreement between the observed data and simulated background estimate; certainly well withing the systematic errors shown using the gray hashed areas. These distributions contain one histogram for each signal, background, and data sample. These template histograms are used as the input for our shape based limit setting procedure, each bin acting as a counting experiment, but with correlated systematics across bins.

I start by reporting an upper bound on $\sigma/\sigma_{\mathrm{SM}}$, which is the ratio of the observed cross section to the SM production cross section, at the 95% confidence level (CL), made by using the modified-frequentist limit setting method with the $\mathrm{CL_S}$ test statistic [193, 194, 195]. Although it is more rigorous to use the toy-based frequentist limit setting procedures, these methods are known to take an exceedingly long time to converge. However, when not in a low statistics regime, the toy-based methods and the asymptotic approximation return roughly equivalent answers. Therefore, we used the asymptotic approximation as we do indeed have copious amounts of background and data in our templates. The computations were done using the Higgs Combine Tool [196], which is a RooStats [197] based limit setting package. A detailed discussion on the computation of $\mathrm{CL_S}$ limits can be found in [110, 198]. The expected and observed upper limits on $\sigma/\sigma_{\mathrm{SM}}$ are shown in fig. 6.2, with the actual values listed in table 6.1.

If the sensitivity of the analysis were to increase, the expected limits (yellow and green bands) should approach and eventually cross the $\sigma/\sigma_{\mathrm{SM}} = 1$ boundary. That boundary denotes the nominal point at which we have the sensitivity to exclude the production of the Higgs boson as predicted by the Standard Model. If the particle we are searching for didn't exist, then the observed value would also cross the boundary at one and we could say the boson was excluded within the Standard Model. However, we have the benefit of of knowing the Higgs boson exists and indeed decays to



(a) (b) (c)

(d) (e) (f)

Figure 6.1: The KinMEBDT distribution in Monte Carlo (filled histograms) and data (black markers). The H → WW signal is shown by red line while the systematic uncertainties are shown by the hashed areas. The plots are ordered by jet bin from left to right, with the leftmost plot being the two-jet bin and the rightmost plot being the greater than or equal to four-jet bin. The top row contains the electron channel plots while the bottom rows contain the muon channel plots.

Category	Observed	Expected
$\geqslant$4 Jets (e)	88.0	$50.5^{+17.1}_{-13.8}$
3 Jets (e)	20.6	$18.9^{+7.5}_{-5.3}$
2 Jets (e)	7.0	$7.4^{+3.0}_{-2.1}$
$\geqslant$4 Jets (μ)	19.4	$12.6^{+5.0}_{-3.5}$
3 Jets (μ)	8.0	$9.3^{+3.7}_{-2.6}$
2 Jets (μ)	11.2	$8.8^{+3.6}_{-2.5}$
Combined	5.4	$3.4^{+1.4}_{-0.9}$

Table 6.1: Observed and median expected and 95% CLs upper limits on μ calculated with the Asymptotic CL_S method. The $\pm 1\sigma$ confidence interval is quoted for the expected limits.

WW.Therefore, as analysis get more sensitive we expect that the background-only expected bands cross the $\sigma/\sigma_{\text{SM}} = 1$ boundary, but the observed limit will lie well above this, making exclusion of the signal using upper limits impossible. At that point it will make sense to measure the result not in upper limits, used to exclude that a particular particle may exist, but to test the strength of the evidence that the alternative hypothesis (i.e. the Higgs boson exists) is valid when compared to the null hypothesis (i.e. that the Higgs boson does not exist). To test the strength of the result we compute the p-value. Under the assumption that the null hypothesis is true, the p-value is the percentage of pseudo-experiments that are at least as extreme (signal like) as what was observed. A p-value which is low means that it is highly improbable that the observation occurred due to a statistical fluctuation. As a general rule in the physics community a p-value of 0.003 is required to claim "evidence of a particle" and a p-value of 3 × 10^{-7} is needed to claim "discovery."[1] We can also convert the p-value to the significance level of the result, which is the number of standard deviations σ from the mean of the null hypothesis, and which quantifies the risk of claiming a significant result when when none exits.[2] Both of these values are listed in table 6.2. Additionally fig. 6.3 has a graphical representation of the p-values.

[1] This corresponds to a 1 in 3.5 million chance that if the Higgs does not exist we would still see a result due to background fluctuations as extreme as we did.

[2] A significance of 3σ is generally considered evidence of a new particle while a 5σ significance is required to claim the existence of a new particle.

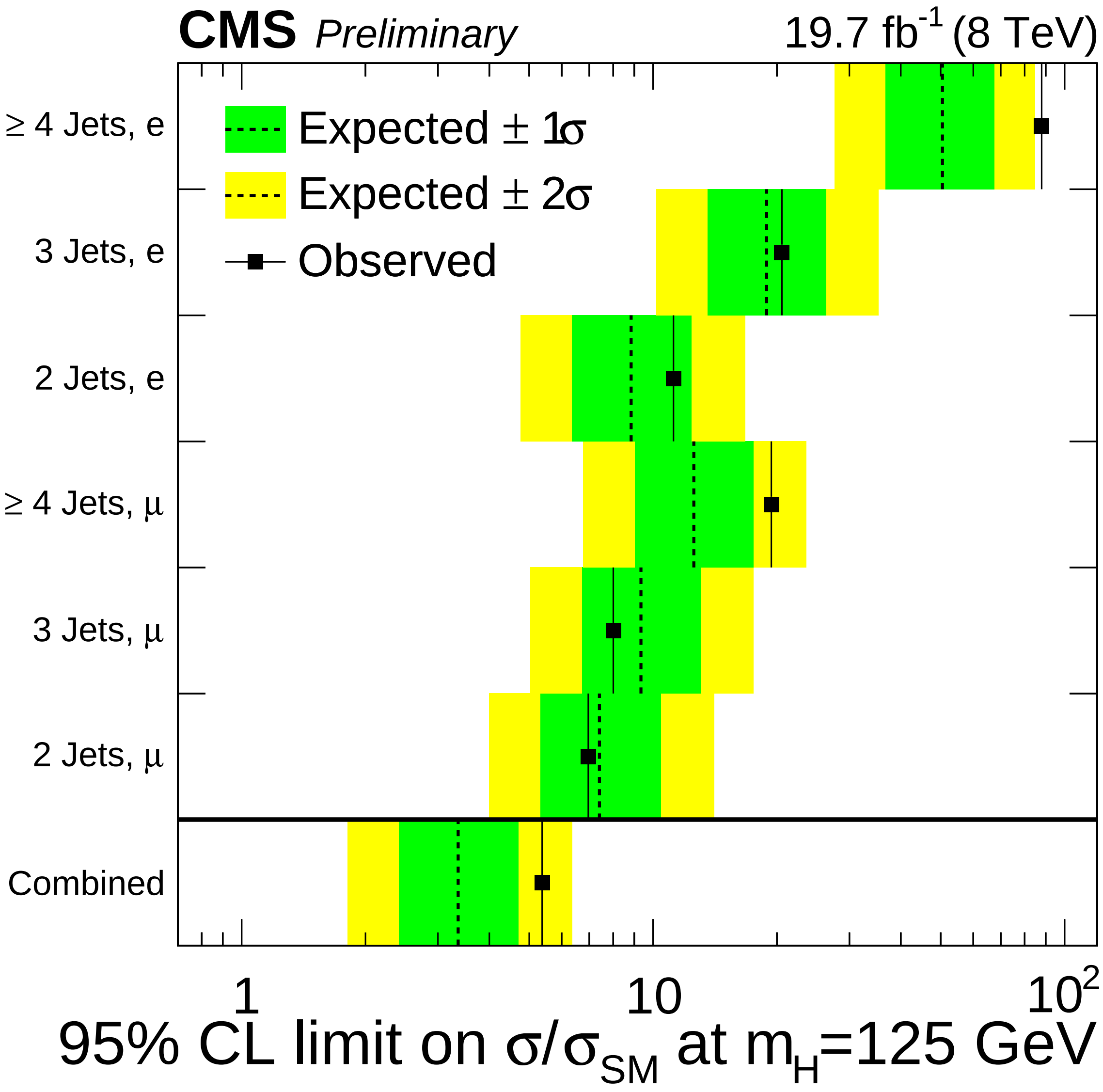


Figure 6.2: Median expected and observed 95% upper confidence level on the cross-section ratio to the expected Standard Model Higgs cross-section (μ). The green and yellow uncertainty bands represent the 68% and 95% CL intervals on the expected limit, respectively. The values were found using the Asymptotic $\mathrm{CL_S}$ approximation.

Category	A-priori Expected	A-posteriori Expected	Observed
$\geqslant$4 Jets (e)	0.045 (0.482)	0.011 (0.496)	2.647 (0.004)
3 Jets (e)	0.104 (0.459)	0.096 (0.462)	2.014 (0.022)
2 Jets (e)	0.178 (0.430)	0.191 (0.424)	0.531 (0.298)
$\geqslant$4 Jets (μ)	0.192 (0.424)	0.153 (0.439)	1.190 (0.117)
3 Jets (μ)	0.218 (0.414)	0.207 (0.418)	0.000 (0.500)
2 Jets (μ)	0.208 (0.418)	0.195 (0.423)	0.000 (0.500)
Combined	0.569 (0.268)	0.547 (0.292)	0.903 (0.183)

Table 6.2: Expected and observed statistical significances as well as their associated p-values. The a-priori expected significances are computed before the background fits to the data. For the two and three jet muon bins the significance is zero because the minimum of the likelihood is for a signal strength $\leqslant 0$.

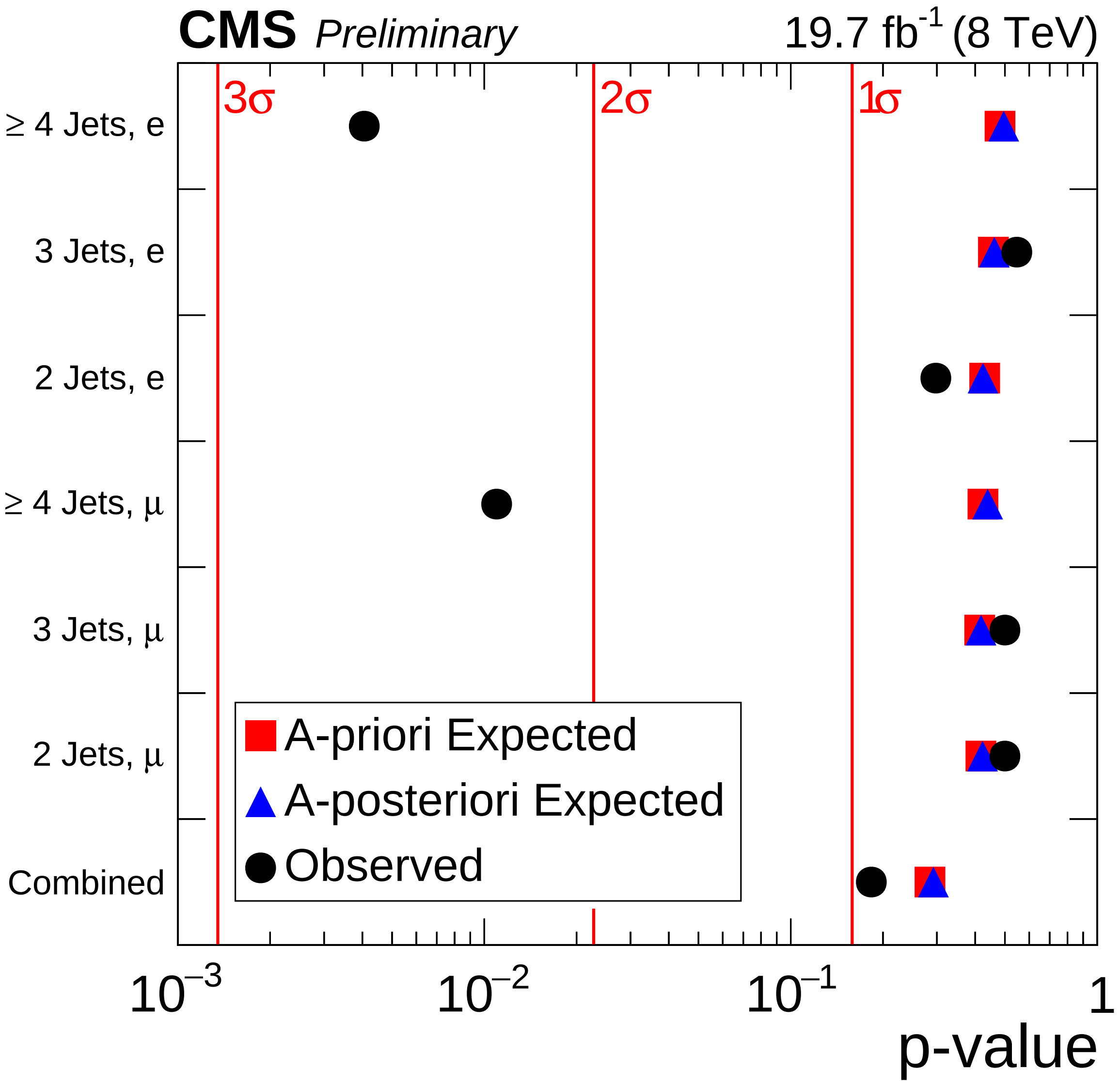


Figure 6.3: The a-priori expected (red square), a-posteriori (blue triangle), and observed (black circle) p-values found in each category.

7. CONCLUSIONS & FUTURE POSSIBILITIES

This dissertation has presented a search for the 125 GeV Standard Model Higgs boson in th the $\mathrm{H} \rightarrow \mathrm{WW} \rightarrow \mathrm{l}\nu\mathrm{jj}$ decay channel. The search used 19.7 fb^{-1} of 8 TeV proton-proton collision data from the CMS experiment collected during the 2012 run of the LHC. The background predictions used in the analysis were derived from both simulation and data-driven techniques and a significant amount of time was put into validating the background modeling. The event selection was chosen based on the signal kinematics, but kept relatively loose to ensure enough of a training ensemble for an analysis using a boosted decision tree (BDT). In addition to a BDT based discriminant, a matrix element method was used to increase the sensitivity of the analysis. No direct observation of the Standard Model Higgs boson can be made at this time in this particular channel, though limits on its production cross section have been made at the 95% confidence level using a modified frequentist approach. A limit of 5.4 times the standard model cross section was set after combining all lepton and jet categories. This limit is the first to be set in the $\mathrm{H} \rightarrow \mathrm{WW} \rightarrow \mathrm{l}\nu\mathrm{jj}$ channel for a Higgs mass of $M_{\mathrm{H}} = 125$ GeV at either the CMS or ATLAS experiments.

Already Run 2 of the LHC has collected significantly more data at the higher center of mass energy of $\sqrt{\mathrm{s}} = 13$ TeV, enabling a future version of this analysis to have significantly greater sensitivity. This increase in energy corresponds to an increase in gluon-gluon fusion Higgs production of approximately 2.4 times [199], while the production cross section of the main background, $\mathrm{W} + \mathrm{jets}$, will only increase by approximately 1.7 times [200]. This will lead to a higher signal fraction, which should be visible given improvements in background modeling and reconstruction techniques. Even now there have been advances in high performance computing which will reduce the time to perform a matrix element analysis by orders of magnitude [201]. Additionally, advances in machine learning will significantly speed up analyses relying on Monte Carlo integration techniques [202].

This analysis now serves as a benchmark for future $\mathrm{H} \rightarrow \mathrm{WW} \rightarrow \mathrm{l}\nu\mathrm{jj}$ analyses and also shows how a matrix element method can be successfully implemented in semi-leptonic channels. While the time investment in performing a similar analysis in the future is large, the benefits of increased discrimination by using ever more advanced analysis techniques could be well worth the wait. I am optimistic that even more stringent measurements of this Higgs decay channel can and will be made in the coming years.

REFERENCES

[1] E. Drexler, "Elementary particle interactions in the standard model." http://creativecommons.org/publicdomain/zero/1.0/deed.enCreative Commons CC0 1.0 Universal Public Domain Dedication, May 2014. http://en.wikipedia.org/wiki/File:Elementary_particle_interactions_in_the_Standard_Model.png.

[2] C. Patrignani *et al.*, "Review of Particle Physics," *Chin. Phys.*, vol. C40, no. 10, p. 100001, 2016.

[3] F. Tanedo, "Why do we expect a Higgs boson? Part I: Electroweak Symmetry Breaking." https://tinyurl.com/ybf3yms7.

[4] K. Andersen and T. Eberle, "The genesis 2.0 project." http://media.vanityfair.com/photos/54cbf6ad1ca1cf0a23ac6c85/master/w_690,c_limit/image.jpg, 2010. Online; accessed November 29, 2016.

[5] F. Marcastel, "CERN's Accelerator Complex. La chaîne des accélérateurs du CERN," Oct 2013. General Photo.

[6] J.-L. Caron, "The LHC injection complex." http://cds.cern.ch/record/841568, May 1993. AC Collection. Legacy of AC. Pictures from 1992 to 2002.

[7] J.-L. Caron, "LHC Layout.. Schema general du LHC.." http://cds.cern.ch/record/841573, Sep 1997. AC Collection. Legacy of AC. Pictures from 1992 to 2002.

[8] S. Dailler, "LHC Dipole." http://cds.cern.ch/record/842253, July 1998. AC Collection. Legacy of AC. Pictures from 1992 to 2002.

[9] "Public CMS Luminosity Information." https://twiki.cern.ch/twiki/bin/view/CMSPublic/LumiPublicResults, Jan 2017.

[10] "SketchUpCMS Gallery." https://twiki.cern.ch/twiki/bin/view/CMSPublic/SketchUpCMSGallery?redirectedfrom=CMS.SketchUpCMSGallery, Jan 2017.

[11] M. Schott and M. Dunford, "Review of single vector boson production in pp collisions at $\sqrt{s} = 7$ TeV," *Eur. Phys. J.*, vol. C74, p. 2916, 2014.

[12] S. Chatrchyan *et al.*, "The CMS experiment at the CERN LHC," *JINST*, vol. 3, p. S08004, 2008.

[13] "The CMS Detector and the Token Bit Manager." https://www.phys.ksu.edu/reu2014/wabehn/, Jan 2017.

[14] CMS Collaboration, "Particle-flow event reconstruction in cms and performance for jets, taus, and met," CMS Physics Analysis Summary CMS-PAS-PFT-09-001, CERN, 2009.

[15] B. Dorney, "Anatomy of a Jet in CMS," 2014.

[16] V. Khachatryan *et al.*, "Jet energy scale and resolution in the CMS experiment in pp collisions at 8 TeV," *JINST*, vol. 12, p. P02014. 92 p, Jul 2016. Replaced with the published version. Added the journal reference and DOI. All the figures and tables can be found at http://cms-results.web.cern.ch/cms-results/public-results/publications/JME-13-004/.

[17] M. Cacciari, G. P. Salam, and G. Soyez, "The anti-k_t jet clustering algorithm," *JHEP*, vol. 04, p. 063, 2008.

[18] "Performance of b tagging at $\sqrt{s} = 8$ TeV in multijet, ttbar and boosted topology events," CMS Physics Analysis Summary CMS-PAS-BTV-13-001, CERN, Geneva, 2013.

[19] "MET Analysis." https://twiki.cern.ch/twiki/bin/view/CMSPublic/WorkBookMetAnalysis, 2017.

[20] F. Siegert, *Monte-Carlo event generation for the LHC.* PhD thesis, Durham U., 2010.

[21] A. D. Martin, W. J. Stirling, R. S. Thorne, and G. Watt, "Parton distributions for the LHC," *Eur. Phys. J.*, vol. C63, pp. 189–285, 2009.

[22] "Public CMS Data Quality Information ." `https://twiki.cern.ch/twiki/bin/view/CMSPublic/DataQuality`, Jun 2013.

[23] Y. Gao, A. V. Gritsan, Z. Guo, K. Melnikov, M. Schulze, and N. V. Tran, "Spin determination of single-produced resonances at hadron colliders," *Phys. Rev. D*, vol. 81, p. 075022, Apr 2010.

[24] P. J. Dong, *Measurement of Electroweak Single Top Quark Production in Proton-Antiproton Collisions at* $1.96\,TeV$. PhD thesis, University of California, Los Angeles, 2008.

[25] T. Aaltonen *et al.*, "Higgs boson studies at the tevatron," *Phys. Rev. D*, vol. 88, p. 052014, Sep 2013.

[26] S. Chatrchyan *et al.*, "Observation of a new boson at a mass of 125 gev with the cms experiment at the lhc," *Phys. Lett. B*, vol. 716, no. 1, pp. 30 – 61, 2012.

[27] V. Khachatryan *et al.*, "Precise determination of the mass of the higgs boson and tests of compatibility of its couplings with the standard model predictions using proton collisions at 7 and 8 tev," *Eur. Phys. J. C*, vol. 75, p. 212, May 2015.

[28] M. Baak *et al.*, "The global electroweak fit at NNLO and prospects for the LHC and ILC," *Eur. Phys. J.*, vol. C74, p. 3046, 2014.

[29] D. Giordano and G. Sguazzoni, "Cms reconstruction improvements for the tracking in large pile-up events," *J. Phys. Conf.*, vol. 396, no. 2, p. 022044, 2012.

[30] G. Aad *et al.*, "Observation of a new particle in the search for the standard model higgs boson with the atlas detector at the lhc," *Phys. Lett. B*, vol. 716, no. 1, pp. 1 – 29, 2012.

[31] G. Aad *et al.*, "Combined Measurement of the Higgs Boson Mass in pp Collisions at $\sqrt{s} = 7$ and 8 TeV with the ATLAS and CMS Experiments," *Phys. Rev. Lett.*, vol. 114, p. 191803, 2015.

[32] "Combination of standard model Higgs boson searches and measurements of the properties of the new boson with a mass near 125 GeV," Tech. Rep. CMS-PAS-HIG-13-005, CERN, Geneva, 2013.

[33] "Search for the Standard Model Higgs boson in the H to WW to lnujj decay channel in pp collisions at the LHC," Tech. Rep. CMS-PAS-HIG-13-027, CERN, Geneva, 2012.

[34] H. Weyl, "Elektron und gravitation. i," *Zeitschrift für Physik*, vol. 56, pp. 330–352, May 1929.

[35] C. N. Yang and R. L. Mills, "Conservation of isotopic spin and isotopic gauge invariance," *Phys. Rev.*, vol. 96, pp. 191–195, Oct 1954.

[36] S. L. Glashow, "Partial-symmetries of weak interactions," *Nuclear Physics*, vol. 22, no. 4, pp. 579 – 588, 1961.

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

[38] A. Salam, "Weak and electromagnetic interactions," in *Elementary particle theory* (N. Svartholm, ed.), pp. 367–377, Almquist & Wiksell.

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

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

[41] P. W. Higgs, "Spontaneous Symmetry Breakdown without Massless Bosons," *Phys. Rev.*, vol. 145, pp. 1156–1163, 1966.

[42] 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.

[43] T. W. B. Kibble, "Symmetry breaking in nonAbelian gauge theories," *Phys. Rev.*, vol. 155, pp. 1554–1561, 1967.

[44] Y. Ne'eman, "Derivation of strong interactions from a gauge invariance," *Nucl. Phys.*, vol. 26, pp. 222–229, 1961.

[45] M. Gell-Mann, "Symmetries of baryons and mesons," *Phys. Rev.*, vol. 125, pp. 1067–1084, 1962.

[46] M. Gell-Mann, "A Schematic Model of Baryons and Mesons," *Phys. Lett.*, vol. 8, pp. 214–215, 1964.

[47] G. Zweig, "An SU(3) model for strong interaction symmetry and its breaking. Version 1," 1964.

[48] H. Fritzsch and M. Gell-Mann, "Current algebra: Quarks and what else?," *eConf*, vol. C720906V2, pp. 135–165, 1972.

[49] O. W. Greenberg, "Spin and unitary-spin independence in a paraquark model of baryons and mesons," *Phys. Rev. Lett.*, vol. 13, pp. 598–602, Nov 1964.

[50] C. Burgess and G. Moore, *The Standard Model: A Primer*. Cambridge University Press, 2007.

[51] I. J. R. Aitchison and A. J. Hey, *Gauge Theories in Particle Physics: A Practical Introduction: From Relativistic Quantum Mechanics to QED*, vol. 2. Taylor & Francis, fourth ed., 2012.

[52] K. J. Barnes, *Group theory for the standard model of particle physics and beyond*. 2010.

[53] S. Dawson, "Introduction to electroweak symmetry breaking," in *Proceedings, Summer School in High-energy physics and cosmology: Trieste, Italy, June 29-July 17, 1998*, pp. 1–83, 1998.

[54] M. Maltoni, T. Schwetz, M. A. Tortola, and J. W. F. Valle, "Status of global fits to neutrino oscillations," *New J. Phys.*, vol. 6, p. 122, 2004.

[55] M. Kobayashi and T. Maskawa, "CP Violation in the Renormalizable Theory of Weak Interaction," *Prog. Theor. Phys.*, vol. 49, pp. 652–657, 1973.

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

[57] H. J. Rothe, *Lattice gauge theories: An Introduction*, vol. 74. World Sci. Lect. Notes Phys., 2005.

[58] G. Arnison *et al.*, “Experimental observation of isolated large transverse energy electrons with associated missing energy at s=540 gev,” *Physics Letters B*, vol. 122, no. 1, pp. 103 – 116, 1983.

[59] G. Arnison *et al.*, “Experimental observation of lepton pairs of invariant mass around 95 gev/c2 at the cern sps collider,” *Phys. Lett. B*, vol. 126, no. 5, pp. 398 – 410, 1983.

[60] K. A. Olive *et al.*, “The review of particle physics,” *Chin. Phys. C*, vol. 38, p. 090001, 2014.

[61] P. van Nieuwenhuizen, “Supergravity,” *Physics Reports*, vol. 68, no. 4, pp. 189 – 398, 1981.

[62] D. Z. Freedman and A. V. Proeyen, *Supergravity*. Cambridge University Press, 2012.

[63] H. Nastase, “Introduction to Supergravity,” 2011.

[64] P. A. R. Ade *et al.*, “Planck 2015 results. XIII. Cosmological parameters,” *Astron. Astrophys.*, vol. 594, p. A13, 2016.

[65] D. Clowe, M. Bradac, A. H. Gonzalez, M. Markevitch, S. W. Randall, C. Jones, and D. Zaritsky, “A direct empirical proof of the existence of dark matter,” *Astrophys. J.*, vol. 648, pp. L109–L113, 2006.

[66] D. E. Morrissey, T. Plehn, and T. M. Tait, “Physics searches at the LHC,” *Phys. Rept.*, vol. 515, p. 1, 2012.

[67] K. Garrett and G. Duda, “Dark Matter: A Primer,” *Adv. Astron.*, vol. 2011, p. 968283, 2011.

[68] G. Bertone, D. Hooper, and J. Silk, “Particle dark matter: Evidence, candidates and constraints,” *Phys. Rept.*, vol. 405, pp. 279–390, 2005.

[69] A. D. Sakharov, “Violation of cp invariance, c asymmetry, and baryon asymmetry of the universe,” *Soviet Physics Uspekhi*, vol. 34, no. 5, p. 392, 1991.

[70] V. Kuzmin, V. Rubakov, and M. Shaposhnikov, “On anomalous electroweak baryon-number non-conservation in the early universe,” *Phys. Lett. B*, vol. 155, no. 1, pp. 36 – 42, 1985.

[71] S. Abe *et al.*, “Precision Measurement of Neutrino Oscillation Parameters with KamLAND,” *Phys. Rev. Lett.*, vol. 100, p. 221803, 2008.

[72] K. Abe *et al.*, “Precise Measurement of the Neutrino Mixing Parameter θ_{23} from Muon Neutrino Disappearance in an Off-Axis Beam,” *Phys. Rev. Lett.*, vol. 112, no. 18, p. 181801, 2014.

[73] N. Agafonova *et al.*, “Observation of tau neutrino appearance in the CNGS beam with the OPERA experiment,” *PTEP*, vol. 2014, no. 10, p. 101C01, 2014.

[74] K. A. Olive *et al.*, “Review of Particle Physics,” *Chin. Phys.*, vol. C38, p. 090001, 2014.

[75] Y. Fukuda *et al.*, “Measurements of the solar neutrino flux from super-kamiokande’s first 300 days,” *Phys. Rev. Lett.*, vol. 81, pp. 1158–1162, Aug 1998.

[76] D. Decamp *et al.*, “A precise determination of the number of families with light neutrinos and of the z boson partial widths,” *Physics Letters B*, vol. 235, no. 3, pp. 399 – 411, 1990.

[77] L. Susskind, “The gauge hierarchy problem, technicolor, supersymmetry, and all that,” *Physics Reports*, vol. 104, no. 2, pp. 181 – 193, 1984.

[78] H.-C. Cheng, “Little Higgs, Non-standard Higgs, No Higgs and All That,” in *SUSY 2007 Proceedings, 15th International Conference on Supersymmetry and Unification of Fundamental Interactions, July 26 - August 1, 2007, Karlsruhe, Germany*, pp. 114–121, 2007.

[79] J. Reuter and M. Tonini, “Can the 125 GeV Higgs be the Little Higgs?,” *JHEP*, vol. 02, p. 077, 2013.

[80] M. Schmaltz and D. Tucker-Smith, “Little Higgs review,” *Ann. Rev. Nucl. Part. Sci.*, vol. 55, pp. 229–270, 2005.

[81] N. Arkani-Hamed, S. Dimopoulos, and G. R. Dvali, “The Hierarchy problem and new dimensions at a millimeter,” *Phys. Lett.*, vol. B429, pp. 263–272, 1998.

[82] H. Miyazawa, “Baryon number changing currents*,” *Progress of Theoretical Physics*, vol. 36, no. 6, pp. 1266–1276, 1966.

[83] H. Miyazawa, “Spinor currents and symmetries of baryons and mesons,” *Phys. Rev.*, vol. 170, pp. 1586–1590, Jun 1968.

[84] J. Wess and B. Zumino, “Supergauge transformations in four dimensions,” *Nucl. Phys. B*, vol. 70, no. 1, pp. 39 – 50, 1974.

[85] Yu. A. Golfand and E. P. Likhtman, “Extension of the Algebra of Poincare Group Generators and Violation of p Invariance,” *JETP Lett.*, vol. 13, pp. 323–326, 1971. [Pisma Zh. Eksp. Teor. Fiz.13,452(1971)].

[86] A. H. Chamseddine, R. Arnowitt, and P. Nath, “Locally supersymmetric grand unification,” *Phys. Rev. Lett.*, vol. 49, pp. 970–974, Oct 1982.

[87] G. L. Kane, C. Kolda, L. Roszkowski, and J. D. Wells, “Study of constrained minimal supersymmetry,” *Phys. Rev. D*, vol. 49, pp. 6173–6210, Jun 1994.

[88] P. Fayet, “Supergauge invariant extension of the higgs mechanism and a model for the electron and its neutrino,” *Nucl. Phys. B*, vol. 90, pp. 104 – 124, 1975.

[89] R. Barbieri, S. Ferrara, and C. Savoy, “Gauge models with spontaneously broken local supersymmetry,” *Phys. Lett. B*, vol. 119, no. 4, pp. 343 – 347, 1982.

[90] L. Hall, J. Lykken, and S. Weinberg, “Supergravity as the messenger of supersymmetry breaking,” *Phys. Rev. D*, vol. 27, pp. 2359–2378, May 1983.

[91] S. P. Martin, “A Supersymmetry primer,” 1997. [Adv. Ser. Direct. High Energy Phys.18,1(1998)].

[92] A. Breskin and R. Voss, *The CERN Large Hadron Collider: Accelerator and Experiments*. Geneva: CERN, 2009.

[93] L. Evans and P. Bryant, “LHC Machine,” *JINST*, vol. 3, p. S08001, 2008.

[94] G. Aad *et al.*, "The ATLAS Experiment at the CERN Large Hadron Collider," *Journal of Instrumentation*, vol. 3, no. 08, p. S08003, 2008.

[95] A. A. Alves, Jr. *et al.*, "The LHCb Detector at the LHC," *JINST*, vol. 3, p. S08005, 2008.

[96] K. Aamodt *et al.*, "The ALICE experiment at the CERN LHC," *JINST*, vol. 3, p. S08002, 2008.

[97] M. Lamont, "Status of the lhc," *Journal of Physics: Conference Series*, vol. 455, no. 1, p. 012001, 2013.

[98] "CMS Web Based Monitoring." `https://cmswbm.web.cern.ch/cmswbm/`, Jan 2017.

[99] S. Chatrchyan *et al.*, "Description and performance of track and primary-vertex reconstruction with the CMS tracker," *JINST*, vol. 9, p. P10009, 2014.

[100] V. Veszpremi, "Operation and performance of the CMS tracker," *JINST*, vol. 9, p. C03005, 2014.

[101] "2012 ECAL detector performance plots," CMS Detector Performance Summary CMS-DP-2013-007, CERN, 2013.

[102] V. Khachatryan *et al.*, "Performance of photon reconstruction and identification with the CMS detector in proton-proton collisions at $\sqrt{s} = 8\,\mathrm{TeV}$," *JINST*, vol. 10, p. P08010, 2015.

[103] S. Chatrchyan *et al.*, "Performance of CMS muon reconstruction in pp collision events at $\sqrt{s} = 7\,\mathrm{TeV}$," *JINST*, vol. 7, p. P10002, 2012.

[104] V. Khachatryan *et al.*, "The CMS trigger system." 2016.

[105] J. Brooke, "Performance of the CMS Level-1 Trigger," *PoS*, vol. ICHEP2012, p. 508, 2013.

[106] D. Trocino, "The CMS High Level Trigger," *J. Phys. Conf.*, vol. 513, p. 012036, 2014.

[107] "CMS Luminosity Based on Pixel Cluster Counting - Summer 2012 Update," CMS Physics Analysis Summary CMS-PAS-LUM-12-001, CERN, Geneva, 2012.

[108] “CMS Luminosity Based on Pixel Cluster Counting - Summer 2013 Update,” CMS Physics Analysis Summary CMS-PAS-LUM-13-001, CERN, Geneva, 2013.

[109] V. Balagura, “Notes on van der Meer Scan for Absolute Luminosity Measurement,” *Nucl. Instrum. Meth. A*, vol. 654, p. 634, 2011.

[110] K. J. Pedro, *Search for Pair Production of Third-Generation Scalar Leptoquarks and R-Parity Violating Top Squarks in Proton-Proton Collisions at $\sqrt{s} = 8\,TeV$.* PhD thesis, University of Maryland, 2014.

[111] CMS Collaboration, “CMS Physics: Technical Design Report Volume 1: Detector Performance and Software,” Tech. Rep. CMS-TDR-8-1, CERN, Geneva, 2006.

[112] K. Rose, “Deterministic annealing for clustering, compression, classification, regression, and related optimization problems,” in *Proceedings of the IEEE*, vol. 86, p. 2210, IEEE, 1998.

[113] “Tracking and Primary Vertex Results in First 7 TeV Collisions,” CMS Physics Analysis Summary CMS-PAS-TRK-10-005, CERN, Geneva, 2010.

[114] CMS Collaboration, “Commissioning of the particle-flow event reconstruction with the first LHC collisions recorded in the CMS detector,” CMS Physics Analysis Summary CMS-PAS-PFT-10-001, 2010.

[115] CMS Collaboration, “Commissioning of the particle-flow reconstruction in minimum-bias and jet events from pp collisions at 7 TeV,” CMS Physics Analysis Summary CMS-PAS-PFT-10-002, CERN, 2010.

[116] CMS Collaboration, “Commissioning of the particle-flow event reconstruction with leptons from J/ψ and W decays at 7 TeV,” CMS Physics Analysis Summary CMS-PAS-PFT-10-003, CERN, 2010.

[117] F. Beaudette, “The CMS Particle Flow Algorithm,” in *Proceedings, International Conference on Calorimetry for the High Energy Frontier (CHEF 2013)*, (Paris), p. 295, École Polytechnique, 2013.

[118] A. M. Sirunyan *et al.*, "Particle-flow reconstruction and global event description with the CMS detector," 2017.

[119] W. Adam, R. Frühwirth, A. Strandlie, and T. Todorov, "Reconstruction of electrons with the gaussian-sum filter in the cms tracker at the lhc," *J. Phys. G*, vol. 31, p. N9, 2005.

[120] C. Collaboration, "Electron reconstruction and identification at sqrt(s) = 7 TeV," CMS Physics Analysis Summary CMS-PAS-EGM-10-004, CERN, Geneva, 2010.

[121] S. Baffioni, C. Charlot, F. Ferri, D. Futyan, P. Meridiani, I. Puljak, C. Rovelli, R. Salerno, and Y. Sirois, "Electron reconstruction in cms," *Eur. Phys. J. C*, vol. 49, no. 4, p. 1099, 2007.

[122] V. Khachatryan *et al.*, "Performance of electron reconstruction and selection with the CMS detector in proton-proton collisions at $\sqrt{s} = 8\,\text{TeV}$," *JINST*, vol. 10, p. P06005, 2015.

[123] "Cut Based Electron ID." `https://twiki.cern.ch/twiki/bin/viewauth/CMS/EgammaCutBasedIdentification`, 2013.

[124] "Multivariate Electron Identification." `https://twiki.cern.ch/twiki/bin/view/CMS/MultivariateElectronIdentification`, 2012.

[125] "Tools for conversion rejection (electron ID) and electron vetoing (photon ID)." `https://twiki.cern.ch/twiki/bin/view/CMS/ConversionTools`, 2012.

[126] "SW Guide Egamma Shower Shape." `https://twiki.cern.ch/twiki/bin/view/CMSPublic/SWGuideEgammaShowerShape`, 2011.

[127] C. Collaboration, "Performance of muon identification in pp collisions at $\sqrt{s}$ = 7 TeV," CMS Physics Analysis Summary CMS-PAS-MUO-10-002, CERN, Geneva, 2010.

[128] G. P. Salam, "Towards jetography," *Eur. Phys. J. C*, vol. 67, p. 637, 2010.

[129] M. Cacciari, G. P. Salam, and G. Soyez, "FastJet user manual," *Eur. Phys. J. C*, vol. 72, p. 1896, 2012.

[130] M. Cacciari, G. P. Salam, and G. Soyez, "The catchment area of jets," *Journal of High Energy Physics*, 2008.

[131] The CMS Collaboration, "Determination of jet energy calibration and transverse momentum resolution in cms," *Journal of Instrumentation*, 2011. http://iopscience.iop.org/1748-0221/6/11/P11002/.

[132] M. Cacciari and G. P. Salam, "Pileup subtraction using jet areas," *Physics Letters B*, 2008.

[133] "Jet Energy Resolution." `https://twiki.cern.ch/twiki/bin/viewauth/CMS/JetResolution`, 2017.

[134] N. Saoulidou, "Particle flow jet identification criteria," CMS Analysis Note CMS-AN-2010-003, CERN, Geneva, Jun 2010.

[135] "Jet identification." `https://twiki.cern.ch/twiki/bin/viewauth/CMS/JetID`, 2012.

[136] J. D. Bjorken, "Properties of hadron distributions in reactions containing very heavy quarks," *Phys. Rev. D*, vol. 17, pp. 171–173, Jan 1978.

[137] C. Weiser, "A Combined Secondary Vertex Based B-Tagging Algorithm in CMS," Tech. Rep. CMS-NOTE-2006-014, CERN, Geneva, Jan 2006.

[138] S. Chatrchyan *et al.*, "Identification of b-quark jets with the CMS experiment," *JINST*, vol. 8, p. P04013, 2013.

[139] "Performance of the missing transverse energy reconstruction by the cms experiment in $\sqrt{s} = 8$ tev pp data." Submitted to *JINST.*, 2014.

[140] "Missing transverse energy performance of the cms detector," *JINST*, vol. 6, p. P09001, 2011.

[141] M. L. Mangano and T. J. Stelzer, "Tools for the simulation of hard hadronic collisions," *Ann. Rev. Nucl. Part. Sci.*, vol. 55, pp. 555–588, 2005.

[142] M. A. Dobbs *et al.*, "Les Houches guidebook to Monte Carlo generators for hadron collider physics," in *Physics at TeV colliders. Proceedings, Workshop, Les Houches, France, May 26-June 3, 2003*, pp. 411–459, 2004.

[143] D. Griffiths, *Introductionto Elementary Particles*. Weinheim, Germany: Wiley-VCH, second, revised ed., 2008.

[144] Y. Kurihara, J. Fujimoto, T. Ishikawa, K. Kato, S. Kawabata, T. Munehisa, and H. Tanaka, “QCD event generators with next-to-leading order matrix elements and parton showers,” *Nucl. Phys.*, vol. B654, pp. 301–319, 2003.

[145] K. Binder, *Monte-Carlo Methods*, pp. 249–280. Wiley-VCH Verlag GmbH & Co. KGaA, 2006.

[146] B. Webber, “A QCD model for jet fragmentation including soft gluon interference,” *Nucl. Phys. B*, vol. 238, no. 3, pp. 492 – 528, 1984.

[147] J.-C. Winter, F. Krauss, and G. Soff, “A Modified cluster hadronization model,” *Eur. Phys. J.*, vol. C36, pp. 381–395, 2004.

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

[149] B. Andersson and M. Ringner, “Bose-Einstein correlations in the Lund model,” *Nucl. Phys.*, vol. B513, pp. 627–644, 1998.

[150] T. Sjöstrand, S. Mrenna, and P. Skands, “Pythia 6.4 physics and manual,” *JHEP*, vol. 2006, no. 05, p. 026, 2006.

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

[152] P. Nason, “A New method for combining NLO QCD with shower Monte Carlo algorithms,” *JHEP*, vol. 11, p. 040, 2004.

[153] S. Alioli, P. Nason, C. Oleari, and E. Re, "A general framework for implementing NLO calculations in shower Monte Carlo programs: the POWHEG BOX," *JHEP*, vol. 06, p. 043, 2010.

[154] Z. Wąs, "Tauola the library for τ lepton decay, and kkmc/koralb/koralz/$\ldots$ status report," *Nucl. Phys. B - Proceedings Supplements*, vol. 98, no. 1, pp. 96 – 102, 2001.

[155] S. Agostinelli *et al.*, "Geant4 a simulation toolkit," *Nucl. Instrum. Meth. A*, vol. 506, no. 3, pp. 250 – 303, 2003.

[156] J. Allison *et al.*, "Geant4 developments and applications," *IEEE Trans. Nucl. Sci.*, vol. 53, pp. 270 –278, Febuary 2006.

[157] "Recommended jet energy corrections and uncertainties for data and mc." `https://twiki.cern.ch/twiki/bin/view/CMS/JECDataMC`, 2017.

[158] R. Brun and F. Rademakers, "ROOT – an object oriented data analysis framework," *Nucl. Instrum. Meth. A*, vol. 389, p. 81, 1997. See also `http://root.cern.ch`.

[159] S. Frixione, P. Nason, and C. Oleari, "Matching NLO QCD computations with parton shower simulations: the POWHEG method," *JHEP*, vol. 11, p. 070, 2007.

[160] S. Alioli, P. Nason, C. Oleari, and E. Re, "NLO single-top production matched with shower in POWHEG: s- and t-channel contributions," *JHEP*, vol. 09, p. 111, 2009. [Erratum: doi:10.1007/JHEP02(2010)011].

[161] E. Re, "Single-top Wt-channel production matched with parton showers using the POWHEG method," *Eur. Phys. J. C*, vol. 71, p. 1547, 2011.

[162] N. Kidonakis, "Differential and total cross sections for top pair and single top production," in *Proc. of XX Int. Workshop on Deep-Inelastic Scattering and Related Subjects*, p. 831, 2012.

[163] K. Melnikov and F. Petriello, "Electroweak gauge boson production at hadron colliders through $\mathcal{O}(\alpha_s^2)$," *Phys. Rev. D*, vol. 74, p. 114017, 2006.

[164] J. M. Campbell, R. K. Ellis, and C. Williams, "Vector boson pair production at the LHC," *JHEP*, vol. 07, p. 018, 2011.

[165] J. R. Andersen *et al.*, "Handbook of LHC Higgs Cross Sections: 3. Higgs Properties," 2013.

[166] S. Chatrchyan *et al.*, "Search for the standard model higgs boson produced in association with a w or a z boson and decaying to bottom quarks," *Phys. Rev. D*, vol. 89, p. 012003, Jan 2014.

[167] "Pileup Studies." `https://twiki.cern.ch/twiki/bin/viewauth/CMS/PileupInformation`, 2017.

[168] "Utilities for Accessing Pileup Information for Data." `https://twiki.cern.ch/twiki/bin/view/CMS/PileupJSONFileforData`, 2017.

[169] N. Bartosik *et al.*, "Calibration of the Combined Secondary Vertex b-Tagging discriminant using dileptonic ttbar and Drell-Yan events," Tech. Rep. CMS-NOTE-2013-130, CERN, Geneva, Nov 2013.

[170] S. Chatrchyan *et al.*, "Measurement of differential top-quark pair production cross sections in pp colisions at $\sqrt{s} = 7$ TeV," *Eur. Phys. J.*, vol. C73, no. 3, p. 2339, 2013.

[171] "pt(top-quark) based reweighting of ttbar MC." `https://twiki.cern.ch/twiki/bin/viewauth/CMS/TopPtReweighting`, 2017.

[172] N. Kidonakis, "Nnll threshold resummation for top-pair and single-top production," *Physics of Particles and Nuclei*, vol. 45, pp. 714–722, Jul 2014.

[173] B. P. Roe, H.-J. Yang, J. Zhu, Y. Liu, I. Stancu, and G. McGregor, "Boosted decision trees, an alternative to artificial neural networks," *Nucl. Instrum. Meth.*, vol. A543, no. 2-3, pp. 577–584, 2005.

[174] P. Speckmayer, A. Höcker, J. Stelzer, and H. Voss, "The toolkit for multivariate data analysis, tmva 4," *J. Phys. Conf. Ser.*, vol. 219, no. 3, p. 032057, 2010.

[175] A. Hocker *et al.*, "TMVA - Toolkit for Multivariate Data Analysis," *PoS*, vol. ACAT, p. 040, 2007.

[176] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *J. Comput. Syst. Sci.*, vol. 55, no. 1, pp. 119 – 139, 1997.

[177] B. A. Dobrescu and J. D. Lykken, "Semileptonic decays of the standard higgs boson," *JHEP*, vol. 2010, p. 83, Apr 2010.

[178] F. Canelli, *Helicity of the W Boson in Single-Lepton Events*. PhD thesis, University of Rochester, 2003.

[179] H. Murayama, I. Watanabe, and K. Hagiwara, "HELAS: HELicity amplitude subroutines for Feynman diagram evaluations," 1992.

[180] P. van Dooren and L. de Ridder, "An adaptive algorithm for numerical integration over an n-dimensional cube," *Journal of Computational and Applied Mathematics*, vol. 2, no. 3, pp. 207 – 217, 1976.

[181] "The cern program library is a package of libraries and modules for use in particle physics analyses.." `https://cernlib.web.cern.ch/cernlib/`, 2017.

[182] A. Genz and A. Malik, "Remarks on algorithm 006: An adaptive algorithm for numerical integration over an n-dimensional rectangular region," *Journal of Computational and Applied Mathematics*, vol. 6, no. 4, pp. 295 – 302, 1980.

[183] R. Brun and F. Rademakers, "ROOT: An object oriented data analysis framework," *Nucl. Instrum. Meth.*, vol. A389, pp. 81–86, 1997. `http://root.cern.ch/`.

[184] T. Hahn, "CUBA: A Library for multidimensional numerical integration," *Comput. Phys. Commun.*, vol. 168, pp. 78–95, 2005.

[185] J. H. Friedman and M. H. Wright, "A nested partitioning procedure for numerical multiple integration," *ACM Trans. Math. Softw.*, vol. 7, pp. 76–92, Mar. 1981. implemented

as CERNLIB algorithm D151, documented at http://wwwasdoc.web.cern.ch/wwwasdoc/shortwrupsdir/d151/top.html.

[186] J. F. Koksma, “Een algemeene stelling uit de theorie der gelijkmatige verdeeling modulo 1,” *Mathematica B (Zutphen)*, vol. 11, pp. 7–11, 1942/43.

[187] E. Hlawka, “Funktionen von beschränkter variatiou in der theorie der gleichverteilung,” *Annali di Matematica Pura ed Applicata*, vol. 54, pp. 325–333, Dec 1961.

[188] F. J. Hickernell, *Koksma–Hlawka Inequality.* John Wiley & Sons, Inc., 2004.

[189] “Standard Model Cross Sections for CMS at 8 TeV.” https://twiki.cern.ch/twiki/bin/viewauth/CMS/StandardModelCrossSectionsat8TeV, 2017.

[190] “SM Higgs production cross sections at $\sqrt{s}\ =\ 8$ TeV (update in CERN Report3).” https://twiki.cern.ch/twiki/bin/view/LHCPhysics/CERNYellowReportPageAt8TeV, 2017.

[191] “Jet Energy Corrections: Official Software Tools for applying JEC Corrections and Uncertainties.” https://twiki.cern.ch/twiki/bin/view/CMSPublic/WorkBookJetEnergyCorrections, 2017.

[192] “Jet energy scale uncertainty sources.” https://twiki.cern.ch/twiki/bin/viewauth/CMS/JECUncertaintySources, 2017.

[193] A. L. Read, “Presentation of search results: the CL_s technique,” *J. Phys. G*, vol. 28, p. 2693, 2002.

[194] T. Junk, “Confidence level computation for combining searches with small statistics,” *Nucl. Instrum. Meth. A*, vol. 434, p. 435, 1999.

[195] ATLAS and CMS Collaborations, LHC Higgs Combination Group, “Procedure for the LHC Higgs boson search combination in Summer 2011,” Tech. Rep. ATL-PHYS-PUB-2011-11, CMS-NOTE-2011-005, CERN, 2011.

[196] “Documentation of the RooStats-based statistics tools for Higgs PAG.” https://twiki.cern.ch/twiki/bin/viewauth/CMS/SWGuideHiggsAnalysisCombinedLimit, 2017.

[197] G. Schott, “RooStats for Searches,” in *Proceedings, PHYSTAT 2011 Workshop on Statistical Issues Related to Discovery Claims in Search Experiments and Unfolding, CERN,Geneva, Switzerland 17-20 January 2011*, (Geneva), pp. 199–208, CERN, CERN, 2011.

[198] J. Stupak, *A Search for First Generation Leptoquarks in* $\sqrt{s} = 7$ *TeV pp Collisions with the ATLAS Detector*. PhD thesis, Aug 2012. Presented 08 May 2012.

[199] J. Baglio, A. Djouadi, and J. Quevillon, “Prospects for Higgs physics at energies up to 100 TeV,” *Rept. Prog. Phys.*, vol. 79, no. 11, p. 116201, 2016.

[200] “Standard Model Cross Sections for CMS at 13 TeV.” https://twiki.cern.ch/twiki/bin/view/CMS/StandardModelCrossSectionsat13TeVInclusive, 2017.

[201] G. Grasseau *et al.*, “Matrix element method for high performance computing platforms,” *J. Phys. Conf. Ser.*, vol. 664, no. 9, p. 092009, 2015.

[202] J. Bendavid, “Efficient Monte Carlo Integration Using Boosted Decision Trees and Generative Deep Neural Networks,” 2017. See also a discussion of the topic at https://indico.cern.ch/event/568875/contributions/2397925/attachments/1459058/2253175/mcgbr-May12-2017.pdf.

[203] B. Bjørken and S. Glashow, “Elementary particles and su(4),” *Phys. Lett.*, vol. 11, no. 3, pp. 255 – 257, 1964.

[204] S. L. Glashow, J. Iliopoulos, and L. Maiani, “Weak interactions with lepton-hadron symmetry,” *Phys. Rev. D*, vol. 2, pp. 1285–1292, Oct 1970.

[205] J. E. Augustin *et al.*, “Discovery of a narrow resonance in e^+e^- annihilation,” *Phys. Rev. Lett.*, vol. 33, pp. 1406–1408, Dec 1974.

[206] J. J. Aubert *et al.*, “Experimental observation of a heavy particle j,” *Phys. Rev. Lett.*, vol. 33, pp. 1404–1406, Dec 1974.

[207] M. Kobayashi and T. Maskawa, “Cp-violation in the renormalizable theory of weak interaction,” *Progress of Theoretical Physics*, vol. 49, no. 2, pp. 652–657, 1973.

[208] S. W. Herb *et al.*, “Observation of a dimuon resonance at 9.5 gev in 400-gev proton-nucleus collisions,” *Phys. Rev. Lett.*, vol. 39, pp. 252–255, Aug 1977.

[209] P. Bagnaia *et al.*, “Evidence for $z^0 \to e^+e^-$ at the cern pp collider,” *Phys. Lett. B*, vol. 129, no. 1, pp. 130 – 140, 1983.

[210] F. Abe *et al.*, “Observation of top quark production in $\overline{p}p$ collisions with the collider detector at fermilab,” *Phys. Rev. Lett.*, vol. 74, pp. 2626–2631, Apr 1995.

[211] S. Abachi *et al.*, “Search for high mass top quark production in $p\overline{p}$ collisions at $\sqrt{s} =$ 1.8 tev,” *Phys. Rev. Lett.*, vol. 74, pp. 2422–2426, Mar 1995.

[212] `http://ckmfitter.in2p3.fr/`.

[213] `http://project-gfitter.web.cern.ch/project-gfitter/`.

[214] `http://zfitter.com/`.

[215] `http://lepewwg.web.cern.ch/LEPEWWG/`.

[216] H. Flacher *et al.*, “Revisiting the Global Electroweak Fit of the Standard Model and Beyond with Gfitter,” *Eur. Phys. J.*, vol. C60, pp. 543–583, 2009. [Erratum: Eur. Phys. J.C71,1718(2011)].

APPENDIX A

HISTORY OF THE STANDARD MODEL

During its tenure, the standard model has provided a remarkably accurate description of results from both accelerator and non-accelerator experiments. In fact, all of the standard model particles shown in 2.1 have been observed and measured, most of these discoveries taking place in the last sixty years. The original quark model proposed by Gell-Mann and Zweig in 1964 only included the up, down, and strange quarks. The up and down quarks were later observed by deep inelastic scattering experiments at the Stanford Linear Accelerator (SLAC), which by extension proved the existence of the strange quark. The charm quark was proposed by Bjørken and Glashow also in 1964 [203], but is credited to Sheldon Lee Glashow, John Iliopoulos, and Luciano Maianiafter they proposed the Glashow–Iliopoulos–Maiani (GIM) mechanism in 1970 [204]. The charm quark was later observed in J/ψ decays by SLAC [205] and Brookhaven National Laboratory (BNL) [206]. The invariant distribution presented in the original BNL paper can be found in fig. A.1a. The bottom or beauty quark was later proposed by Kobayashi and Maskawa in 1973 [207] and observed by the E288 experiment led by Leon Lederman at the Fermi National Accelerator Laboratory (FNAL) in 1977 [208]. Kobayashi and Maskawa were trying to describe CP violation in the weak interaction, finally earning a Nobel prize for their work in 2008.

Following this flurry of quark discoveries, the W and Z bosons were observed at CERN in 1983 in proton-antiproton collisions of $\sqrt{s} = 540\,\text{GeV}$ at the Super Proton Synchroton (SPS). This was research lead by Carlo Rubbia using the UA1 experiment [59] and Pierre Darriulat on the UA2 experiment [209]. The invariant mass of the Z boson as seen by UA2 is shown in fig. A.1b. While an insufficient number of W bosons were observed to make precision measurements, this was accomplished using Large Electron-Positron Collider (LEP) experiment, also at CERN, where the W and Z masses were measured to be:

$$\begin{aligned} M_{\mathrm{Z}} &= 91.1875 \pm 0.0021\,\mathrm{GeV} \\ M_{\mathrm{W}} &= 80.376 \pm 0.0033\,\mathrm{GeV} \end{aligned} \tag{A.1}$$

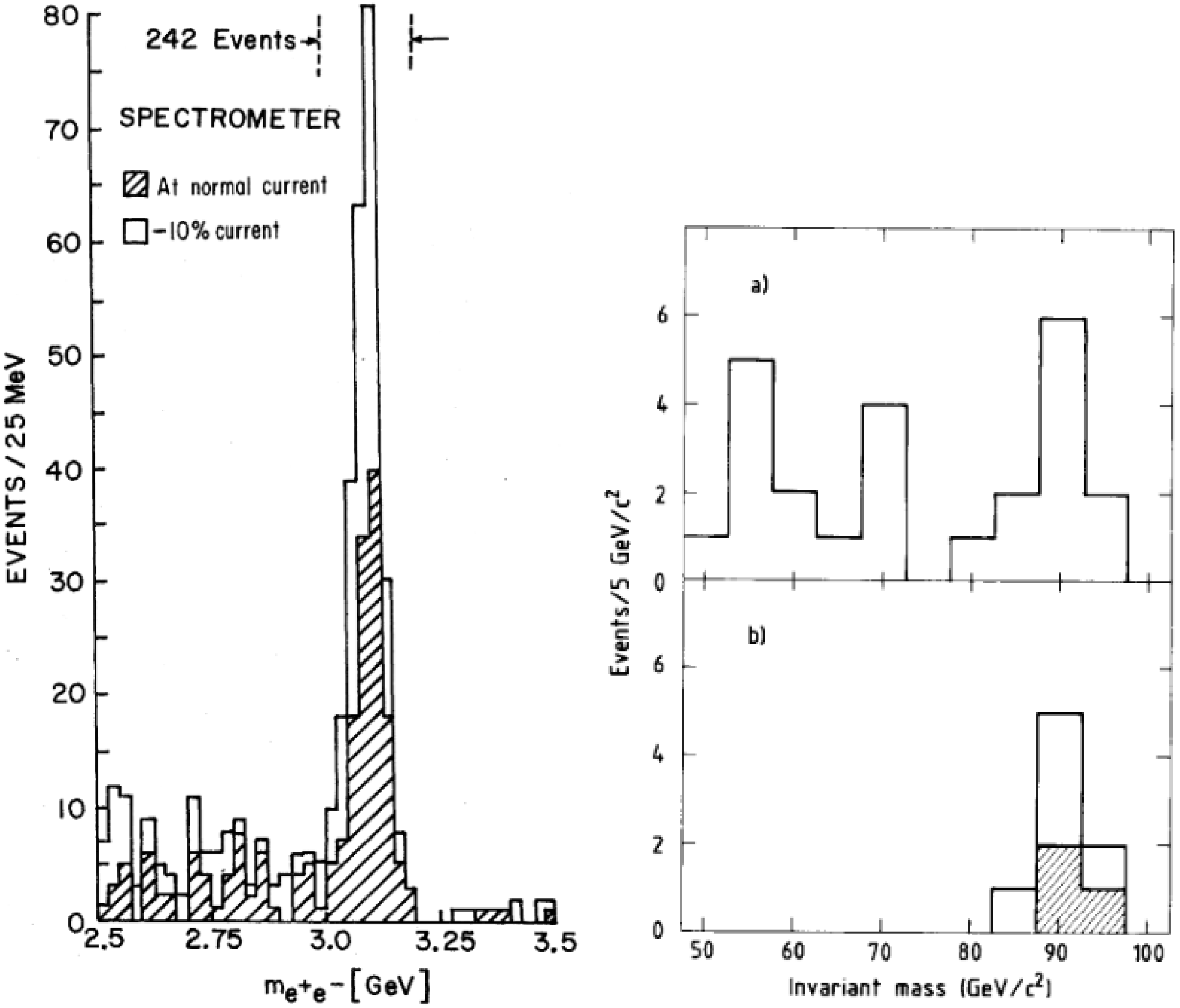
Finishing off an amazing 30 years of discoveries and completing the third and final generation of quarks predicted by Kobayashi and Maskawa, the top quark was jointly discovered in 1995 by the CDF [210] and D0 [211] experiments at FNAL using the $\sqrt{\mathrm{s}} = 1.4\,\mathrm{TeV}$ Tevatron accelerator. Its mass was measured to be $M_{\mathrm{t}} \sim 176$ GeV.

At this point in time, the Higgs boson was the final particle left to be discovered. Both LEP and the Tevatron failed to observe the particle, though CDF and D0 were able to exclude all masses for the Higgs boson except in the ranges $115 < M_{\mathrm{H}} < 155$ GeV and $M_{\mathrm{H}} > 176$ GeV as seen in fig A.2a [25]. The 2012 Higgs boson discovery was jointly announced by the CMS and ATLAS collaborations at CERN [26, 30]. By combining the 5.1 fb^{-1} of 7 TeV data and 19.7 fb^{-1} of 8 TeV data, CMS was able to uses the $\mathrm{H}\to\gamma\gamma$ and $\mathrm{H}\to\mathrm{ZZ}^*\to4\ell$ channels to measure the mass to be $125.3^{+0.26}_{-0.27}\,(\mathrm{stat.})^{+0.15}_{-0.15}\,(\mathrm{syst.})$ GeV as shown in fig. A.2b [27]. Figs. A.2c and A.2d show the invariant mass distributions for the diphoton and four-lepton systems obtained by the CMS experiment. The cross section σ was found to be consistent with that of the standard model such that the signal strength at the measured mass was found to be

$$\frac{\sigma}{\sigma_{SM}} = 1.00 \pm 0.09\,(\mathrm{stat.})^{+0.08}_{-0.07}\,(\mathrm{theory}) \pm 0.07\,(\mathrm{syst.}) \tag{A.2}$$

A graphical representation of this can be found in fig. A.3. Measurements of other properties such as spin, parity, production rates, and the ratio of couplings to fermions and vector bosons are discussed in [27].

In the past, the experimental measurements of electroweak precision observables at LEP, SLAC, the Tevatron, and the LHC have been paired with very accurate theoretical predictions. The benefit of these observables is that they can probe energy scales beyond what is capable through direct

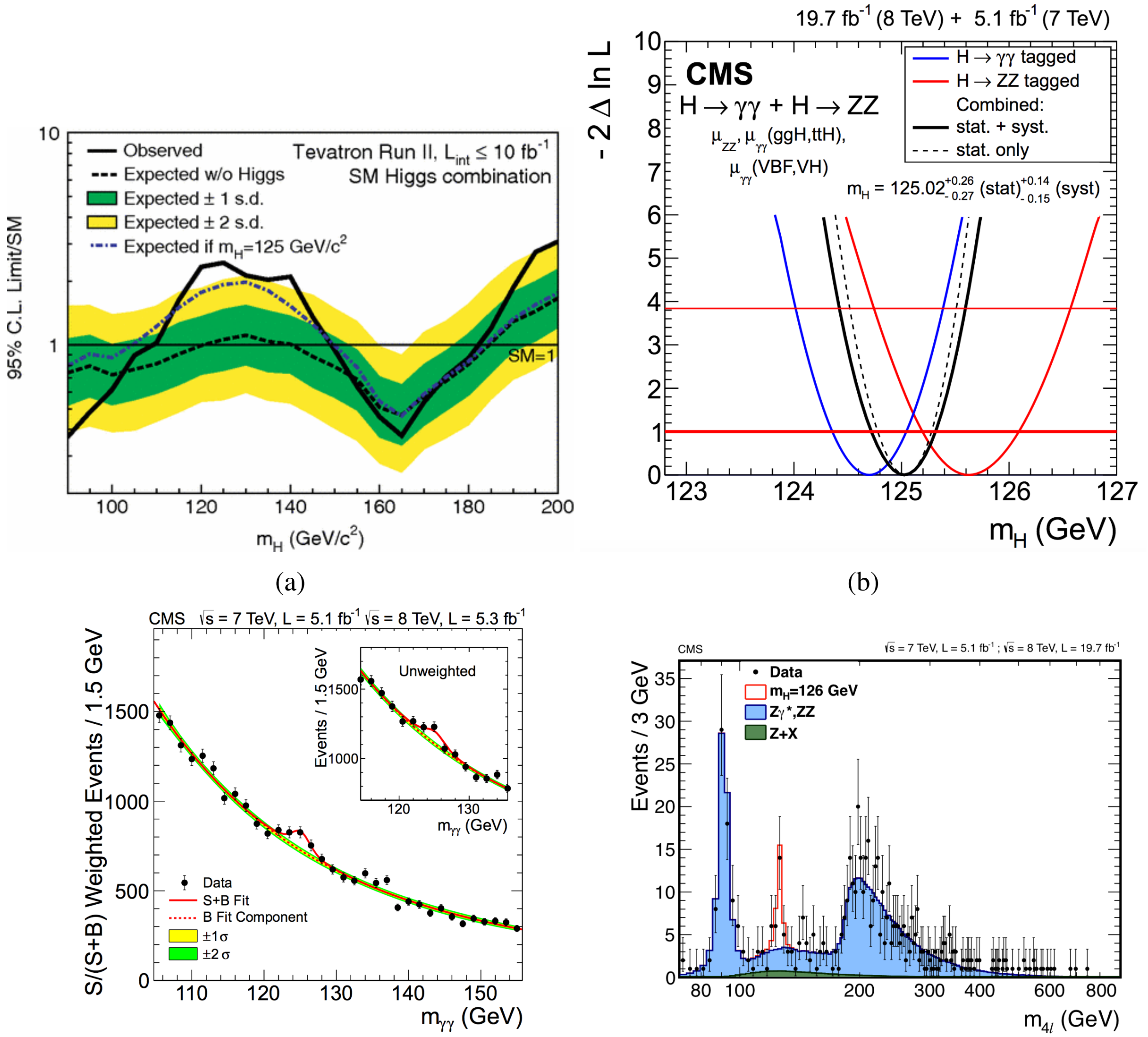


(a) The invariant mass spectrum of the J/ψ particle discovered at BNL. Reprinted from [206].

(b) The invariant mass spectrum of the Z→e^+e^- decay as seen by the UA2 collaboration. The upper half of the figure shows the number of events with a calorimeter cluster while the lower half shows the eight events that mad it past all selection criteria. Reprinted from [209].

Figure A.1: Invariant mass distributions from the discoveries of the J/ψ meson and Z boson.

measurements by accounting for the effects of higher order corrections. Free parameters in the Standard Model could be constrained by doing global fits of the electroweak sector. Now that the Higgs boson has been found, and assuming this is the SM Higgs boson, the fit is over-constrained because all parameters used in the fit are known. Instead of constraining the free parameters we are now able to test the consistency of the Standard Model and even predict some parameters to



(a)

(b)

(c)

(d)

Figure A.2: Key figures showing the Higgs boson discovery in two high-resolution channels. (a) The combined CDF and D0 exclusion plot for the Higgs mass before the discovery, reprinted from [25]. (b) The best fit mass results from the $\gamma\gamma$ and ZZ decay channels at CMS. (c) The diphoton invariant mass distribution. The black markers represent the data, the solid and dashed red lines represent the fitted signal and background, and the colored bands represent the ± 1 and ± 2 standard deviation uncertainties in the background estimate. The major canvas shows each event weighted by the $\frac{S}{S+B}$ value of its selection category. (d) The four-lepton invariant mass distribution where the black markers are the data, the filled histograms show the background estimates, and the open histogram shows the background plus signal expectation for a Higgs boson mass of $M_{\rm H} = 125\,\mathrm{GeV}$. Figs. (b)-(c) are reprinted from [26].

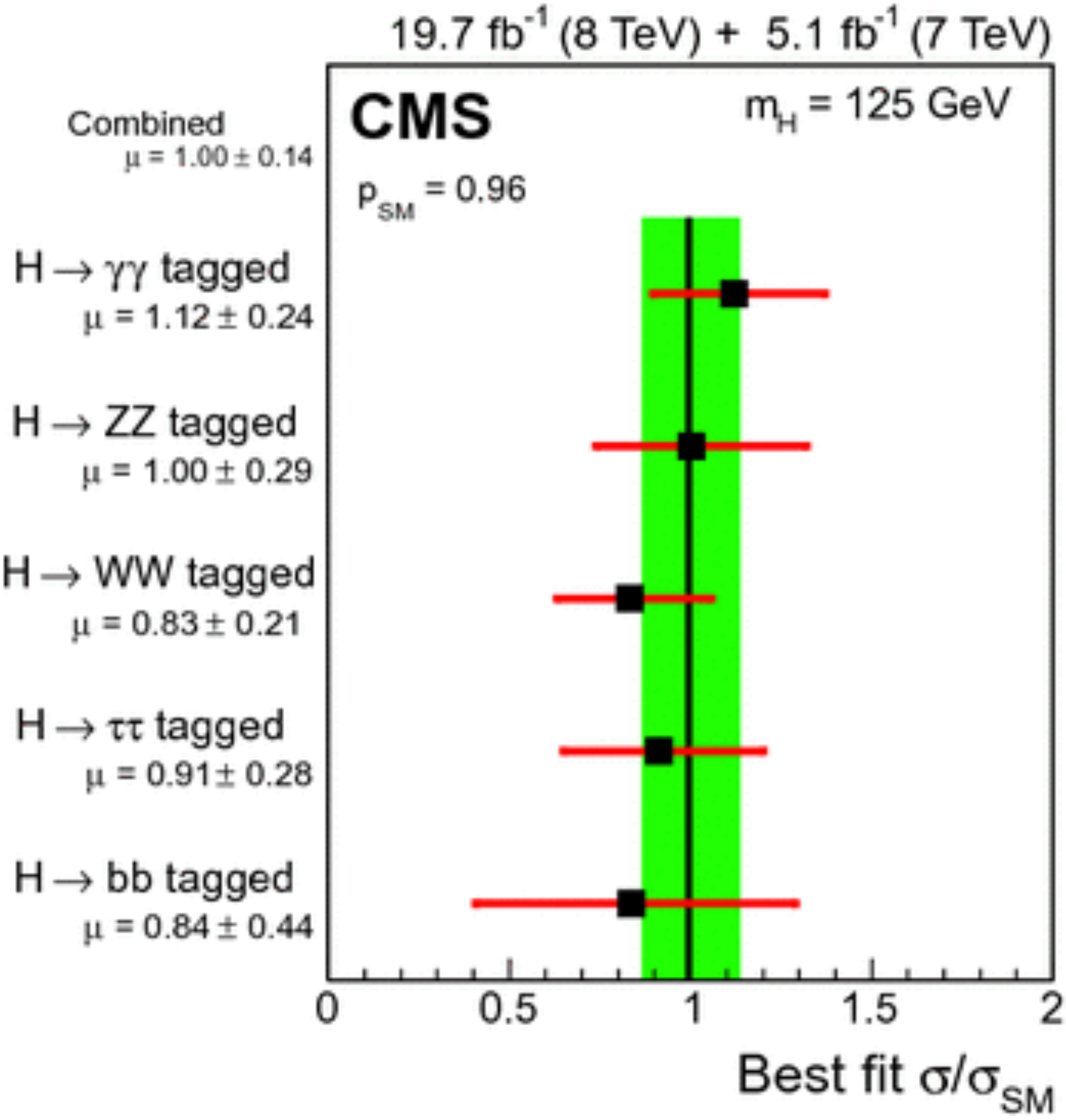


Figure A.3: Best-fit σ/σ_{SM} grouped by predominant decay mode. The vertical band is the overall combined analysis value and the horizontal bars show the ±1 uncertainties (statistical and systematic). Reprinted from [27].

higher precision than we are currently able to measure.

These complicated fits are performed by several groups [212, 213, 214, 215], but only the results from the GFitter group [28, 216] will be used here. Some of the measurements included in the fits are of the mass of the Higgs boson, the mass and widths of the W and Z bosons, the masses of the top, botton, and charm quarks, the strong coupling constant, the weak mixing angle, among others. Fig. A.4 shows the comparison of the fit results with the direct measurements of the parameters, all of which agree to within 3σ. A common test of the Standard Model is to independently measure the top quark and W boson masses. Fig. A.5a shows the 68% and 95% confidence level intervals obtained for M_{W} versus M_{t} for the case where the direct Higgs mass measurement is included (blue) and excluded (grey). In both cases the fits agree with the direct measurements shown in the green bands and ellipses. Fig. A.5b shows the corresponding plot for the W boson mass and the effective weak mixing angle. In all cases the fit procedure agrees with the direct measurements, showing the consistency of the Standard Model within current experimental precision.

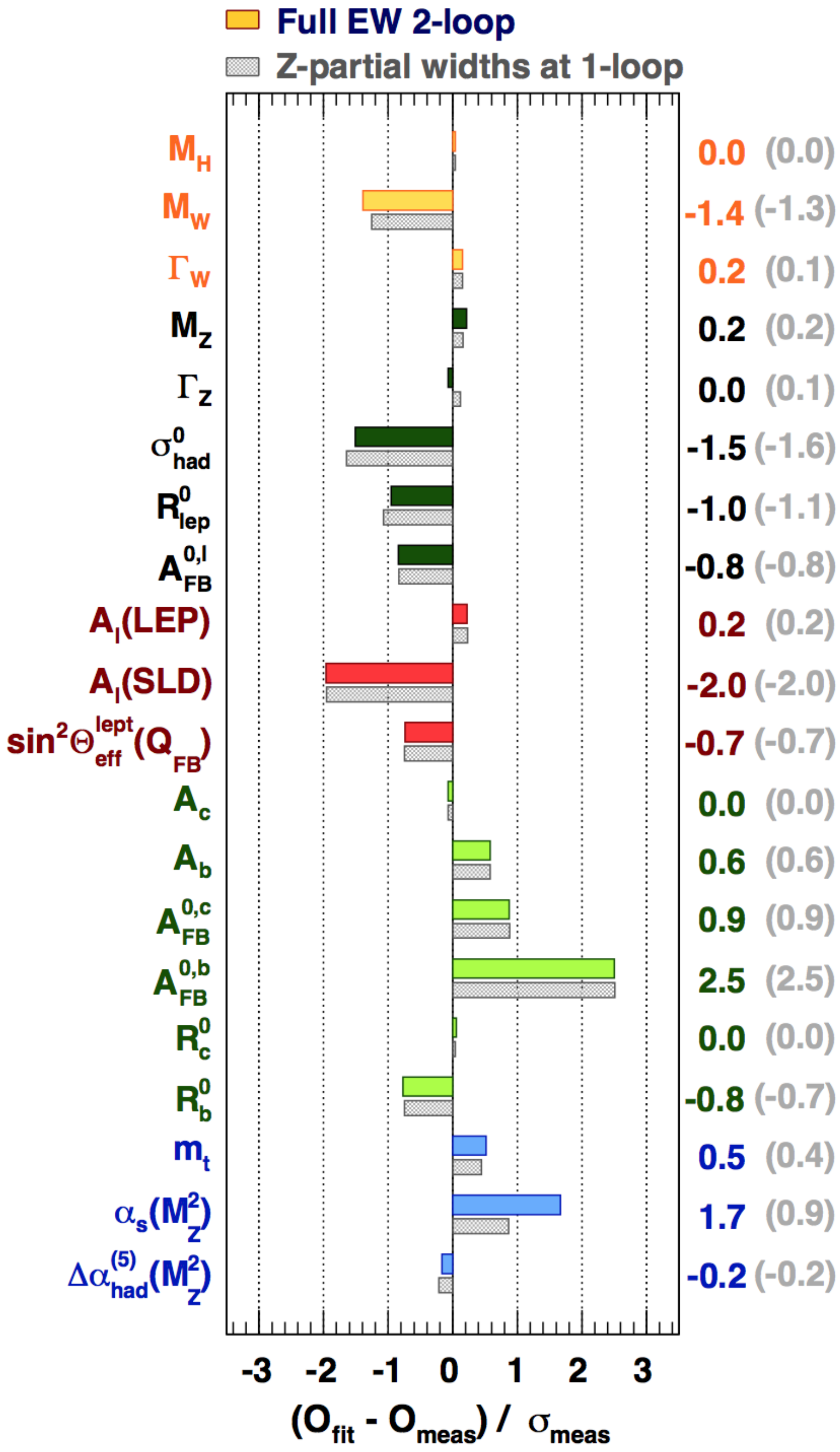
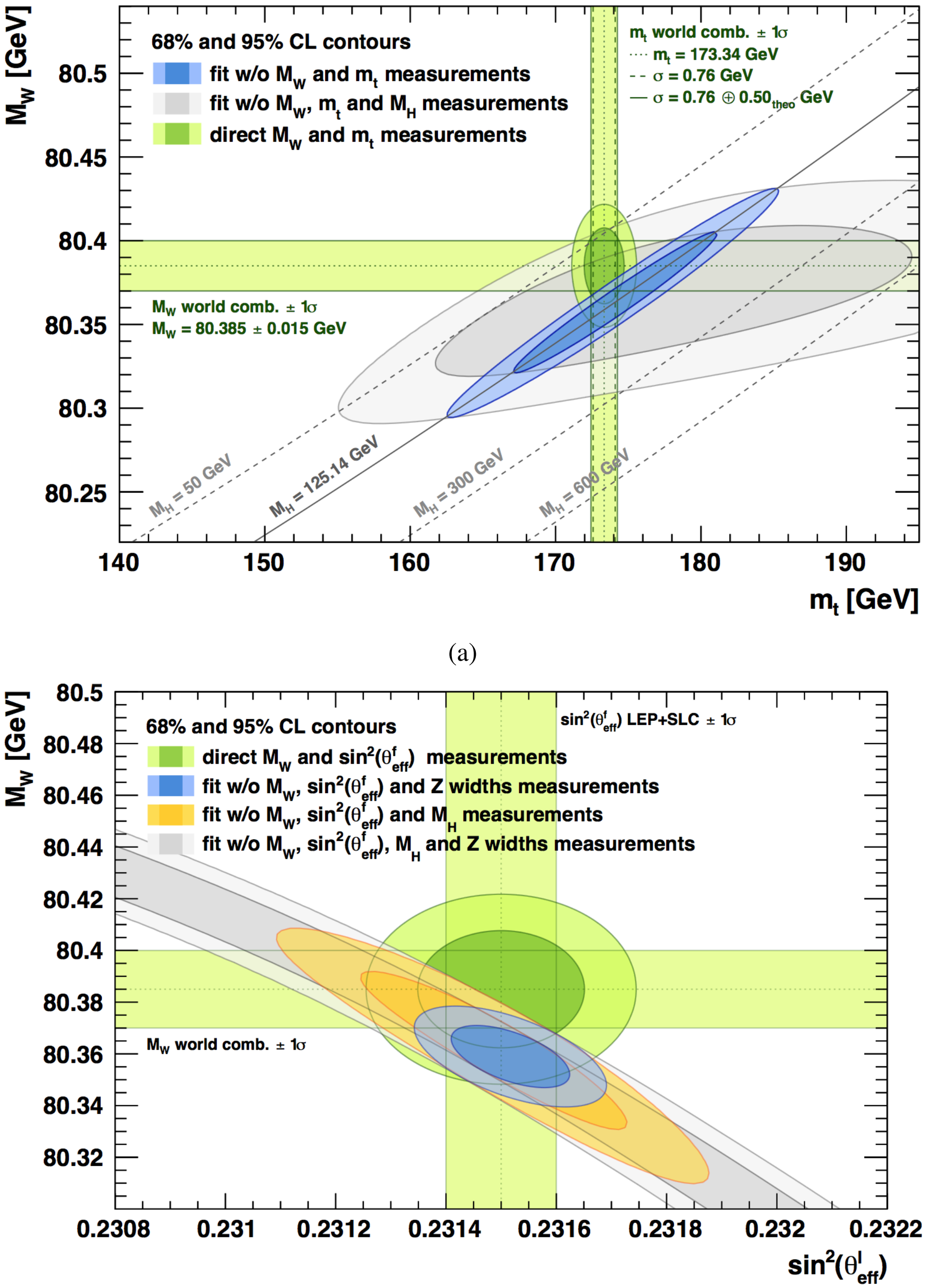


Figure A.4: Comparison of the GFitter fit results with the direct measurements in units of the experimental uncertainty. Reprinted from [28].



(a)

(b)

Figure A.5: Contours at 68% and 95% confidence level obtained from scans of $M_{\rm W}$ versus $M_{\rm t}$ (top) and $M_{\rm W}$ versus $\sin^2\left(\theta^l_{eff}\right)$ (bottom), for a fit including $M_{\rm H}$ (blue) and excluding $M_{\rm H}$ (grey), as compared to the direct measurements (vertical and horizontal green bands and ellipses). In both figures, the corresponding direct measurements are excluded from the fit. Figure and caption from [28].

APPENDIX B

$\vec{\not{E}}_{\mathrm{T}}$ PERFORMANCE AND CORRECTIONS

B.1 Type-0 $\vec{\not{E}}_{\mathrm{T}}$ Correction

Pileup interactions typically produce visible particles, with only a few processes, like neutrinos from Kaon decays, producing invisible particles. If CMS were able to perfectly measure all of the visible particles then pileup would have little effect on the $\vec{\not{E}}_{\mathrm{T}}$ reconstruction. However, as discussed in section 4.7, the $\vec{\not{E}}_{\mathrm{T}}$ reconstruction does degrade as the number of pileup interactions increases. The type-0 correction is an attempt to remove this pileup effect for the $\vec{\not{E}}_{\mathrm{T}}$ calculated using PF candidates, as opposed to calorimeter towers or tracks.

In essence, the type-0 correction is an application of CHS (see 4.5 for a discussion of CHS), but also removes a portion of the $\vec{\not{E}}_{\mathrm{T}}$ estimated to come from neutral pileup. The neutral pileup estimate is necessary because removing only charged particles might cause the $\vec{\not{E}}_{\mathrm{T}}$ to move further from its true value. In this section the pileup particles will be broken up as being neutral (neuPU) or charges (chPU). Furthermore, the correction makes three assumptions about the pileup particles as spelled out in equation B.1. The first assumption is that the sum of p_{T} for the neutral and charged components of the $\vec{\not{E}}_{\mathrm{T}}$ due to pileup are equal and opposite. At the truth level this cancellation is very nearly exact. The part of B.1 says that the charged particles can be measured exactly, which is also a good assumption for low $\vec{p}_{\mathrm{T}}$ tracks. The last assumption says that the direction of the neutral pileup can be measured exactly, but that the energy is off by the same amount for each particle. The directionality is measured using the position of the calorimeter cells, but the energy measurement calibration was done using high $\vec{p}_{\mathrm{T}}$ particles so that the system systematically mismeasures low $\vec{p}_{\mathrm{T}}$

particles.

$$\begin{aligned}\sum_{i\in\text{neuPU}} \vec{p}_{\mathrm{T,i}}^{\,\text{true}} + \sum_{i\in\text{chPU}} \vec{p}_{\mathrm{T,i}}^{\,\text{true}} &= 0 \\ \sum_{i\in\text{chPU}} \vec{p}_{\mathrm{T,i}}^{\,\text{true}} &= \sum_{i\in\text{chPU}} \vec{p}_{\mathrm{T,i}} \\ \sum_{i\in\text{neuPU}} \vec{p}_{\mathrm{T,i}} &= R^0 \sum_{i\in\text{neuPU}} \vec{p}_{\mathrm{T,i}}\end{aligned} \tag{B.1}$$

The assumptions can then be combined into equation B.2.

$$\sum_{i\in\text{neuPU}} \vec{p}_{\mathrm{T,i}} = -R^0 \sum_{i\in\text{chPU}} \vec{p}_{\mathrm{T,i}} \tag{B.2}$$

The raw $\vec{\not\!\!E}_{\mathrm{T}}$ components can be broken up as coming from either the hard scatter (HS) vertex or from pileup (PU) interactions. The pileup can then be further boken down into the neutral and charged components as previously specified. This categorization is shown in equation B.3.

$$\begin{aligned}\vec{\not\!\!E}_{\mathrm{T}}^{\text{raw}} &= -\sum_{i\in\text{HS}} \vec{p}_{\mathrm{T,i}} - \sum_{i\in\text{PU}} \vec{p}_{\mathrm{T,i}} \\ &= -\sum_{i\in\text{HS}} \vec{p}_{\mathrm{T,i}} - \sum_{i\in\text{neuPU}} \vec{p}_{\mathrm{T,i}} - \sum_{i\in\text{chPU}} \vec{p}_{\mathrm{T,i}}\end{aligned} \tag{B.3}$$

CHS is able to remove the third sum, but is not able to separate the first and second sums.

The type-0 corrections is the estimate of the neutral pileup shown in equation B.2 plus the sum over the charged particles from pileup.

$$\vec{C}_{\mathrm{T}}^{\,Type-0} = \left(1 - R^0\right) \sum_{i\in\text{chPU}} \vec{p}_{\mathrm{T,i}} \tag{B.4}$$

This corrections added to the raw $\vec{\not\!\!E}_{\mathrm{T}}$ yields the type-0 corrected $\vec{\not\!\!E}_{\mathrm{T}}$. To also propogate the JEC to the pileup corrected $\vec{\not\!\!E}_{\mathrm{T}}$ one can add type-1 correction to the type-0 corrected $\vec{\not\!\!E}_{\mathrm{T}}$. This process can be seen in equation B.5.

$$\begin{aligned}\vec{\not\!\!E}_{\mathrm{T}}^{\text{Type}-0} &= \vec{\not\!\!E}_{\mathrm{T}}^{\text{raw}} + \vec{C}_{\mathrm{T}}^{\text{Type}-0} \\ \vec{\not\!\!E}_{\mathrm{T}}^{\text{Type}-0-1} &= \vec{\not\!\!E}_{\mathrm{T}}^{\text{Type}-0} + \vec{C}_{\mathrm{T}}^{\text{Type}-1}\end{aligned} \tag{B.5}$$

B.2 $\vec{\not{E}}_{\mathrm{T}}$ Filters

Besides interesting physics processes, high values of $\not{E}_{\mathrm{T}}$ can be caused by cosmic rays, detector noise, and particles from the beam-halo. In addition to the previous corrections used to make sure the $\vec{\not{E}}_{\mathrm{T}}$ is reconstructed correctly, CMS has also developed several algorithms for identifying and removing sources of fake $\vec{\not{E}}_{\mathrm{T}}$. False $\vec{\not{E}}_{\mathrm{T}}$ is a problem because is causes a discrepancy between the data and MC, where the sources of fake $\vec{\not{E}}_{\mathrm{T}}$ are not explicitly simulated. After several of these filters are used this agreement will typically improve.

APPENDIX C

COMPARISON PLOTS

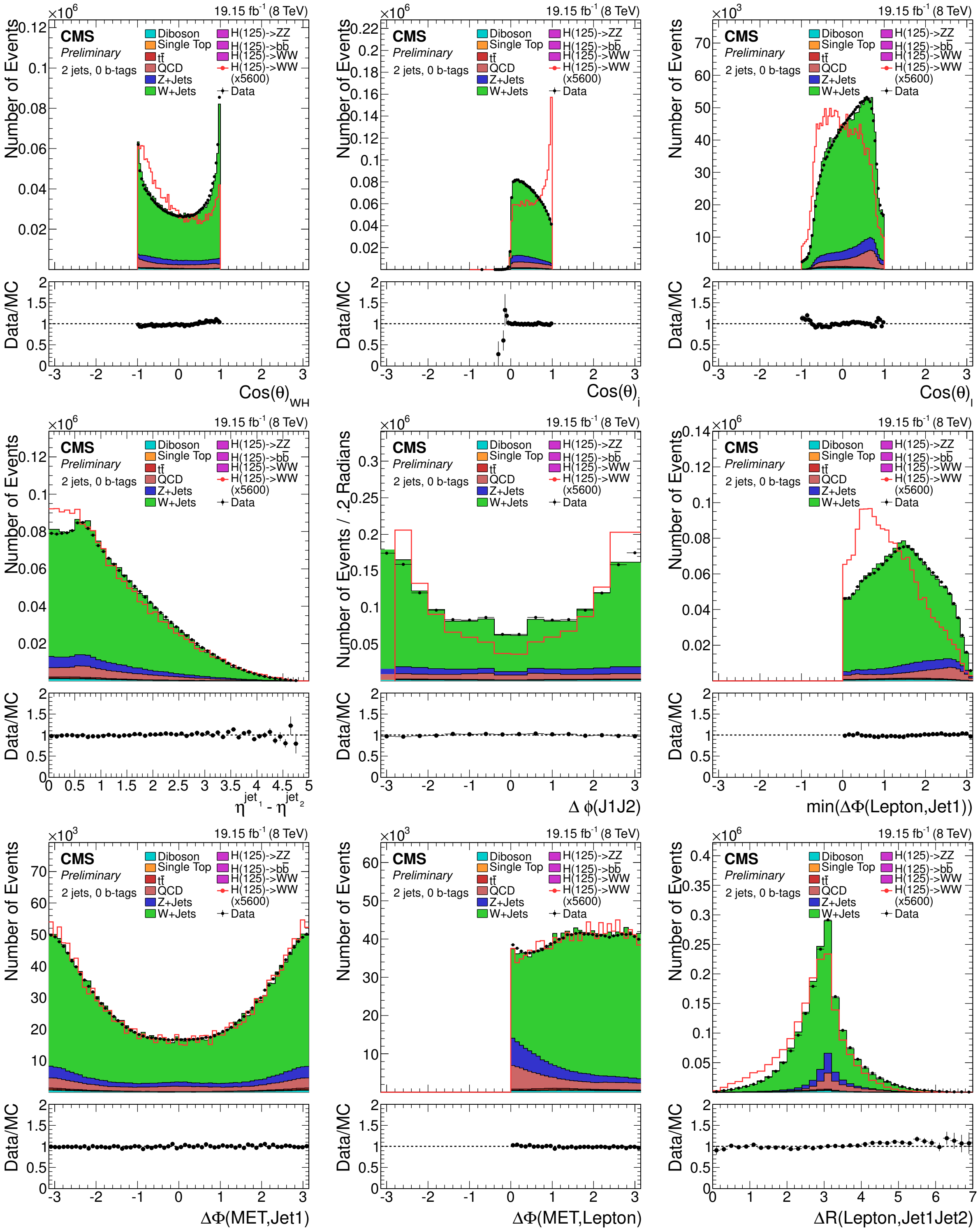


Figure C.1: Data-to-MC comparison plots for the 2-jet electron channel.

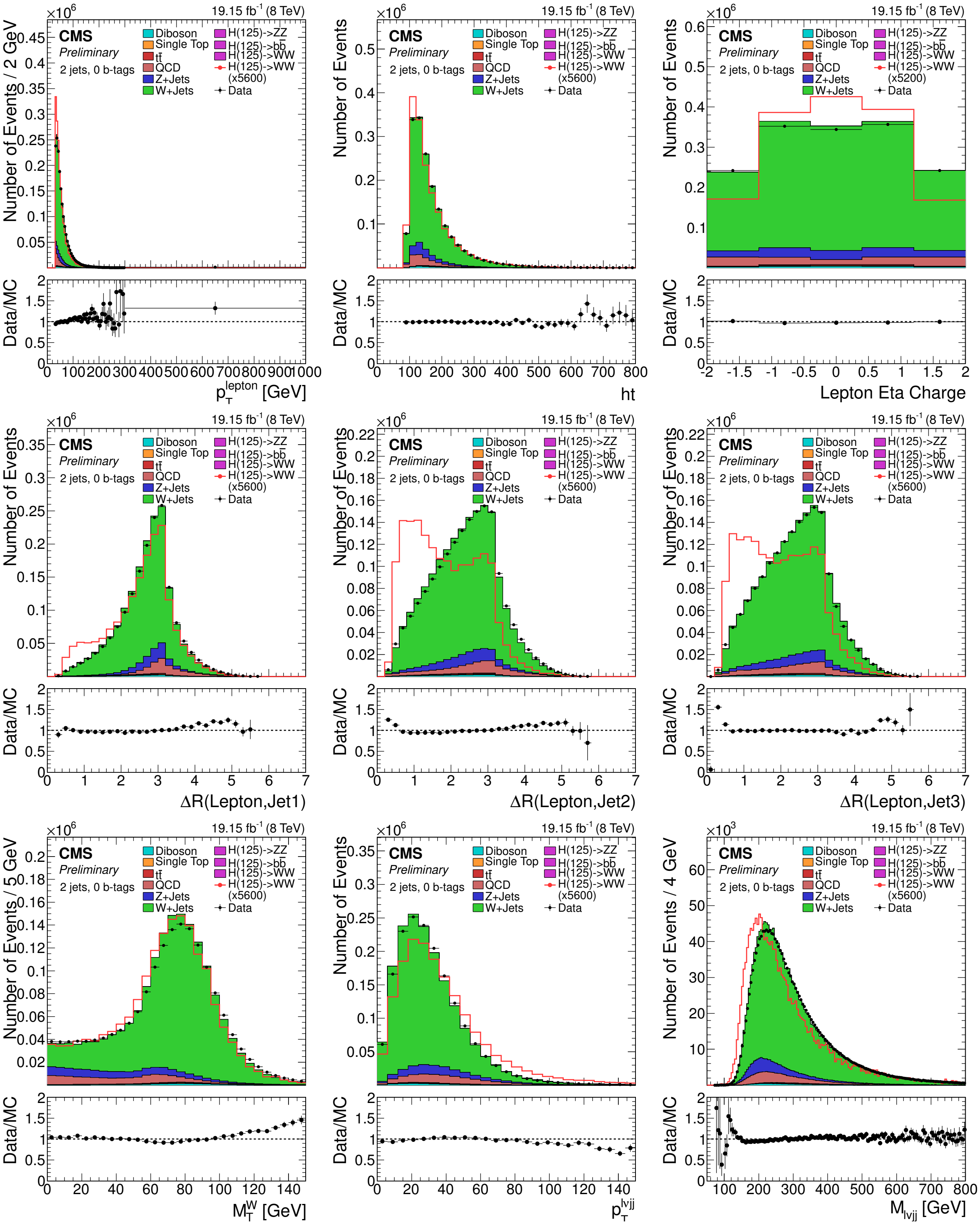


Figure C.2: Data-to-MC comparison plots for the 2-jet electron channel.

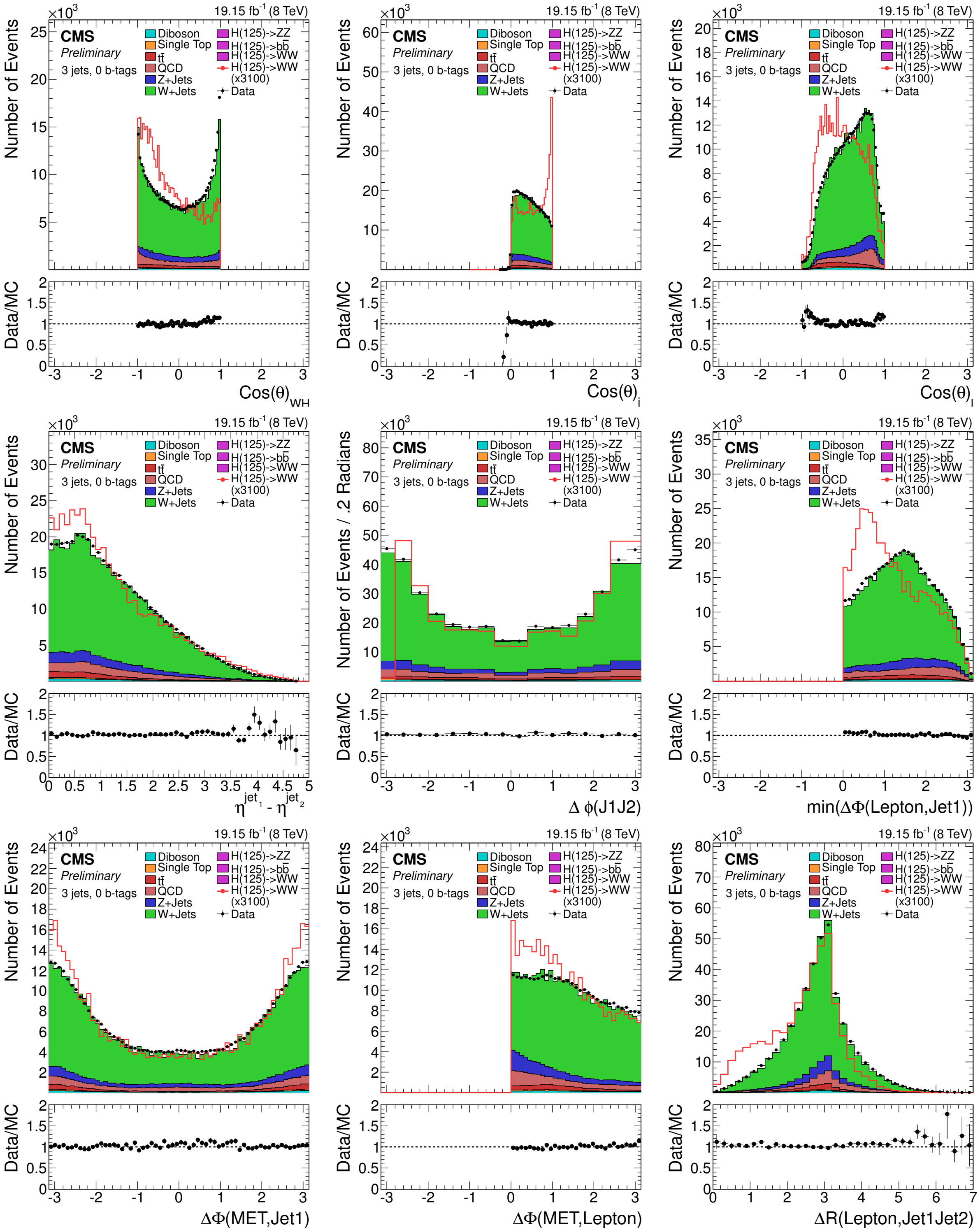


Figure C.3: Data-to-MC comparison plots for the 3-jet electron channel.

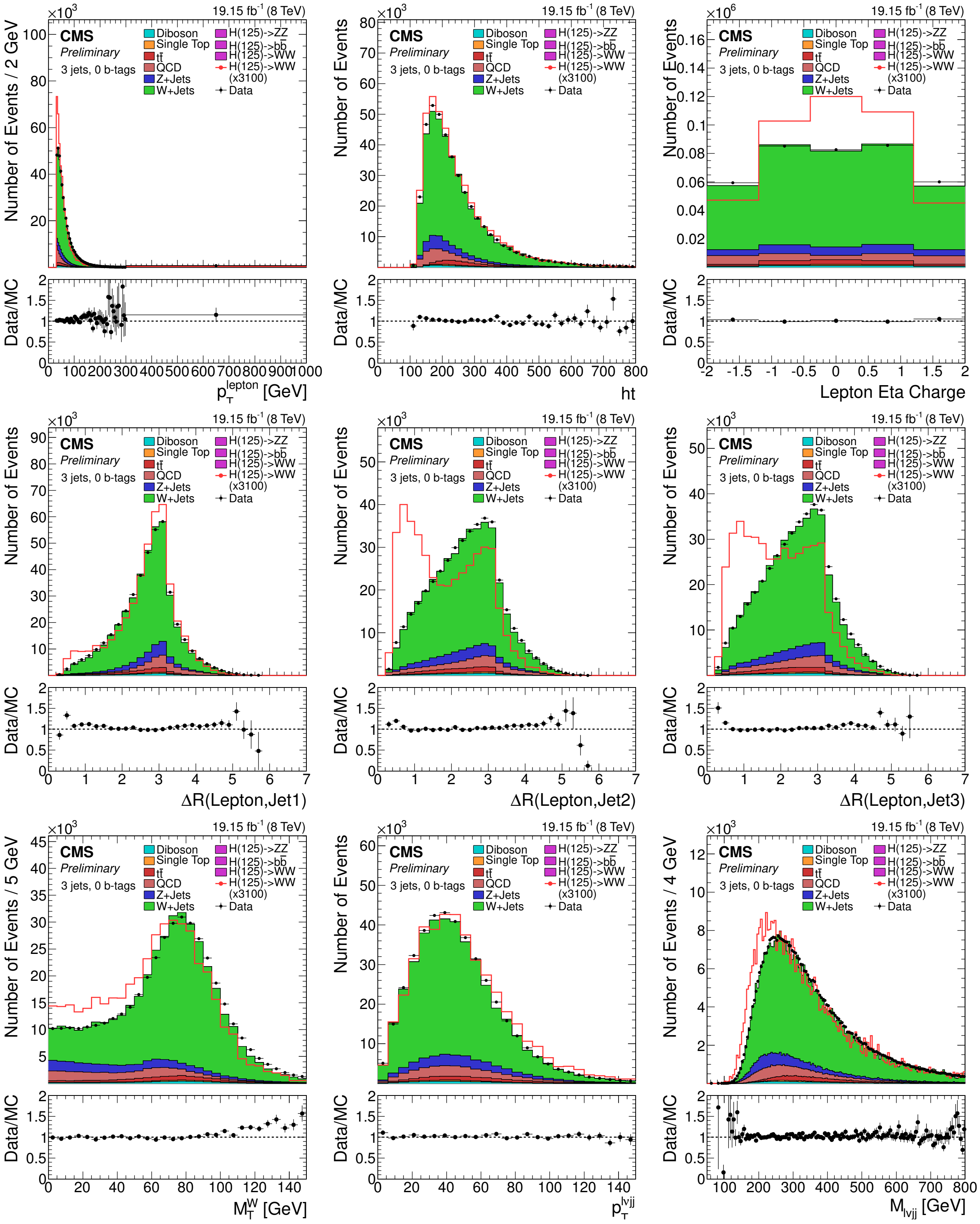


Figure C.4: Data-to-MC comparison plots for the 3-jet electron channel.

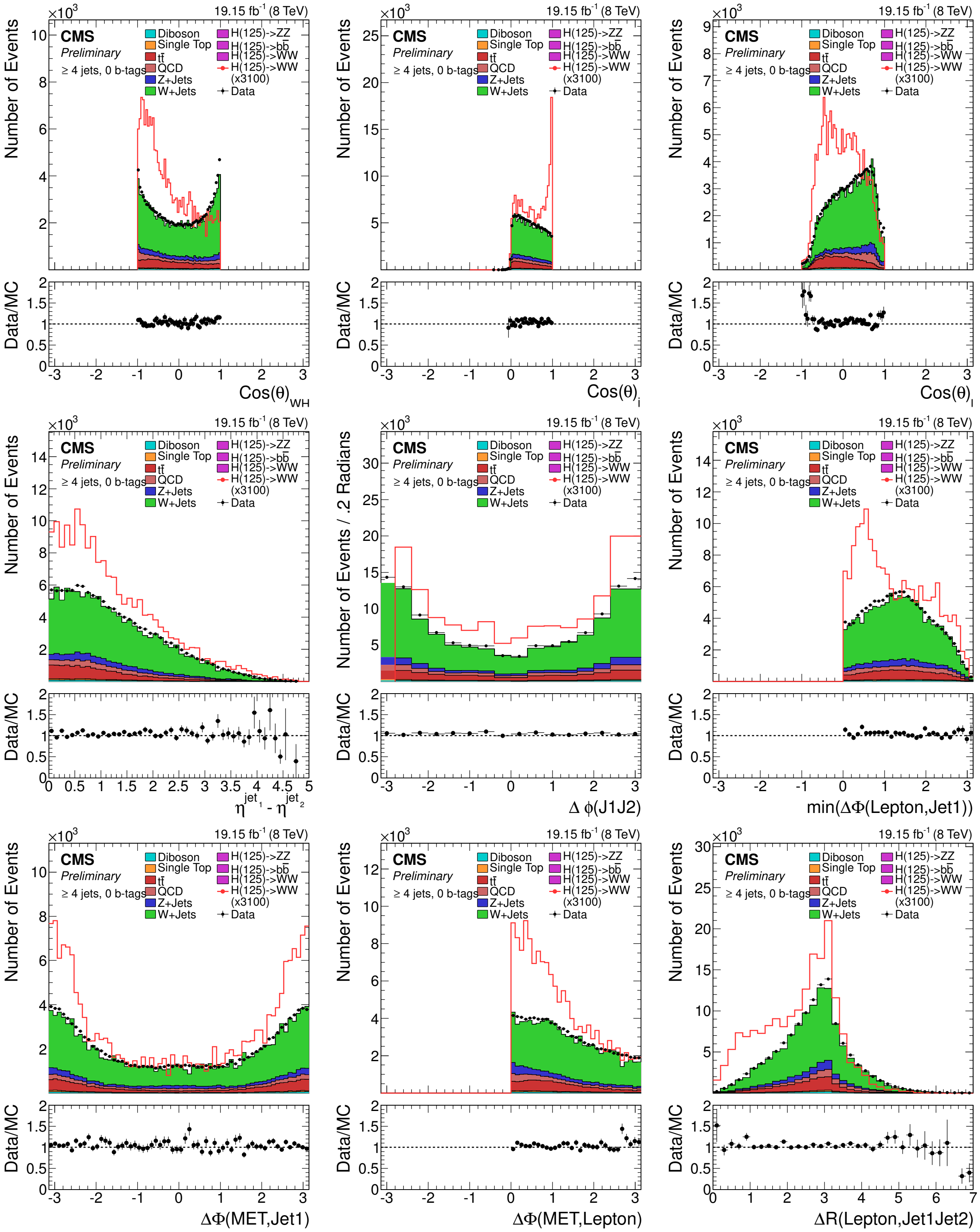


Figure C.5: Data-to-MC comparison plots for the ⩾4-jet electron channel.

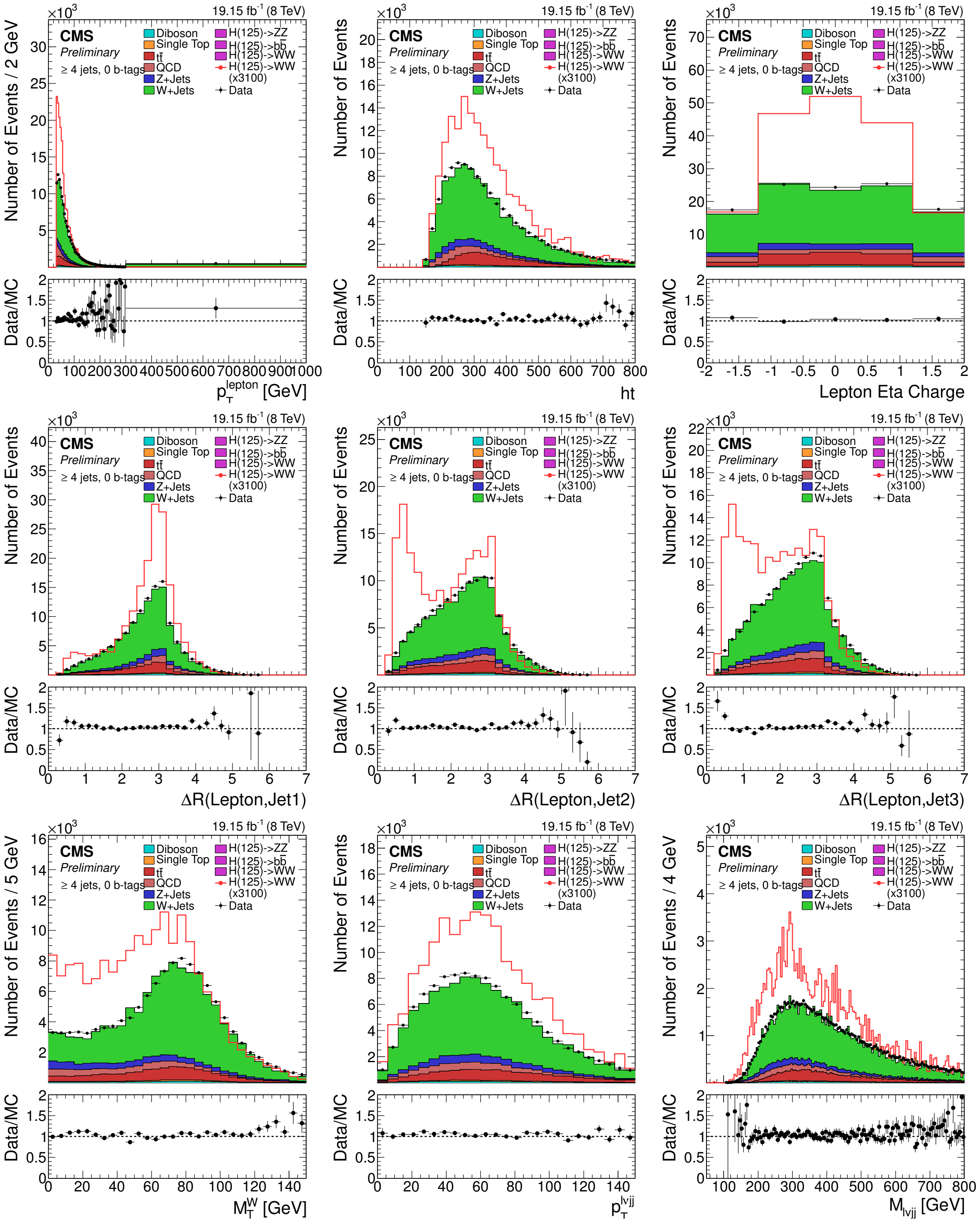


Figure C.6: Data-to-MC comparison plots for the ⩾4-jet electron channel.

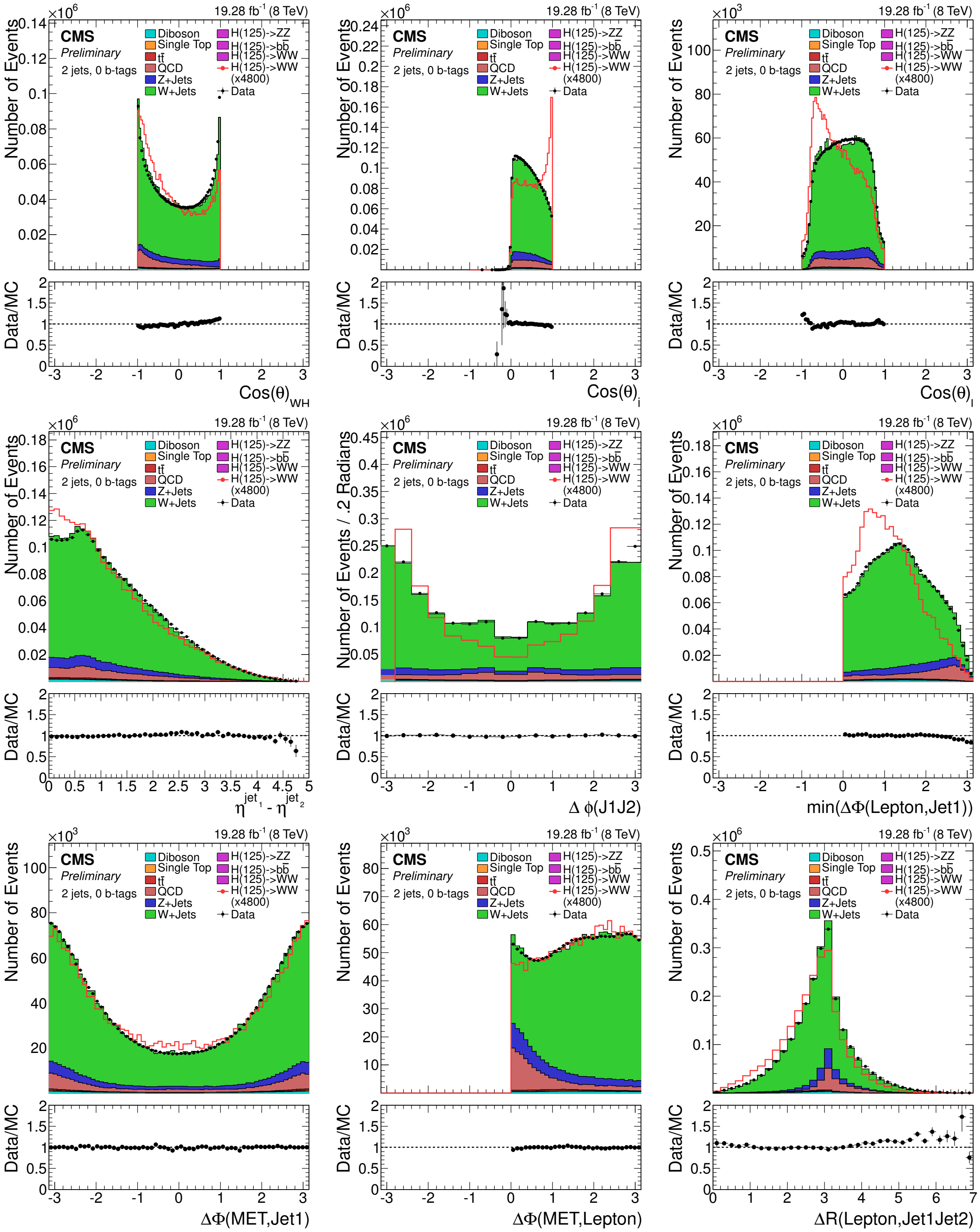


Figure C.7: Data-to-MC comparison plots for the 2-jet muon channel.

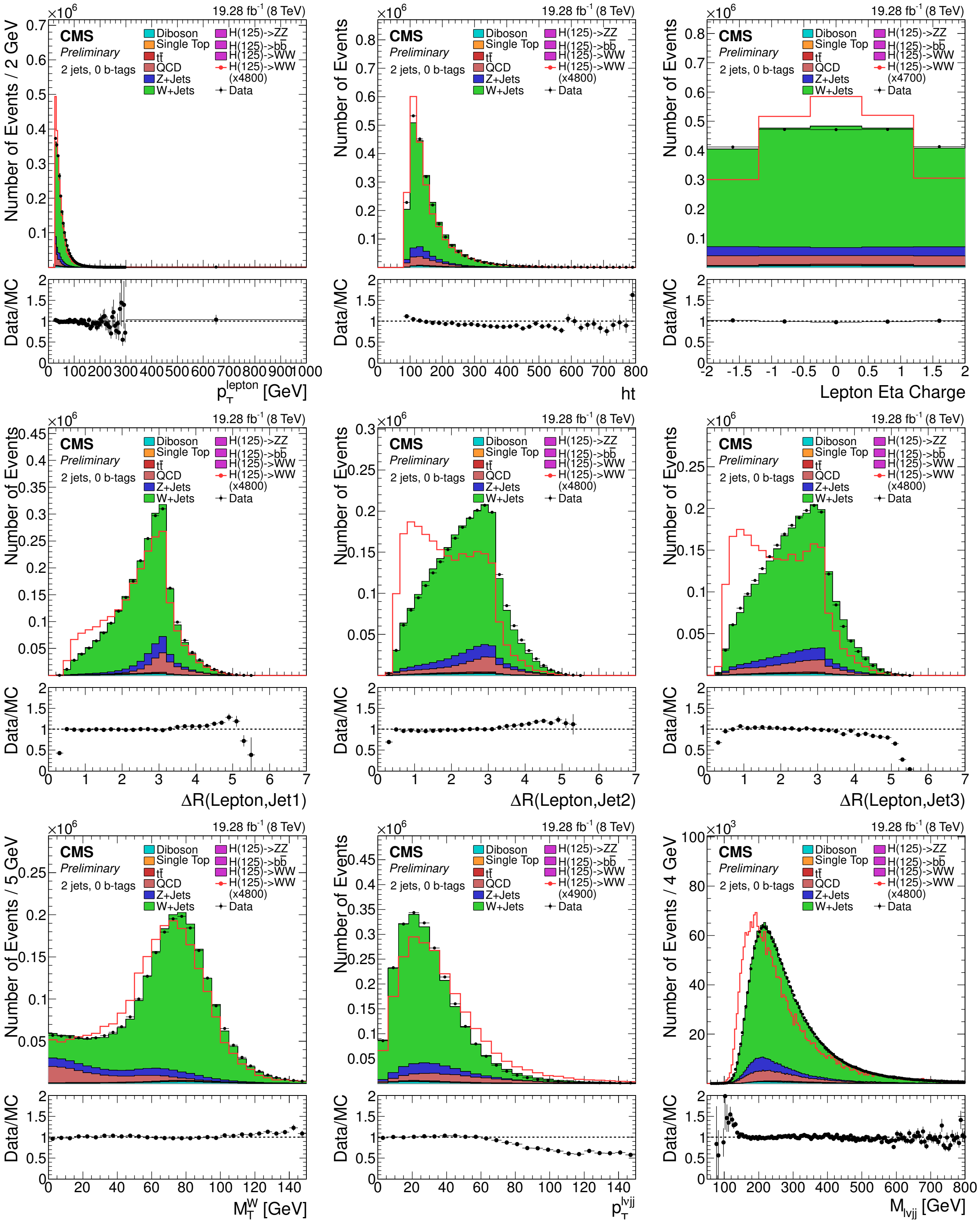


Figure C.8: Data-to-MC comparison plots for the 2-jet muon channel.

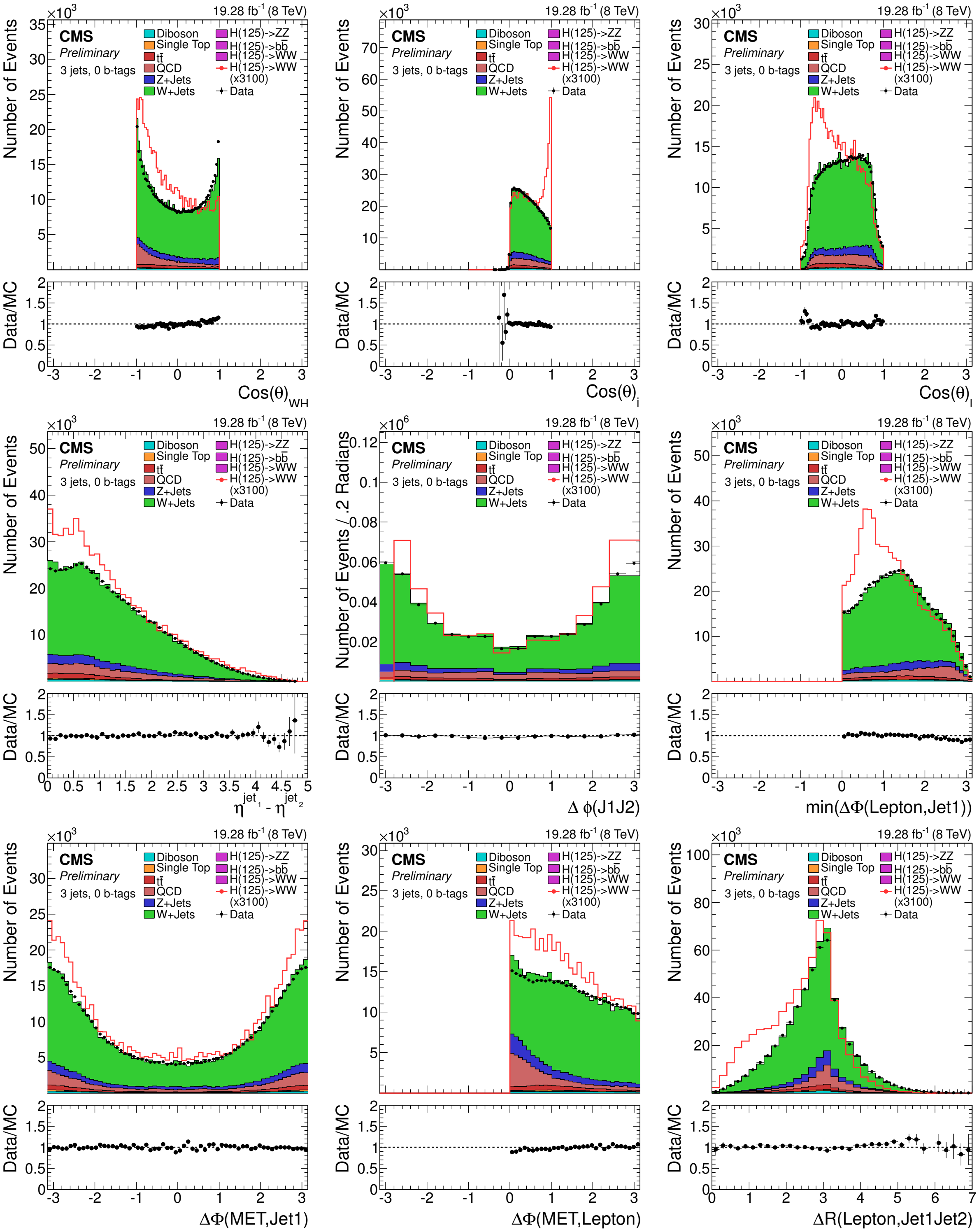


Figure C.9: Data-to-MC comparison plots for the 3-jet muon channel.

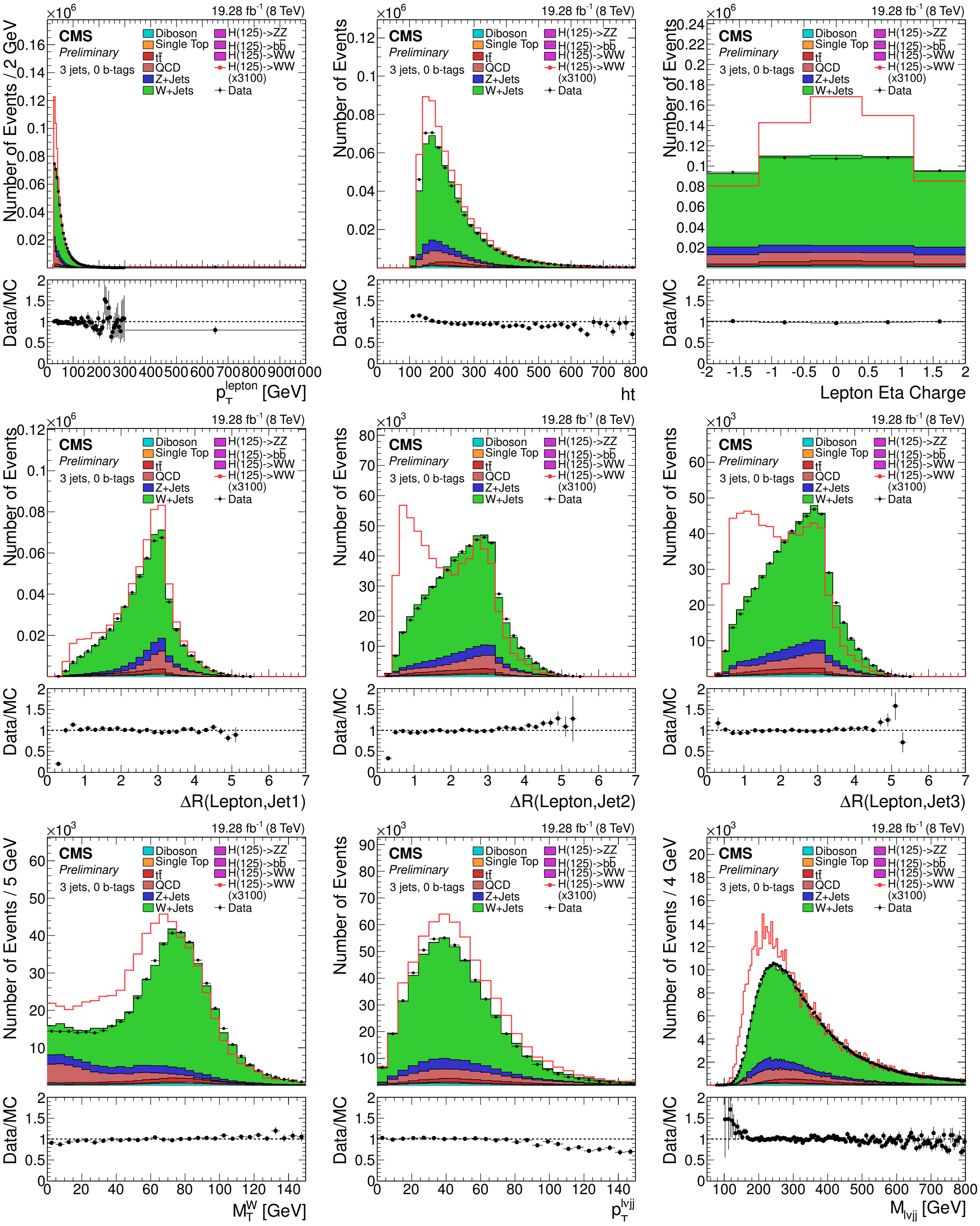


Figure C.10: Data-to-MC comparison plots for the 3-jet muon channel.

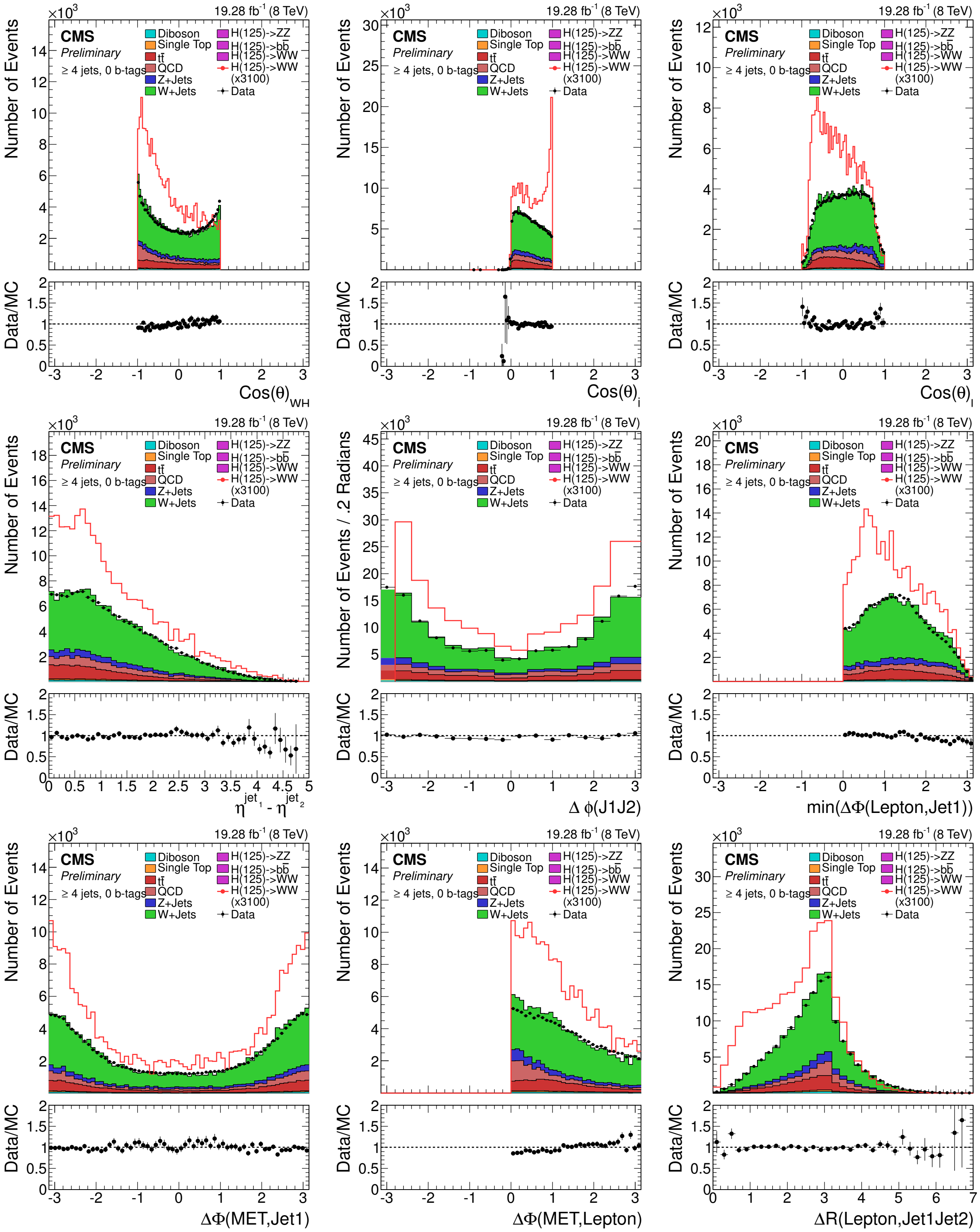


Figure C.11: Data-to-MC comparison plots for the ⩾4-jet muon channel.

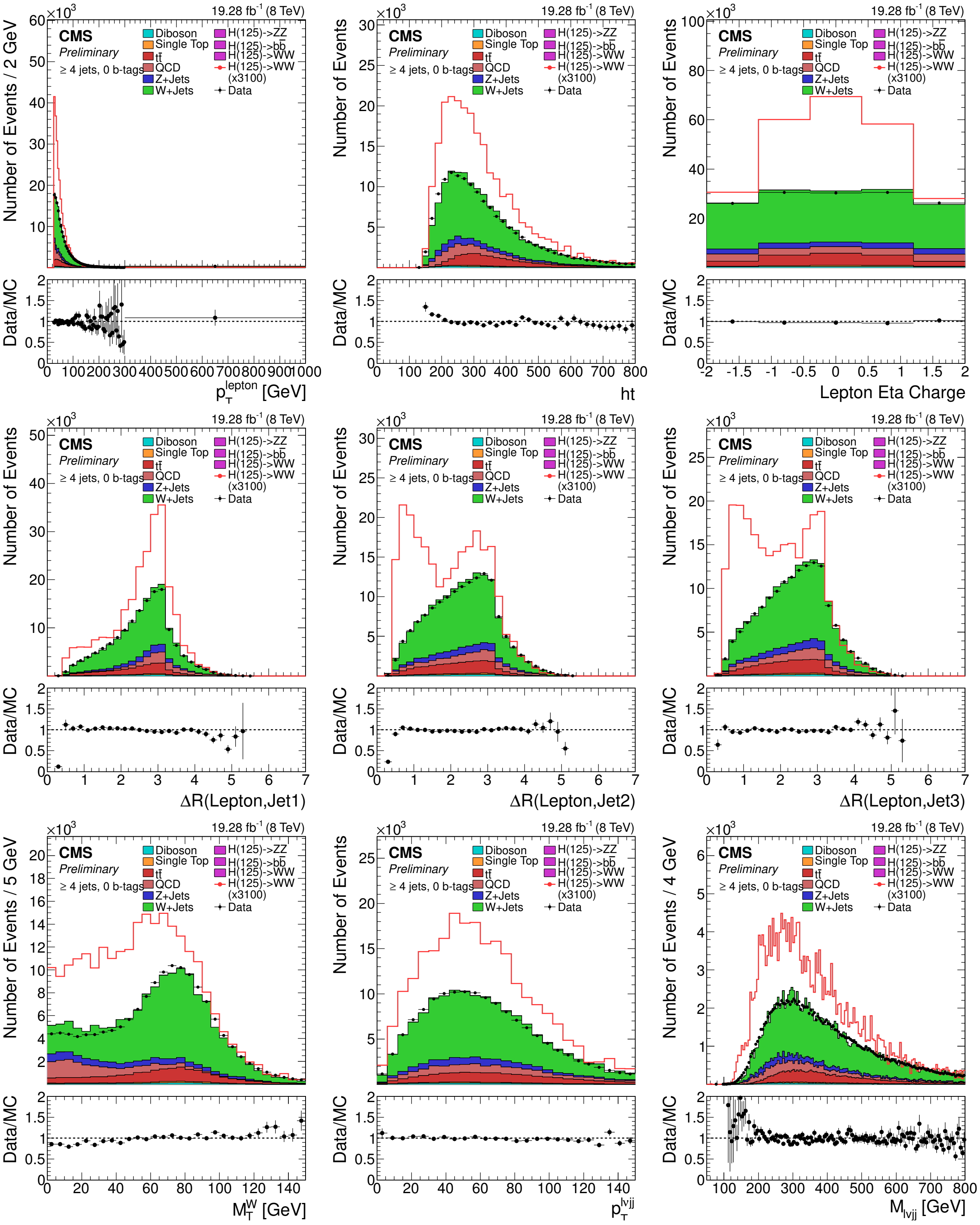
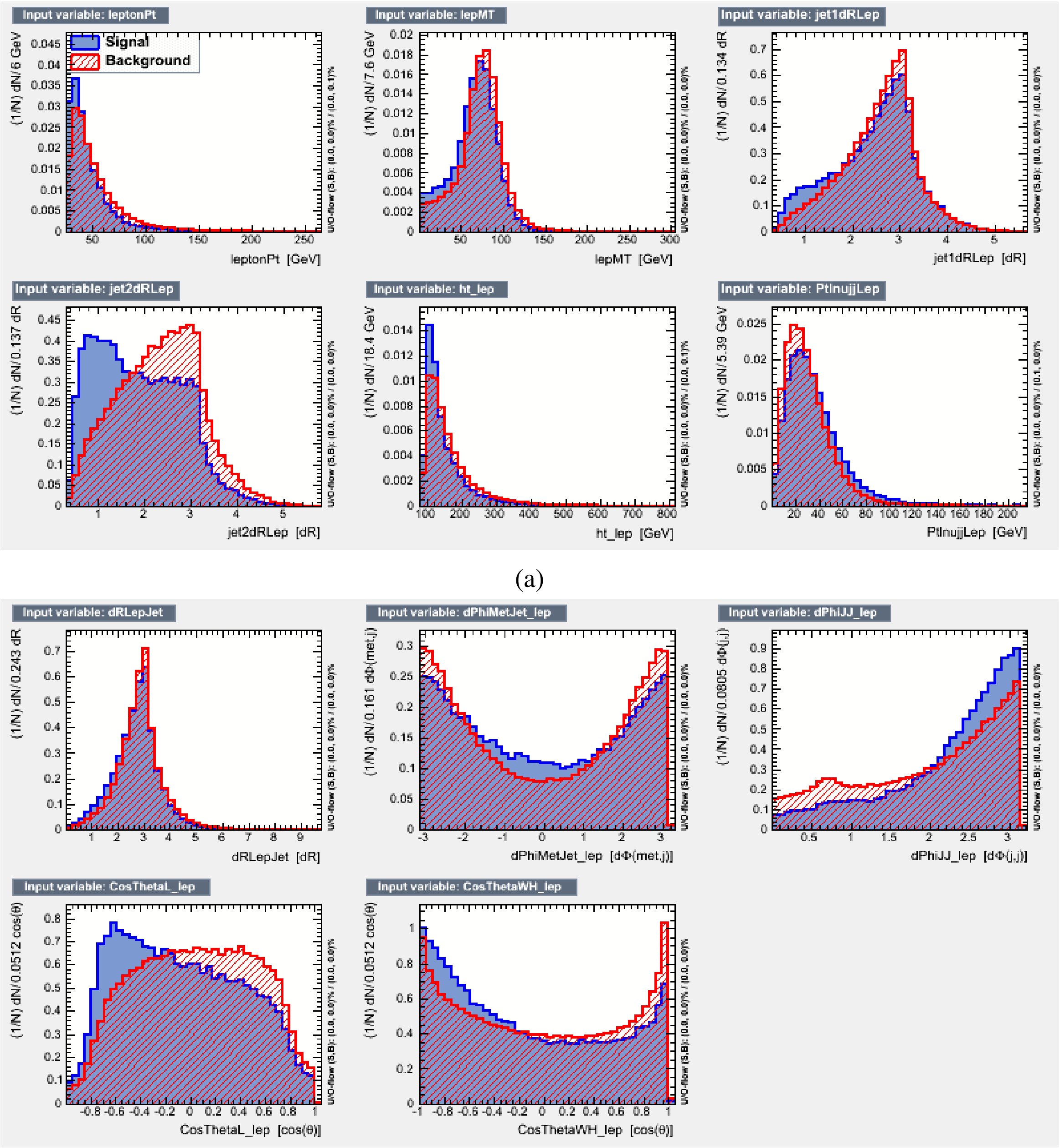


Figure C.12: Data-to-MC comparison plots for the ⩾4-jet muon channel.

APPENDIX D

BOOSTED DECISION TREES

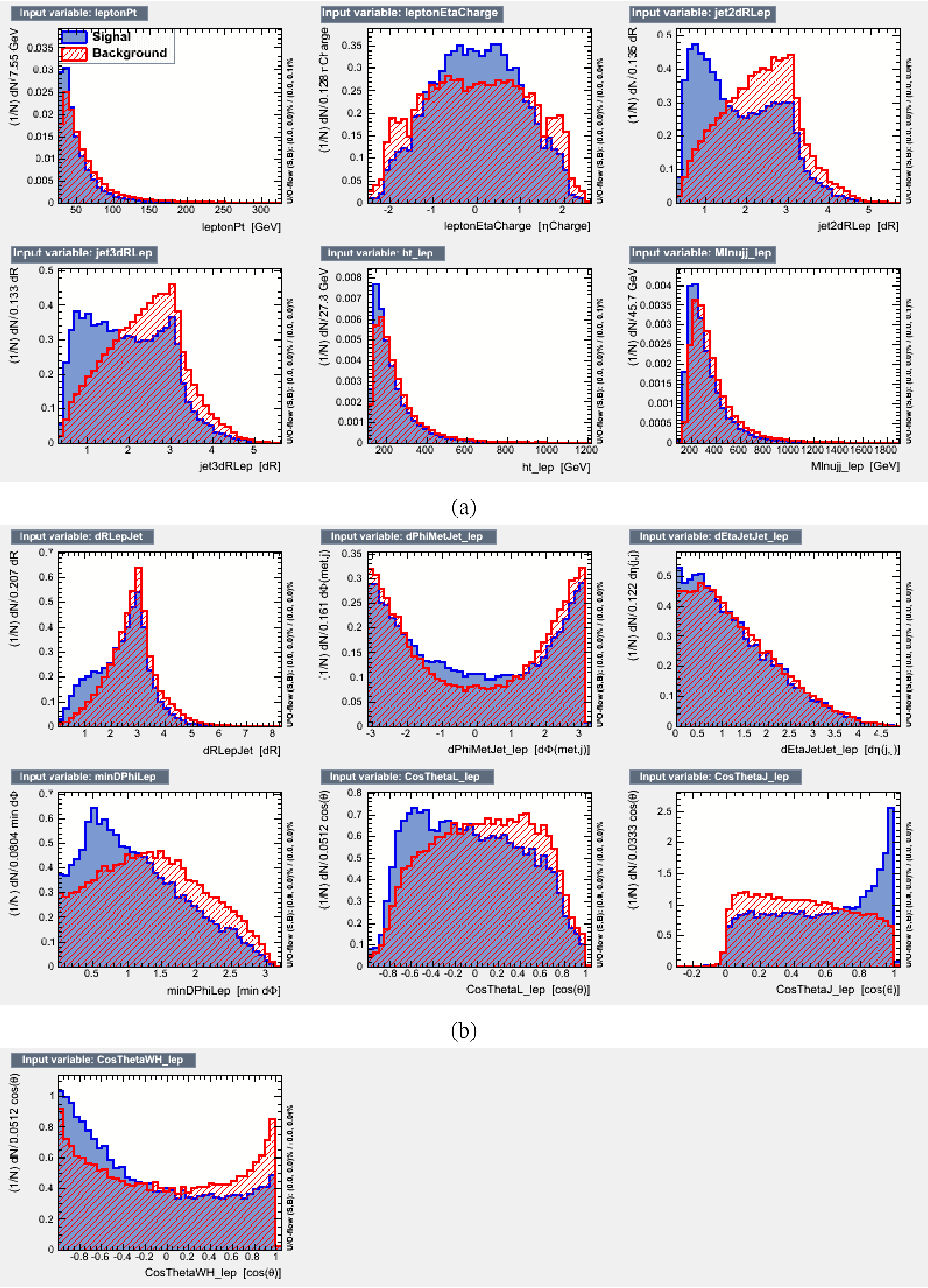
D.1 Inputs



(a)

(b)

Figure D.1: Inputs used to train the BDTs with kinematic variables in the 2 jets bin.

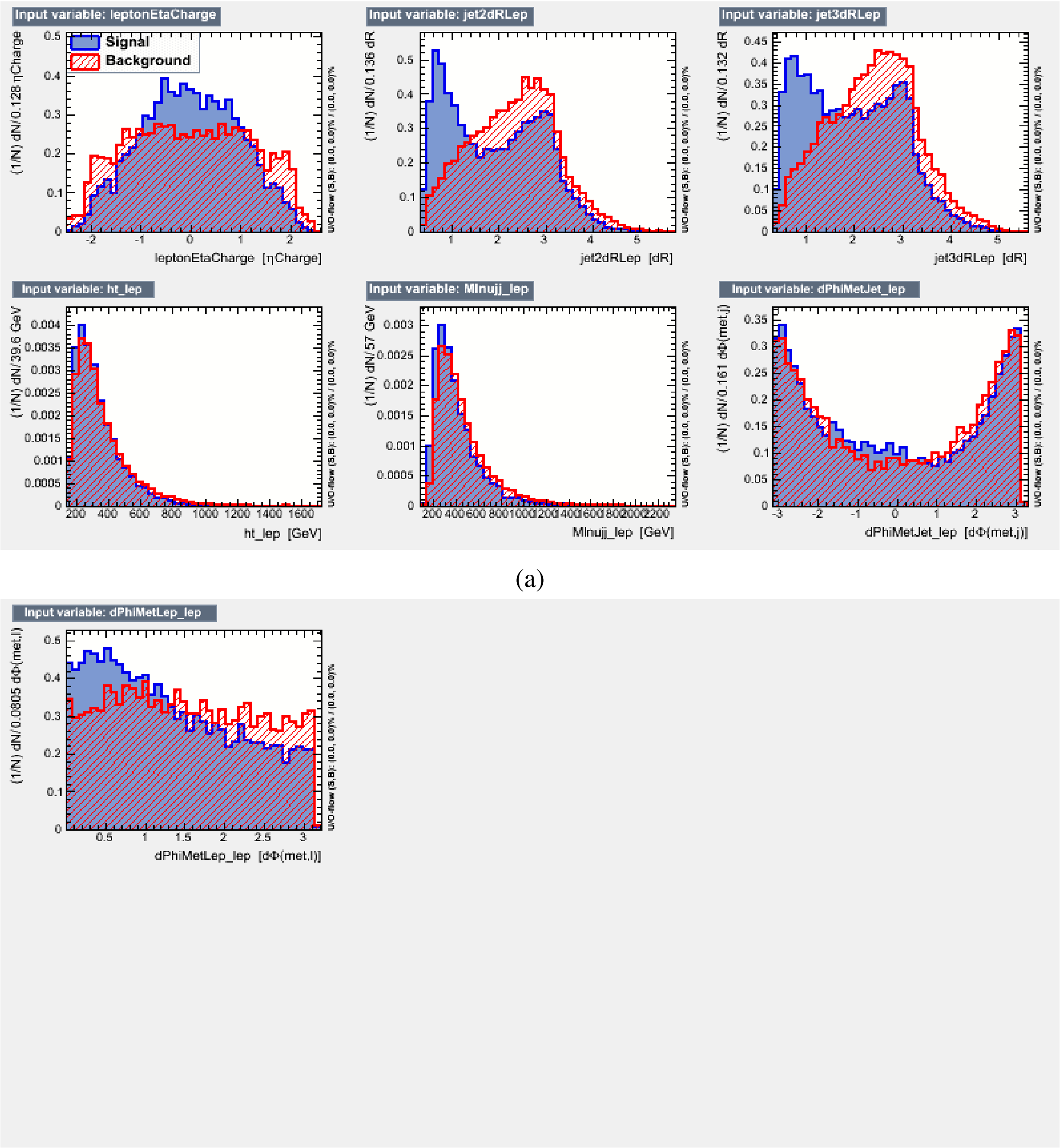


(a)

(b)

(c)

Figure D.2: Inputs used to train the BDTs with kinematic variables in the 3 jets bin.

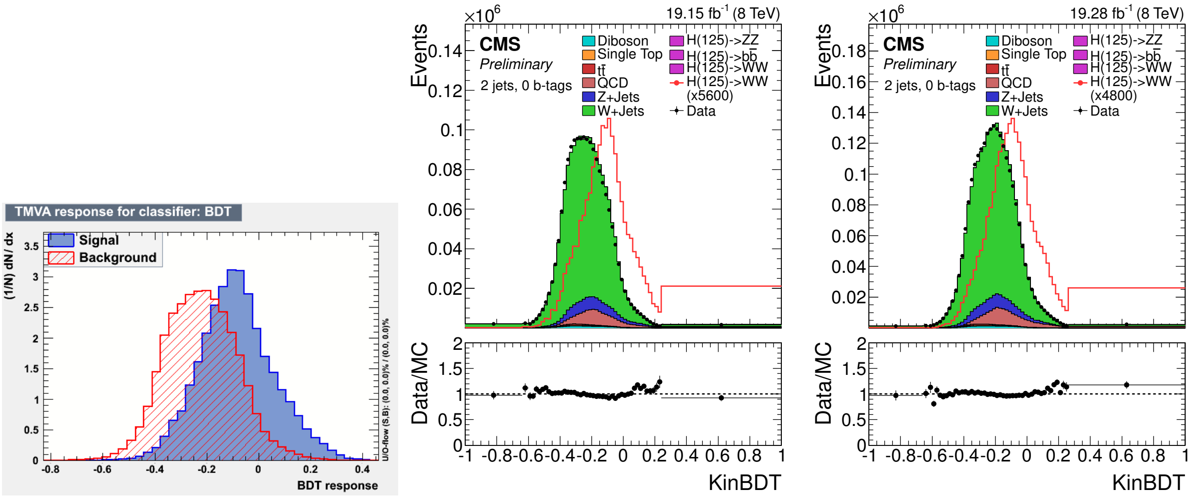


(a)

(b)

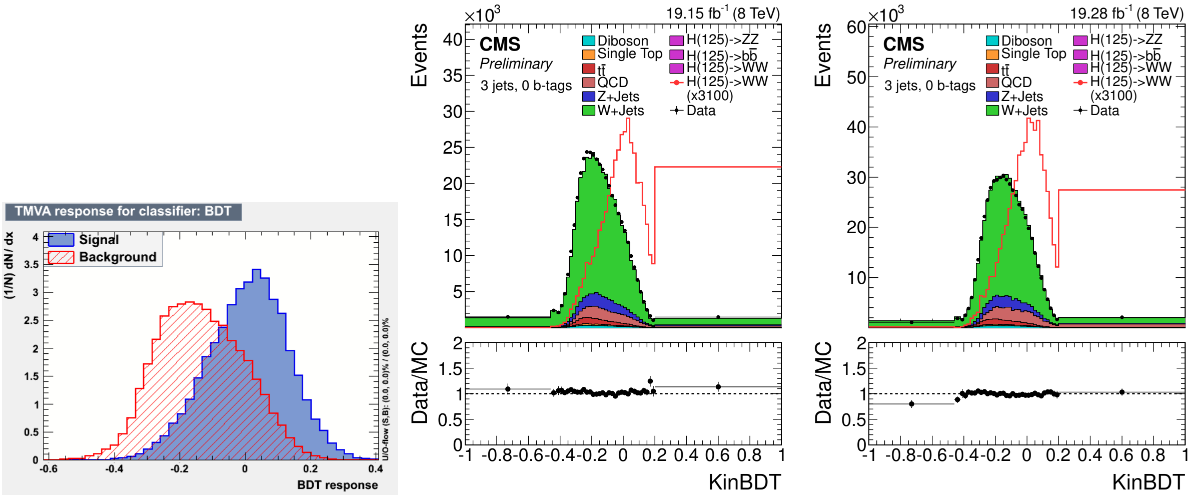
Figure D.3: Inputs used to train the BDTs with kinematic variables in the ⩾4 jets bin.

D.2 Outputs



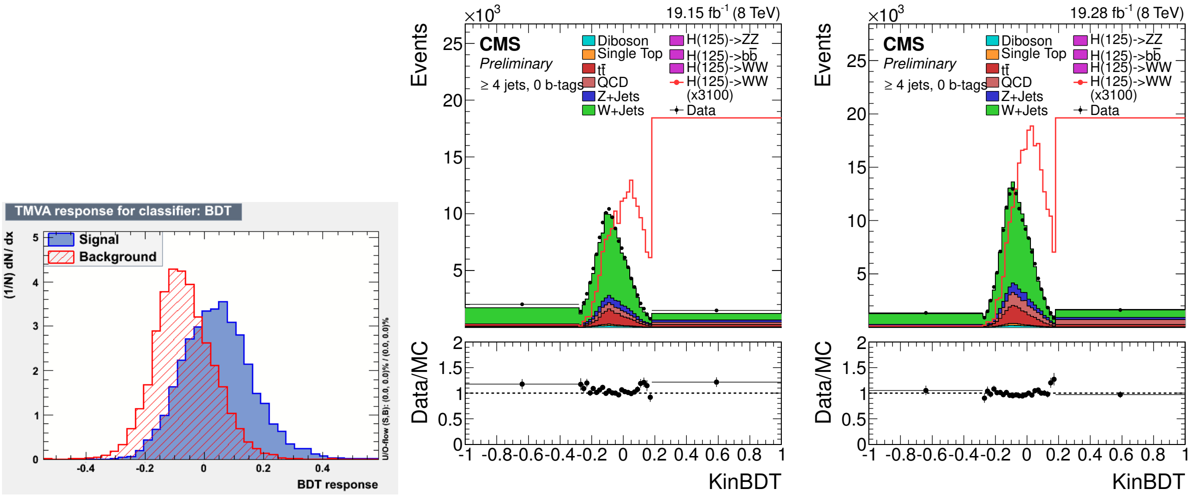
(a) (b) (c)

Figure D.4: (a) The BDT response plot from TMVA for the training with only kinematic variables in the 2 jet bin for the combined lepton channel. Validation plot for the BDT in the 2 jet bin for the (b) electron and (c) muon channels. Only statistical uncertainties are shown in the validation plots.



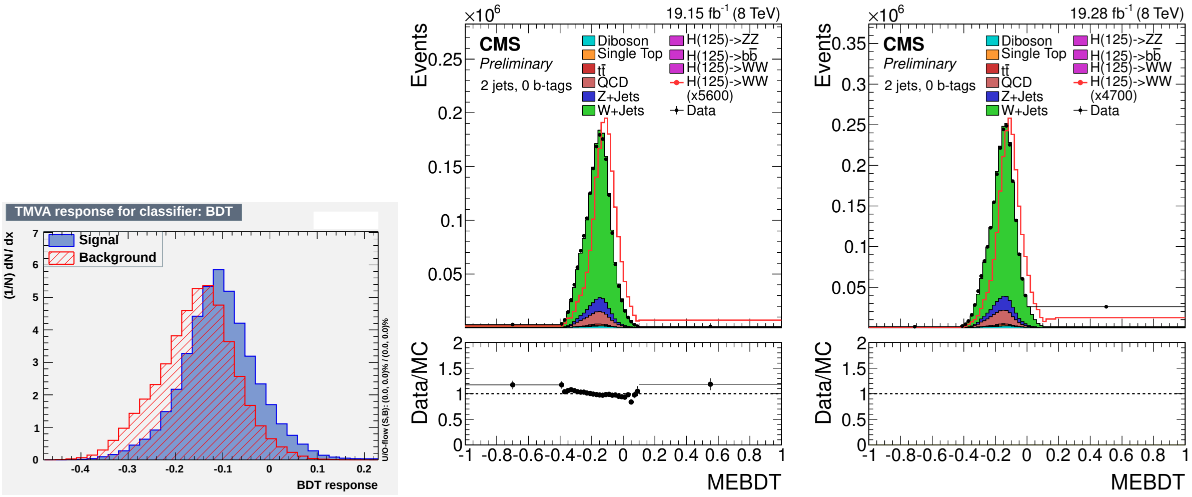
(a) (b) (c)

Figure D.5: (a) The BDT response plot from TMVA for the training with only kinematic variables in the 3 jet bin for the combined lepton channel. Validation plot for the BDT in the 3 jet bin for the (b) electron and (c) muon channels. Only statistical uncertainties are shown in the validation plots.



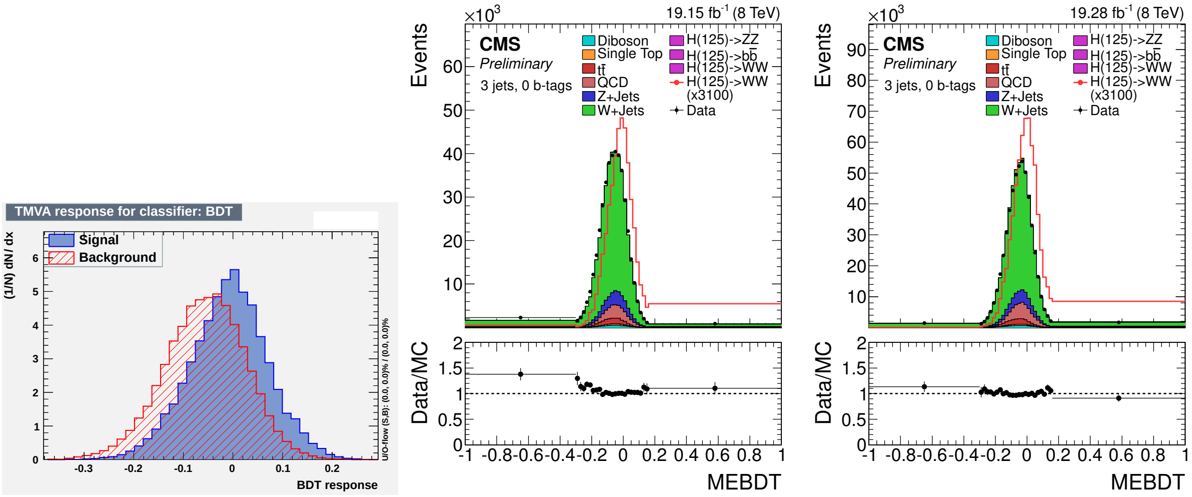
(a) (b) (c)

Figure D.6: (a) The BDT response plot from TMVA for the training with only kinematic variables in the ⩾4 jet bin for the combined lepton channel. Validation plot for the BDT in the ⩾4 jet bin for the (b) electron and (c) muon channels. Only statistical uncertainties are shown in the validation plots.



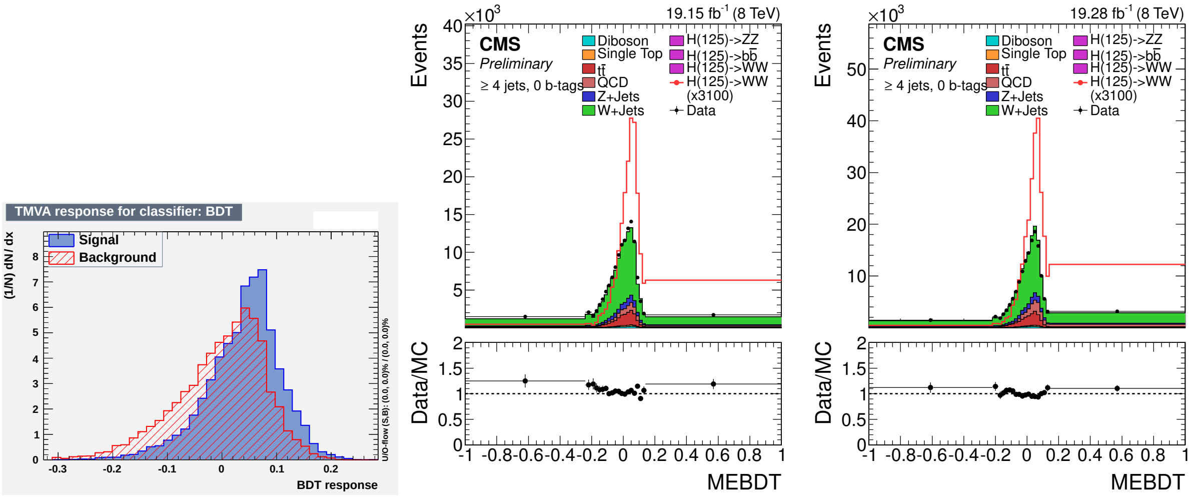
(a) (b) (c)

Figure D.7: (a) The BDT response plot from TMVA for the training with only matrix element probabilities in the 2 jet bin for the combined lepton channel. Validation plot for the BDT in the 2 jet bin for the (b) electron and (c) muon channels. Only statistical uncertainties are shown in the validation plots.



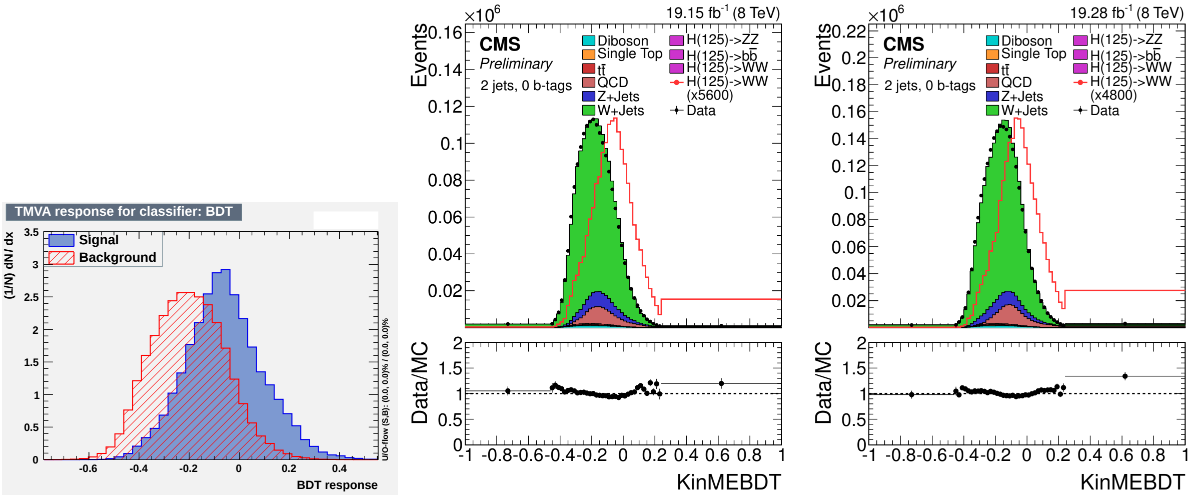
(a) (b) (c)

Figure D.8: (a) The BDT response plot from TMVA for the training with only matrix element probabilities in the 3 jet bin for the combined lepton channel. Validation plot for the BDT in the 3 jet bin for the (b) electron and (c) muon channels. Only statistical uncertainties are shown in the validation plots.



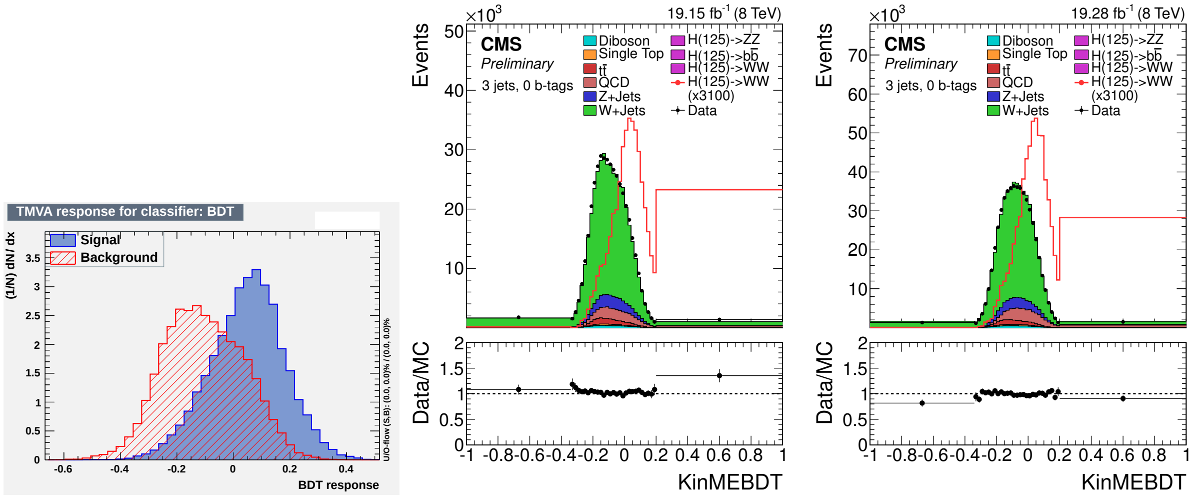
(a) (b) (c)

Figure D.9: (a) The BDT response plot from TMVA for the training with only matrix element probabilities in the ⩾4 jet bin for the combined lepton channel. Validation plot for the BDT in the ⩾4 jet bin for the (b) electron and (c) muon channels. Only statistical uncertainties are shown in the validation plots.



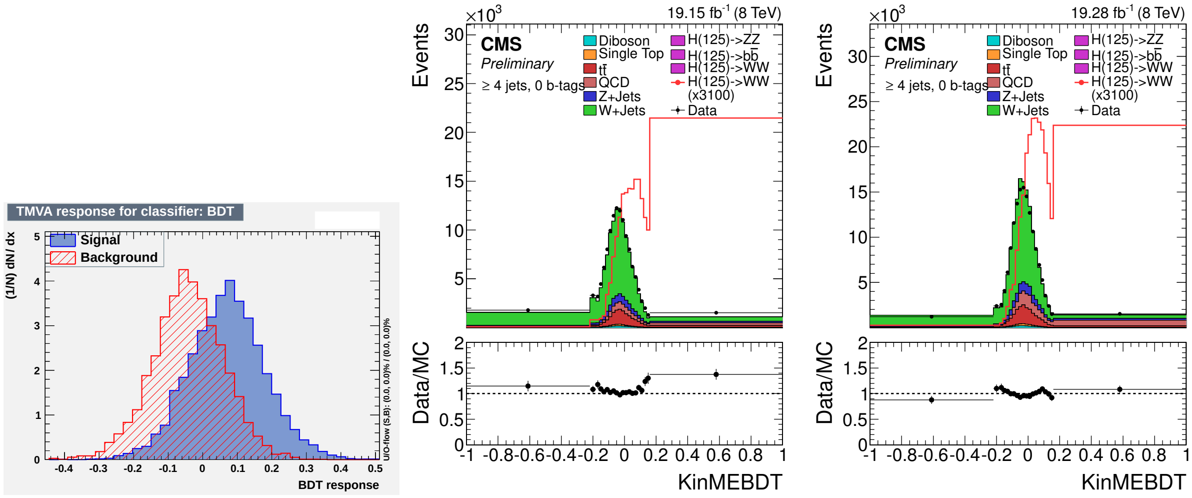
(a) (b) (c)

Figure D.10: (a) The BDT response plot from TMVA for the training with the kinematic variables and the ME BDT in the 2 jet bin for the combined lepton channel. Validation plot for the BDT in the 2 jet bin for the (b) electron and (c) muon channels. Only statistical uncertainties are shown in the validation plots.



(a) (b) (c)

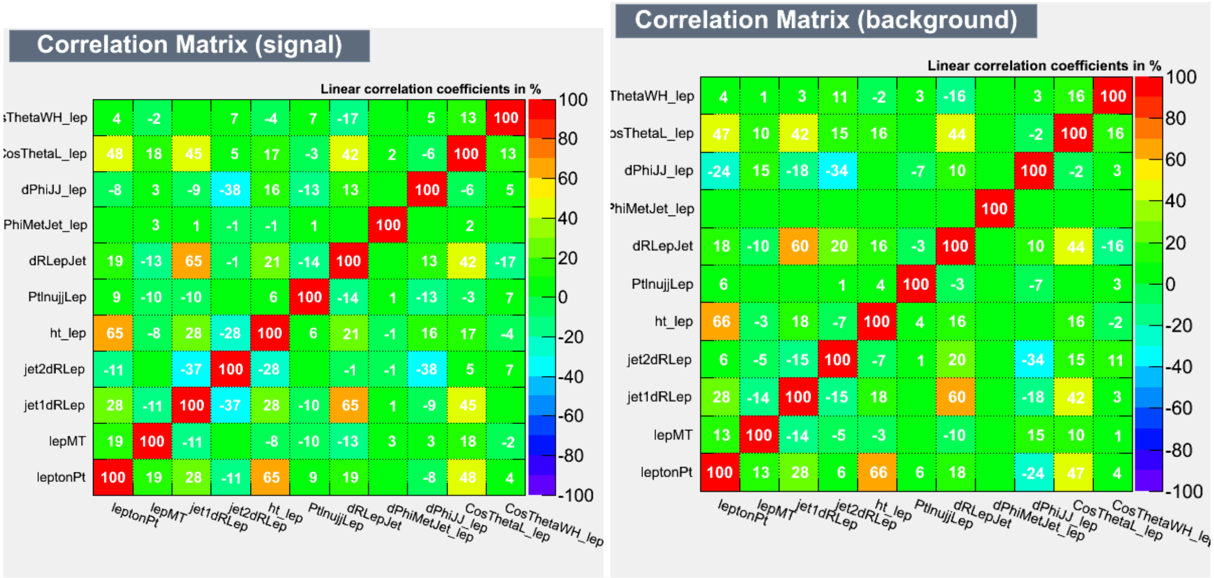
Figure D.11: (a) The BDT response plot from TMVA for the training with the kinematic variables and the ME BDT in the 3 jet bin for the combined lepton channel. Validation plot for the BDT in the 3 jet bin for the (b) electron and (c) muon channels. Only statistical uncertainties are shown in the validation plots.



(a) (b) (c)

Figure D.12: (a) The BDT response plot from TMVA for the training with the kinematic variables and the ME BDT in the ⩾4 jet bin for the combined lepton channel. Validation plot for the BDT in the ⩾4 jet bin for the (b) electron and (c) muon channels. Only statistical uncertainties are shown in the validation plots.

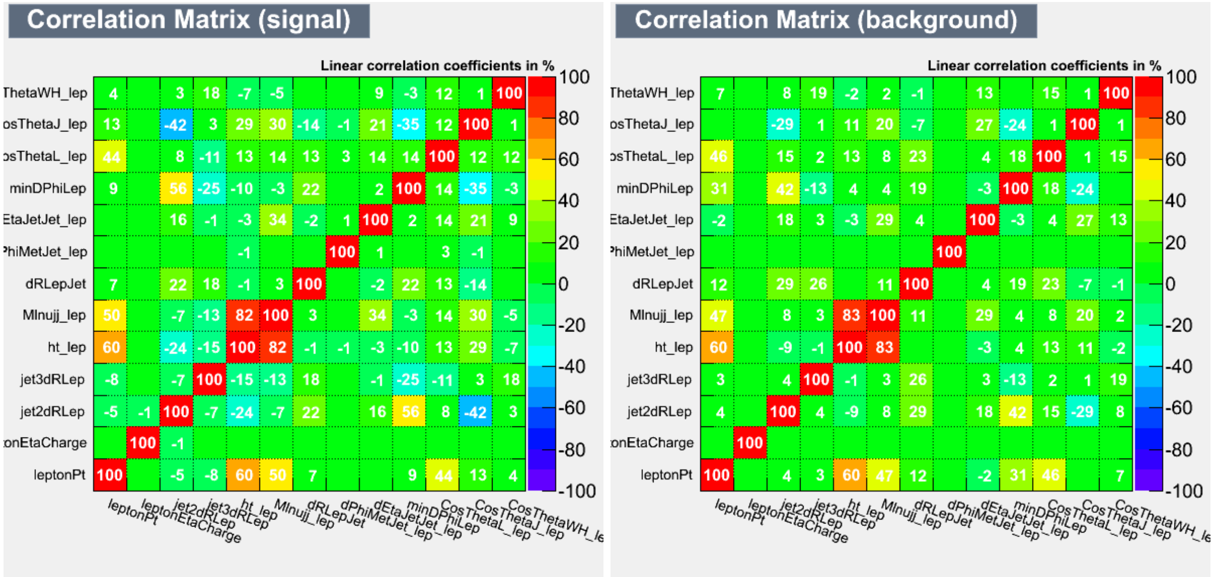
D.3 Correlations



(a)

(b)

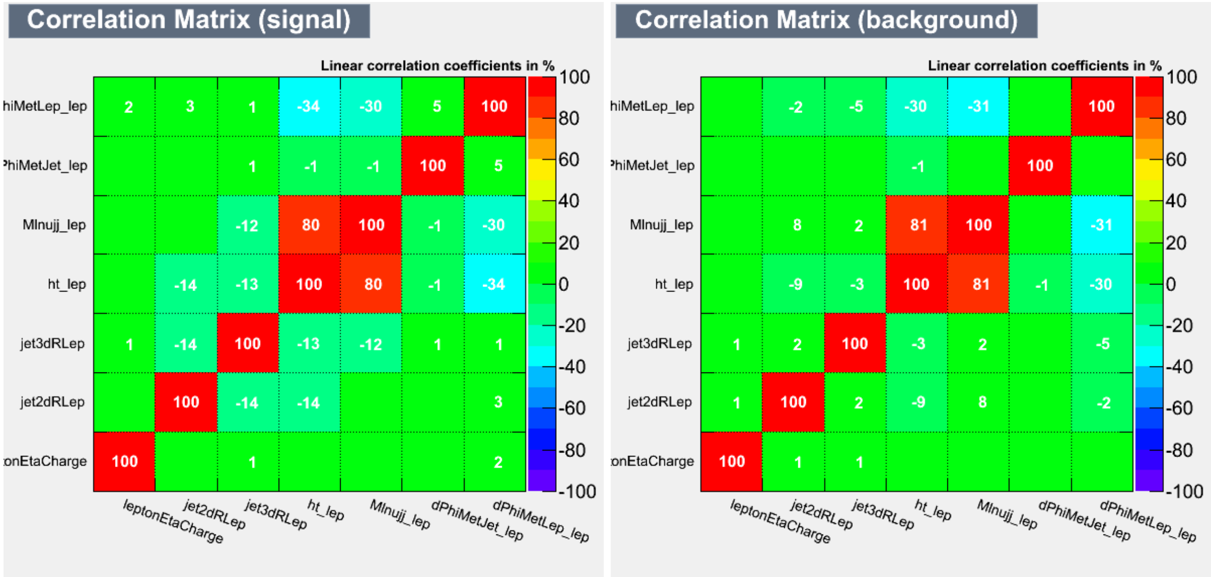
Figure D.13: Correlation plots for (a) signal and (b) background for the BDT trained with only kinematic variables in the 2 jet bin.



(a)

(b)

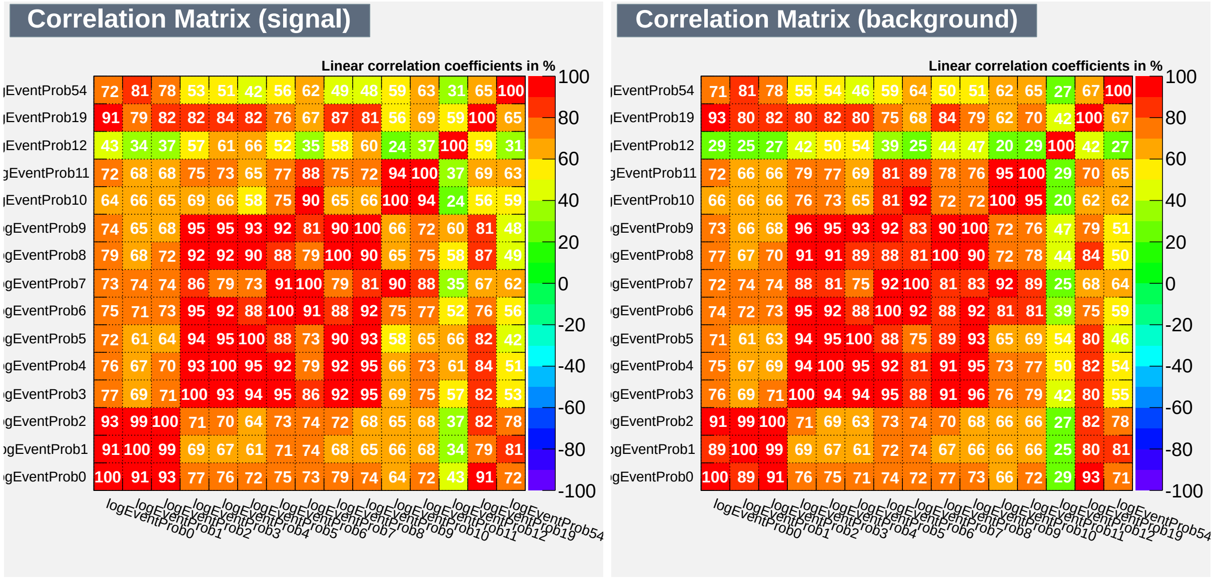
Figure D.14: Correlation plots for (a) signal and (b) background for the BDT trained with only kinematic variables in the 3 jet bin.



(a)

(b)

Figure D.15: Correlation plots for (a) signal and (b) background for the BDT trained with only kinematic variables in the ⩾4 jet bin.



Correlation Matrix (signal)

Linear correlation coefficients in %

	logEventProb0	logEventProb1	logEventProb2	logEventProb3	logEventProb4	logEventProb5	logEventProb6	logEventProb7	logEventProb8	logEventProb9	logEventProb10	logEventProb11	logEventProb12	logEventProb19	logEventProb54
gEventProb54	72	81	78	53	51	42	56	62	49	48	59	63	31	65	100
gEventProb19	91	79	82	82	84	82	76	67	87	81	56	69	59	100	65
gEventProb12	43	34	37	57	61	66	52	35	58	60	24	37	100	59	31
gEventProb11	72	68	68	75	73	65	77	88	75	72	94	100	37	69	63
gEventProb10	64	66	65	69	66	58	75	90	65	66	100	94	24	56	59
gEventProb9	74	65	68	95	95	93	92	81	90	100	66	72	60	81	48
gEventProb8	79	68	72	92	92	90	88	79	100	90	65	75	58	87	49
gEventProb7	73	74	74	86	79	73	91	100	79	81	90	88	35	67	62
gEventProb6	75	71	73	95	92	88	100	91	88	92	75	77	52	76	56
gEventProb5	72	61	64	94	95	100	88	73	90	93	58	65	66	82	42
gEventProb4	76	67	70	93	100	95	92	79	92	95	66	73	61	84	51
gEventProb3	77	69	71	100	93	94	95	86	92	95	69	75	57	82	53
gEventProb2	93	99	100	71	70	64	73	74	72	68	65	68	37	82	78
gEventProb1	91	100	99	69	67	61	71	74	68	65	66	68	34	79	81
gEventProb0	100	91	93	77	76	72	75	73	79	74	64	72	43	91	72

(a)

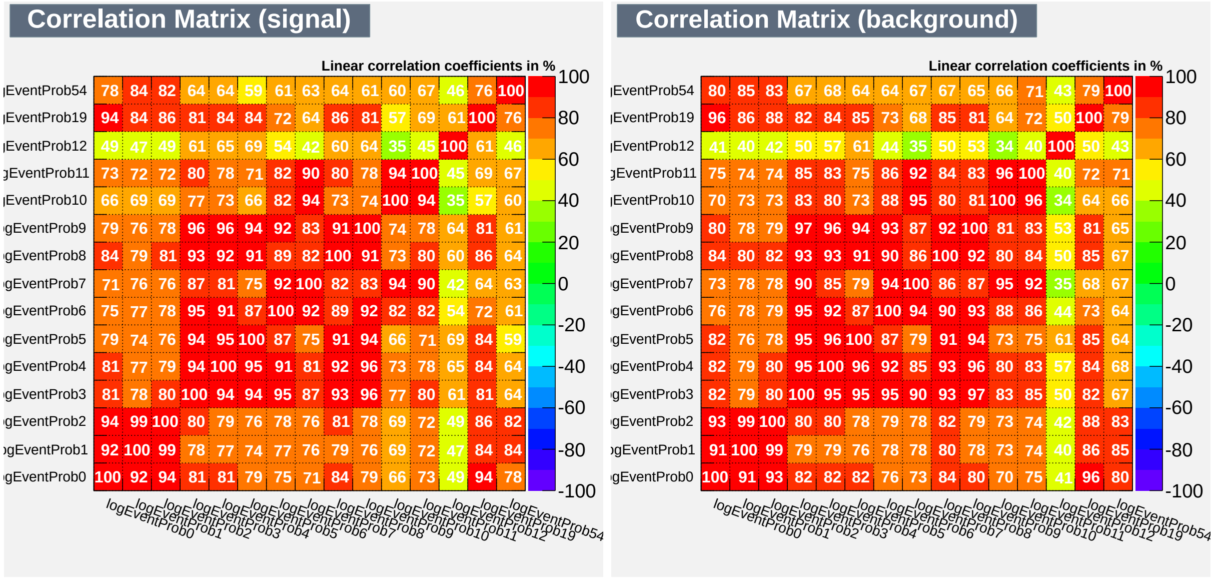
Correlation Matrix (background)

Linear correlation coefficients in %

	logEventProb0	logEventProb1	logEventProb2	logEventProb3	logEventProb4	logEventProb5	logEventProb6	logEventProb7	logEventProb8	logEventProb9	logEventProb10	logEventProb11	logEventProb12	logEventProb19	logEventProb54
gEventProb54	71	81	78	55	54	46	59	64	50	51	62	65	27	67	100
gEventProb19	93	80	82	80	82	80	75	68	84	79	62	70	42	100	67
gEventProb12	29	25	27	42	50	54	39	25	44	47	20	29	100	42	27
gEventProb11	72	66	66	79	77	69	81	89	78	76	95	100	29	70	65
gEventProb10	66	66	66	76	73	65	81	92	72	72	100	95	20	62	62
gEventProb9	73	66	68	96	95	93	92	83	90	100	72	76	47	79	51
gEventProb8	77	67	70	91	91	89	88	81	100	90	72	78	44	84	50
gEventProb7	72	74	74	88	81	75	92	100	81	83	92	89	25	68	64
gEventProb6	74	72	73	95	92	88	100	92	88	92	81	81	39	75	59
gEventProb5	71	61	63	94	95	100	88	75	89	93	65	69	54	80	46
gEventProb4	75	67	69	94	100	95	92	81	91	95	73	77	50	82	54
gEventProb3	76	69	71	100	94	94	95	88	91	96	76	79	42	80	55
gEventProb2	91	99	100	71	69	63	73	74	70	68	66	66	27	82	78
gEventProb1	89	100	99	69	67	61	72	74	67	66	66	66	25	80	81
gEventProb0	100	89	91	76	75	71	74	72	77	73	66	72	29	93	71

(b)

Figure D.16: Correlation plots for (a) signal and (b) background for the BDT trained with only matrix elements variables in the 2 jet bin.



Correlation Matrix (signal)

Linear correlation coefficients in %

	logEventProb0	logEventProb1	logEventProb2	logEventProb3	logEventProb4	logEventProb5	logEventProb6	logEventProb7	logEventProb8	logEventProb9	logEventProb10	logEventProb11	logEventProb12	logEventProb19	logEventProb54
gEventProb54	78	84	82	64	64	59	61	63	64	61	60	67	46	76	100
gEventProb19	94	84	86	81	84	84	72	64	86	81	57	69	61	100	76
gEventProb12	49	47	49	61	65	69	54	42	60	64	35	45	100	61	46
gEventProb11	73	72	72	80	78	71	82	90	80	78	94	100	45	69	67
gEventProb10	66	69	69	77	73	66	82	94	73	74	100	94	35	57	60
gEventProb9	79	76	78	96	96	94	92	83	91	100	74	78	64	81	61
gEventProb8	84	79	81	93	92	91	89	82	100	91	73	80	60	86	64
gEventProb7	71	76	76	87	81	75	92	100	82	83	94	90	42	64	63
gEventProb6	75	77	78	95	91	87	100	92	89	92	82	82	54	72	61
gEventProb5	79	74	76	94	95	100	87	75	91	94	66	71	69	84	59
gEventProb4	81	77	79	94	100	95	91	81	92	96	73	78	65	84	64
gEventProb3	81	78	80	100	94	94	95	87	93	96	77	80	61	81	64
gEventProb2	94	99	100	80	79	76	78	76	81	78	69	72	49	86	82
gEventProb1	92	100	99	78	77	74	77	76	79	76	69	72	47	84	84
gEventProb0	100	92	94	81	81	79	75	71	84	79	66	73	49	94	78

(a)

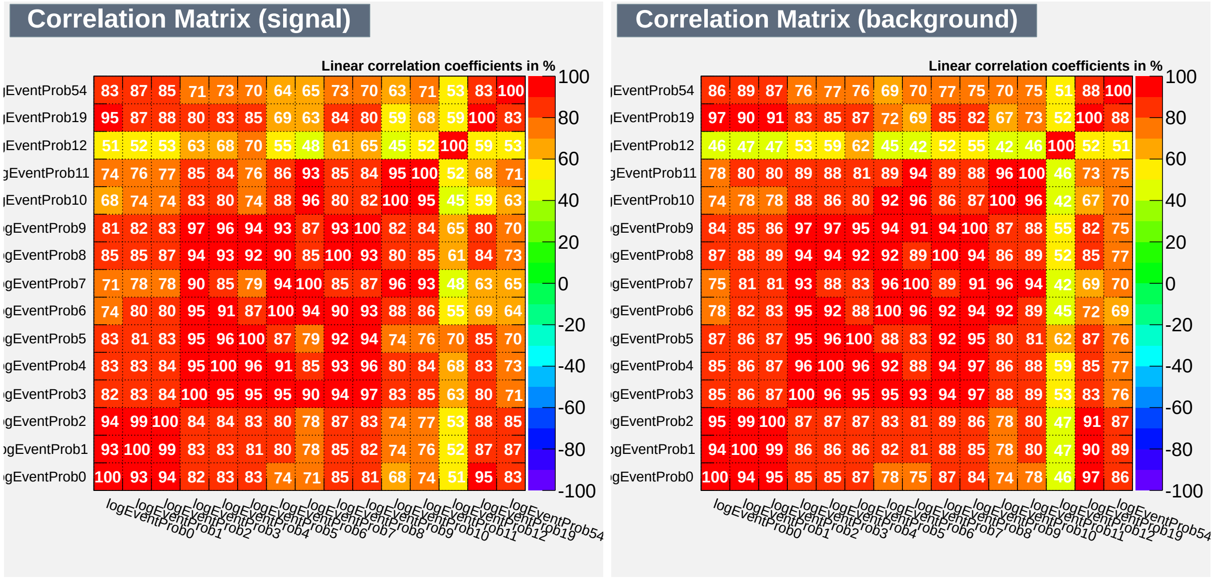
Correlation Matrix (background)

Linear correlation coefficients in %

	logEventProb0	logEventProb1	logEventProb2	logEventProb3	logEventProb4	logEventProb5	logEventProb6	logEventProb7	logEventProb8	logEventProb9	logEventProb10	logEventProb11	logEventProb12	logEventProb19	logEventProb54
gEventProb54	80	85	83	67	68	64	64	67	67	65	66	71	43	79	100
gEventProb19	96	86	88	82	84	85	73	68	85	81	64	72	50	100	79
gEventProb12	41	40	42	50	57	61	44	35	50	53	34	40	100	50	43
gEventProb11	75	74	74	85	83	75	86	92	84	83	96	100	40	72	71
gEventProb10	70	73	73	83	80	73	88	95	80	81	100	96	34	64	66
gEventProb9	80	78	79	97	96	94	93	87	92	100	81	83	53	81	65
gEventProb8	84	80	82	93	93	91	90	86	100	92	80	84	50	85	67
gEventProb7	73	78	78	90	85	79	94	100	86	87	95	92	35	68	67
gEventProb6	76	78	79	95	92	87	100	94	90	93	88	86	44	73	64
gEventProb5	82	76	78	95	96	100	87	79	91	94	73	75	61	85	64
gEventProb4	82	79	80	95	100	96	92	85	93	96	80	83	57	84	68
gEventProb3	82	79	80	100	95	95	95	90	93	97	83	85	50	82	67
gEventProb2	93	99	100	80	80	78	79	78	82	79	73	74	42	88	83
gEventProb1	91	100	99	79	79	76	78	78	80	78	73	74	40	86	85
gEventProb0	100	91	93	82	82	82	76	73	84	80	70	75	41	96	80

(b)

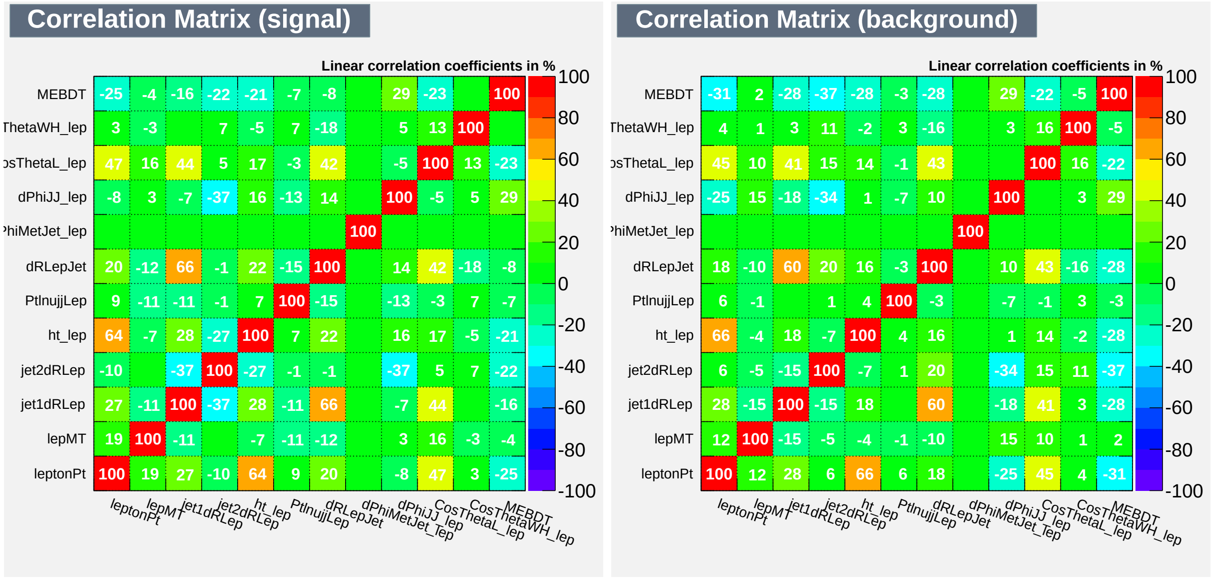
Figure D.17: Correlation plots for (a) signal and (b) background for the BDT trained with only matrix elements variables in the 3 jet bin.



(a)

(b)

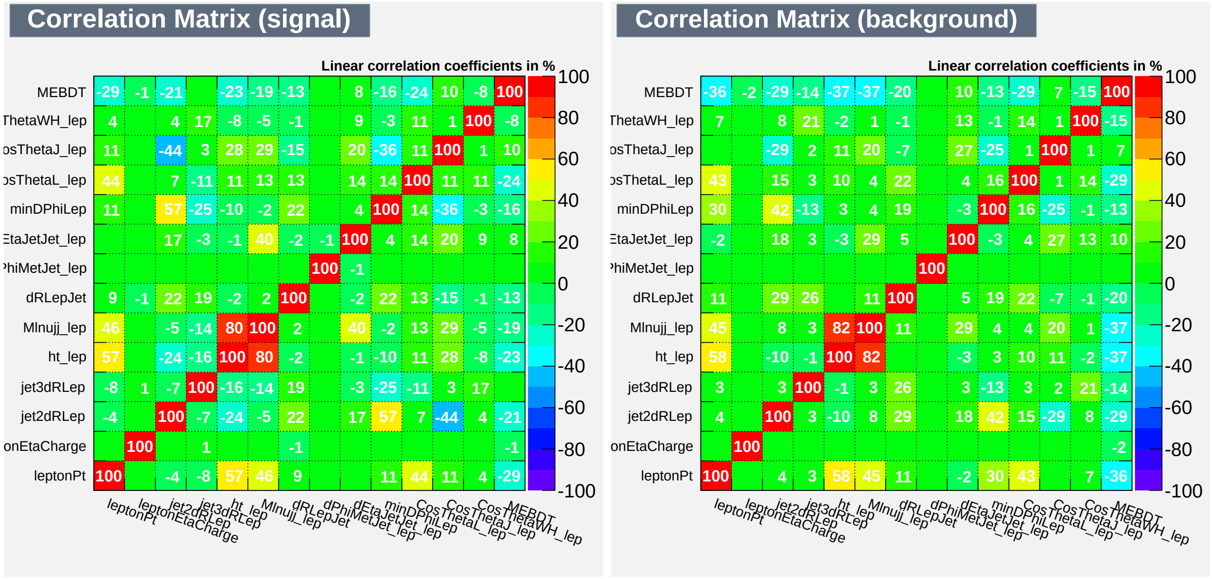
Figure D.18: Correlation plots for (a) signal and (b) background for the BDT trained with only matrix elements variables in the ⩾4 jet bin.



(a)

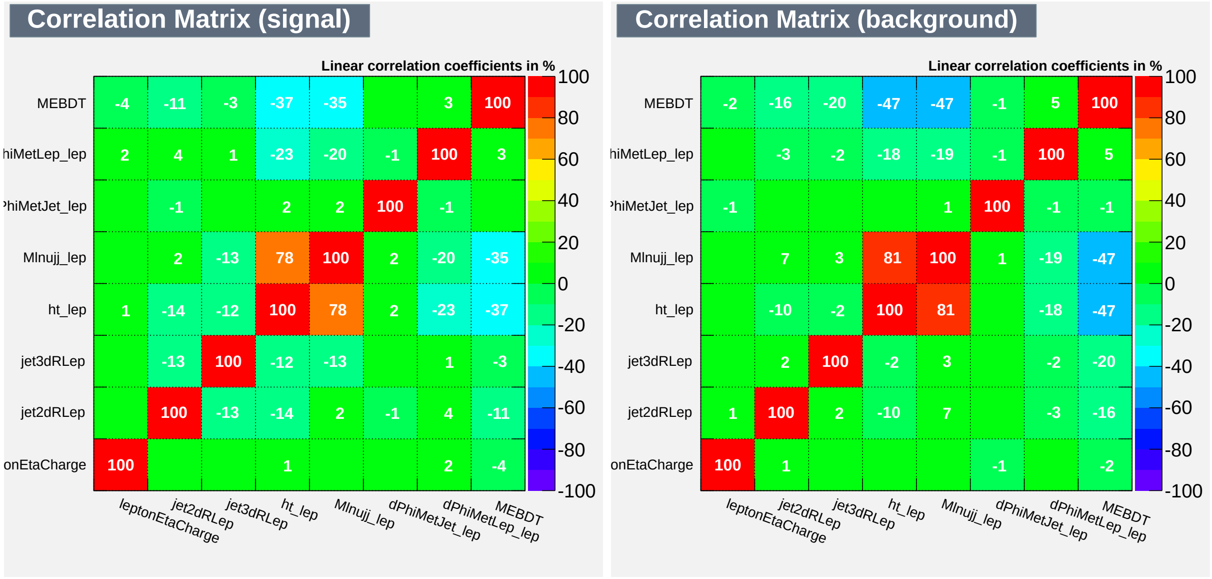
(b)

Figure D.19: Correlation plots for (a) signal and (b) background for the BDT trained with both the kinematic variables and the ME BDT in the 2 jet bin.



(a) (b)

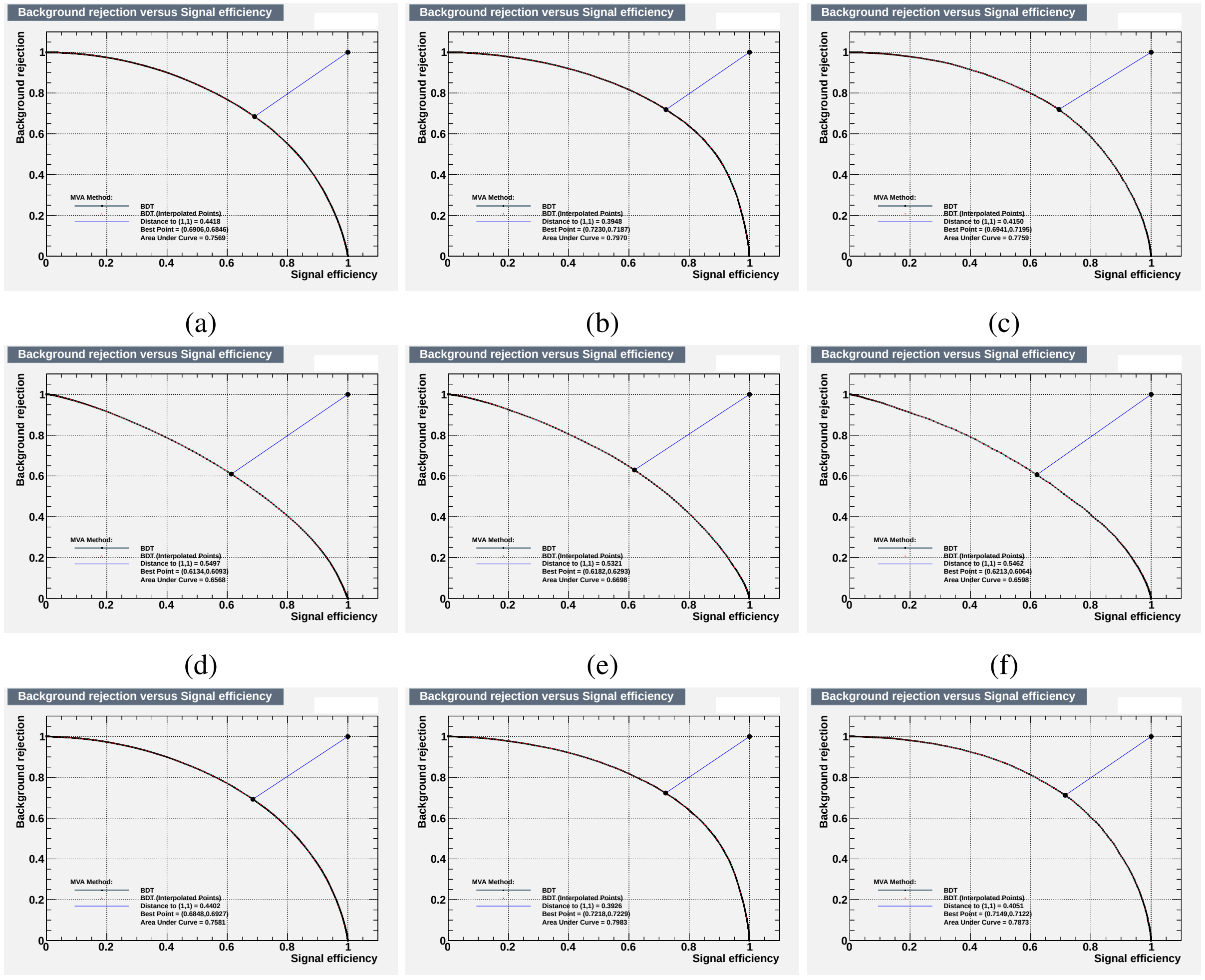
Figure D.20: Correlation plots for (a) signal and (b) background for the BDT trained with both the kinematic variables and the ME BDT in the 3 jet bin.



(a) (b)

Figure D.21: Correlation plots for (a) signal and (b) background for the BDT trained with both the kinematic variables and the ME BDT in the ⩾4 jet bin.

D.4 ROC Curves



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Figure D.22: The receiver operating characteristic (ROC) curves for the various BDT trainings. The plots are ordered by jet bin from left to right, with the leftmost plot being the two-jet bin and the rightmost plot being the greater than or equal to four-jet bin. The top row contains the KinBDT plots while the middle and bottom rows contain the MEBDT and KinMEBDT plots, respectively.