A Search For the Standard Model Higgs Boson Produced in Association with Top Quarks in the Lepton + Jets Channel at CMS

DISSERTATION

Presented in Partial Fulfillment of the Requirements for the Degree Doctor of Philosophy in the Graduate School of The Ohio State University

By

Geoffrey N. Smith, B.S., M.S.

Graduate Program in Physics

The Ohio State University

2014

Dissertation Committee:

R. Hughes, Advisor

B. Winer

M. Lisa

E. Braaten

Abstract

A search for the Standard Model Higgs boson in the $t\bar{t}H$ production mode is presented. The search uses $19.3\,\text{fb}^{-1}$ of data collected at $\sqrt{s} = 8\,\text{TeV}$ by the Compact Muon Solenoid detector at the Large Hadron Collider. The focus of the analysis is the semi-leptonic decay of the $t\bar{t}$ pair, accompanied by the decay of the Higgs boson to b-quarks. The search sets a 95% confidence upper limit of $\mu < 5.1$, where μ is the ratio of the $t\bar{t}H$ cross-section to the $t\bar{t}H$ cross-section predicted by the Standard Model.

To my parents.

Vita

August 1982	Born – New London, CT
June 2001	Diploma, Stonington High School, Stonington, Connecticut
December 2006	B.S., University of Connecticut, Storrs, Connecticut
September 2008 – May 2011	Graduate Teaching Associate, The Ohio State University, Columbus, Ohio
June 2011 – May 2014	Graduate Research Associate, The Ohio State University, Columbus, Ohio
September 2011	M.S. Physics, The Ohio State University, Columbus, Ohio

Publications

CMS Collaboration, Measurements of the $t\bar{t}$ charge asymmetry using the dilepton decay channel in pp collisions at $\sqrt{s}$=7 TeV, 2014. arXiv:1402.3803.

CMS Collaboration, Search for W' to tb decays in the lepton + jets final state in pp collisions at $\sqrt{s}$ = 8 TeV, 2014. arXiv:1402.2176.

CMS Collaboration, Measurement of the production cross sections for a Z boson and one or more b jets in pp collisions at $\sqrt{s}$ = 7 TeV, 2014. arXiv:1402.1521.

CMS Collaboration, Measurement of inclusive W and Z boson production cross sections in pp collisions at $\sqrt{s}$ =8 TeV, 2014. arXiv:1402.0923.

CMS Collaboration, Evidence for the direct decay of the 125 GeV Higgs boson to fermions, 2014. arXiv:1401.6527.

CMS Collaboration, Evidence for the 125 GeV Higgs boson decaying to a pair of τ leptons, 2014. arXiv:1401.5041.

CMS Collaboration, Studies of dijet pseudorapidity distributions and transverse momentum balance in pPb collisions at $\sqrt{s_{NN}}$=5.02 TeV, 2014. arXiv:1401.4433.

CMS Collaboration, Observation of the associated production of a single top quark and a W boson in pp collisions at $\sqrt{s}$ = 8 TeV, 2014. arXiv:1401.2942.

CMS Collaboration, Measurement of the $t\bar{t}$ production cross section in the dilepton channel in pp collisions at $\sqrt{s}$=8 TeV, 2014. arXiv:1312.7582.

CMS Collaboration, Measurement of the production cross section for a W boson and two b jets in pp collisions at $\sqrt{s}$ = 7 TeV, 2013. arXiv:1312.6608.

CMS Collaboration, Measurement of four-jet production in proton-proton collisions at $\sqrt{s}$=7 TeV, 2013. arXiv:1312.6440.

CMS Collaboration, Measurement of the muon charge asymmetry in inclusive pp $\rightarrow$ W + X production at $\sqrt{s}$=7 TeV and an improved determination of light parton, 2013. arXiv:1312.6283.

CMS Collaboration, Event activity dependence of Y(nS) production in $\sqrt{s_{NN}}$=5.02 TeV pPb and $\sqrt{s}$=2.76 TeV pp collisions, 2013. arXiv:1312.6300.

CMS Collaboration, Study of double parton scattering using W + 2-jet events in proton-proton collisions at $\sqrt{s}$ = 7 TeV, 2013. arXiv:1312.5729.

CMS Collaboration, Measurement of the properties of a Higgs boson in the four-lepton final state, 2013. arXiv:1312.5353.

CMS Collaboration, Evidence of b-jet quenching in PbPb collisions at $\sqrt{s_{NN}}$ = 2.76 TeV, 2013. arXiv:1312.4198.

CMS Collaboration, Search for flavor-changing neutral currents in top-quark decays $t \rightarrow Zq$ in pp collisions at $\sqrt{s}$=8 TeV, 2013. arXiv:1312.4194.

CMS Collaboration, Search for stop and higgsino production using diphoton Higgs boson decays, 2013. arXiv:1312.3310.

CMS Collaboration, Search for top-quark partners with charge 5/3 in the same-sign dilepton final state, 2013. arXiv:1312.2391.

CMS Collaboration, Studies of azimuthal dihadron correlations in ultra-central PbPb collisions at $\sqrt{s_{NN}}$ = 2.76 TeV, 2013. arXiv:1312.1845.

CMS Collaboration, Measurement of Higgs boson production and properties in the WW decay channel with leptonic final states, 2014. arXiv:1312.1129.

CMS Collaboration, Inclusive search for a vector-like T quark with charge 2/3 in pp collisions at $\sqrt{s}$=8 TeV, 2014. arXiv:1311.7667.

CMS Collaboration, Search for new physics in events with same-sign dileptons and jets in pp collisions at $\sqrt{s}$=8 TeV, 2014. arXiv:1311.6736.

CMS Collaboration, Measurement of the triple-differential cross section for photon+jets production in proton-proton collisions at $\sqrt{s}$=7 TeV, 2013. arXiv:1311.6141.

CMS Collaboration, Probing color coherence effects in pp collisions at $\sqrt{s}$ = 7 TeV, 2013. arXiv:1311.5815.

CMS Collaboration, Search for pair production of excited top quarks in the lepton+jets final state, 2013. arXiv:1311.5357.

CMS Collaboration, Search for supersymmetry in pp collisions at $\sqrt{s}$ = 8 TeV in events with a single lepton, large jet multiplicity, and multiple b jets, 2013. arXiv:1311.4937.

CMS Collaboration, Measurements of $t\bar{t}$ spin correlations and top-quark polarization using dilepton final states in pp collisions at $\sqrt{s}$ = 7 TeV, 2013. arXiv:1311.3924.

CMS Collaboration, Searches for light- and heavy-flavour three-jet resonances in pp collisions at $\sqrt{s}$ = 8 TeV, 2014. arXiv:1311.1799.

CMS Collaboration, Measurement of higher-order harmonic azimuthal anisotropy in PbPb collisions at a nucleon-nucleon center-of-mass energy of 2.76 TeV Measurement of higher-order harmonic azimuthal anisotropy in PbPb collisions at $\sqrt{s_{NN}}$ = 2.76 TeV, 2013. arXiv:1310.8651.

CMS Collaboration, Measurement of the differential and double-differential Drell-Yan cross sections in proton-proton collisions at $\sqrt{s}$ = 7 TeV, 2013. arXiv:1310.7291.

CMS Collaboration, Jet and underlying event properties as a function of charged-particle multiplicity in proton-proton collisions at $\sqrt{s}$ = 7 TeV, 2013. arXiv:1310.4554.

CMS Collaboration, Search for the standard model Higgs boson produced in association with a W or a Z boson and decaying to bottom quarks, 2014. arXiv:1310.3687.

CMS Collaboration, Rapidity distributions in exclusive Z + jet and photon + jet events in pp collisions at $\sqrt{s}$=7 TeV Rapidity distributions in exclusive Z + jet and γ + jet events in pp collisions at $\sqrt{s}$=7 TeV, 2013. arXiv:1310.3082.

CMS Collaboration, Search for baryon number violation in top quark decays, 2013. arXiv:1310.1618.

CMS Collaboration, Measurement of associated W + charm production in pp collisions at $\sqrt{s} = 7$ TeV, 2013. arXiv:1310.1138.

CMS Collaboration, Measurement of the cross section and angular correlations for associated production of a Z boson with b hadrons in pp collisions at $\sqrt{s} = 7$ TeV, 2013. arXiv:1310.1349.

CMS Collaboration, Modification of jet shapes in PbPb collisions at $\sqrt{s_{NN}} = 2.76$ TeV, 2014. arXiv:1310.0878.

CMS Collaboration, Observation of a peaking structure in the J/$\psi\phi$ mass spectrum from $B^{\pm} \to J/\psi\phi K^{\pm}$ decays, 2013. arXiv:1309.6920.

CMS Collaboration, Searches for new physics using the $t\bar{t}$ invariant mass distribution in pp collisions at $\sqrt{s}$=8 TeV, 2013. arXiv:1309.2030.

CMS Collaboration, Measurement of the production cross section for $Z\gamma \to \nu\bar{\nu}\gamma$ in pp collisions at $\sqrt{s} = 7$ TeV and limits on $ZZ\gamma$ and $Z\gamma\gamma$ triple gauge boson couplings, 2013. arXiv:1309.1117.

CMS Collaboration, Search for a new bottomonium state decaying to $\Upsilon(1S)\pi^+\pi^-$ in pp collisions at $\sqrt{s} = 8$ TeV, 2013. arXiv:1309.0250.

CMS Collaboration, Measurement of the $W\gamma$ and $Z\gamma$ inclusive cross sections in pp collisions at $\sqrt{s} = 7$ TeV and limits on anomalous triple gauge boson couplings, 2013. arXiv:1308.6832.

CMS Collaboration, Measurement of the W-boson helicity in top-quark decays from $t\bar{t}$ production in lepton+jets events in pp collisions at $\sqrt{s}$=7 TeV, 2013. arXiv:1308.3879.

CMS Collaboration, Angular analysis and branching fraction measurement of the decay $B^0 \to K^{*0}\mu^+\mu^-$, 2013. arXiv:1308.3409.

CMS Collaboration, Search for top-squark pair production in the single-lepton final state in pp collisions at $\sqrt{s} = 8$ TeV, 2013. arXiv:1308.1586.

CMS Collaboration, Measurement of the prompt J/ψ and ψ(2S) polarizations in pp collisions at $\sqrt{s} = 7$ TeV, 2013. arXiv:1307.6070.

CMS Collaboration, Search for a Higgs boson decaying into a Z and a photon in pp collisions at $\sqrt{s} = 7$ and 8 TeV, 2013. arXiv:1307.5515.

CMS Collaboration, Measurement of the $B_s^0 \to \mu^+\mu^-$ branching fraction and search for $B^0 \to \mu^+\mu^-$ with the CMS Experiment, 2013. arXiv:1307.5025.

CMS Collaboration, Measurement of the top-quark mass in all-jets $t\bar{t}$ events in pp collisions at $\sqrt{s}$=7 TeV, 2013. arXiv:1307.4617.

CMS Collaboration, Study of the production of charged pions, kaons, and protons in pPb collisions at $\sqrt{s_{NN}} = 5.02$ TeV, 2013. arXiv:1307.3442.

CMS Collaboration, Determination of the top-quark pole mass and strong coupling constant from the $t\bar{t}$ production cross section in pp collisions at $\sqrt{s} = 7$ TeV, 2013. arXiv:1307.1907.

CMS Collaboration, Search for top squarks in R-parity-violating supersymmetry using three or more leptons and b-tagged jets, 2013. arXiv:1306.6643.

CMS Collaboration, Energy calibration and resolution of the CMS electromagnetic calorimeter in pp collisions at $\sqrt{s} = 7$ TeV, 2013. arXiv:1306.2016.

CMS Collaboration, Measurement of the W^+W^- cross section in pp collisions at $\sqrt{s} = 7$ TeV and limits on anomalous WW_γ and WWZ couplings, 2013. arXiv:1306.1126.

CMS Collaboration, Measurement of the hadronic activity in events with a Z and two jets and extraction of the cross section for the electroweak production of a Z with two jets in pp collisions at $\sqrt{s} = 7$ TeV, 2013. arXiv:1305.7389.

CMS Collaboration, Measurement of neutral strange particle production in the underlying event in proton-proton collisions at $\sqrt{s} = 7$ TeV, 2013. arXiv:1305.6016.

CMS Collaboration, Study of exclusive two-photon production of W^+W^- in pp collisions at $\sqrt{s}$ =7 TeV and constraints on anomalous quartic gauge couplings, 2013. arXiv:1305.5596.

CMS Collaboration, Search for gluino mediated bottom- and top-squark production in multijet final states in pp collisions at 8 TeV, 2013. arXiv:1305.2390.

CMS Collaboration, Multiplicity and transverse momentum dependence of two- and four-particle correlations in pPb and PbPb collisions, 2013. arXiv:1305.0609.

CMS Collaboration, Searches for long-lived charged particles in pp collisions at $\sqrt{s}$=7 and 8 TeV, 2013. arXiv:1305.0491.

CMS Collaboration, Measurement of the ratio of the inclusive 3-jet cross section to the inclusive 2-jet cross section in pp collisions at $\sqrt{s} = 7$ TeV and first determination of the strong coupling constant in the TeV range, 2013. arXiv:1304.7498.

CMS Collaboration, Measurement of the Λ_b^0 lifetime in pp collisions at $\sqrt{s} = 7$ TeV, 2013. arXiv:1304.7495.

CMS Collaboration, Measurement of masses in the $t\bar{t}$ system by kinematic endpoints in pp collisions at $\sqrt{s}$=7 TeV, 2013. arXiv:1304.5783.

CMS Collaboration, Search for a standard-model-like Higgs boson with a mass in the range 145 to 1000 GeV at the LHC, 2013. arXiv:1304.0213.

CMS Collaboration, Search for microscopic black holes in pp collisions at $\sqrt{s} = 8$ TeV, 2013. arXiv:1303.5338.

CMS Collaboration, Observation of a new boson with mass near 125 GeV in pp collisions at $\sqrt{s} = 7$ and 8 TeV, 2013. arXiv:1303.4571.

CMS Collaboration, Search for supersymmetry in hadronic final states with missing transverse energy using the variables α_T and b-quark multiplicity in pp collisions at $\sqrt{s} = 8$ TeV, 2013. arXiv:1303.2985.

CMS Collaboration, Search for the standard model Higgs boson produced in association with a top-quark pair in pp collisions at the LHC, 2013. arXiv:1303.0763.

CMS Collaboration, Search for narrow resonances using the dijet mass spectrum in pp collisions at $\sqrt{s} = 8$ TeV, 2013. arXiv:1302.4794.

CMS Collaboration, Study of the underlying event at forward rapidity in pp collisions at $\sqrt{s}$ = 0.9, 2.76, and 7 TeV, 2013. arXiv:1302.2394.

CMS Collaboration, Measurement of W^+W^- and ZZ production cross sections in pp collisions at $\sqrt{s}$=8 TeV, 2013. arXiv:1301.4698.

CMS Collaboration, Study of the Mass and Spin-Parity of the Higgs Boson Candidate via Its Decays to Z Boson Pairs, 2013. arXiv:1212.6639.

CMS Collaboration, Search for narrow resonances and quantum black holes in inclusive and b-tagged dijet mass spectra from pp collisions at $\sqrt{s} = 7$ TeV, 2013. arXiv:1210.2387.

Fields of Study

Major Field: Physics

Studies in Experimental High Energy Physics: Professor Richard Hughes

Table of Contents

Page

Abstract . . . ii
Dedication . . . iii
Vita . . . iv
List of Figures . . . xiii
List of Tables . . . xix

Chapters

1 Introduction **1**

2 Theory **2**
2.1 Overview of the Standard Model . . . 2
2.1.1 Particle Spectrum . . . 2
2.1.2 History . . . 5
2.2 Mathematical Description . . . 9
2.2.1 The Electroweak Lagrangian . . . 13
2.2.2 The Higgs Mechanism . . . 16
2.3 The Standard Model Higgs Boson . . . 19
2.3.1 Higgs Production . . . 22
2.3.2 Higgs Decay . . . 23
2.3.3 Higgs Boson Status . . . 27

3 CMS and the LHC **30**
3.1 The LHC . . . 30
3.2 The CMS Detector . . . 31
3.2.1 Tracker . . . 34
3.2.2 Calorimetry . . . 35
3.2.3 Muon System . . . 38
3.2.4 Trigger and Data Acquisition . . . 40

4 Event Selection and Object Identification **42**
4.1 Data Samples . . . 42
4.1.1 Triggers . . . 42
4.1.2 Event Cleaning . . . 44
4.2 Event Reconstruction . . . 45

4.3 Objects . . . 47
4.3.1 Leptons . . . 47
4.3.2 Jets . . . 49
4.3.3 Missing Energy . . . 52
4.3.4 B-Tagging . . . 52
4.4 Categories . . . 53

5 Data Modeling **56**
5.1 Generation of Simulated Data . . . 56
5.2 MC Samples . . . 57
5.2.1 Signal . . . 57
5.2.2 Background . . . 58
5.3 Corrections . . . 61
5.3.1 Lepton Trigger, Isolation and ID Efficiencies . . . 61
5.3.2 PU Reweighting . . . 61
5.3.3 JE Correction . . . 62
5.3.4 Top-p_T Reweighting . . . 63
5.3.5 CSV reweighting . . . 64
5.4 Data-MC Comparison . . . 67

6 Signal Extraction **70**
6.1 Discriminating Variables . . . 71
6.1.1 Selection of Variables by Category . . . 72
6.2 BDT Configuration . . . 75
6.3 Training Procedure . . . 79
6.3.1 $t\bar{t}H/t\bar{t}$ BDTs . . . 79
6.3.2 $t\bar{t}H/t\bar{t}+b\bar{b}$ BDTs . . . 81
6.4 Validation . . . 82
6.5 Data-MC Comparison of BDT Outputs . . . 82

7 Uncertainties **87**
7.1 Overview . . . 87
7.2 Systematics . . . 87
7.2.1 Luminosity and Pileup . . . 89
7.2.2 Monte Carlo Cross-Section, Q^2 and Statistical Uncertainties . . . 89
7.2.3 Lepton ID . . . 90
7.2.4 Jet Energy Corrections . . . 91
7.2.5 Top p_{T} Reweighting . . . 92
7.2.6 B-tagging . . . 92

8 Results **96**
8.1 Statistical Method . . . 96
8.2 Results of This Analysis . . . 98
8.3 Combined $t\bar{t}H$ Results . . . 99

9 Conclusion **103**

Appendices

A Data/Monte-Carlo Comparison of Input Variables **109**

List of Figures

Figure **Page**

2.1 The particles of the Standard Model and some of their basic properties (see text for full description). The Higgs boson shown here is the boson recently discovered at the LHC, whose properties have so far been consistent with the Higgs boson predicted by the Standard Model, within experimental uncertainty. 3
2.2 The couplings between different particles of the standard model [38]. . . . 3
2.3 Fit to Standard Model observables, compared to measured values. Results are shown with and without the inclusion of the measurement $m_H = 125.7 \pm 0.4$GeV [36]. 10
2.4 The potential of the Abelian Higgs model, shown to illustrate the quartic shape and the rotationally degenerate minimum. The non-Abelian (Standard Model) version has the same quartic structure, but its $SU(2)$ symmetry is harder to visualize [58]. 18
2.5 Predicted Standard Model Higgs branching ratios, as a function of m_H. Left: branching ratios across a wide range of m_H; right: detail of branching ratios roughly within $\pm 5\,\mathrm{GeV}/c^2$ of the observed Higgs boson mass. In general, the branching ratios are influenced by two competing features: the tendency of the Higgs to couple to more massive particles, and the fact that each decay channel is only "turned on" as the Higgs becomes heavy enough for its daughters to be on shell [47]. 20
2.6 Production cross-sections at specific energies, for a $125\,\mathrm{GeV}/c^2$ Standard Model Higgs boson [42]. Values shown at 1.96 TeV are for $p\bar{p}$ collisions; all other values are for pp collisions. 21
2.7 Feynman diagrams of the various production mechanisms [42]. 21
2.8 Summary plot showing 95% confidence limits on Higgs boson production at the Tevatron, as a function of m_H. Also shown are regions where the Higgs had been excluded by various experiments, up to June 2012. 28
2.9 Higgs Boson mass measurements using the combined 7 TeV and 8 TeV datasets [42]. 29
2.10 Signal strengths by Higgs boson decay channel [42]. 29

3.1 Illustration showing stages of the LHC accelerator chain, and the position of Intersection Point 5 on the LHC. The flow of protons are indicated by arrows. Note that the purpose of the illustration is to give the approximate sizes and relative locations of the accelerators, and is not strictly to scale. 31
3.2 Peak instantaneous luminosity per day, 2010-2012[10]. 32
3.3 Integrated luminosity delivered to CMS by year. More than 90 percent of this data was recorded: 5.55 fb^{-1} in 2011 and 21.79 fb^{-1} in 2012[10]. . . . 32
3.4 A schematic view of the CMS detector. 33
3.5 A simplified view of a longitudinal quadrant of CMS, showing the major systems and their coverage in η. 34
3.6 Cutaway view of the CMS tracker, in the $r-z$ plane [56]. The labeled regions are the pixel detector, the tracker inner barrel (TIB), tracker inner disks (TID), tracker outer barrel (TOB), and tracker end cap (TEC). See text for further description. 36
3.7 From left to right: the transverse momentum, transverse impact parameter and longitudinal impact parameter resolutions of the tracker as a function of η. The values shown are for muons with $p_T = 1$, 10 and 100 GeV. 36
3.8 Primary vertex resolution of the tracker as a function of number of tracks used in the reconstruction, in x (left) and z (right). 37
3.9 ECAL energy resolution as a function of energy as measured from a test beam. 38
3.10 The jet transverse energy resolution as a function of the simulated jet transverse energy for barrel jets ($|\mu| < 1.4$), endcap jets ($1.4< |\mu| <3.0$) and very forward jets. 39
3.11 Muon p_T resolution in the barrel (left) and endcap (right) regions. Resolution is shown for muons reconstructed using the tracker only, muon system only, and combined tracker and muon system measurements. 40

4.1 The lepton + jets mode of $t\bar{t}H$. 43
4.2 A cartoon of a transverse slice of the barrel region of the detector, illustrating the role of the different sub-detectors in identifying different particles. In the particle-flow reconstruction algorithm, particles are identified and reconstructed using information in a coordinated way from all the detector components. 46
4.3 Diagram showing isolation cones surrounding the reconstructed muon [17]. 49
4.4 A 3D view in detector $\eta-\phi$ space showing an electron+jets candidate event in data. The circles represent jets identified by the anti-kt algorithm ($R = 0.5$). The red and blue bars are ECAL and HCAL energy deposits, respectively, and their height is proportional to the amount of energy deposited. Tracks are not shown, but contribute to the PF objects used in the jet reconstruction. 51

4.5 CSV discriminant output distribution for MC-truth b-jets, c-jets and light-flavor (other) jets. This plot was produced from the $t\bar{t}H$ signal MC after baseline analysis selection with $\geq$ 4 jets and $\geq$ 2 b-tagged jets, so the distributions are biased towards high b-tagging efficiency here; however, the purpose of the plot is simply to show the relative shapes of the different distributions. For the efficiencies, see the text and figure 4.6. 54
4.6 Efficiency comparison of b-tagging algorithms at CMS, measured in simulated multijet events [55]. Left(Right): probability to misidentify udsg(c) jets, as a function of identification efficiency of real b-jets. 54
4.7 Number of jets (left) and number of b-tagged jets (right) in $t\bar{t}H$ and $t\bar{t}+$jets simulated events passing the full event selection. The plots are normalized to the number of entries for comparison. 55

5.1 Examples of basic Feynman diagrams for the various backgrounds. The bottom four are possible diagrams for the sub-dominant backgrounds. Clockwise from bottom left, they are: Diboson production, V+jets or Drell-Yan+jets, single top production, and $t\bar{t}Z$. The top diagram represents the dominant background, $t\bar{t}+$jets. 60
5.2 Left (right): combined muon (electron) ID, isolation selection and trigger efficiency scale factors in bins of p_T and η. 62
5.3 Comparison of number of reconstructed vertices for data (black) and the sum of all background MC samples before (red) and after (blue) pileup reweighting. After pileup reweighting, the MC agrees well with the data. 63
5.4 Top p_{T} reweighting function fit to the SFs from the CMS top POG. . . . 64
5.5 The p_{T} of the highest-p_{T} jet in a ==4 jet, ==2 tag control region, before (left) and after (right) the top-p_{T} reweighting procedure, for sum background MC and $t\bar{t}H$ signal. Note the presence of a systematic data/background MC disagreement in the left plot, that is eliminated in the right plot. 65
5.6 Various stages in the process of determining the HF CSV scale factor, in the p_{T} region $40\,\mathrm{GeV} \leq p_{\mathrm{T}} < 60\,\mathrm{GeV}$. Left: Data/MC comparison of initial, HF-enriched distributions. Center: CSV distributions after LF contamination subtracted from data and removed from MC. Right: final SFs with polynomial fit. [25] 67
5.7 Plots showing overall data/MC agreement per category, data/MC in p_{T} and in the distribution of the CSV output. 69

6.1 Examples of discriminating variables for a variety of categories. The plots are normalized to the number of entries to show the difference in shape between $t\bar{t}H$ and the different $t\bar{t}+$jets components. Several types of variables are shown, including shape, kinematic and CSV variables. Some variables, such as the "third-highest CSV output," are mainly good at separating LF from HF jets, while others, such as "$\Sigma jetp_T/\Sigma jetE$," more uniformly separate the $t\bar{t}+$jets components from $t\bar{t}H$. 74

6.2 Illustration of the structure of a basic decision tree, with 7 nodes in 3 layers. This tree uses a series of cuts on three variables in an attempt to separate the two classes of events, represented by the red and blue shades. The cuts are selected to increase the purity of blue or red events, or both. The initial cut on Variable 1 splits the events into a sample that is more pure for blue events, and another that is more pure for red events. These samples are further divided to increase the purity of the red and blue events, respectively. One can imagine that the cut on Variable 2 was made to optimize the purity of the red events in Node 4, resulting in less blue purity in Node 5. In that case, red events that were misclassified as blue in Node 5 might be "boosted," or given a greater weight when selecting the cuts for the next tree in the forest. 78

6.3 Plots showing the range of typical KS-test values in different scenarios. For each scenario, two histograms were each filled with 1000 Gaussian- distributed, random entries (this number was chosen to approximate the statistics used to train the BDTs). A KS test was then performed comparing the two distributions. This was repeated 1000 times, and a histogram was filled with the results of the KS tests. The bottom row shows the histograms filled with the KS-test values, and the top row shows a representative pair of Gaussian distributions from a given trial. The top and bottom plots are grouped together so that each column shows a different scenario. Left: the two Gaussians being compared come from the same parent distribution; center: one Gaussian is displaced by 0.2σ w.r.t. the other; right: one Gaussian is 30% wider than the other. This study demonstrates that the KS test is sensitive to statistically significant differences in both shape and location that may be too small to be visually obvious. 80

6.4 Two possible Feynman diagrams for $t\bar{t}+b\bar{b}$ (left) and $t\bar{t}H$ (right), illustrating the similarity between the two processes. 82

6.5 Normalized output distributions of the final BDTs for $t\bar{t}H$ and the various flavors of $t\bar{t}+$jets. Top row: $\geq$5 jets + $\geq$4 b-tags; center row: $\geq$6 jets + 3 b-tags; bottom row: $\geq$6 jets + $\geq$4 b-tags. Left column: BDT trained without use of $t\bar{t}H/t\bar{t}+b\bar{b}$ variable. Right column: BDT trained using $t\bar{t}H/t\bar{t}+b\bar{b}$ variable. 83

6.6 Overtraining checks for the final BDTs in each category. The KS test result is shown as a measure of the agreement between training and testing samples. The top, middle and, bottom rows are events with 4, 5, and $\geq$6 jets, respectively, while the left, middle, and right-hand columns are events with 2, 3, and $\geq$4 b-tags, respectively. The background (red) and signal (blue) are shown for the testing (solid line) and training (points) samples. 84

6.7 Overtraining checks for the $t\bar{t}+b\bar{b}/t\bar{t}H$ BDTs, in the $\geq$6 jets + 3 b-tags (left), 5 jets + $\geq$4 b-tags (center), and $\geq$6 jets + $\geq$4 b-tags (right) categories. The background (red) and signal (blue) are shown for the testing (solid line) and training (points with errors) samples. 84

6.8 Data/MC comparisons for all final BDTs. The top, middle and, bottom rows are events with 4, 5, and $\geq$6 jets, respectively, while the left, middle, and right-hand columns are events with 2, 3, and $\geq$4 b-tags, respectively. . . . 85

6.9 Data/MC comparisons for the $t\bar{t}H/t\bar{t}+b\bar{b}$ BDTs in the ≥6 jets + 3 b-tags (left), 5 jets + ≥4 b-tags (center), and ≥6 jets + ≥4 b-tags (right) categories. See figure 6.8 for legend. 86

7.1 Comparison of the BDT output when shifting the Q^2 scale up and down by its uncertainties. Shown are the shift upwards (red) and downwards (blue) relative to the nominal (black) shape for the $t\bar{t}+b\bar{b}$(left) $t\bar{t}+b$ (center) and $t\bar{t}$+LF (right) background samples.. The plots are normalized to unit area. 90

7.2 Comparison of the final BDT output in ≥ 6 jets ≥ 4 tags, for JES shift upwards (red) and downwards (blue) relative to the nominal (black) shape for the $t\bar{t}H(125)$ signal (left) and the $t\bar{t}+b\bar{b}$ background (right). The plots are normalized to unit area. 91

7.3 Example plots showing the up-and-down variations of selected b-tag shape systematics (top), and the resulting change in shape of the final BDT in the ≥ 6 jet ≥ 4 category (bottom). In a given column, the variations depicted in the top plot are reflected in the change in the shape of the distribution in the bottom plot. The bottom plots all show the change in the $t\bar{t}$ +LF distribution, after varying the HF contamination in the determination of the LF SF (left), and after varying the linear (center) and quadratic (right) distortions that determine the statistical uncertainty of the LF SF extraction. The All plots are normalized to unit area. These are the largest proportional shape variations due to b-tag uncertainties among all the flavors of $t\bar{t}+$ jets in this category. 94

7.4 Example plots showing the up-and-down variations of selected b-tag shape systematics (top), and the resulting change in shape of the final BDT in the ≥ 6 jet ≥ 4 category (bottom). In a given column, the variations depicted in the top plot are reflected in the change in the shape of the distribution in the bottom plot. The bottom plots all show the change in the $t\bar{t}$ +HF distribution, after varying the LF contamination in the determination of the HF SF (left), and after varying the linear (center) and quadratic (right) distortions that determine the statistical uncertainty of the HF SF extraction. The All plots are normalized to unit area. 95

8.1 Background-only fit to the data in the 5 jets + ≥4 b-tags (left), ≥6 jets + 3 b-tags (center), and ≥6 jets + ≥4 b-tags (right) categories. The post-fit uncertainties are constrained relative to the uncertainties before the fit, as can be seen by comparing to figure 6.8. 97

8.2 The observed and expected 95% confidence upper limits on the signal strength modifier μ, for $t\bar{t}H$ production in the lepton+jets channel, as a function of m_H. The solid black line is the observed limit. The dashed line is the median expected limit, and the green (yellow) regions are the 68% (95%) error bands on the expected limit. 100

8.3 The best-fit to the signal strength modifier μ, with $\pm1\sigma$ error bars. Results are shown separately for the different $t\bar{t}H$ channels, as well as the result for a combined fit using all channels. The Standard Model value is shown as a black vertical line. 101

8.4 The observed and expected 95% confidence upper limits on the signal strength modifier μ, in a combined search for $t\bar{t}H$ production at CMS. The solid black line is the observed limit. The dashed line is the median expected limit, and the green (yellow) regions are the 68% (95%) error bands on the expected limit. The red dashed line represents the median expected limit calculated with signal injected at the Standard Model rate. Top: limits separated by channel, for $m_H = 125.6\,\text{GeV}$. Bottom: the combined limit as a function of m_H. 102

A.1 Data/MC comparisons for events with one lepton and ≥6 jets + 2 b-tags. The uncertainty band includes statistical and systematic uncertainties that affect both the rate and shape of the background distributions. 110
A.2 Data/MC comparisons for events with one lepton and 4 jets + 3 b-tags. The uncertainty band includes statistical and systematic uncertainties that affect both the rate and shape of the background distributions. 111
A.3 Data/MC comparisons for events with one lepton and 5 jets + 3 b-tags. The uncertainty band includes statistical and systematic uncertainties that affect both the rate and shape of the background distributions. 112
A.4 Data/MC comparisons for events with one lepton and ≥6 jets + 3 b-tags. The uncertainty band includes statistical and systematic uncertainties that affect both the rate and shape of the background distributions. 113
A.5 Data/MC comparisons for events with one lepton and ≥6 jets + 3 b-tags. The uncertainty band includes statistical and systematic uncertainties that affect both the rate and shape of the background distributions. 114
A.6 Data/MC comparisons for events with one lepton and 4 jets + 4 b-tags. The uncertainty band includes statistical and systematic uncertainties that affect both the rate and shape of the background distributions. 115
A.7 Data/MC comparisons for events with one lepton and 5 jets +≥4 b-tags. The uncertainty band includes statistical and systematic uncertainties that affect both the rate and shape of the background distributions. 116
A.8 Data/MC comparisons for events with one lepton and 5 jets +≥4 b-tags. The uncertainty band includes statistical and systematic uncertainties that affect both the rate and shape of the background distributions. 117
A.9 Data/MC comparisons for events with one lepton and ≥6 jets + ≥4 b-tags. The uncertainty band includes statistical and systematic uncertainties that affect both the rate and shape of the background distributions. 118
A.10 Data/MC comparisons for events with one lepton and ≥6 jets + ≥4 b-tags. The uncertainty band includes statistical and systematic uncertainties that affect both the rate and shape of the background distributions. 119

List of Tables

Table **Page**

2.1 Gauge symmetries of the Standard Model and their conserved quantities. Here, $X^j = \lambda^j/2$, where λ^j are the Gell-Mann matrices, and the $T^i = \sigma^i/2$, where σ^i are the Pauli spin matrices (see text). 13
2.2 Weak isospin and hypercharge assignments by field [2]. 16

4.1 Triggers used to collect the data for this analysis. 43
4.2 Summary of the data analyzed in this dissertation. 44
4.3 Cuts for selecting tight and loose muons. 48
4.4 Tight and loose electron selection cuts. 50

5.1 List of signal MC masses with corresponding cross sections, number of generated MC events and number of MC events passing the 1 tight lepton, $\geq$4 jets, $\geq$2 b-tagged jets event selection. 58
5.2 From left to right: list of background MC datasets used in the analysis, software used to generate events, cross sections used for normalization, number of generated MC events and number of MC events passing the 1 tight lepton, $\geq$4 jets, $\geq$2 b-tagged jets event selection (the term "jets" in the leftmost column denotes generated jets, and not jets as defined in the event selection). 59
5.3 Predicted signal and background event yields in each of the jet-tag categories, after all event selection criteria and corrections to MC have been applied. The number of observed data events are also shown. The errors on the predicted signal and background are the combined statistical and systematic uncertainties, which will be discussed in chapter 7. 68

6.1 Event variables used in the boosted decision trees and their descriptions. . 73

6.2 BDT input variable assignments for the final BDTs in each category. The variables used in each category were selected following the procedure outlined in the text. However, once selected, it is possible to identify trends in the types of variables that offer the best separating power in each of the categories. In general, a mix of different kinematic, shape, and b-tagging variables are used in each BDT, so that a variety of event information is available during training. In the categories that contain greater numbers of jets and tagged jets, more of the CSV variables are used since they offer separating power between events containing a certain number of correctly tagged jets and events containing some jets that were not correctly tagged (either mistagged or incorrectly not tagged). In the categories with lower numbers of jets and tags (such as 4 jets + 3 tags), the best discriminating information is provided by kinematic variables. In categories with the greatest combinatorics, specialized algorithms (such as the "best Higgs mass") and event shape variables are useful in distilling complex event information. 76

6.3 List of variables used as inputs in each of the $t\bar{t}H/t\bar{t}+b\bar{b}$ BDTs. In contrast to the final BDTs trained in the same categories, b-tagging information does not provide much discriminating power between $t\bar{t}H$ and $t\bar{t}+b\bar{b}$ since both processes nominally contain the same number of real b-jets. Furthermore, $t\bar{t}H$ and $t\bar{t}+b\bar{b}$ are kinematically similar, necessitating the extensive research of event shape variables. In particular, we found differences in the η distributions of objects to be useful, especially the difference in η between the reconstructed tops and the bb-pair assigned to the Higgs by the best Higgs mass algorithm. 77

7.1 Summary of the systematic uncertainties considered in the inputs to the limit calculation. Except where noted, each row in this table is treated as a single, independent nuisance parameter. 88

7.2 This table summarizes the effect of each of the independent LF and HF b-tag nuisance parameters on the yields of different samples, in the ≥ 6 jet ≥ 4 category. Variations due to JES are not shown. "Stat. Err. 1" and "Stat. Err. 2" refer to the linear and nonlinear components of the respective statistical uncertainties. The light SF purity for $t\bar{t}$+LF events is affected the most; this uncertainty also has the largest effect on the shape of $t\bar{t}$+LF events (see figure 7.3). 93

8.1 The observed and expected 95% confidence upper limits on μ for $t\bar{t}H$ production in the lepton+jets channel, at $m_H = 125.6\,\text{GeV}$. 99

Chapter 1
Introduction

This dissertation describes a search for the Standard Model Higgs boson in the $t\bar{t}H$ production mode, in the lepton + jets or LJ channel. In this channel, the Higgs boson decays to a pair of b quarks, and the top-quark pair follows the decay: $t\bar{t} \to b\bar{b}q\bar{q}l\nu$. The search is performed using $19.3\,\mathrm{fb}^{-1}$ of data from the Compact Muon Solenoid (CMS) detector, at a center-of-mass energy of $\sqrt{s} = 8\,\mathrm{TeV}$. This search compliments a previous analysis that searched for the Higgs boson in the same channel, using $5\,\mathrm{fb}^{-1}$ of data at $\sqrt{s} = 7\,\mathrm{TeV}$ [27]. The work of this dissertation has also been combined with other $t\bar{t}H$ searches at CMS; this combination is discussed in chapter 8.

The dissertation is structured as follows: first, a review of the Standard Model of particle physics is given, with a focus on the physics of the Higgs boson. We next give an overview of the LHC and the CMS detector. The description of the analysis begins with the process of selecting the dataset to be analyzed, followed by a discussion of the modeling of the data. A multivariate technique that enhances the sensitivity of the analysis is then described. Finally, the handling of uncertainties is reviewed, and the results of the analysis are presented.

Throughout the dissertation, the "natural" or "energy" units are used, with $c = \hbar = 1$.

Chapter 2
Theory

2.1 Overview of the Standard Model

The Standard Model of particle physics is a relativistic, quantum field theory that describes the fundamental constituents of matter, and how they interact via the strong, weak and electromagnetic forces. Through the Higgs field and electroweak symmetry-breaking, it explains the mechanism by which the fundamental particles obtain mass. The Standard Model is the result of many decades of theoretical development proceeding in parallel with experimental observation, and is the most precisely tested and successful model to date that accounts for the properties of matter and their interactions.

2.1.1 Particle Spectrum

Figure 2.1 summarizes the particles of the Standard Model and some of their properties, including their masses, charges and spins. Figure 2.2 illustrates the couplings between the particles. There are 61 fundamental particles and antiparticles in total, each with a unique set of features, resulting in the extremely rich phenomenology found in nature. Broadly speaking, they are divided into "matter" particles called fermions that have an intrinsic spin of 1/2; force-carrying particles called vector bosons, with a spin of 1; and a mass-imparting, scalar boson (the Higgs boson) that has zero spin.

In general, all fermions have half-integer intrinsic spin, and all bosons have integer spin. This label applies not only to the fundamental particles, but extends to composite particles such as atoms, where the bosonic or fermionic nature of the combination is determined

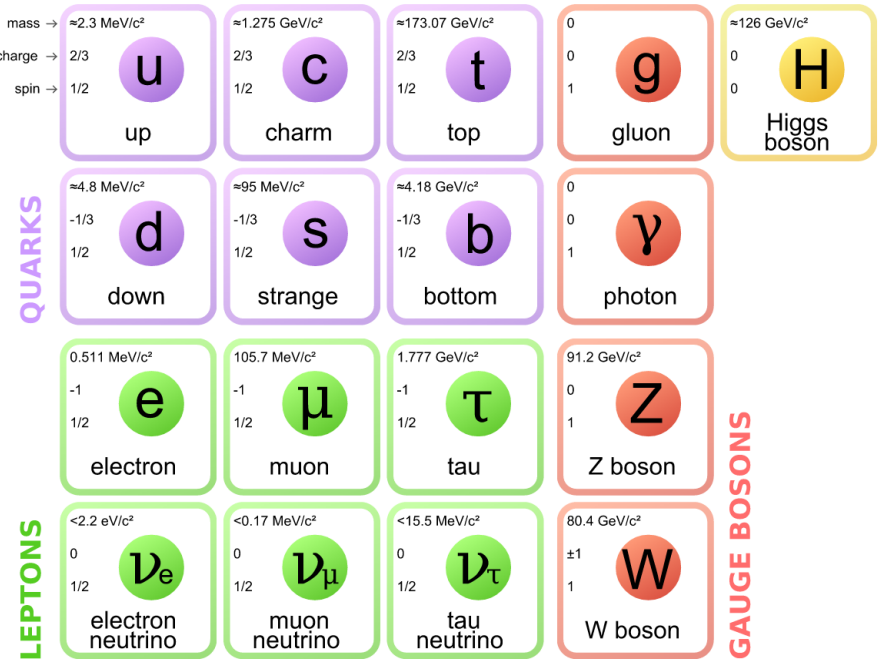


Figure 2.1: The particles of the Standard Model and some of their basic properties (see text for full description). The Higgs boson shown here is the boson recently discovered at the LHC, whose properties have so far been consistent with the Higgs boson predicted by the Standard Model, within experimental uncertainty.

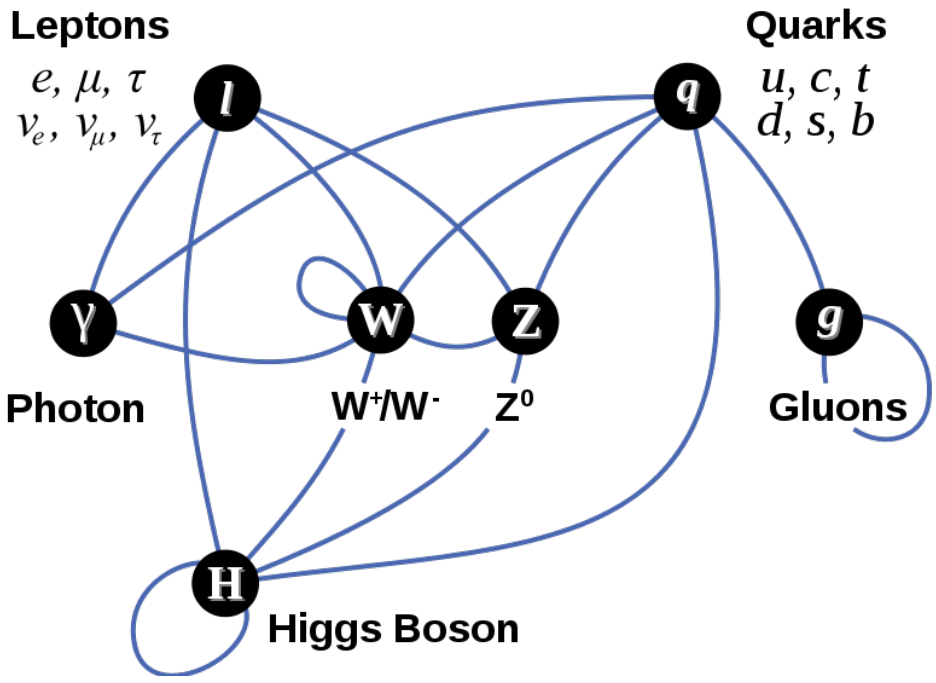


Figure 2.2: The couplings between different particles of the standard model [38].

by the addition of the spins of the individual particles. Fermions are described by Fermi statistics, which means they respect the Pauli exclusion principal so that only one fermion may occupy a given quantum state. Bosons are not subject to the same exclusion principle: any number of bosons may occupy a given quantum state. When many bosons occupy the same state, macroscopic quantum effects may result – examples include lasers (photons), super-conductors (Cooper pairs), and super-fluids (bosonic atoms such as He^4).

The fundamental fermions consist of quarks and leptons. Quarks possess color charge, which causes them to engage in the strong interaction, mediated by the gluon. Since they also interact weakly, and carry charge and mass, they are the only particles to experience all four forces. Due to a phenomenon called color confinement, quarks cannot exist in isolation, and instead must combine to form strongly-bound states called hadrons. A hadron containing 3 quarks is called a baryon. The nuclei of atoms are made up of baryons consisting of two up quarks and one down quark (protons) and baryons that contain two down quarks and one up quark (neutrons). Mesons (such as pions) are hadrons containing two quarks: a quark and an anti-quark.[1] There are six flavors of quarks that make up 3 generations: the first generation includes up and down quarks, the second includes charm and strange, and the third consists of top and bottom quarks. The up-type (anti)quarks all have an electric charge of (-)+2/3 e, and the bottom-type (anti)quarks have an electric charge of (+)-1/3 e. Additional quantum numbers describe the flavor properties of the quarks. There are hundreds of ways for quarks and anti-quarks to combine into hadrons, and many of these have been observed in high-energy experiments [37].

Leptons are also organized into three generations: the electron and electron neutrino, the muon and muon neutrino, and the tau and tau neutrino. They do not participate in strong interactions, and are not confined in the same way as quarks. However, both quarks and leptons engage in the weak interaction, mediated by the W and Z bosons. Fermions may change flavor via weak processes mediated by the $W^{\pm}$ bosons; for example, β decay

[1]This is a simplified description. The 3 (2) quarks that make up baryons (mesons) are actually valence quarks: these are the quarks that determine the quantum numbers of the hadron. Virtual quarks also occupy the bound state, and are known as "sea" quarks, since they make up a sea of quark/anti-quark pairs that are produced through gluon splitting, and that quickly annihilate within the interior of the hadron.

occurs when a d quark in a neutron changes to an u quark by emitting a virtual W^-, thereby changing the neutron to a proton. The neutrinos are all electrically neutral, while the electron, muon and tau all have a charge -e. Thus, the neutrinos only experience weak interactions. As a result, it is difficult to detect them directly as they typically pass through matter undisturbed. The electron is the only stable charged lepton, and therefore is the only lepton to form stable bound states in atoms.

All charged particles interact electromagnetically, via the photon. The photon is the only stable boson, and is therefore capable of mediating electromagnetic interactions at great distances. Electromagnetic phenomena are extensive [44], and include the electrostatic attraction/repulsion between charged particles or substances, magnetostatic phenomena, as well as dynamically varying fields and electromagnetic radiation. Along with gravity, these make up the macroscopic phenomena with which we are most familiar. All the allowed interactions of the photon are shared by the neutral Z^0 boson, but the comparatively small coupling constant of the weak force, combined with the short range/high mass of the Z^0, dramatically reduces the relative likelihood of neutral current weak interactions. At high energies, however, the electromagnetic and weak forces become comparable enough that they are united by a single electroweak description. The distinction between the massive vector bosons that mediate the weak force, and the massless photon that mediates the electromagnetic force, is due to the existence of the only scalar (spin 0) boson of the Standard Model: the Higgs boson. This relationship will be discussed in greater detail later in this chapter.

2.1.2 History

The development of modern physics began with electromagnetism. The term that Maxwell added to Ampere's law in 1865 described magnetic fields that arose from time-varying electric fields, in the same way as time-varying magnetic fields lead to electric fields in Faraday's law. The result of this addition was that the two differential equations now gave a description of electromagnetic waves, propagating at the speed of light [44]. Radio waves were later observed by Hertz, confirming the prediction. The Michelson-Morley experiment further

showed that the propagation of these waves was not due to the presence of any media.

In 1897, J. J. Thompson discovered the electron in an experiment with cathode rays. By deflecting the rays with perpendicular electric and magnetic fields, he was able to determine that the rays were actually streams of particles, and calculated their charge-to-mass ratio. After the Rutherford scattering experiment showed that the nuclei of atoms were hard, compact, heavy and positively charged, Bohr proposed a model for Hydrogen in which a single electron orbited a nucleus that consisted of a single proton. This allowed some rudimentary but accurate quantum mechanical calculations for the emission spectrum of Hydrogen.

In 1900, Planck offered an explanation for the observed spectrum of black-body radiation. Classical statistical mechanics predicted an "ultraviolet catastrophe" of infinite radiated power. Planck circumvented this problem by assuming energy came in packets or "quanta" of energy that were proportional to the frequency of the radiation. The constant of this proportionality became known as Planck's constant, and the unit of light quantization was the photon.

Einstein then proposed that this light quantization was a feature of the electromagnetic field itself. He offered a quantum-mechanical explanation of the photoelectric effect: when light strikes a metal surface, the energy of the emitted electrons is proportional to the frequency of the incident radiation. Einstein's contributions to physics were extensive; his theory of General Relativity is still the model used today to explain the force of gravity. The concept of spacetime is present in both the Standard Model and General Relativity, but there is no mention of the geometry of spacetime in the Standard Model. General Relativity explicitly uses gravity to explain spacetime curvature (and vice-versa).

The basic form of quantum mechanics took shape during the 1920s. Compton's 1923 scattering experiments demonstrated that light behaves like a particle on a subatomic scale. The particles (photons) had zero mass, and an energy given by Planck's equation $E = h\nu$. Nonrelativistic quantum mechanics was developed from 1923-1926, and included the work of Heisenberg, de Broglie, Pauli, Schrodinger and others, and culminated in the description of the electron given by Schrodinger's equation. In 1927, Dirac extended this description

to include relativistic particles. His equation admitted negative energy solutions, which contradicted the belief that the vacuum was the lowest possible energy state. Dirac solved the problem with his concept of an infinite "sea" of electrons that filled these negative energy states. A fluctuation could promote an electron from this sea to positive energy, thereby creating a "hole" particle with the same mass but opposite charge as the electron. A particle matching Dirac's description, the positron, was discovered in 1931 in an experiment by Anderson. Feynman and others later recast Dirac's negative-energy positron as an antiparticle with positive energy solutions [37].

In 1930 Pauli proposed the existence of a new, light, neutral particle to explain the continuous spectrum of β decay. Fermi incorporated this particle, which he called the neutrino, into his 1934 theory of β decay (Chadwick had already taken the name "neutron" for the neutral version of the proton he discovered in 1932). In 1934, Yukawa proposed the strong force as a means to keep protons and neutrons together in an atomic nucleus. In Yukawa's model, nucleons were attracted by a massive quantized field, and Yukawa calculated this mass to be somewhere between the electron and proton, leading to the name "meson," which means middle-weight. The meson label was given to cosmic particles detected in 1937 that were in the same mass range as Yukawa's meson. However, experiments in 1946 showed that these particles interacted only weakly with atomic nuclei, indicating they could not be strong force mediators. In 1947, a distinction was made between cosmic rays consisting of π and lighter μ particles.

Neutrinos were first observed at the Savannah River nuclear reactor in South Carolina in 1950, via inverse β decay using the large neutrino flux from the reactor. Subsequent experiments established that the same reaction with antineutrinos did not occur, so that neutrinos and antineutrinos were distinct particles – this led to the idea of conservation of lepton number, where $L = 1$ for leptons, and $L = -1$ for anti-leptons.

Throughout the 1950s and 1960s, scores of mesons and heavy baryons were discovered. The quantity "strangeness" was introduced in 1961 by Murray Gell-Mann to help describe the increasingly diverse spectrum of hadrons. His method of categorization, the Eightfold Way, was successful in predicting the existence of the Ω^- baryon, observed in 1964. This

success led to the proposal of the quark model by Gell-Mann and others, which offered an explanation for the patterns of the Eightfold Way. The discovery of the long-lived J/Ψ meson in 1974, and the so-called November Revolution that followed, implied the existence of a fourth quark which was given the name "charm." Charmed baryons were observed soon after. A third generation was added to the leptons with the discovery of the tau in 1975, and to the quarks with the observation of the $\Upsilon(b\bar{b})$ in 1977. Evidence of the gluon was seen in three-jet events in electron-positron collisions at DESY In 1978 and 1979. The up-type quark of the third generation, the t, was not observed until 1995 at Fermilab's Tevatron collider. This quark was too short-lived to produce any bound states, and could only be detected by analyzing its decay products [37].

The 1960s and 1970s also saw the development of the electroweak theory by Glashow, Weinberg, and Salam. Evidence of weak neutral currents in neutrino scattering was seen at the Gargamelle experiment at CERN in 1973. However, unlike the particle zoo of hadrons and leptons, no massive vector bosons had been observed at all up to this point. Finally, in 1983, the near-simultaneous observation of the W and Z bosons occurred at UA1 and UA2, CERN [37]. The discovery gave validity to the electroweak theory, and hence to the emerging picture of the Standard Model. A scalar boson consistent with the Higgs boson, the last particle predicted by the electroweak theory, was observed at CERN in July 2012 by the CMS[13] and ATLAS[4] collaborations. Limits have been placed on existence of additional generations of fermions by measuring the lifetime of the Z – specifically, the number of light neutrinos has been measured to be 2.99 ± 0.06 [37]. The complexity of the Standard Model leads to a large number of empirical parameters, yet there is striking agreement between combined observations and a global fit to the standard model prediction (see figure 2.3). However, for all its successes, the Standard Model is still not the complete picture. Neutrinos are not assumed to be massive in the Standard Model, but evidence of neutrino oscillations in 1998 at the Super-Kamiokande neutrino detector shows that at least one of the neutrinos must be massive [37]. While the electroweak theory does predict the masses of the W and Z (and requires the photon to be massless), the masses of the rest of the particles are not predetermined by theory, and must be added in a semi-ad-hoc manner

through Yukawa couplings to the Higgs field. Dark matter does not behave like any known Standard Model particle, and Dark Energy is completely unexplained by any known physical process.

A number of theoretical extensions to the Standard Model exist, most notably Super-symmetry (SUSY), which predicts the existence of a super-symmetric partner for every particle in the Standard Model. Each super-symmetric particle differs from its Standard Model partner by a half-integer in spin. Some versions of SUSY are useful in explaining how the strong and electroweak theories might be unified at high energy into a "Grand Unified Theory" of particle physics, and may also lend insight into the hierarchy problem, which is the difference in strength between gravity and the Standard Model forces at even higher energies. Evidence for TeV-scale SUSY was expected shortly after the LHC began taking data; however, current searches have yielded no such evidence [48].

2.2 Mathematical Description

The mathematical description of the Standard Model is deeply linked to the concept of symmetries, and how these relate to conserved quantities. These symmetries consist of both discrete and continuous varieties.

Discretely, charge (C), parity (P) and time reversal (T) transformations are considered. The standard model preserves a combined CPT symmetry. This combination is a stronger requirement than the individual C, P, or T, symmetries, all of which are violated in one way or another by particles of the Standard Model. For example, the weak interactions violate C-symmetry: charged particles do not interact identically to their respective oppositely-charged antiparticles. The weak force also violates CP-symmetry (particularly in the decays of kaons and B-mesons), and due to CPT invariance, this implies that T-symmetry is also violated [37]. The Standard Model also obeys a set of continuous symmetries. These symmetries can be mathematically described by the $SU(3)\times SU(2)\times U(1)$ Lie group. Through the use of gauge fields, a function (called a Lagrangian) is constructed that encodes all the physics of the Standard Model, and is invariant under a continuous group of local transformations that

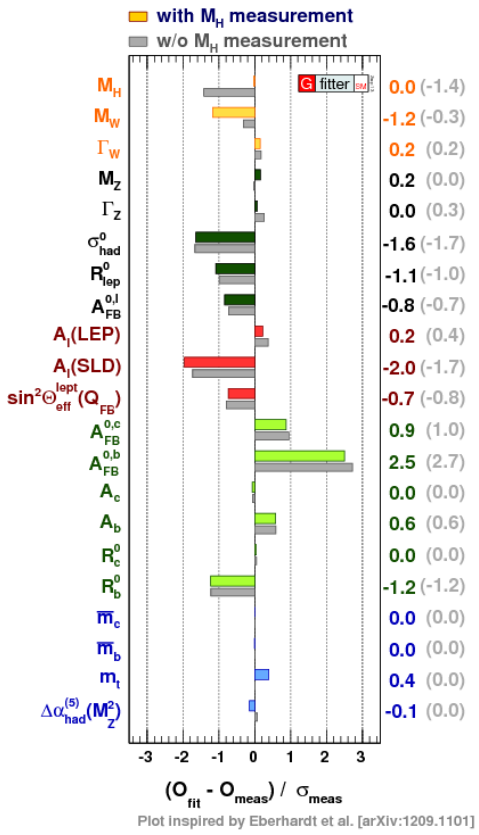


Figure 2.3: Fit to Standard Model observables, compared to measured values. Results are shown with and without the inclusion of the measurement $m_H = 125.7 \pm 0.4$GeV [36].

collectively reflect the SU(3)×SU(2)×U(1) symmetry. The symmetries of the portions of the Lagrangian that make up Quantum Chromodynamics (QCD) are described by the SU(3) group, while the weak and electromagnetic forces are combined into a single electroweak theory that respects a SU(2)×U(1) symmetry. The electroweak symmetry is broken in a nontrivial way (that will be discussed in the next section), producing the distinction between electromagnetic and weak forces. The symmetries of the different groups can be demonstrated by performing appropriate gauge transformations, and demanding that the Lagrangian of the theory be invariant under these transformations.

Noether's theorem states that for each continuous symmetry of a physical system, some

physical property is conserved. For example, Maxwell's equations:

$$\begin{aligned} \boldsymbol{\nabla} \cdot \boldsymbol{E} &= \rho/\epsilon_0 & \boldsymbol{\nabla} \cdot \boldsymbol{B} &= 0 \\ \boldsymbol{\nabla} \times \boldsymbol{E} &= -\partial \boldsymbol{B}/\partial t & \boldsymbol{\nabla} \times \boldsymbol{B} &= \mu_0 \boldsymbol{J} + \partial \boldsymbol{E}/c^2 \partial t \end{aligned} \tag{2.1}$$

are invariant when the potentials undergo the gauge transformation

$$\begin{aligned} \boldsymbol{A} &\rightarrow \boldsymbol{A} + \boldsymbol{\nabla} f \\ V &\rightarrow V - \partial f/\partial t \end{aligned}, \tag{2.2}$$

where $\boldsymbol{E} = -\boldsymbol{\nabla} V - \partial \boldsymbol{A}/\partial t$, $\boldsymbol{B} = \boldsymbol{\nabla} \times \boldsymbol{A}$, and f is some scalar function. Accordingly, Maxwell's equations imply a conservation law; one can see that taking the divergence of the first and last equations of (2.1) results in

$$\boldsymbol{\nabla} \cdot \boldsymbol{J} = -\partial \rho/\partial t, \tag{2.3}$$

which is simply the local conservation of electric charge.

This concept can be extended to the gauge field theory of the Standard Model, which consists of fermionic matter fields, and gauge fields which interact with the matter fields. Each of the gauge fields corresponds to the generator of a given symmetry group, and for each of the local gauge symmetries, a physical quantity is conserved. Table 2.1 lists these conserved quantities, along with their associated generators and symmetries.

The requirement of local gauge invariance has powerful implications for the quantum theory of the standard model. For example, the free Dirac Lagrangian for relativistic quantum mechanics is written as:

$$\mathcal{L} = \bar{\psi}(i\gamma^\mu \partial_\mu - m)\psi. \tag{2.4}$$

Here, ψ is the fermionic field, γ are the Dirac matrices, and m is the mass of the field. Eq. (2.4) has the U(1) symmetry of rotational invariance under the global transformation $\psi \rightarrow \ e^{i\theta}\psi$. However, the stronger requirement of a local U(1) gauge invariance, with symmetry under $\psi \rightarrow \ e^{i\theta(x)}\psi$, is not respected by this Lagrangian. Instead, making this

local transformation results in a change $\Delta\mathcal{L}$:

$$\Delta\mathcal{L} = -\bar{\psi}(\gamma^\mu \partial_\mu \theta(x))\psi. \tag{2.5}$$

In order to make the Lagrangian invariant under local gauge transformations, we must somehow produce an additional term to cancel the extra term in eq. (2.5). This happens when we add a term to the Lagrangian (2.4), involving a new field A_μ:

$$\mathcal{L} = \bar{\psi}(i\gamma^\mu \partial_\mu - m)\psi + e\bar{\psi}\gamma^\mu A_\mu \psi, \tag{2.6}$$

where A_μ transforms as:

$$A_\mu \to A_\mu + \frac{1}{e}(\partial_\mu \theta(x)). \tag{2.7}$$

This is equivalent to replacing the derivative ∂_μ with a covariant derivative:

$$\partial_\mu \to D_\mu \equiv \partial_\mu - ieA_\mu. \tag{2.8}$$

The new field A_μ also requires its own kinetic term. The resulting (now locally gauge invariant) Lagrangian can be written as:

$$\mathcal{L} = \bar{\psi}(i\gamma^\mu D_\mu - m)\psi - \frac{1}{4}F_{\mu\nu}F^{\mu\nu}, \tag{2.9}$$

where $F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu$.

We see that eq. (2.9) does not contain a mass term for A – indeed, the addition of a mass term would spoil the local gauge invariance. The gauge boson describing this field must be massless. This massless boson is in fact the photon: remarkably, the requirement of local U(1) invariance of the Dirac Lagrangian generates all of electrodynamics, and specifies how Dirac fields (electrons and positrons) couple to electromagnetic fields. In other words, eq. (2.9) is the Lagrangian for quantum electrodynamics. The conserved current is

$$J^\mu = e(\bar{\psi}\gamma^\mu \psi), \tag{2.10}$$

and the coupling constant associated with the local U(1) symmetry is the electric charge, e [33].

Continuous Symmetry	Generator(s)	Gauge Field(s)	Conserved Quantity
$U(1)_{EM}$	Q	A_μ	electric charge
$U(1)_Y$	Y	B_μ	weak hypercharge
$SU(2)_L$	$T^i;\ i = 1, 2, 3$	W^i_μ	weak isospin (T^3)
$SU(3)_C$	$X^j;\ j = 1, 2, ...8$	G^j_μ	color charge

Table 2.1: Gauge symmetries of the Standard Model and their conserved quantities. Here, $X^j = \lambda^j/2$, where λ^j are the Gell-Mann matrices, and the $T^i = \sigma^i/2$, where σ^i are the Pauli spin matrices (see text).

The addition of a covariant derivative containing a gauge field is a common device among the Standard Model gauge field theories for turning a globally invariant Lagrangian into a locally invariant one. Next, we will demonstrate how the requirement of local gauge invariance is applied to electroweak unification.

2.2.1 The Electroweak Lagrangian

The idea of electroweak unification is to unite the electrodynamic and weak interactions through the weak hypercharge:

$$Y = 2(Q - T^3), \tag{2.11}$$

as a new conserved quantity, where Q is the electric charge, and T^3 is the third component of the weak isospin vector. The components of T are related to the Pauli spin matrices by $T^i = \sigma^i/2$. For further discussion, it is convenient to represent the fermion fields via:

$$\psi_L = \begin{pmatrix} \nu_{eL} \\ e_L \end{pmatrix}, \begin{pmatrix} u_L \\ d_L \end{pmatrix}, \ ... \tag{2.12}$$

$$\psi_R = e_R, u_R, d_R, \ ... \tag{2.13}$$

where ψ_L and ψ_R are the left-handed and right-handed helicity states, respectively, and the ellipses denote the other generations. This notation is due to the fact that the weak force is parity-violating, and the helicity states of the matter fields differ between left-handed and right-handed fermions. The left-handed states are composed of isospin doublets – one

doublet per generation – whereas the right-handed singlet states exist only for the quarks and charged leptons; there is no right-handed component for the neutrinos.

Table 2.2 shows the Y and T^3 values for the different components of the fermion fields. Both left- and right-handed fields conserve hypercharge. However, only the left-handed fields conserve isospin. Thus, the gauge group for electroweak symmetry is often written as $SU(2)_L \times U(1)_Y$, where the subscripts highlight the helicity distinction. The electroweak gauge transformation is [33]:

$$\psi_L \rightarrow e^{iT^j \Lambda^j(x) + iY_L \theta(x)/2} \psi_L \tag{2.14}$$

$$\psi_R \rightarrow e^{iY_L \theta(x)/2} \psi_R \tag{2.15}$$

To produce the quantum electroweak gauge theory we would now like to follow an analogous procedure to that in the previous section, where we introduced local gauge invariance to the Dirac equation via the covariant derivative. However, there are two problems with this approach. First, we cannot include mass terms for the fermions in a similar fashion to equation (2.4), because this would lead to terms of the form [2]:

$$-m(\bar{\psi_L}\psi_R + \bar{\psi_R}\psi_L) \tag{2.16}$$

Where the left- and right-handed fields are coupled. This is not invariant under the transform (2.14); mass terms such as (2.16) will explicitly break $SU(2)_L$ [2]. The only solution is to leave out the mass terms for now, and continue just with the kinetic fermion terms of form $i\bar{\psi}\gamma^\mu \partial_\mu \psi$. Following the approach as before, we introduce the gauge fields B and W^i via the covariant derivative

$$D_\mu = \partial_\mu - ig_Y Y B_\mu - ig_L W^i_\mu T^i, \tag{2.17}$$

where g_Y and g_L are couplings to the $U(1)_Y$ and $SU(2)_L$ groups, respectively. This results in the locally gauge-invariant Lagrangian

$$\mathcal{L}^e_{EW} = \bar{\phi}_{eL} i\gamma^\mu D_\mu \phi_{eL} + \bar{\phi}_{eR} i\gamma^\mu D_\mu \phi_{eR} + \ ... \tag{2.18}$$

where only the electron right-handed singlet and the (e, ν_e) left-handed doublet are shown as an example, and the ellipses denote similar terms for the other fermions.

We have now successfully obtained an equation describing the interaction of the fermion fields with the electroweak gauge fields W^i and B, but at the cost of making the fermions massless. This brings us to the second problem, when we turn our attention to the gauge fields themselves. First, we can separate out the charged and neutral components of the W^i and B fields by rewriting them in terms of the linear combinations:

$$W_\mu^\pm = (W_\mu^1 \mp \ iW_\mu^2)/\sqrt{2} \tag{2.19}$$

$$Z_\mu = -sin\theta_W B_\mu + cos\theta_W W_\mu^3 \tag{2.20}$$

$$A_\mu = cos\theta_W B_\mu + sin\theta_W W_\mu^3 \tag{2.21}$$

The $W^\pm$, Z, and A fields now correspond to the charged-current weak mediators $W^\pm$, the neutral-current weak Z and the electromagnetic field A. Here, θ_W is the weak mixing angle which determines how the neutral components B and W^3 are orthogonally combined into Z and A. Now, we would like to add kinetic, interaction and mass terms for the new gauge fields, but we see a similar feature as when we constructed the QCD Lagrangian: namely, we can add kinetic and interaction terms for the fields, but cannot add mass terms as this will spoil the gauge invariance. Since the photon is massless, this did not matter for the QCD Lagrangian, but here three of the four gauge bosons must somehow acquire mass. Thus, we can proceed no further than the following without a solution for producing the mass terms:

$$\mathcal{L}_{EW} = \mathcal{L}_{fermions} - \frac{1}{4} W_{\mu\nu}^i W^{\mu\nu i} - \frac{1}{4} B_{\mu\nu} B^{\mu\nu}, \tag{2.22}$$

where the field strength tensors $W_{\mu\nu}^i$ and $B_{\mu\nu}$ are:

$$\begin{aligned} W_{\mu\nu}^i &= \partial_\mu W_\nu^i - \partial_\nu W_\mu^i + g_L \epsilon^{ijk} W_\mu^j W_\nu^k \\ B_{\mu\nu} &= \partial_\mu B_\nu - \partial_\nu B_\mu \end{aligned} \ . \tag{2.23}$$

Fields	T^3	Y	Q
$\nu_{eL}, \nu_{\mu L}, \nu_{\tau L}$	1/2	-1	0
e_L, μ_L, τ_L	-1/2	-1	-1
e_R, μ_R, τ_R	0	-2	-1
u_L, c_L, t_L	1/2	1/3	2/3
d_L, s_L, b_L	-1/2	1/3	-1/3
u_R, c_R, t_R	0	4/3	2/3
d_R, s_R, b_R	0	-2/3	-1/3
ϕ^+	1/2	1	1
ϕ^0	-1/2	1	0

Table 2.2: Weak isospin and hypercharge assignments by field [2].

Using the relations (2.19)-(2.21), along with:

$$\begin{aligned} W^{\pm}_{\mu\nu} &= \partial_\mu W^{\pm}_\nu - \partial_\nu W^{\pm}_\mu \\ Z_{\mu\nu} &= \partial_\mu Z_\nu - \partial_\nu Z_\mu \end{aligned} , \tag{2.24}$$

(2.22) can be expressed in terms of the (for now, massless) vector bosons:

$$\mathcal{L}_{EW} = \mathcal{L}_{fermions} - \frac{1}{2} W^+_{\mu\nu} W^{\mu\nu -} - \frac{1}{4} Z_{\mu\nu} Z^{\mu\nu} - \frac{1}{4} F_{\mu\nu} F^{\mu\nu} + \mathcal{L}_{WZA}, \tag{2.25}$$

where $\mathcal{L}_{WZA}$ describes the interaction between the vector bosons.

2.2.2 The Higgs Mechanism

As shown above, the requirement of (non-Abelian) local gauge invariance seems to require the existence of a number of massless gauge vector bosons equal to the number of generators of the gauge group [35]. Therefore if we are to connect these gauge fields with physical states (as in eqs. (2.19)-(2.21)), we must find a way to produce the correct masses by spontaneously breaking the symmetry, in a way that does not spoil the gauge invariance of the theory.

Just such a solution was proposed in the early 1960s by Peter Higgs [40], and independently by Brout and Englert [31] and Guralnik, Hagen, and Kibble [39]. The approach begins by adding the following term to the electroweak Lagrangian (2.22):

$$\mathcal{L}_{Higgs} = D_\mu \Phi^\dagger D^\mu \Phi - V(\Phi) \tag{2.26}$$

where D is the same covariant derivative as given above in (2.17), and Φ is the complex $SU(2)_L$ scalar doublet:

$$\Phi = \begin{pmatrix} \phi^+ \\ \phi^0 \end{pmatrix}, \tag{2.27}$$

where ϕ^+ and ϕ^0 are the charged and neutral components of the field, respectively. Since ϕ^+ and ϕ^0 are complex, we may write them in terms of their components, with $\phi^+ = \phi_1 + i\phi_2$ and $\phi^0 = \phi_3 + i\phi_4$. The potential V is of the form

$$V(\Phi) = -\xi^2 \Phi^\dagger \Phi + \lambda (\Phi^\dagger \Phi)^2. \tag{2.28}$$

The addition of $\mathcal{L}_{Higgs}$ does not affect the overall $SU(2)_L \times \ U(1)_Y$ symmetry of the electroweak Lagrangian.

The crucial point comes when we examine the quadratic term in (2.28). If the quadratic term were positive, the minimum of the potential would correspond to a vacuum expectation value (VEV) of $\langle 0|\Phi|0\rangle = 0$. However, since we are subtracting the quadratic term, the potential has a global minimum at:

$$|\Phi| = \sqrt{\xi^2/2\lambda} = \frac{v}{\sqrt{2}}. \tag{2.29}$$

The shape of this potential is illustrated in figure 2.4. Here, Φ has a non-zero VEV of $v = \xi/\sqrt{\lambda}$, and $|\Phi| = 0$ is an unstable local maximum. Note that the vacuum is continuously degenerate in the parameter space of Φ, so that any (real) values may be assigned to the components ϕ_1, ϕ_2, ϕ_3 and ϕ_4, as long as (2.29) is satisfied. Due to the $SU(2)$ gauge invariance, we are free to arbitrarily choose any of these degenerate vacuum states as the physical vacuum without affecting the theory. However, once the choice is made, the symmetry of the potential is lost. We make the choice:

$$\langle \phi_1 \rangle = 0,\ \langle \phi_2 \rangle = 0,\ \langle \phi_4 \rangle = 0;\ \langle \phi_3 \rangle = v \tag{2.30}$$

so that

$$\Phi(x) = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ \nu + h(x) \end{pmatrix}, \tag{2.31}$$

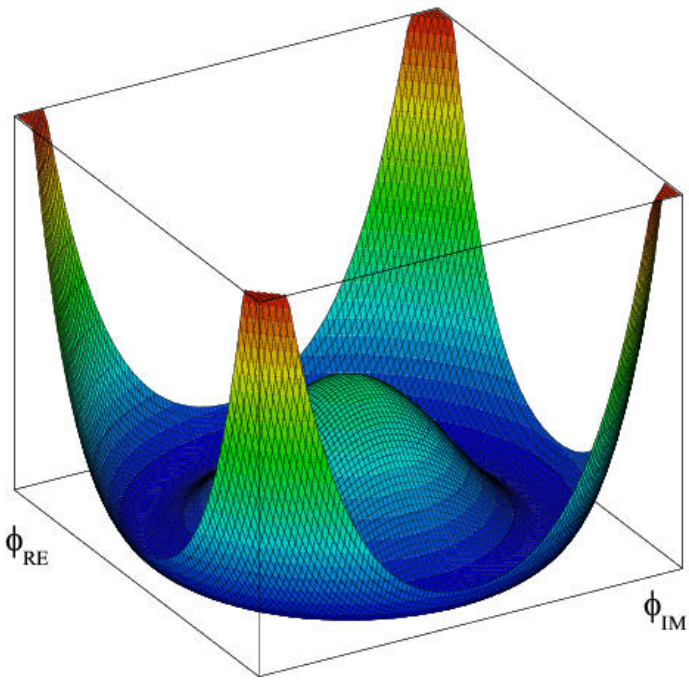


Figure 2.4: The potential of the Abelian Higgs model, shown to illustrate the quartic shape and the rotationally degenerate minimum. The non-Abelian (Standard Model) version has the same quartic structure, but its $SU(2)$ symmetry is harder to visualize [58].

where $h(x)$ is a (real-valued) perturbation about the ground state. We now plug the above back into (2.26), and evaluate the first term:

$$\begin{aligned} D_\mu \Phi^\dagger D^\mu \Phi = \tfrac{1}{2}\partial_\mu h \partial^\mu h + \tfrac{1}{4} g_L^2 v^2 W_\mu^- W^{+\mu} + \tfrac{1}{8}(g_L^2 + g_Y^2) v^2 Z_\mu Z^\mu \\ + \left(\tfrac{2h}{v} + \tfrac{h^2}{v^2}\right)\left(\tfrac{1}{4} g_L^2 v^2 W_\mu^- W^{+\mu} + \tfrac{1}{8}(g_L^2 + g_Y^2) v^2 Z_\mu Z^\mu\right) \end{aligned} \quad . \tag{2.32}$$

We can finally identify:

$$\begin{aligned} m_Z^2 &= \tfrac{1}{4}(g_L^2 + g_Y^2) v^2 \\ m_W^2 &= \tfrac{1}{4} g_L^2 v^2 \end{aligned} \quad . \tag{2.33}$$

Thus, the desired masses have been imparted to the W and Z bosons through the mechanism of spontaneous symmetry breaking. The degrees of freedom that were lost through the specification of the vacuum state of Φ are said to have been absorbed or "eaten" by the W and Z boson fields, which gain a longitudinal polarization (mass) [37]. Furthermore, it is important to note that the photon field A remains massless as a result of this choice of gauge.

The potential is evaluated as:

$$V(after\ S.S.B.) = \xi^2 h^2 + (h.o.\ terms) \tag{2.34}$$

where the higher-order terms involve the self-coupling of the higgs field. Here,

$$m_H^2 = \sqrt{2}\xi^2 = \sqrt{2}\lambda\ v^2. \tag{2.35}$$

Evidently, the scalar field has its own mass term. This mass belongs to a new neutral spin-0 particle that was generated by the broken symmetry, the Standard Model Higgs boson. The expectation value of the Higgs field, v, is given in terms of the Fermi coupling constant $v = (\sqrt{2}G_F) \approx 246\,\mathrm{GeV}/c^2$, and G_F is determined with a 0.6 ppm precision from muon decay measurements [42]. However, the variable λ is a free parameter; thus, the value of m_H is not predicted by the theory, and must be determined experimentally.

2.3 The Standard Model Higgs Boson

The Higgs boson is the quantum excitation of the Higgs field, which gives mass to the weak vector bosons as outlined in the previous section. The spontaneous symmetry breaking of the Higgs Mechanism does not automatically generate the fermion masses, so the Higgs field is generally assumed to give mass to the fermions via Yukawa couplings. Indeed, this must be the case if no other mechanism for generating the masses is discovered.

The Yukawa interactions are:

$$\mathcal{L}_{Yukawa} = -\alpha_{d_{ij}} \bar{q}_{L_i} \Phi d_{R_j} - i\alpha_{u_{ij}} \bar{q}_{L_i} \sigma_2 \Phi^* u_{R_j} - \alpha_{l_{ij}} \bar{l}_{L_i} \Phi e_{R_j} \tag{2.36}$$

where the $\alpha_{f_{ij}}$ are the couplings. The u_R, d_R and e_R represent the right-handed fermion singlets, and the q_L and l_L are the left-handed fermion doublets, and each term is parametrized by a 3×3 matrix in generation space [42]. Once the Higgs acquires a VEV, the Higgs-fermion interactions are diagonalized, so that $\alpha_{f_{ij}} \rightarrow \alpha_{f_{ii}}$. The fermion masses are then $m_f = \alpha_f v/\sqrt{2}$, with the important result that the Higgs coupling is proportional to the fermion mass. Since there are no right-handed neutrinos in the Standard Model, they do

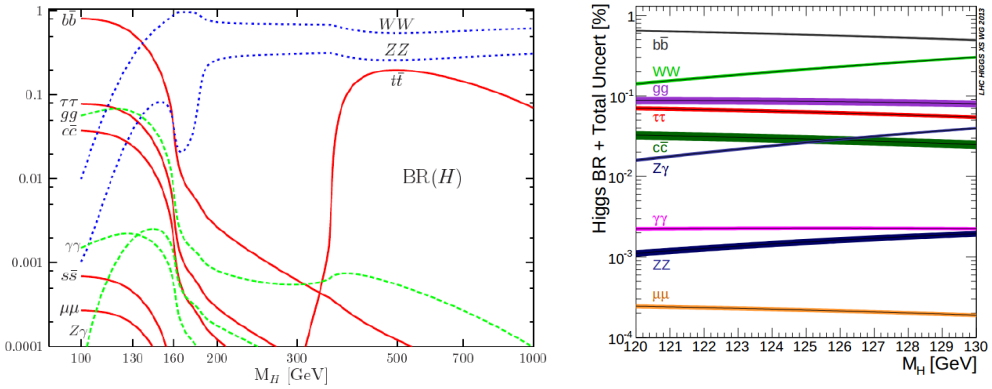


Figure 2.5: Predicted Standard Model Higgs branching ratios, as a function of m_H. Left: branching ratios across a wide range of m_H; right: detail of branching ratios roughly within $\pm 5\,\mathrm{GeV}/c^2$ of the observed Higgs boson mass. In general, the branching ratios are influenced by two competing features: the tendency of the Higgs to couple to more massive particles, and the fact that each decay channel is only "turned on" as the Higgs becomes heavy enough for its daughters to be on shell [47].

not participate in the Yukawa interactions and thus cannot obtain mass in this manner. As mentioned earlier, this is one of the limitations of the Standard Model, since neutrino oscillations have shown that they are indeed massive. The Higgs couplings to the W and Z are quadratic, as seen in the spontaneous symmetry breaking interaction terms of eq. (2.32). Given kinematic constraints, the couplings of the Standard Model Higgs allow a prediction of its branching ratios as a function of m_H. These branching ratios are given in figure 2.5. The production cross-sections are given in table 2.6 as a function of center-of-mass energy, and Feynman diagrams for the various modes are shown in figure 2.7.

The sensitivity of any Higgs analysis is dictated by the production rate of the Higgs in the channel of interest, compared to the rate of production of the relevant backgrounds. Channels that combine high production cross sections with distinctive final states are generally the most sensitive. At the LHC, QCD multijet production poses the largest and most difficult background for Higgs analyses, so searches where the final states consist of only jets and no other objects have the least sensitivity. Analyses that require the presence of at least one electron, muon, or photon in the final state, or make a cut on the amount of missing energy in the event, are able to reduce the QCD background and therefore improve their sensitivity.

$\sqrt{s}$ (TeV)	Production cross section (in pb) for $m_H = 125\,\text{GeV}$					
	ggF	VBF	WH	ZH	$t\bar{t}H$	total
1.96	$0.95^{+17\%}_{-17\%}$	$0.065^{+8\%}_{-7\%}$	$0.13^{+8\%}_{-8\%}$	$0.079^{+8\%}_{-8\%}$	$0.004^{+10\%}_{-10\%}$	1.23
7	$15.1^{+15\%}_{-15\%}$	$1.22^{+3\%}_{-2\%}$	$0.58^{+4\%}_{-4\%}$	$0.33^{+6\%}_{-6\%}$	$0.09^{+12\%}_{-18\%}$	17.4
8	$19.3^{+15\%}_{-15\%}$	$1.58^{+3\%}_{-2\%}$	$0.70^{+4\%}_{-5\%}$	$0.41^{+6\%}_{-6\%}$	$0.13^{+12\%}_{-18\%}$	22.1
14	$49.8^{+20\%}_{-15\%}$	$4.18^{+3\%}_{-3\%}$	$1.50^{+4\%}_{-4\%}$	$0.88^{+6\%}_{-5\%}$	$0.61^{+15\%}_{-28\%}$	57.0

Figure 2.6: Production cross-sections at specific energies, for a $125\,\text{GeV}/c^2$ Standard Model Higgs boson [42]. Values shown at 1.96 TeV are for $p\bar{p}$ collisions; all other values are for pp collisions.

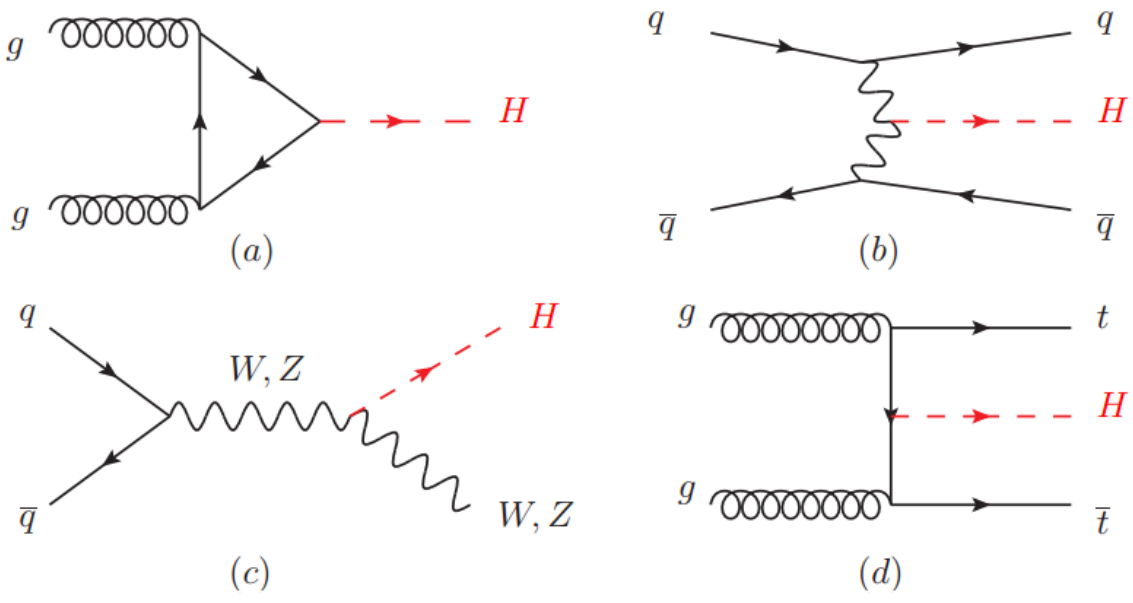


Figure 2.7: Feynman diagrams of the various production mechanisms [42].

Below, we give a review of some of the features of the production mechanisms and decay modes of the Standard Model Higgs, and discuss Higgs analyses at the LHC in the context of these channels.

2.3.1 Higgs Production

Gluon Fusion ($gg \rightarrow H$)

As is shown in Figure 2.6, the dominant SM Higgs production mechanism is gluon-gluon fusion. Because the gluon is massless, this process must be mediated by a virtual quark loop, and since the Higgs coupling is proportional to mass, the top quark is the main contributor to this loop. Here, the Higgs is produced in isolation, so that only its decay products can be used to identify the event. A $H \rightarrow bb$ search in this mode is extremely difficult due to the high QCD background. However, this production mechanism has been successfully utilized in searches where the Higgs decay products aid in identification, such as the $H \rightarrow \gamma\gamma$, $H \rightarrow ZZ$, and $H \rightarrow WW$ decays.

Vector Boson Fusion ($qq \rightarrow qqH$)

The production mechanism with the next-highest cross section is vector boson fusion. It involves the radiation of a W^+ and W^- by a quark and an antiquark, or the radiation of Z^0's by a qq, $q\bar{q}$, or $\bar{q}\bar{q}$ pair. The Higgs boson is then produced by the collisions of the vector bosons. Due to the relatively small momentum transferred to the quarks by the radiation of the vector bosons, these events tend to produce jets in the forward direction, with few central jets. The background can be reduced based on these selection requirements. However, the remaining high QCD background still tends to favor using the same decay products as the $gg \rightarrow H$ analyses.

Vector Boson Associated Production or "Higgs-strahlung" ($qq \rightarrow ZH/WH$)

An alternative is to consider associated production of the Higgs with weak vector bosons, namely $qq \rightarrow ZH$ and $qq \rightarrow WH$. At LHC energies and at $m_H \approx 125\,\text{GeV}$, the combined

cross section for this process is comparable to (though somewhat less than) the cross section for vector boson fusion. Here, a quark-antiquark pair annihilate to produce an off-shell W or Z, which goes on-shell by radiating a Higgs. The on-shell vector boson in the final state aids in the identification of this process, as the W and Z are easily identified by their decay products.

Associated Production with Top Quark Pairs ($gg/qq \rightarrow t\bar{t}H$)

This process occurs in one of two ways at the LHC. The first is the mechanism shown in figure 2.7, in which the collision of two gluons creates a top quark/antiquark pair, with a top or anti-top radiating a Higgs. A second possible diagram is one in which the collision of the two gluons produces a single virtual gluon that decays to a top/anti-top pair, and one of these radiates a Higgs. Although the $t\bar{t}H$ production mechanism has the lowest cross section, its decay products form a very distinctive and versatile final state. Analyses involving this production mode are quite complex, with events split into different categories depending on the number of jets, b-tagged jets, and number and flavor of leptons. Due to its relatively low electroweak and very low QCD backgrounds, the $t\bar{t}H$ mode also offers the opportunity to search for the Higgs inclusively across all its decays, independently from analyses that use other production mechanisms. Once enough data is collected, it will also be possible to measure the Higgs coupling to the top quark in ttH (it is not possible to do this accurately in $gg \rightarrow H$, since there may be additional significant contributions to the loop besides top quarks, if there are heavier quarks in nature).

2.3.2 Higgs Decay

$H \rightarrow \gamma\gamma$

Despite its small branching ratio, this channel played a key role in the search for a low mass (110 to 140 GeV) Higgs at the LHC, and was one of the main Higgs discovery modes. Here, the Higgs decays to two photons via a top quark or vector boson loop. This two-photon decay led to the conclusion that the newly observed particle was a boson with spin different

from one [13]. The sensitivity of the channel is driven by two factors: first, it is very "clean," in that the two photons are the only Higgs decay products and are easily identified; thus, the background is lower than in other channels, and the high-cross section $gg \to H$ mechanism may be exploited. The second advantage is that the excellent diphoton mass resolution (1-2%) of both the CMS and ATLAS detectors allows for a sharp invariant mass peak that stands out above the remaining background. The reducible background consists of jet-jet and jet-γ events where one jet is misidentified as a photon, and the irreducible backgrounds are from prompt $\gamma\gamma$ emission, as well as quark bremsstrahlung [43]. At CMS, in this channel alone, the observed significance above background at the time of the 2012 discovery was 4.1σ[13], and has since decreased slightly to 3.9σ[29] for the cut-based analysis. ATLAS, however, currently observes a 7.4σ excess in this channel, compared to 4.5σ at the time of the discovery [4].

$H \to ZZ \to 4l$ ($4e$, 4μ and $2e2\mu$)

This is the so-called golden decay, where the Higgs couples directly to two Z bosons, with each Z decaying to a $\mu\bar{\mu}$ or $e\bar{e}$ pair. The presence of 4 leptons in the final state allows precise reconstruction of the Higgs invariant mass at CMS and ATLAS, due to the excellent identification and reconstruction of muons and electrons at both detectors. If the mass of the Higgs boson had turned out to be greater than 180 GeV, it would have been discovered almost immediately after the LHC began taking data. In that case, two separate, on-shell Z bosons would have been easily identified via their sharp invariant mass peaks. At $m_H = 200\,\text{GeV}$, for example, a 5σ discovery at CMS would have been possible in the combined $H \to ZZ \to 4l$ channels with just over 2 fb^{-1} of data [18]. At $m_H = 125\,\text{GeV}$, however, only one of the Z bosons may be on shell, and a matrix element approach is used to construct a kinematic discriminant for each $4l$ event [42]. The dominant background is non-resonant $ZZ^{(*)}$ from $q\bar{q}$ annihilation and gg fusion. Other backgrounds include $Zb\bar{b}$, $t\bar{t}$ and Z + jets, which are reduced by making lepton isolation and impact parameter requirements. The quoted significance in this channel was 3.2σ in the CMS 2012 discovery paper[13], and has since climbed to 7.2σ with the full 7 TeV and 8 TeV datasets.

$H \rightarrow WW \rightarrow l\nu l\nu$

Here, the Higgs decays to a W^+W^- pair, each of which decays to a lepton and a neutrino. The neutrinos are not detectable, so their presence can only be inferred by measuring $\not{E}_t$ (missing transverse energy or "MET"). As a result, the invariant mass resolution of the Higgs is poor (about 20% m_H). The analysis therefore entails counting the events and carefully comparing to background, which includes irreducible WW, WZ and ZZ diboson processes, as well as $t\bar{t}$, W+jets and others. These can be suppressed to some extent: backgrounds such as $t\bar{t}$ can be reduced by applying a jet veto, and the MET requirement reduces the Drell-Yan and multi-jet backgrounds. In general, however, an accurate estimate of all the backgrounds is necessary in order to measure an excess. Although this channel did not contribute significantly to the initial Higgs discovery, it now has an excess of 3.8σ (4.0σ) for ATLAS (CMS), in the 0- and 1-jet categories [42].

$H \rightarrow b\bar{b}$

This decay of the Standard Model Higgs has the highest branching ratio, but the most problematic backgrounds of all the decay modes. The final state consists of just 2 b-jets, making it impossible to distinguish against the high QCD background when produced in isolation. $H \rightarrow b\bar{b}$ is therefore only generally searched for in conjunction with a distinctive production mode, such as ZH, WH or ttH, where the presence of leptons can be used to reduce the multijet background. B-tagging algorithms are also used to help identify the b-quark/antiquark pair from the Higgs, or reduce combinatorics in events with more than 2 b-jets. However, ZZ, WZ, $W/Z + jets$, and $t\bar{t}$ processes all contribute to backgrounds that include b quarks and leptons in the final state, and sophisticated classification techniques must be used to reduce this remaining background. For example, the VH($H \rightarrow b\bar{b}$) channels at CMS use multivariate classifiers trained on event kinematic, topological and b-tagging information to separate the Higgs boson signal in different p_T categories, and at different values of m_H [42]. The output of these MVAs are then binned by signal/background ratio, and the output of all channels is combined. An excess of events at $m_H = 125$ GeV is

observed in bins with the largest signal/background ratios, at a significance of 2.1σ. ATLAS performs a cut-based analysis that sees no significant excess over the predicted Standard Model background [42].

$H \rightarrow \tau^+\tau^-$

$H \rightarrow \tau^+\tau^-$ has the next highest fermionic branching ratio after $H \rightarrow b\bar{b}$, but τs are harder to identify than other leptons because of their wide range of decay modes, including hadronic decays. Searches are generally performed in the VBF, VH or ttH production modes. Backgrounds are similar to other analyses that use these modes, but are dominated by the Drell-Yan $Z \rightarrow \tau\tau$ production. The $\tau^+\tau^-$ invariant mass may be reconstructed from the visible τ decay products and a fit to the missing energy in the event, but this results in a relatively poor 15% resolution. Therefore, searches look for a broad excess over the $m_{\tau\tau}$ distribution. At $m_H = 125$ GeV, CMS (ATLAS) observes an excess above the background with a local significance of 3.4σ (4.1σ) in the full dataset. Both of these measurements provide substantial evidence of the Higgs boson coupling to leptons [42].

Other Decays

$H \rightarrow Z\gamma$ is similar to $H \rightarrow \gamma\gamma$, in that it makes up for its relatively low branching ratio by having a clean signal, with a $m_{ll\gamma}$ resolution of about 1-3%. Analyses search for a narrow peak over a continuous background that includes final state radiation from Drell-Yan decays, Z+γ and Z+jets where a jet is misidentified as a γ. No excess of signal events are observed. At $m_H = 125$ GeV, CMS (ATLAS), sets 95% CL upper limits of $9.5\times\sigma_{SM}$ ($18.2\times\sigma_{SM}$) for this channel.

Other modes are severely limited by statistics. $H \rightarrow e\bar{e}/\mu\bar{\mu}$ should be easily identifiable via its narrow invariant mass peak, but has a very low branching ratio. $H \rightarrow cc$ and the light-flavor $H \rightarrow qq$ channels have the same high QCD background problem as $H \rightarrow b\bar{b}$, but with lower branching ratios.

2.3.3 Higgs Boson Status

As stated in section 2.1, a new boson was observed jointly by the CMS and ATLAS collaborations in 2012, using 4.8 (5.1) fb^{-1} of data at 7 TeV and 5.9 (5.3) fb^{-1} at 8 TeV by ATLAS [4] and CMS [13], respectively. These observations were quantified by a p-value, which indicates the probability of a background-only measurement. A p-value of 3×10^{-7}, for example, corresponds to roughly a 5-standard-deviation (5σ) excess over the background. The observation reported by ATLAS had a local significance of 5.9σ at $m_H = 126.5$ GeV, and CMS reported a 4.9σ excess at $m_H = 125.5$ GeV.

Prior to the discovery, limits had been placed on the possible values of m_H for the Standard Model Higgs. Global fits to electroweak observables using data from LEP, the Tevatron and elsewhere, suggested $m_H = 89^{+22}_{-18}\ GeV$, while direct searches by LEP experiments yielded a lower bound of $m_H > 114.4$ GeV at 95% confidence. Combined data from direct searches by the CDF and D0 Tevatron experiments excluded m_H in the ranges 90 GeV to 109 GeV and 149 GeV to 182 GeV using 10 fb^{-1} of data. The Tevatron also observed a broad excess between 115 GeV and 140 GeV, with a 3σ local significance at $m_H = 125$ GeV, as seen in figure 2.8. This observation was not significant enough to claim discovery [42].

In 2013, CMS and ATLAS updated their previous results to include the full 8 TeV dataset (roughly 20 fb$^-$1 for each experiment). The sensitivity in the individual analyses was increased, with results listed above for each of the channels. The mass of the new boson has been determined to be 125.6±0.3 GeV (see figure 2.9), and the hypothesis that its spin is other than 0^+ (positive parity) has been rejected by various measurements at 95% confidence or greater [11]. Increases in the sensitivity of the various Higgs searches, together with the developing consensus of the identity of the new particle, has led to the transition from expressing results in terms of 95% confidence upper limits or significance of an excess, to measurements in terms of the observed signal strength parameter μ:

$$\mu = (\sigma \cdot BR)_{obs}/(\sigma \cdot BR)_{SM}, \tag{2.37}$$

which is the observed product of the Higgs boson cross-section and branching ratio for

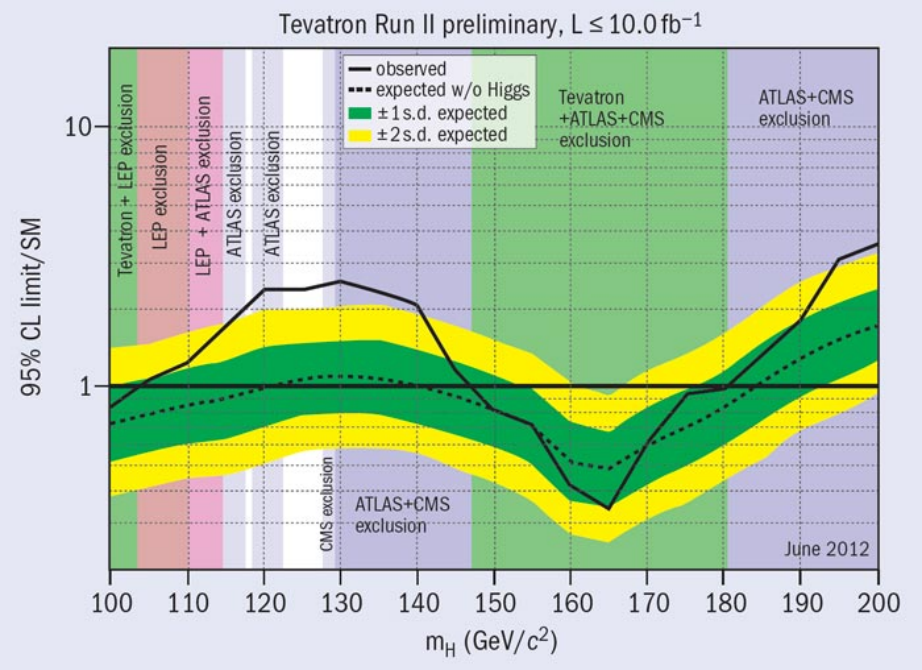


Figure 2.8: Summary plot showing 95% confidence limits on Higgs boson production at the Tevatron, as a function of m_H. Also shown are regions where the Higgs had been excluded by various experiments, up to June 2012.

a given channel, in units of the Standard Model prediction. Figure 2.10 summarizes the current signal strengths observed in combined ATLAS and CMS analyses, in the different decay modes of the Higgs boson [42].

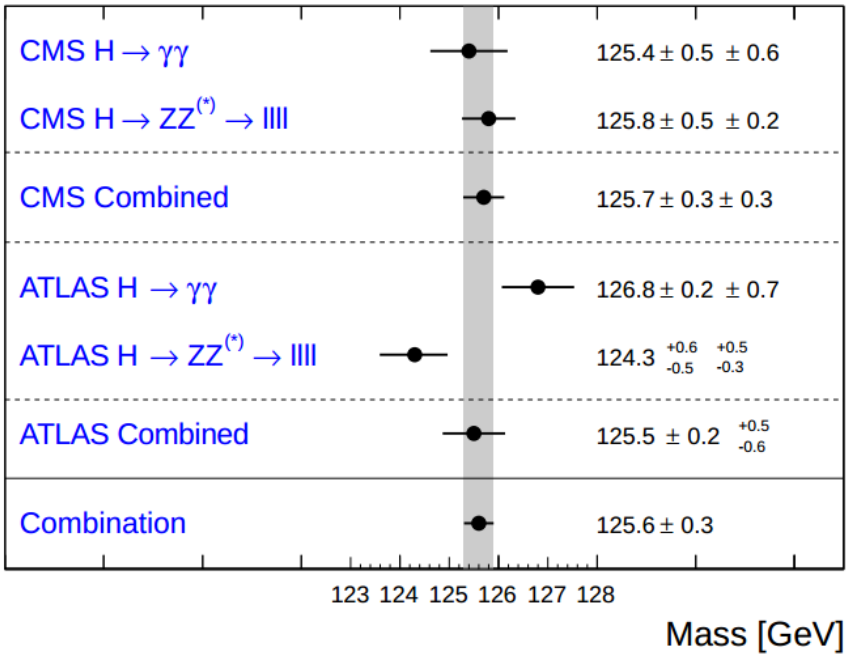


Figure 2.9: Higgs Boson mass measurements using the combined 7 TeV and 8 TeV datasets [42].

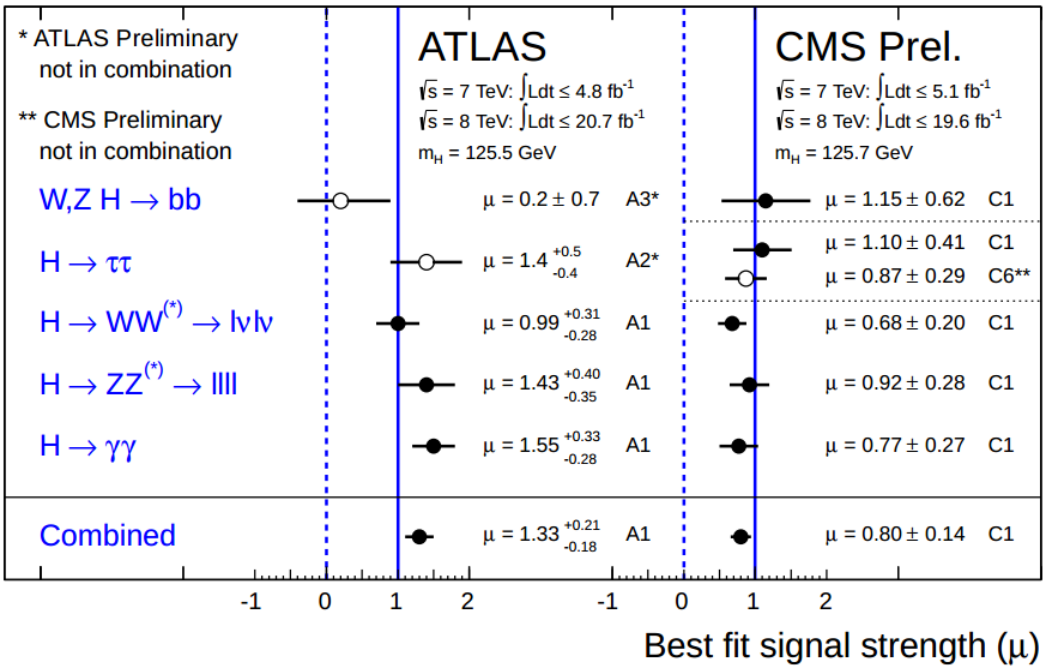


Figure 2.10: Signal strengths by Higgs boson decay channel [42].

Chapter 3
CMS and the LHC

3.1 The LHC

The Large Hadron Collider (LHC) at CERN in Geneva, Switzerland is the world's highest-energy particle accelerator. Two counter-circulating beams of protons collide at four points around the 27 km-circumference ring, at the locations of the ATLAS, CMS, ALICE and LHCb detectors, respectively. A chain of accelerators inject protons into the LHC, and are illustrated in figure 3.1. The protons begin their journey in a simple tank of Hydrogen gas. The gas is ionized by a duoplasmatron, and the resulting protons are accelerated through a linear accelerator to 50 MeV. These protons are then successively accelerated to 1.4 GeV in the Proton Synchrotron Booster (PBS), and then to 25 GeV in the Proton Synchrotron (PS). Here, the beam size and final bunch spacing are established. The protons are then accelerated to 450 GeV in the Super Proton Synchrotron (SPS) before being injected into the LHC. There, they are gradually accelerated to the final beam energy, and held at this energy for several hours during collisions[9]. During 2010-2011, the proton-proton center-of-mass collision energy was 7 TeV, and was increased to 8 TeV in 2012.

While the energy of colliding protons during the first long run was kept at roughly half the LHC design value of 14 TeV, various adjustments to beam parameters caused the instantaneous luminosity to gradually approach the designed $10^{34}\,\mathrm{cm}^{-2}\mathrm{s}^{-1}$. A nominal bunch spacing of 50 ns was used throughout 2010-2012, which was half the LHC-designed spacing of 25 ns. However, an increase in the bunch population of up to a factor of 1.5 with respect to the designed 1.15×10^{11}, as well as a reduction in emmitance and β^* at

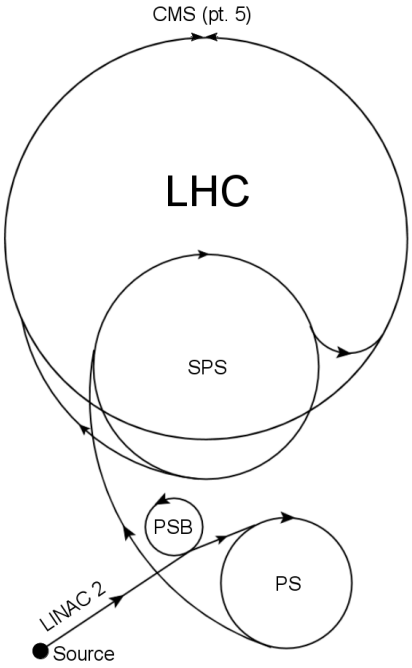


Figure 3.1: Illustration showing stages of the LHC accelerator chain, and the position of Intersection Point 5 on the LHC. The flow of protons are indicated by arrows. Note that the purpose of the illustration is to give the approximate sizes and relative locations of the accelerators, and is not strictly to scale.

the interaction points, resulted in the delivery of approximately $30\,\mathrm{fb}^{-1}$ of data over the 2010-2012 period [46]. Figure 3.2 shows the maximum instantaneous luminosity delivered to CMS per day, and figure 3.3 shows the integrated luminosity as a function of time.

3.2 The CMS Detector

The Compact Muon Solenoid (CMS) detector is located at Intersection Point 5 of the LHC. Along with ATLAS (A Toroidal LHC Apparatus), it is one of the LHC's two general-purpose detectors. CMS shares Point 5 with the Total Elastic and diffractive cross section Measurement (TOTEM) experiment, a forward-physics detector that studies the total proton-proton cross section and proton structure[57]. The CMS detector is cylindrical in shape. It is 21.6 m long, 14.6 m in diameter and weighs 12 500 t, making it the heaviest (but not the largest) detector at the LHC. CMS gets its name from the 4 T superconducting solenoid which laterally encloses its tracking and calorimetry systems.

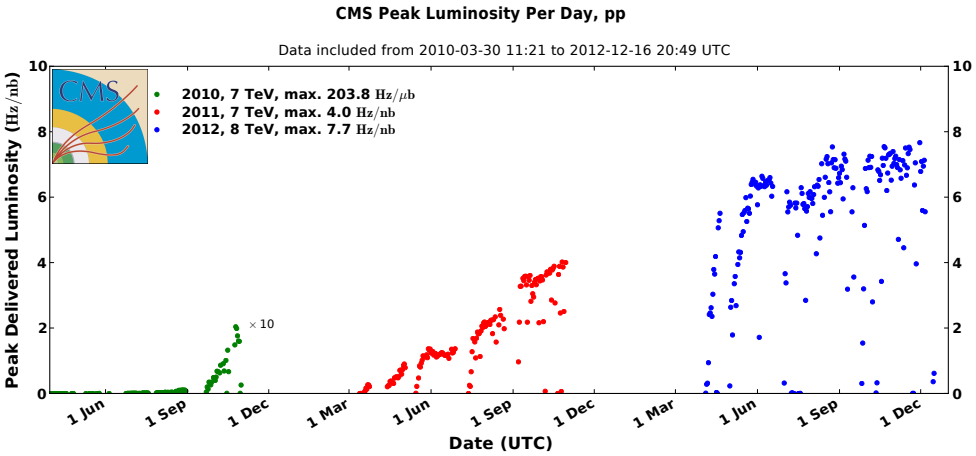


Figure 3.2: Peak instantaneous luminosity per day, 2010-2012[10].

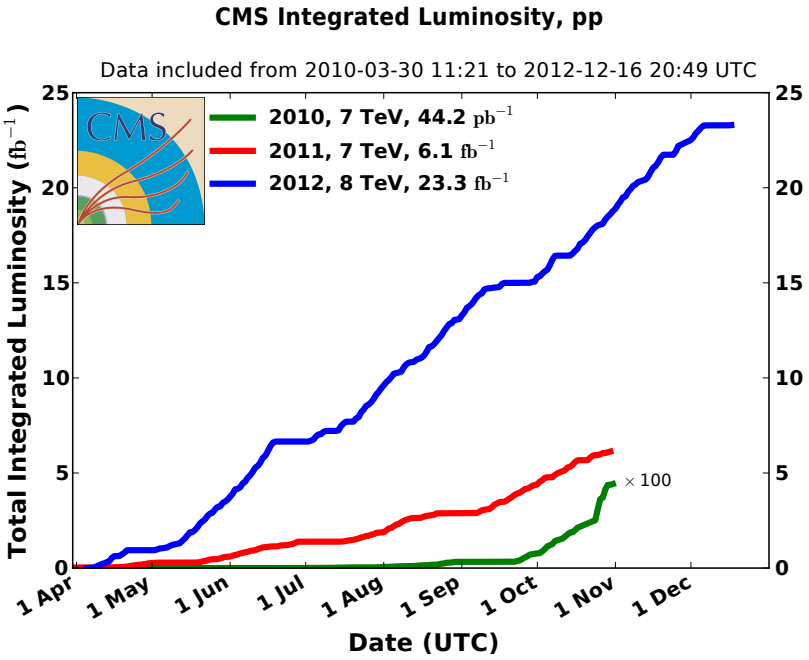


Figure 3.3: Integrated luminosity delivered to CMS by year. More than 90 percent of this data was recorded: 5.55 fb^{-1} in 2011 and 21.79 fb^{-1} in 2012[10].

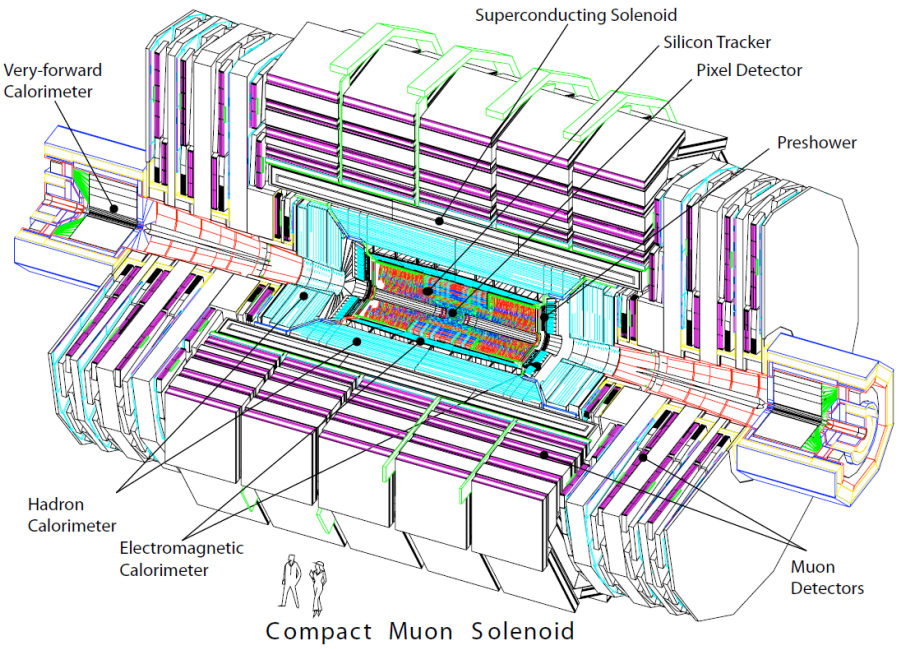


Figure 3.4: A schematic view of the CMS detector.

CMS follows a Cartesian coordinate system, with the origin located at the interaction point. The positive x direction points toward the center of the LHC ring, the positive y direction points vertically up, and the positive z direction points tangent to the beam to the west. The r, θ, and ϕ coordinates have the usual definitions, and the pseudorapidity coordinate η is defined as:

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

CMS is made up of several sub-detectors, arranged in concentric cylindrical shells. Working out from the interaction point, these include a pixel tracker, silicon strip trackers, an electromagnetic calorimeter, and a hadronic calorimeter. Beyond the solenoid, muon drift chambers are interspersed within the magnet's iron return yoke. The detector is also divided into endcap and barrel regions that can be separated for maintenance. Figure 3.4 shows an overall schematic view of the detector, and figure 3.5 shows the η coverage of the detector components.

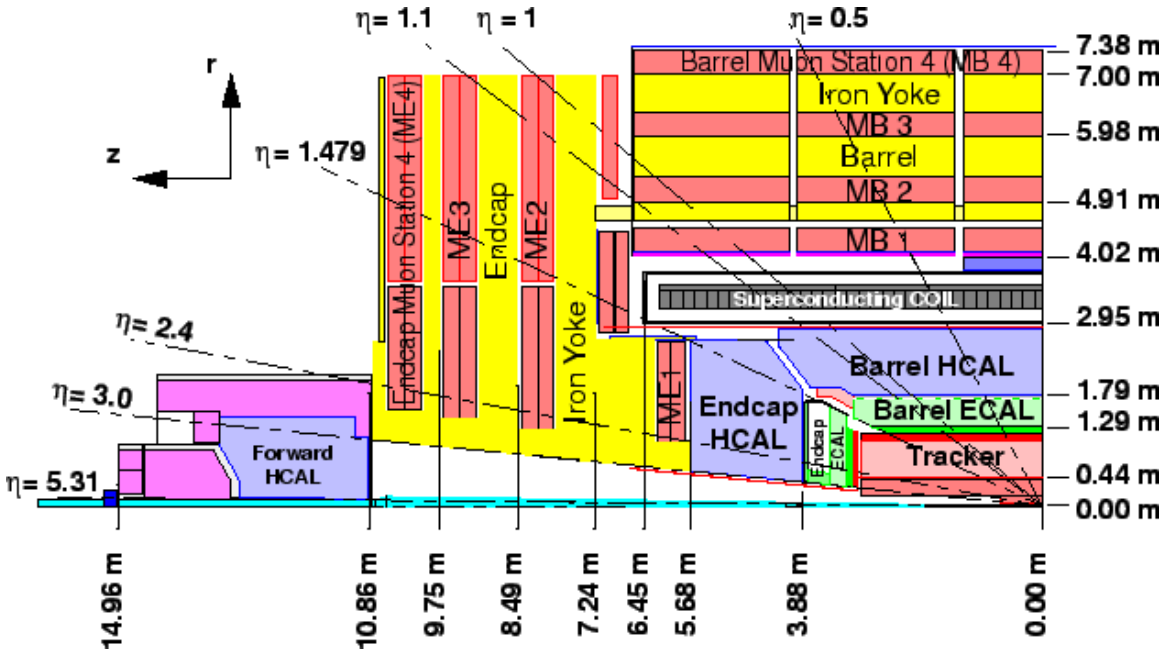


Figure 3.5: A simplified view of a longitudinal quadrant of CMS, showing the major systems and their coverage in η.

3.2.1 Tracker

The CMS tracker is the largest all-silicon tracking system ever constructed. It is made up of two types of detectors: pixel and silicon strip detectors, which cover a total effective area of over 200 m^2[56]. A longitudinal view of the tracker is shown in figure 3.6.

The pixel tracker is located closest to the interaction vertex, and is made up of 3 layers at 4.5 cm–10 cm radius, as well as two end disks on each side at $z = \pm 34.5$ cm and ± 46.5 cm. The 10^7/s particle flux at these radii is balanced by a small 100×150 $\mu\mathrm{m}^2$ pixel size, resulting in an average occupancy of 10^{-4} per bunch crossing per pixel. The fine granularity of the pixel detector requires 66 million readout channels.

In the barrel region, the strip detector is divided into the Tracker Inner Barrel and Disks (TIB/TID) and Tracker Outer Barrel (TOB). At 20 cm $< r <$ 55 cm, the TIB contains four layers of 10 cm $\times$ (80 µm–120 µm) silicon strips. The strips in the first two of these layers are made up of "stereo" (double-sided) modules that provide a two-dimensional single-point measurement that is accurate to within a few tens of microns[17]. There are 3 TID disks

on each end, where the strip pitch varies between 100 µm–140 µm. The first two rings of the TID contain also stereo modules. The TOB laterally surrounds the TIB/TID and has 6 layers, extending out to a radius of 115 cm. The strips in the TOB are 25 cm × (120 µm–180 µm). Again, the first two layers of the TOB are composed of stereo modules. The Tracker End Cap (TEC) comprises 9 disks that extend into the region 120 cm $< |z| <$ 280 cm. The innermost 2 rings and the fifth ring of the TEC have stereo modules, with radial strips of 97 µm to 184 µm average pitch[17].

The strip detector has a total of 9.3 million channels. The choice to use silicon strips at larger radii was sufficient for the lower-flux environment of the outer regions of the tracker, with occupancy in the range of 1-3%. Depending on η, the strip tracker has the ability to make 8-14 measurements per trajectory, 4-6 of which are "stereo" measurements[56].

Figure 3.7 shows some of the performance characteristics of the entire tracker. The impact parameter resolution is dominated by the fine spatial granularity of the pixel detector, while the momentum resolution is due to the lever arm of the combined pixel and strip trackers. Further improvement in the momentum resolution of muons is provided by the muon system (see section 3.2.3). The tracker has excellent primary vertex resolution (figure 3.8), and the efficiency to identify primary vertices is nearly 100% with two or more tracks[54]. Given that typical impact parameters for tracks from B hadron decays are on the order of a few hundred µm, the tracker is also able to offer precise reconstruction of secondary vertices[55]. Offline software is used to further refine the identification of jets from bottom quark hadronization.

3.2.2 Calorimetry

The Electromagnetic Calorimeter (ECAL) and Hadronic Calorimeter (HCAL) together form a complete calorimetry system for the measurement of the energy of photons, electrons, hadronic jets and missing transverse energy (MET).

The electromagnetic calorimeter is divided into a barrel section that covers a pseudo-rapidity range $|\eta| < 1.5$, and an endcap region that covers $1.5 < |\eta| < 3.0$. It is composed of a total of 68524 lead tungstate (PbWO4) scintillating crystals, coupled to avalanche photodiodes (barrel) or vacuum photodiodes (endcaps). This arrangement has allowed the

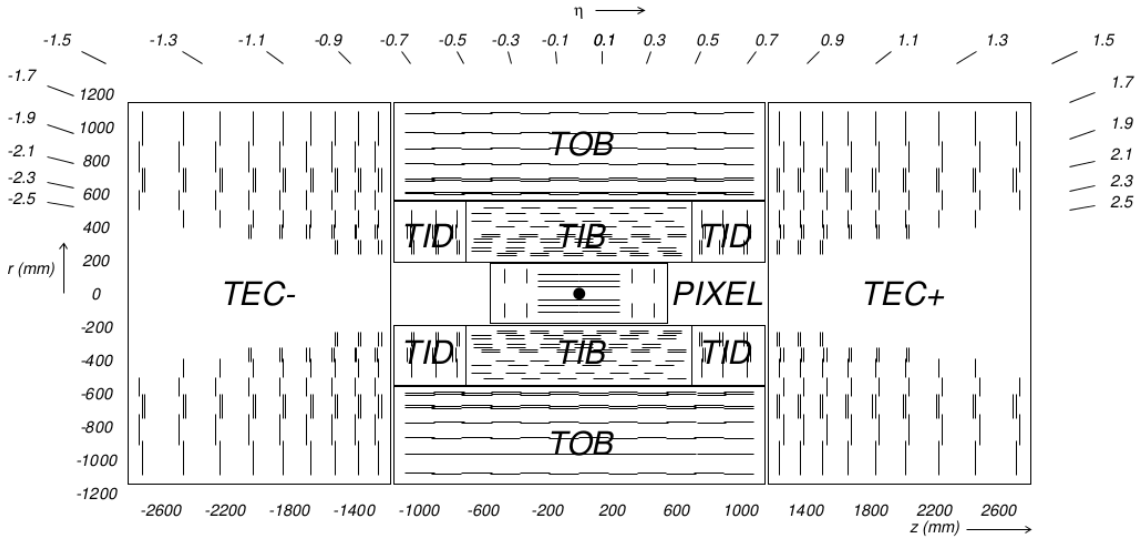


Figure 3.6: Cutaway view of the CMS tracker, in the $r - z$ plane [56]. The labeled regions are the pixel detector, the tracker inner barrel (TIB), tracker inner disks (TID), tracker outer barrel (TOB), and tracker end cap (TEC). See text for further description.

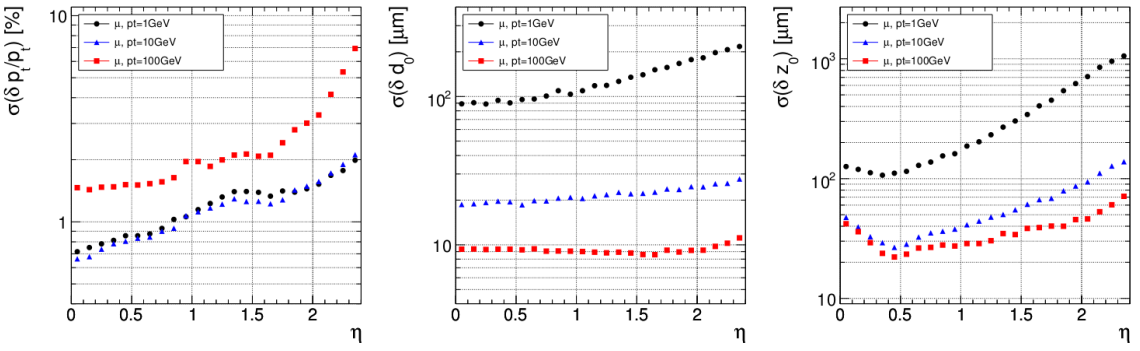


Figure 3.7: From left to right: the transverse momentum, transverse impact parameter and longitudinal impact parameter resolutions of the tracker as a function of η. The values shown are for muons with $p_T = 1$, 10 and 100 GeV.

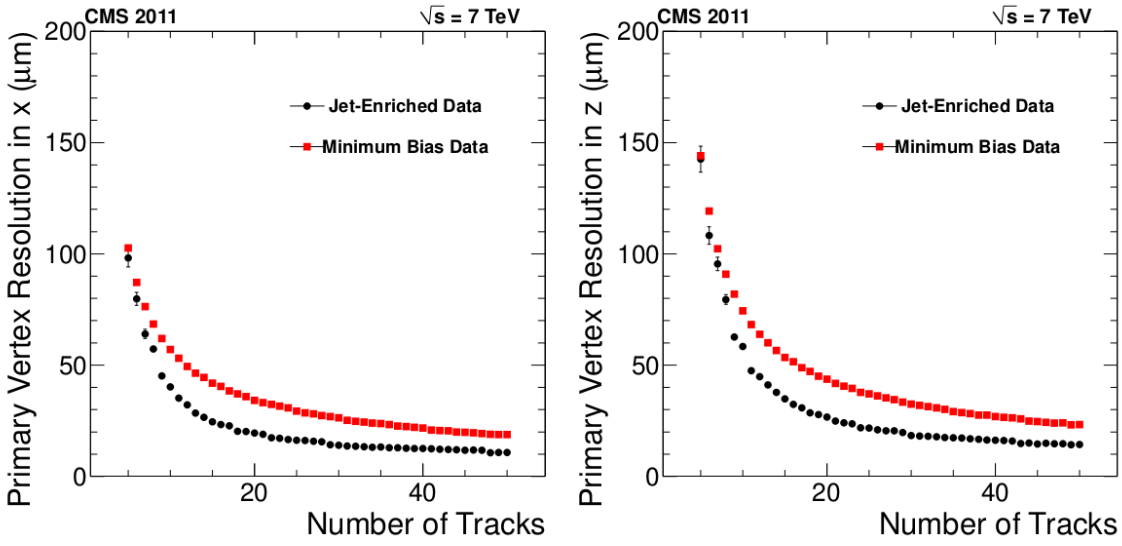


Figure 3.8: Primary vertex resolution of the tracker as a function of number of tracks used in the reconstruction, in x (left) and z (right).

design of a compact calorimeter inside the solenoid that is fast, has fine granularity (with a front face of about $2\,\text{cm} \times 2\,\text{cm}$ in the barrel and $3\,\text{cm} \times 3\,\text{cm}$ in the endcaps), and is radiation resistant. The crystal length is 23.0 cm in the barrel and 22.0 cm in the endcaps, covering many radiation lengths with $X_0 = 0.9\,\text{cm}$ in the lead tungstate crystals [17]. In addition, a preshower device is placed in front of the ECAL in the endcaps. The CMS ECAL provides very good electron and photon energy resolution (see figure 3.9).

The HCAL is divided into two general regions: the barrel and forward sections are located within the solenoid, while the endcap and outer sections are outside the solenoid. The constraints imposed by the solenoid led to a relatively short absorption length of 7.2λ at $\eta = 0$ for the barrel section, so the outer section was added as a complementary "tail catcher" to aid in calorimetry and to help prevent leakage into the muon system. The HCAL is organized into towers that radiate away from the center of the detector. These are composed of alternating layers of brass or steel absorbers and plastic scintillator tiles. The HCAL energy resolution for single jets (R=0.5) is given in figure 3.10. The good hermeticity of this system around the interaction point facilitates the measurement of missing transverse energy. In the absence of energy clustering corrections, the MET resolution is approximately

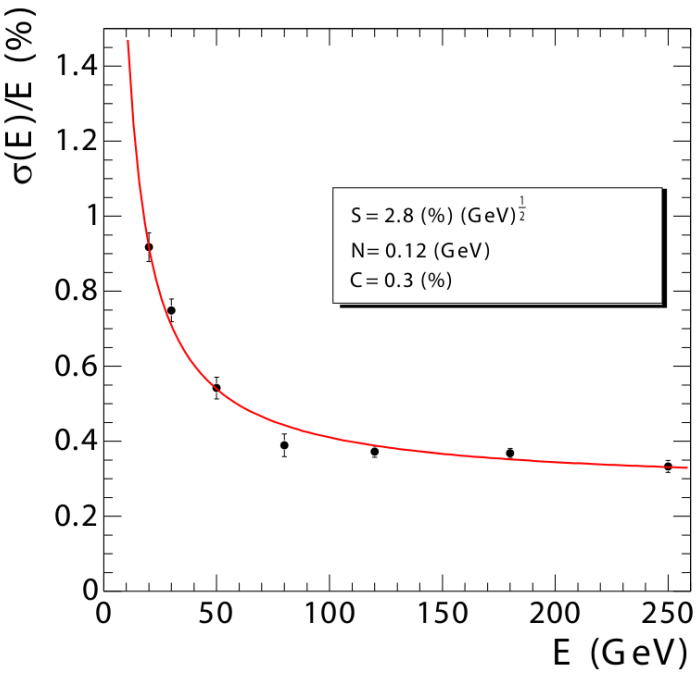


Figure 3.9: ECAL energy resolution as a function of energy as measured from a test beam.

$\sigma(MET) \approx 1.25\sqrt{\Sigma E_T}$[17].

3.2.3 Muon System

The desire to accurately identify, trigger and reconstruct muons was one of the driving design considerations for CMS. The CMS muon system is located outside the solenoid and has been integrated into the return yoke where a 2T magnetic field is present. Three types of gaseous detectors are used: in the barrel region, the neutron-induced background is small, so the muon rate and residual magnetic field is low, and drift tube chambers are used. In the two endcaps, where the muon rate as well as the neutron induced background rate is high, and the magnetic field is also high and non-uniform, cathode strip chambers are deployed and cover the region up to $|\eta| < 2.4$. In addition, resistive plate chambers are used in both the barrel and endcap regions. The coarse spatial resolution but good time resolution of the RPCs compliments the good spatial resolution of the CSCs and DTs, so that the correct bunch crossing may be identified. The HCAL "tail catcher," as well as the solenoid itself

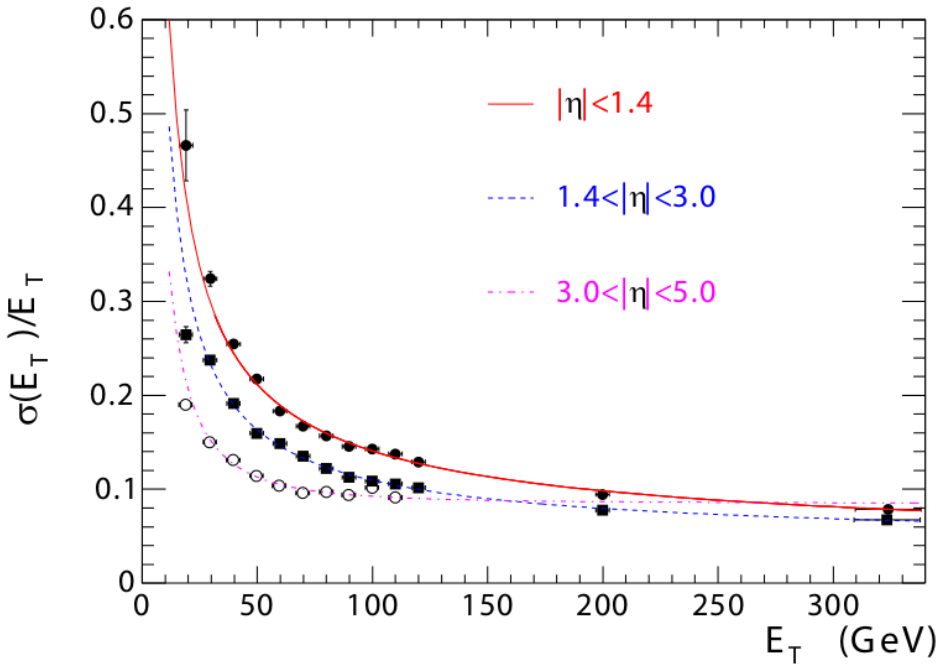


Figure 3.10: The jet transverse energy resolution as a function of the simulated jet transverse energy for barrel jets ($|\mu| < 1.4$), endcap jets ($1.4 < |\mu| < 3.0$) and very forward jets.

further aid identification by absorbing hadrons not captured by the inner HCAL. Figure 3.11 shows the improvement in muon momentum resolution obtained by using a combined fit to tracker and muon detector measurements [17].

In the barrel, the muon chambers are arranged in coaxial layers, interleaved with the return yoke. These layers are called "stations," and there are 4 stations in each of the 5 wheels of the yoke, located at a distance ranging from 4.0 m–7.4 m from the beam. Each wheel is divided into 12 sectors, with chambers in different stations staggered so that high-p_T muons cross at least 3 stations. In each of the endcaps, the CSCs and RPCs are arranged in 4 disks (also called stations) perpendicular to and concentric with the beam. Here, the trapezoidal CSCs are arranged in concentric rings that overlap in ϕ: 3 rings in the innermost station and 2 rings in the other stations. The position resolution provided by each chamber is roughly 200 µm, and the angular resolution of the muon direction is about 1(10) mrad for the DTs(CSCs). In total, the muon system covers roughly 25 000 m^2 of detector surface

area and has almost 1 million channels [17].

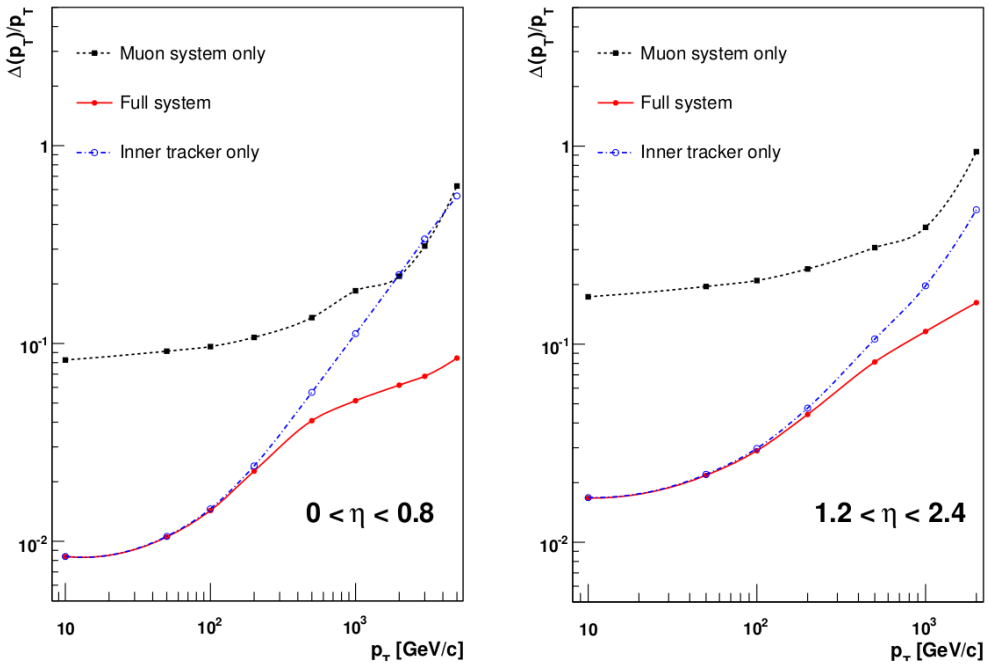


Figure 3.11: Muon p_T resolution in the barrel (left) and endcap (right) regions. Resolution is shown for muons reconstructed using the tracker only, muon system only, and combined tracker and muon system measurements.

3.2.4 Trigger and Data Acquisition

During 2011-2012 proton-proton collisions, the bunch-crossing rate being delivered to CMS was roughly 15 MHz [46]. Given that each event from the detector is approximately 1.5 MB [17] of information, recording every single event would have meant a raw data rate on the order of 20 TB/s. At design luminosity and high pileup, this rate climbs to order 1 PB/s. Blindly recording every event would therefore be both technologically unfeasible, as well as undesirable from an analysis standpoint, as the number of uninteresting events would be overwhelming and unmanageable. Instead, CMS uses a tiered trigger system to filter the data and record only events of interest for further offline analysis.

The Level 1 (L1) Trigger functions as part of the detector hardware to reduce the initial rate of tens of MHz to about 100 kHz. The data that passes L1 is filtered by the High-Level Trigger (HLT), which further reduces the rate to a few hundred Hz for offline storage and processing. Reduced-granularity data from the detector, called "trigger primitives," are sampled by the L1 Trigger, which uses custom-designed, programmable electronics to quickly reach a decision regarding whether to discard the event or to pass it on to the HLT. The L1 Trigger electronics are located partly on the detector itself, and partly in an underground room 90 m from the detector. The constraint of the bunch-crossing rate, added to the time required for data transmission, means that the total time allotted to L1 Trigger calculations is less than 1 µs[17].

Events that pass L1 are sent to a computer farm that runs the HLT software. In contrast to L1, the full event information is available at HLT. The online HLT is run in the same software framework as is used for offline reconstruction and analysis. The HLT code is organized into modules, each representing an algorithm that performs a specific physics object selection. In order to pass a given trigger, an event must pass a selection of these modules in sequence – this sequence is called a path. Modules are organized along the path so that events are rejected as quickly as possible if they fail basic criteria (such as a calorimeter energy deposit threshold). More complex algorithms are reserved for the end of the path, and may include full or partial event reconstructions, including tracking[8],[17].

Chapter 4
Event Selection and Object Identification

The remainder of this dissertation is devoted to a search for the Standard Model Higgs boson in the $t\bar{t}H$ associated production mode. The specific channel considered by this $t\bar{t}H$ search involves the semi-leptonic decay of the $t\bar{t}$ pair. This scenario is illustrated in figure 4.1. The top quark decay to a bottom quark and a W boson happens in an overwhelming majority of cases ($> 99\%$). The remaining decay products in the event are produced when one of the W bosons decays leptonically, and the other decays hadronically. The focus of this analysis is the $H \to b\bar{b}$ decay of the Higgs boson, which occurs about 58% of the time at $m_H = 125\,\text{GeV}$, relative to other Higgs decays. Thus, the overall process being sought is $t\bar{t}H \to b\bar{b}b\bar{b}q\bar{q}l\nu$. The procedure for selecting collision events, and identification and selection of objects corresponding to this process is described below.

4.1 Data Samples

4.1.1 Triggers

As discussed in chapter 3, it undesirable for the CMS detector to attempt to record each collision event. Therefore, a system is needed to determine which events are recorded and which are discarded: this is the role of the trigger. There exist hundreds of different trigger algorithms that may cause an event to be recorded, and these correspond to different types of physical processes that may be of interest. The different triggers and the physical processes

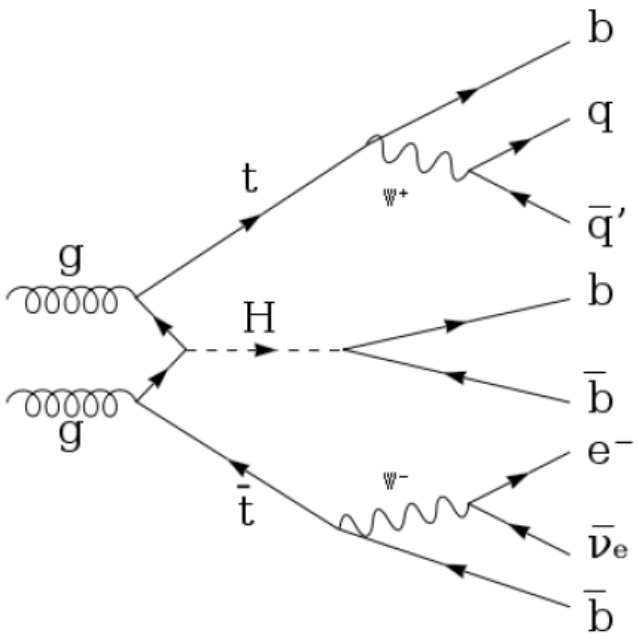


Figure 4.1: The lepton + jets mode of $t\bar{t}H$.

Dataset	Trigger Name	Description
SingleMu	HLT_IsoMu24_eta2p1	At least one isolated muon with $p_T > 24\,\text{GeV}$ and $\lvert\eta\rvert < 2.1$
SingleEle	HLT_Ele27_WP80	At least one high-quality electron with $p_T > 27\,\text{GeV}$

Table 4.1: Triggers used to collect the data for this analysis.

they attempt to capture are classified broadly into different datasets. The software (HLT) triggers used to collect the data for this analysis are given in table 4.1. The table also gives a brief description of the trigger requirements, and the dataset name corresponding to each trigger.

This analysis uses a total of $19.3\,\text{fb}^{-1}$ of data collected in 2012 at CMS, in pp collisions at 8 TeV. Table 4.2 summarizes the runs comprising each dataset, and the amount of data collected during each run period.

Dataset	Run Range	Int. Luminosity
/SingleMu/Run2012A-13Jul2012-v1/AOD	190456–193621	0.81 fb^{-1}
/SingleMu/Run2012A-recover-06Aug2012-v1/AOD	190782–190949	0.08 fb^{-1}
/SingleMu/Run2012B-13Jul2012-v1/AOD	193834–196531	4.40 fb^{-1}
/SingleMu/Run2012C-24Aug2012-v1/AOD	198022–198523	0.50 fb^{-1}
/SingleMu/Run2012C-PromptReco-v2/AOD	198941–203746	6.39 fb^{-1}
/SingleMu/Run2012D-PromptReco-v1/AOD	203768–208686	7.27 fb^{-1}
Total SingleMu	**190645–208686**	**19.3 fb^{-1}**
/SingleElectron/Run2012A-13Jul2012-v1/AOD	190456–193621	0.81 fb^{-1}
/SingleElectron/Run2012A-recover-06Aug2012-v1/AOD	190782–190949	0.08 fb^{-1}
/SingleElectron/Run2012B-13Jul2012-v1/AOD	193834–196531	4.40 fb^{-1}
/SingleElectron/Run2012C-24Aug2012-v1/AOD	198022–198523	0.50 fb^{-1}
/SingleElectron/Run2012C-PromptReco-v2/AOD	198941–203746	6.40 fb^{-1}
/SingleElectron/Run2012D-PromptReco-v1/AOD	203768–208686	7.27 fb^{-1}
Total SingleElectron	**190645–208686**	**19.3 fb^{-1}**

Table 4.2: Summary of the data analyzed in this dissertation.

4.1.2 Event Cleaning

In the data, we require that every event pass the following filters [15]:

- CSC tight beam halo filter (noise from beam setting off endcap muon cambers),
- HBHE noise filter (HCAL electronics noise),
- HCAL laser filter (HCAL calibration laser firing during collisions),
- ECAL dead cell filter,
- Tracking failure (events with too few tracks),
- Noisy SCs in EE (ECAL electronics noise),
- Beam-scraping filter ($\geq 25\%$ high-purity tracks).

These are standard event cleaning filters, mainly designed to reduce known sources of detector noise and noise from the beam. In addition, every data event must contain at least one reconstructed primary vertex (PV) that passes the following selection:

- The number of degrees of freedom used to find the PV must be larger than 4,

- The absolute value of the z-coordinate of the PV must be smaller than 24 cm,
- The absolute value of the ρ-coordinate of the PV must be smaller than 2 cm,
- The PV must not be identified as fake.

This set of requirements on the PV is an additional sanity check to ensure that the event contains at least one real pp hard-scattering process that originated within the pixel detector, close to the beam axis.

4.2 Event Reconstruction

At CMS, event reconstruction proceeds in a centrally-organized fashion. Reconstruction begins during data collection: starting from raw detector information, the HLT performs partial reconstruction of objects in the process of evaluating the various trigger algorithms, and determining whether to discard the event or send it to the next level of processing. If the event is kept, it is sent in raw form to a dedicated cluster called the Tier 0 (T0), located at CERN, and is immediately archived to tape. At the T0, the data also undergo full, prompt reconstruction, as well as reformatting into the more user-accessible Analysis Object Dataset (AOD) format. Conditions tags are applied during reconstruction to account for data-taking conditions that vary from run-to-run, such as location and size of the beam spot, non-active parts of the detector, and any modifications to the trigger menu. If data are later re-reconstructed, a new tag may be applied due to a change in software or to reflect evolving understanding of conditions.

The Tier 1 (T1) and Tier 2 (T2) sites receive datasets from the T0 for storage and additional processing. These sites also generate the monte-carlo simulation samples, and (in the case of T2s) provide resources for individuals and institutions to access data, run jobs and store output in a variety of formats for specific analyses. Commonly, AOD and RECO formats are further simplified to the Physics Analysis nTuple format (PAT), which keeps only needed collections and is somewhat more user-friendly. The ntuples used in this analysis are derived from PAT-formatted data.

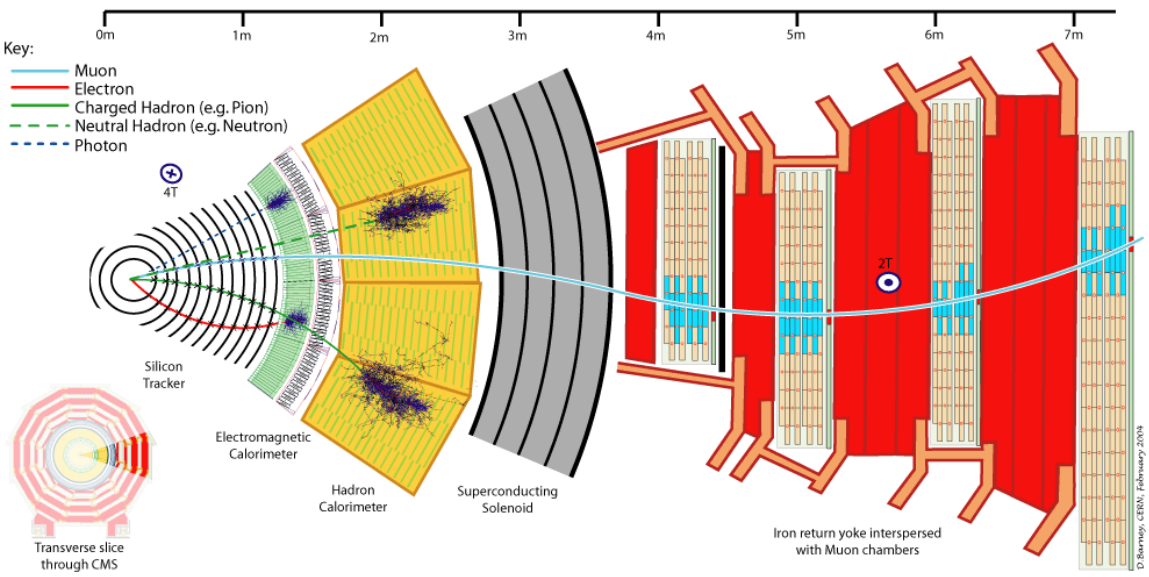


Figure 4.2: A cartoon of a transverse slice of the barrel region of the detector, illustrating the role of the different sub-detectors in identifying different particles. In the particle-flow reconstruction algorithm, particles are identified and reconstructed using information in a coordinated way from all the detector components.

Many different algorithms exist for reconstructing and identifying the physical particles in a collision event. In reconstructed data, the results of these different algorithms are stored in parallel as separate collections of objects. For this analysis, we use objects that were reconstructed as part of the Particle-Flow (PF) algorithm [19]. This algorithm simultaneously incorporates information from all the sub-detectors to reconstruct objects in a coordinated fashion. The role of each sub-detector is illustrated in figure 4.2. Briefly, each event is processed as follows: first, tracks are reconstructed using a high-efficiency, low-fake-rate iterative method [19]. High-quality muons (and corresponding tracks) are identified and removed from the set of candidates. Next, electrons are reconstructed and identified, and their tracks and associated ECAL deposits are removed from the algorithm. The remaining tracks and calorimeter deposits may then be linked and identified as charged hadrons. Finally, unmatched HCAL deposits are attributed to neutral hadrons, and unmatched ECAL deposits are identified as photons [19]. Additional details on the methods used to reconstruct the individual objects are given below.

4.3 Objects

This section describes the identification and selection of the objects used in the analysis. For all objects, we begin with the particle flow reconstruction, and make additional requirements.

4.3.1 Leptons

Muons

Muons are generally the easiest particles to identify in a CMS event, and are thus the first PF objects to be reconstructed. Muons may be reconstructed by using information soley from the tracker or DTs/CSCs, or by fitting tracks to hits in the muon chambers by extrapolating the DT and CSC hits back to the tracker, and performing a minimum-χ^2 match to the tracks [17]. Muons reconstructed using the latter method are called global muons. Muons of sufficient p_{T} traverse the entire detector almost unimpeded, leaving only minimum-ionizing deposits in the calorimeter, and a combined s-shape trajectory in the tracker and muon system. Thus, the signature of a high-p_{T}, well-reconstructed global muon is very hard to fake by other charged particles that do not survive beyond the ECAL or HCAL. As discussed in chapter 3, the combined use of the muon system and tracker also leads to superior muon momentum resolution for high-p_T muons.

We select muons based on the requirements given in table 4.3. Here, we make a distinction between "tight" and "loose" muons. For tight muons, we require a global muon with p_{T} above the trigger threshold, where the track trajectory has been matched to the muon chamber hits to within a certain precision. To identify the muon as prompt (i.e., originating from the initial hard scattering process), we make relative isolation and impact parameter requirements. The impact parameter requirements associate the muon to the event primary vertex. The relative isolation is calculated by defining two cones around the muon trajectory (as shown in figure 4.3): the energy measured inside the smaller cone is subtracted from that of the larger cone (where the axis of the larger cone is defined by the direction of the muon at the primary vertex), and the result is divided by the p_{T} of the muon. By placing a limit on this ratio, we veto muons that may have been produced in hadronic showers, which

Tight μ	Loose μ
$p_T > 30\,\text{GeV}$	$p_T > 10\,\text{GeV}$
PFRelIso($R = 0.4$)<0.12	PFRelIso($R = 0.4$)<0.2
$\lvert\eta\rvert$<2.1	$\lvert\eta\rvert$<2.5
Global Muon	Global Muon or Tracker Muon
hits in > 5 tracker layers	
χ^2/N_{DOF} of track fit <10	
> 0 hits in pixel tracker	
> 1 muon stations hit	
$\lvert d0(\text{PV})\rvert < 0.2$ cm	
$\lvert dZ(\text{PV})\rvert < 0.5$ cm	

Table 4.3: Cuts for selecting tight and loose muons.

will typically consist of jets located close to or surrounding the muon. The loose muons are still fairly isolated, but may not have been measured as accurately or may be due to pileup (multiple pp collisions occurring in the same bunch-crossing). We demand that all single-muon events passing our selection contain exactly one tight muon, and no loose muons.

Electrons

As with the muons, electrons are selected from the objects reconstructed by the particle flow algorithm. Here, the particle flow identification procedure is somewhat more complex. The large amount of tracker material causes significant bremsstrahlung of electrons, which results in trajectory discontinuities and a wide angular spread of particles in the ECAL. Dedicated electron track reconstruction and bremsstrahlung recovery algorithms account for these effects. The appropriate tracks and ECAL energy deposits are then linked, and the output of a boosted decision tree (BDT) is used for the final electron identification [5]. Conversions of prompt photons in the tracker are identified and handled separately by the particle flow.

Our electron selection criteria are given in table 4.4. The impact parameter and relative isolation requirements for tight electrons are applied for similar reasons as the corresponding requirements for muons. In addition, electrons must be above a certain value of the particle

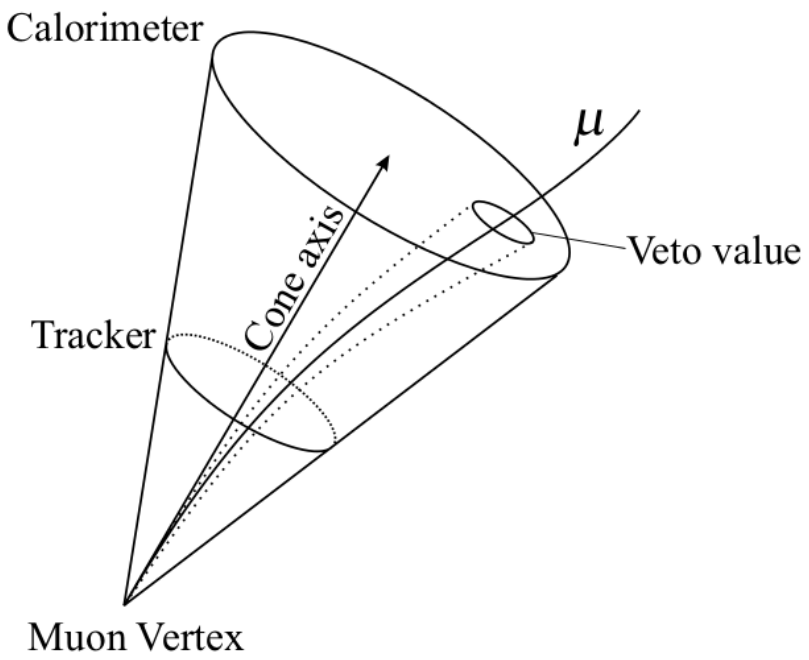


Figure 4.3: Diagram showing isolation cones surrounding the reconstructed muon [17].

flow electron ID BDT, not be identified as a converted photon by the PF, and must also not have any missing inner track hits (an additional check against converted photons). Tight electrons must be above the threshold of the single electron trigger. Loose electrons have relaxed impact parameter and isolation requirements, as well as a p_{T} cutoff below the trigger threshold. We demand that single-electron events must have exactly one tight electron, and no loose electrons.

4.3.2 Jets

Due to color confinement, all quarks (except for the top) quickly hadronize after being produced in isolation. Quark-antiquark pairs are pulled from the vacuum to form colorless combinations of quarks (hadrons). In high-energy experiments, this process is observed as a spray of particles in the direction of the original quark, called a jet. At CMS, jets may be reconstructed soley from deposits in the HCAL ("calojets"), particle-flow objects ("PFjets"), or with a variety of other object collections. In addition, there are several available standard jet clustering algorithms [7].

Tight e	Loose e
$p_T > 30\,\text{GeV}$	$p_T > 10\,\text{GeV}$
PFRelIso(R=0.3)< 0.1	PFRelIso(R=0.3)< 0.2
$\lvert\eta\rvert < 2.5$	$\lvert\eta\rvert < 2.5$
$!(1.442 < \lvert\eta\rvert < 1.566)$	$!(1.442 < \lvert\eta\rvert < 1.566)$
output of PF ID BDT > 0.5	output of PF ID BDT > 0.5
pass PF conversion veto	pass PF conversion veto
expected inner track hits ≤ 0	expected inner track hits ≤ 0
$\lvert d0(\text{PV})\rvert < 0.02$cm	$\lvert d0(\text{PV})\rvert < 0.04$cm
$\lvert dZ(\text{PV})\rvert < 1$cm	

Table 4.4: Tight and loose electron selection cuts.

In this analysis, jet reconstruction begins with the set of individual particles reconstructed with the particle-flow algorithm. The jets are subsequently formed from these objects using the anti-kt jet clustering algorithm [7]. The anti-kt method is based on two competing "distances" between candidate objects d_{ij} and d_i, given by:

$$\begin{aligned} d_{ij} &= \min\left(\tfrac{1}{k_{ti}^2},\ \tfrac{1}{k_{tj}^2}\right)\tfrac{\Delta_{ij}^2}{R^2} \\ d_i &= \tfrac{1}{k_{ti}^2} \end{aligned}, \tag{4.1}$$

where R is the cone radius, k_{ti} is the transverse momentum of a particle or jet candidate, and

$$\Delta_{ij}^2 = (\phi_i - \phi_j)^2 + (y_i - y_j)^2, \tag{4.2}$$

where y is the rapidity [7]. This analysis uses a cone radius of $R = 0.5$. The clustering proceeds by determining the smallest of all possible d_i and d_{ij} among the set of particles and/or jet candidates. If $d_{ij} < d_i$, the i and j particles/candidates are merged into a new jet candidate, and the process begins again. If $d_{ij} > d_i$, the ith candidate is called a jet, and is removed from the list. The algorithm continues until all candidate objects have been assigned to a jet [7]. Figure 4.4 shows an example of a CMS event containing anti-kt-clustered jets. The following requirements are made on the reconstructed jets:

- $p_T > 30\,\text{GeV}$ for all jets,
- $p_T > 40\,\text{GeV}$ for 3 highest-p_T jets,

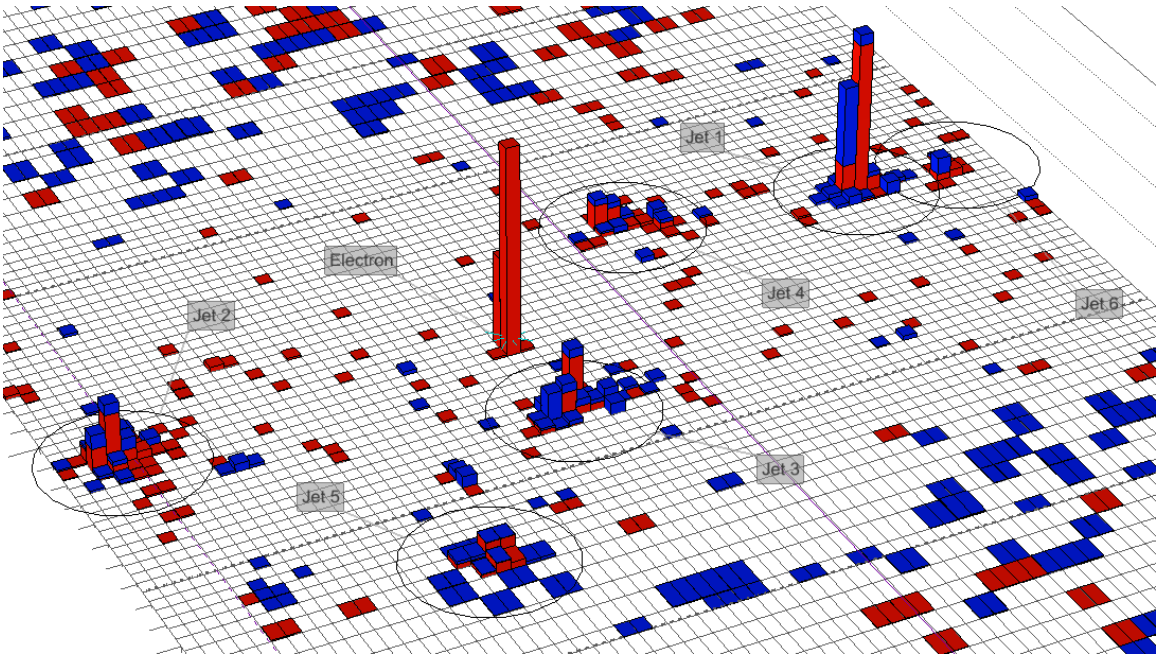


Figure 4.4: A 3D view in detector $\eta - \phi$ space showing an electron+jets candidate event in data. The circles represent jets identified by the anti-kt algorithm ($R = 0.5$). The red and blue bars are ECAL and HCAL energy deposits, respectively, and their height is proportional to the amount of energy deposited. Tracks are not shown, but contribute to the PF objects used in the jet reconstruction.

- $|\eta| < 2.4$.

Events that contain fewer than 4 jets meeting the above criteria are discarded. The specific p_{T} requirements were originally introduced in an effort to improve agreement between data and simulation at low jet p_{T}; the modeling has since improved (as will be discussed in chapter 5), and these cuts may be relaxed at the next iteration of the analysis, subject to further studies.

4.3.3 Missing Energy

The event missing transverse energy (MET) is also a PF object – it is calculated by performing a vector sum of the 4-momentum of all the measured PF objects in the event, finding the direction and magnitude of the result, and subtracting 180 degrees in ϕ from that result. Due to the good hermicity of the calorimeters, it is expected that this quantity should give a reasonable estimate of transverse component of the momentum of any neutrinos present. We make no explicit requirement on MET in the event selection; however, MET (and MET-ϕ) is used later in the analysis.

4.3.4 B-Tagging

For each jet, the Combined Secondary Vertex (CSV) algorithm is used to estimate whether or not the jet was produced as a result of b-quark hadronization. Jets positively identified by the algorithm are called "b-tagged" jets. Bottom quarks typically hadronize into B hadrons, whose relatively long lifetime causes them to decay hundreds of microns to several millimeters away from the primary vertex. The point at which this decay occurs is called a secondary vertex, and may be accurately resolved with respect to the primary vertex by the pixel tracker. The CSV algorithm therefore looks for a secondary vertex, and combines several variables involving this vertex into a final discriminant. The variables used include [55]:

- Significance of the transverse distance between the primary and secondary vertices,
- Vertex mass (from associated tracks),

- Number of tracks associated to the secondary vertex,
- Number of tracks in the jet,
- Energy of tracks associated to vertex w.r.t. energy of all tracks in the jet,
- Pseudorapidities of secondary vertex tracks w.r.t. the jet axis,
- 2D impact parameter (IP) significance of first track that raises the vertex mass above 1.5 GeV (charm threshold),
- 3D IP significance for each track in the jet.

In the case where no secondary vertex was reconstructed, only the fourth and last variables in the preceding list are used. The variables are combined into a set of likelihood ratios, which attempt to discriminate separately between b-, c-, and light-flavor (LF) jets. These are then combined into a final discriminant, whose output is shown in figure 4.5. By making cuts on the output of this discriminant, one can achieve a given efficiency for real b-jets, at the expense of contamination by other flavors. This situation is illustrated in figure 4.6. We identify b-jets by placing a cut at a CSV output of 0.679, which corresponds to an approximate efficiency of 60-70% for real b-jets, 10-20% for c-jets and 1-2% for light-flavor (LF) jets [55]. Further improvements are obtained by incorporating the output of the CSV discriminant itself into the analysis.

4.4 Categories

The above requirements on triggers, event quality, and reconstructed objects constitute the basic selection criteria for the data analyzed in this dissertation. Each of the events that pass the selection must contain exactly one lepton – either a tight muon or tight electron – as well as four or more jets, at least two of which must be b-tagged. After the basic selection is performed, the events are separated into categories based on the number of jets that they contain, as well as the number of jets that have passed the medium working point of the CSV b-tagging algorithm. The aim of this categorization is to give several regions of varying

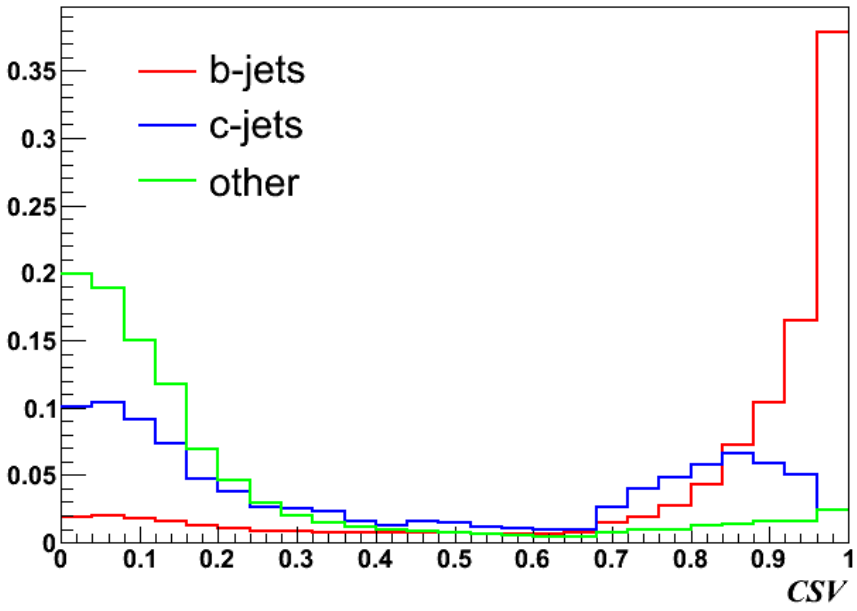


Figure 4.5: CSV discriminant output distribution for MC-truth b-jets, c-jets and light-flavor (other) jets. This plot was produced from the $t\bar{t}H$ signal MC after baseline analysis selection with ≥ 4 jets and ≥ 2 b-tagged jets, so the distributions are biased towards high b-tagging efficiency here; however, the purpose of the plot is simply to show the relative shapes of the different distributions. For the efficiencies, see the text and figure 4.6.

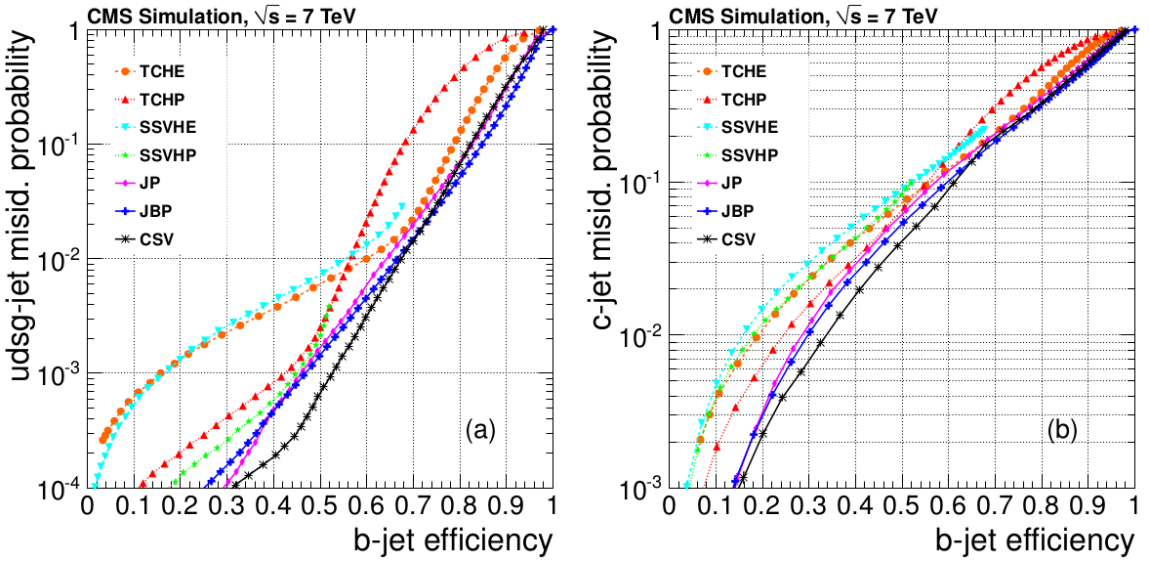


Figure 4.6: Efficiency comparison of b-tagging algorithms at CMS, measured in simulated multijet events [55]. Left(Right): probability to misidentify udsg(c) jets, as a function of identification efficiency of real b-jets.

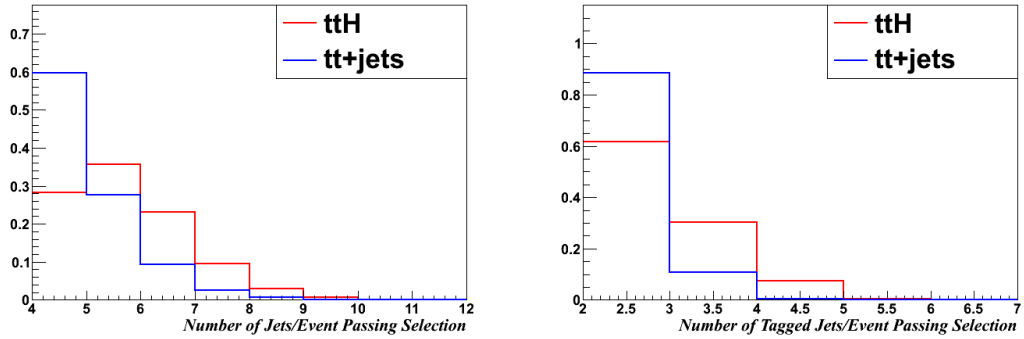


Figure 4.7: Number of jets (left) and number of b-tagged jets (right) in $t\bar{t}H$ and $t\bar{t}+$jets simulated events passing the full event selection. The plots are normalized to the number of entries for comparison.

signal and background concentrations, so that some categories serve as control regions, while others provide greater sensitivity to the $t\bar{t}H$ signal. As figure 4.7 illustrates, $t\bar{t}H$ events contain more jets and b-tagged jets on average than $t\bar{t}+$jets events; therefore, the more signal-rich categories are those with higher numbers of jets and b-tags. We use 7 categories, containing ≥ 6 jets + 2 b-tags, 4 jets + 3 b-tags, 5 jets + 3 b-tags, ≥ 6 jets + 3 b-tags, 4 jets + 4 b-tags, 5 jets + ≥ 4 b-tags, and ≥ 6 jets + ≥ 4 b-tags, respectively. An explicit yield table for each of the categories will be shown after a discussion of the backgrounds in the next chapter.

Chapter 5
Data Modeling

In the preceding chapter, we made certain requirements that defined which data events would be included in the analysis. The goal of this selection was to try to isolate as many $t\bar{t}H$ events as possible while eliminating as many other kinds of events as possible. However, the $t\bar{t}H$ process for which we are searching still makes up only a small fraction of events passing the basic selection. The events present in our sample that are not $t\bar{t}H$ events are called background events, and the desired $t\bar{t}H$ events are called signal events. In order to accurately distinguish between the background events and any $t\bar{t}H$ signal events that may be present, we must understand the relative concentration of signal and background in our selection, as well as the relative contributions of the different backgrounds. In addition, we must assess how accurately the simulated measurement of physical objects reflects measurements of objects in the data, and make adjustments for any discrepancies that would affect the analysis.

5.1 Generation of Simulated Data

Signal and background processes are modeled with Monte Carlo (MC) simulation. Event-generation software such as PYTHIA [52] or MADGRAPH [32] is used to perform numerical computations of Standard Model processes. The dynamics and kinematics of hard scatter (pp) events are modelled using statistical methods that simulate the probabilistic nature of high-energy quantum-mechanical interactions. The subsequent decays of the particles are also simulated, including quark hadronization. Effects from additional pp interactions in the

same bunch crossing (pileup) are modeled by adding simulated minimum-bias events to the generated hard interactions.

The particles produced in the event are then handed to another software program which simulates the detector response; we use the GEANT [1] software package to simulate the CMS detector. This program calculates the effects of the detector environment on the particles, such as the alteration of particle trajectory due to the large magnetic field, and interaction with the detector material, including showering. The simulated response of the detector electronics is obtained, and the events proceed through the same reconstruction steps as collision data. They are stored in the same ntuple formats and analyzed with the same software framework as the real data. We subject every MC sample used in the analysis to the same event selection requirements as the data.

5.2 MC Samples

5.2.1 Signal

The $t\bar{t}H$ signal MC is generated using PYTHIA at 9 separate presumed Higgs masses, and is listed in table 5.1. Since we are searching for the leading-order $t\bar{t}H$ process (and not "$t\bar{t}H$ +jets"), we do not need to generate additional partons, and the use of MADGRAPH to generate the signal samples is not necessary. The decay of the Higgs boson is not forced, but is allowed to decay according to the branching ratios predicted by the standard model at a given m_H. There are approximately 1 million raw MC events for each mass point in the signal samples before event selection. After event selection at $m_H = 125\,\text{GeV}$, about 8% of the original MC events remain. The cross section and integrated luminosity are used to give a flat normalization weight to all events, so that the number of expected events in the data is calculated as:

$$N_{exp} = \frac{L \cdot \sigma}{N_{gen}} N_{sel}. \tag{5.1}$$

This is also done for each of the simulated background processes. Individual events receive additional corrections, which alters N_{exp} depending on the sample. This is described later in section 5.3.

Higgs Mass	Cross Sect.	N_{gen}	N_{pass}
110 GeV	0.1887 pb	977880	82272
115 GeV	0.1663 pb	1000000	83443
120 GeV	0.1470 pb	999508	82737
122.5 GeV	0.1383 pb	999400	81800
125 GeV	0.1302 pb	995697	80226
127.5 GeV	0.1227 pb	1000000	79144
130 GeV	0.1157 pb	933970	72697
135 GeV	0.1031 pb	996800	74954
140 GeV	0.09207 pb	1000000	72829

Table 5.1: List of signal MC masses with corresponding cross sections, number of generated MC events and number of MC events passing the 1 tight lepton, $\geq$4 jets, $\geq$2 b-tagged jets event selection.

5.2.2 Background

The dominant background $t\bar{t}+$jets sample, as well as $t\bar{t}W$, $t\bar{t}Z$, W + jets, and Drell-Yan processes, are generated with MADGRAPH to leading order. PYTHIA is then used to calculate the parton shower. Single-top production is modeled with the next-to-leading order (NLO) generator POWHEG [3], and is combined with PYTHIA in the same manner as the MADGRAPH samples. Electroweak diboson processes (WW, WZ, and ZZ) are simulated entirely with PYTHIA. The $t\bar{t}+$jets MADGRAPH sample is generated inclusively, with tree-level diagrams for up to $t\bar{t}+3$ extra partons. These extra partons include both b and c quarks. We chose MADGRAPH primary because of its ability to accurately model these extra partons, which make up the most important part of the $t\bar{t}H$ background.

Examples of Feynman diagrams for some of the background processes are given in figure 5.1. Broadly speaking, the backgrounds consist of processes involving combinations of strong and electroweak interactions. Electroweak interactions that result in either leptons or jets in the final state (such as Drell-Yan or electroweak quark pair production), or processes involving only strong interactions (such as multi-jet QCD) are strongly suppressed by our event selection requirements. Instead, our backgrounds are primarily made up of processes that are able to produce a combination of both jets and isolated leptons. This leads to the inclusion of V+jets and diboson events, which may contain multiple jets and leptons,

Sample	Generator	Cross Sect.	N_{gen}	N_{pass}
$t\bar{t}$+ jets				
$t\bar{t} \rightarrow$ jets	madgraph	112.33 pb	31111456	419
$t\bar{t} \rightarrow \ell\nu + 4$ jets	madgraph	107.66 pb	25327478	1411535
$t\bar{t} \rightarrow \ell\nu\ell\nu + 2$ jets	madgraph	25.81 pb	12100452	259733
$t\bar{t} + W$	madgraph	0.249 pb	195396	10907
$t\bar{t} + Z$	madgraph	0.208 pb	209512	11447
$W + 1$ jet	madgraph	6440.4 pb	23134881	2
$W + 2$ jets	madgraph	2087.2 pb	33933328	31
$W + 3$ jets	madgraph	619.0 pb	15463420	114
$W + 4$ jets	madgraph	255.2 pb	13365439	6818
Z/γ^*+ jets				
10 GeV/c^2 $< M_{\ell\ell} <$50 GeV/c^2	madgraph	14702 pb	37828841	6
$Z/\gamma^* + 1$ jet ($M_{\ell\ell} > 50$ GeV/c^2)	madgraph	666.7 pb	24032562	7
$Z/\gamma^* + 2$ jets ($M_{\ell\ell} > 50$ GeV/c^2)	madgraph	215.1 pb	2350806	11
$Z/\gamma^* + 3$ jets ($M_{\ell\ell} > 50$ GeV/c^2)	madgraph	66.07 pb	10753491	836
$Z/\gamma^* + 4$ jets ($M_{\ell\ell} > 50$ GeV/c^2)	madgraph	27.38 pb	6370630	5730
Single t				
s-channel	powheg	3.79 pb	259657	424
t-channel	powheg	56.4 pb	3744404	3359
tW	powheg	11.1 pb	496918	4110
Single $\bar{t}$				
s-channel	powheg	1.76 pb	139835	187
t-channel	powheg	30.7 pb	1933504	1832
tW	powheg	11.1 pb	492779	4241
WW	pythia	54.8 pb	9955089	532
WZ	pythia	32.3 pb	9931257	1191
ZZ	pythia	7.7 pb	9755621	864

Table 5.2: From left to right: list of background MC datasets used in the analysis, software used to generate events, cross sections used for normalization, number of generated MC events and number of MC events passing the 1 tight lepton, $\geq$4 jets, $\geq$2 b-tagged jets event selection (the term "jets" in the leftmost column denotes generated jets, and not jets as defined in the event selection).

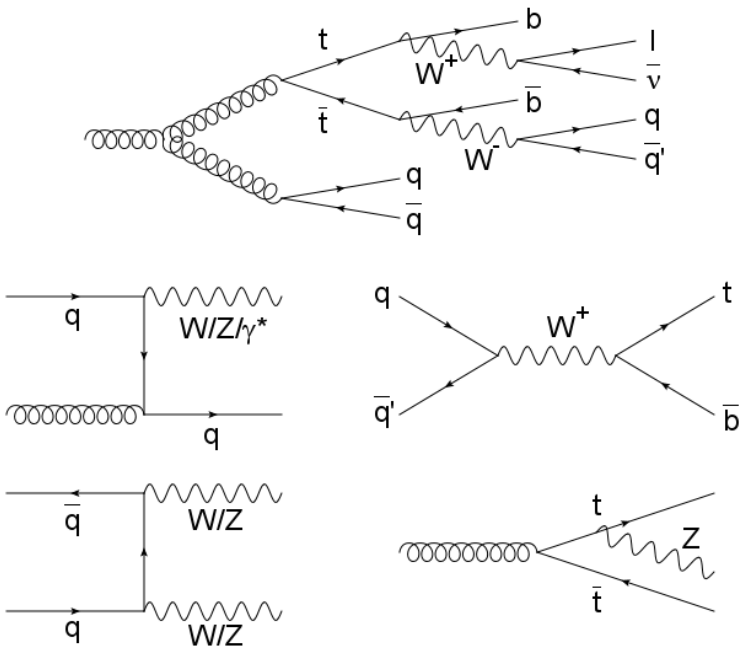


Figure 5.1: Examples of basic Feynman diagrams for the various backgrounds. The bottom four are possible diagrams for the sub-dominant backgrounds. Clockwise from bottom left, they are: Diboson production, V+jets or Drell-Yan+jets, single top production, and $t\bar{t}Z$. The top diagram represents the dominant background, $t\bar{t}+$jets.

as well as real or fake b-jets. However, by far the dominant background ($> 99\%$ of events across the categories) are events involving the decay of the top quark. Because $t\bar{t}H$ contains top quarks, these unavoidably enter our initial selection, and are addressed by the signal extraction methods later in the analysis.

Separation of $t\bar{t}$+jets by Jet Flavor

This analysis treats the $t\bar{t}+$jets background differently from the other backgrounds. The extra jets associated with the $t\bar{t}$ pair (but not originating from a top quark decay) may be composed of a variety of quark flavors, and there are significantly different uncertainties on the production of additional light-flavor (LF) jets as opposed to heavy-flavor (HF) jets. Therefore, we separate the $t\bar{t}+$jets sample into subsamples based on the quark flavor associated with the reconstructed jets in the event. Using the MC truth information, events

containing at least two reconstructed jets matched to extra b quarks are labeled as $t\bar{t} + b\bar{b}$ events. If only a single jet is matched to a b quark, the event is labeled as $t\bar{t} + b$. These cases typically occur because the second extra b quark in the event is either too far forward ($|\eta| > 2.4$) or too soft (p_{T} below the cutoff) to be reconstructed as a jet, or the two extra b quarks have merged into a single jet. If at least one reconstructed jet is matched to a c quark, the event is labeled as $t\bar{t} + c\bar{c}$. All remaining $t\bar{t}$ + jets events are labelled $t\bar{t} + LF$. In all, there are four separate sub-designations of $t\bar{t}$ + jets events: $t\bar{t} + b\bar{b}$, $t\bar{t} + b$, $t\bar{t} + c\bar{c}$ and $t\bar{t} + LF$. The relative contributions from the subsets after selections is reflected in table 5.3.

5.3 Corrections

In addition to overall normalization by cross-section, several other weights are applied to simulated events to improve their modelling of the data.

5.3.1 Lepton Trigger, Isolation and ID Efficiencies

We apply p_{T}- and η- dependant weights to single-lepton events in the MC passing the trigger, isolation and ID requirements. These event weights are called lepton scale-factors. They have been measured by the Muon Physics Object Group (POG) at CMS for the IsoMu24 trigger efficiency, and for the single tight muon isolation and ID efficiencies [28]. We perform our own measurement for electrons following the same procedure, using a tag-and-probe method in a Z-boson-enriched sample [20]. The product of the lepton trigger, isolation and ID scale factors is shown in figure 5.2, for both single muon and single electron events.

5.3.2 PU Reweighting

When the MC was generated, the number of interactions per bunch crossing that would occur during data-taking at 8 TeV was not known, so a distribution was used that roughly covered but did not exactly match the observed conditions. Therefore, we reweight the MC events based on the number of generated primary vertices to match the luminosity profile of the observed pp collisions. We do not match to the number of reconstructed vertices

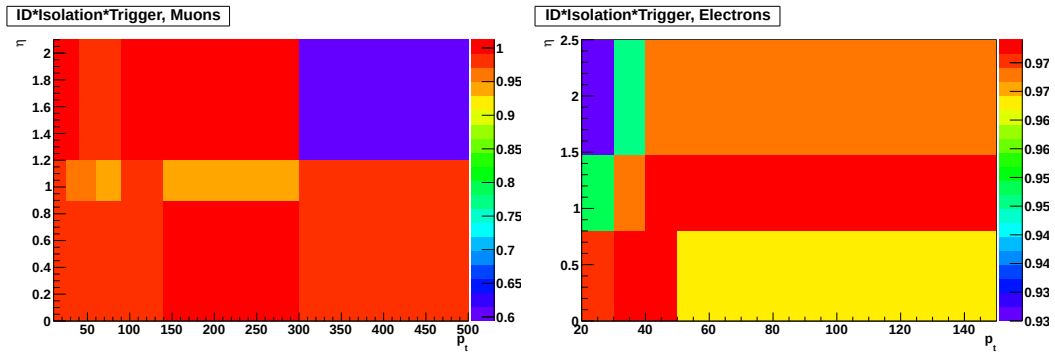


Figure 5.2: Left (right): combined muon (electron) ID, isolation selection and trigger efficiency scale factors in bins of p_T and η.

because of differences in the underlying event in data vs. MC, as well as biases that may be introduced during event selection. Instead, the distribution of primary vertex multiplicity in the data is determined by multiplying the per-bunch-crossing instantaneous luminosity for a given lumi section with the total pp inelastic cross section, and weighting the result by the per-bunch-crossing-per-lumi section integrated luminosity. This is done for all the lumi-sections to obtain the pileup distribution. We obtain the scale factors by normalizing the respective primary vertex (PV) distributions in data and MC, and dividing the data distribution by the MC. The effect of the reweighting by PV is shown in figure 5.3. A value of 69.4 mb is used for the inelastic cross-section.

5.3.3 JE Correction

We apply a series of standard corrections [53] to improve how well the energy of reconstructed jets matches the energy of the particle that produced the jet, and how well the measurement of jets agrees between data and MC. The jet energy scale (JES) calibration applies an absolute correction to the jet energy as a function of p_{T} and η. The JES correction is applied separately for data and MC. Additionally, the energy resolution of jets in MC is calculated by comparing reconstructed jets to generator-level particles, and is corrected to match the

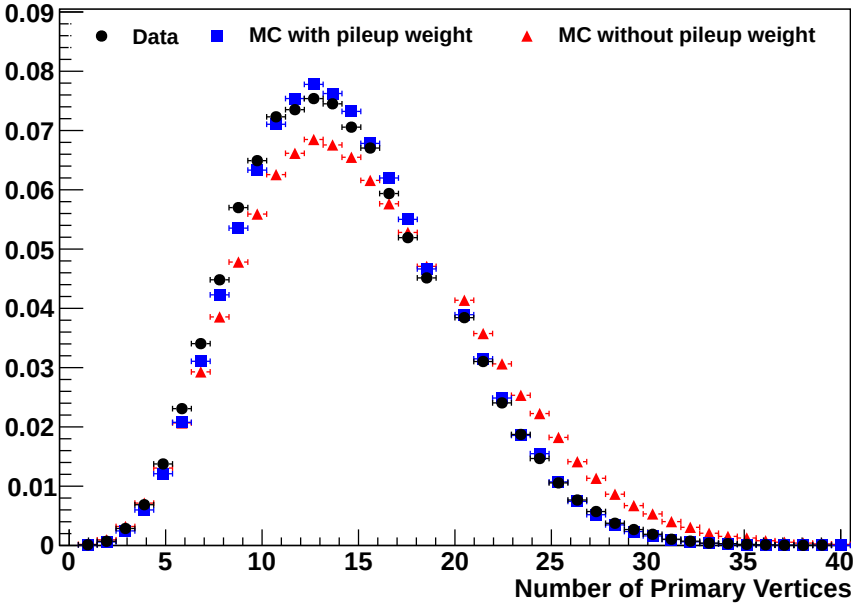


Figure 5.3: Comparison of number of reconstructed vertices for data (black) and the sum of all background MC samples before (red) and after (blue) pileup reweighting. After pileup reweighting, the MC agrees well with the data.

resolution measured in γ+jets events in data [53]. For a given jet, the JER correction is:

$$p_T^{'} = max[0, p_T^{gen} + c(p_T^{reco} - p_T^{gen})], \tag{5.2}$$

where $p_T^{'}$ is the corrected p_{T} and c is the correction factor, which is a function of η.

5.3.4 Top-p_T Reweighting

We performed studies comparing Monte Carlo p_{T} distributions to p_{T} distributions in data, and noticed that the p_{T} spectra of our jets were not well modeled, as shown in figure 5.5. The problem was diagnosed by the CMS top group as being due to mismodeling of the top quark p_{T}, and that top quarks in our MC were given a systematically higher p_{T} than was measured by top differential cross section measurements in 8 TeV data [22, 23]. We derived scale factors to correct the top quark p_{T} distributions in our MADGRAPH samples to match the data [22, 23], and fit the scale factors to a second order polynomial to get a

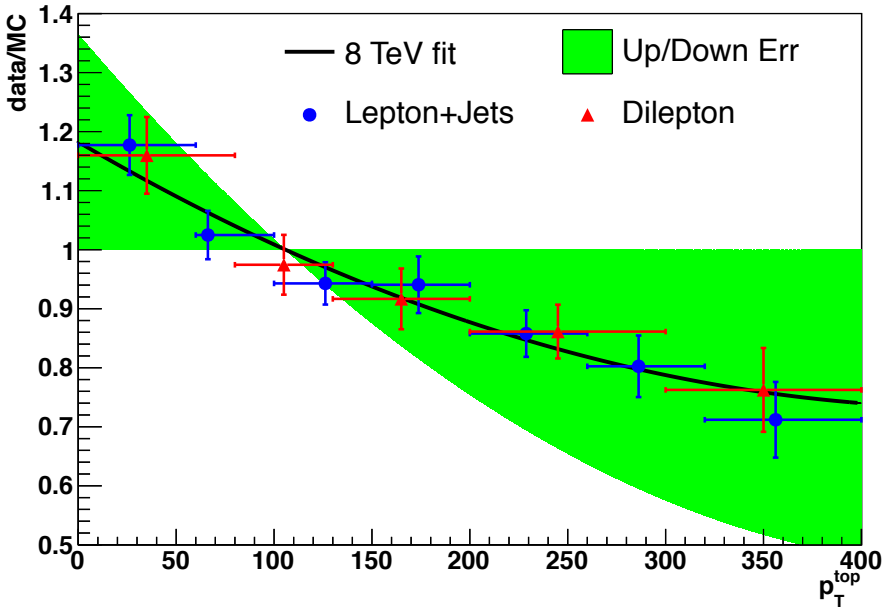


Figure 5.4: Top p_{T} reweighting function fit to the SFs from the CMS top POG.

continuous distribution. The scale factors, along with the fitted polynomial are shown in fig. 5.4. The fitted function is:

$$SF = 1.18246 + 2.10061 \times 10^{-6} \, (p_{\mathrm{T}} - 2 \cdot 463.312) \, p_{\mathrm{T}}. \tag{5.3}$$

For $p_{\mathrm{T}} > 463.312\,\mathrm{GeV}$, a constant scale factor of 0.732 is used. We apply this reweighting function to all top quark MC samples.

5.3.5 CSV reweighting

The CSV tagging algorithm plays a dual role in the analysis. We place a cut on the CSV output that determines which jets are b-tagged, and therefore which events fall into given jet-tag categories. We also utilize the shape of the output distribution of the CSV discriminant as an input to the $t\bar{t}H$ signal extraction BDTs. Therefore, we must ensure that the efficiency for tagging various jet flavors agrees between data and simulation, and we must also simultaneously correct for any disagreement in the shape of the CSV distribution

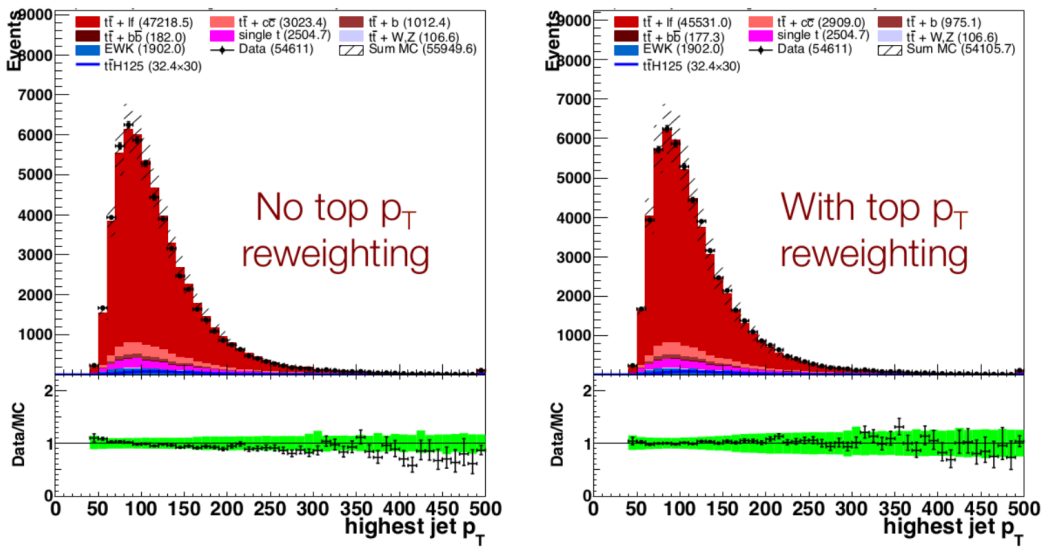


Figure 5.5: The p_{T} of the highest-p_{T} jet in a ==4 jet, ==2 tag control region, before (left) and after (right) the top-p_{T} reweighting procedure, for sum background MC and $t\bar{t}H$ signal. Note the presence of a systematic data/background MC disagreement in the left plot, that is eliminated in the right plot.

between data and MC. In order to accomplish this, we could not follow the standard methods[21, 24, 12] for applying b-tag scale factors, but designed our own method specific to the $t\bar{t}H$ analysis.

The method is outlined briefly here, and is described in more detail elsewhere [25]. The approach was developed by members of the $t\bar{t}H$ group working on a related $t\bar{t}$ ($H \to bb$) analysis, where the search is done in the $t\bar{t}H \to b\bar{b}b\bar{b}l\nu l\nu$ channel. This dilepton analysis is complimentary to and closely coordinated with the work of this dissertation. The method utilizes a tag-and-probe approach to isolate separate high-purity samples of heavy-flavor (HF) and light-flavor (LF) jets. Control samples are obtained from the data consisting of events which contain exactly two jets, and are derived from the full 8 TeV DoubleMu, DoubleElectron and MuEG datasets. The scale factors for light-flavor (LF) and heavy-flavor (HF) jets are separately determined. Requirements are placed on the lepton pair, MET, and the tag jet to select a sub-sample of the data that is either enriched in $t\bar{t}$ (for the HF scale factor) or $Z + jets$ (for the LF scale factor). For the HF-enriched sample, the CSV

distribution of the probe jet is compared to a similarly-obtained distribution from (primarily) $t\bar{t}$ + jets MC, and for the LF-enriched sample, the CSV distribution of the probe jet is compared to $Z + jets$-dominated MC. The $t\bar{t}$ + jets and $Z + jets$ events account for more than 90% of events in the HF and LF samples, respectively [25].

To determine the HF scale factor (SF), the total HF MC is first normalized to the HF-enriched data. The MC is then divided into a sample where the probe jet is actually a b-jet (from MC truth), and a sample where the probe jet is not a b-jet. The contribution from the sample containing the non-b probe is then subtracted from the data, to reduce LF contamination. The distribution of the remaining data is then compared to the MC containing the real b-jet probe, which determines the SF for a given bin in the CSV distribution. The HF SF is also determined as a function of binned p_{T} regions. Finally, a polynomial is fit to the bin-by-bin SF to obtain a smooth SF as a function of CSV. This is done separately for the different p_{T} regions. Figure 5.6 shows CSV distributions for data and MC at various stages in the HF SF determination for a given p_{T} region, as well as the final fitted polynomial [25].

The LF SF is determined in a similar manner. The LF MC is normalized to the LF-enriched data. The MC is divided into a real LF component and a non-LF component. Here, c-jets are grouped with the non-LF component. The non-LF component is subtracted from the data and the scale factor is the ratio of data/MC for each of the CSV bins. The LF SF is determined as a function of bins in η as well as p_{T}. Here as well, polynomials are fit to the bin-by-bin LF SFs, which is done separately for the p_{T} and η regions [25].

In the case of c-jets, we apply a flat SF of unity. The convention is to use the HF SF for c-jets, but with twice the uncertainty[12]; however, we found this to be inadequate for our needs. We need to correct the shape of the CSV distribution, but the CSV output shapes for b-jets in data is quite different from what the MC predicts for c-jets (as can be seen in figure 4.5). If we were to apply our derived HF SFs to charm jets, the tagging rate for these jets would change by a large amount, and the CSV distributions for c-jets would be affected in a non-negligible way. In the absence of a data-driven calibration sample for charm jets, the most reasonable solution is to set the SF==1, and retain the relative uncertainty from

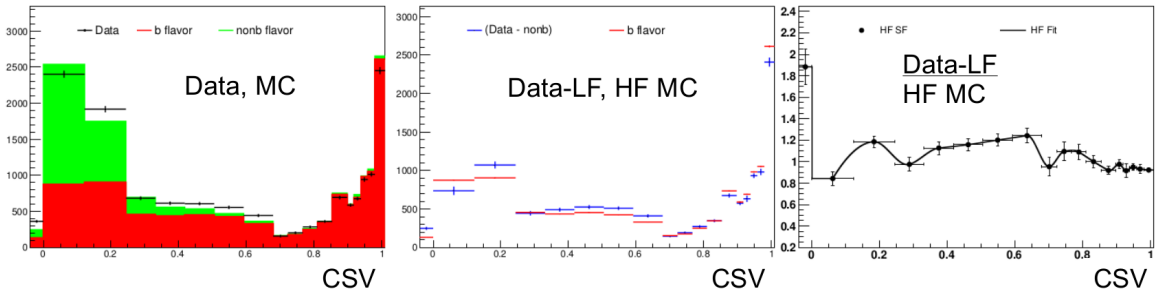


Figure 5.6: Various stages in the process of determining the HF CSV scale factor, in the p_{T} region $40\,\mathrm{GeV} \leq p_{\mathrm{T}} < 60\,\mathrm{GeV}$. Left: Data/MC comparison of initial, HF-enriched distributions. Center: CSV distributions after LF contamination subtracted from data and removed from MC. Right: final SFs with polynomial fit. [25]

the b-jet calibration.

To apply the CSV SFs in this analysis, we loop through each jet in each MC event, and assign a HF SF in the case of a real b-jet, and a LF SF in the case of a LF-jet, and a SF of unity in the case of a c-jet. The total weight for each event is the product of the SFs of the individual jets. The uncertainties associated with the CSV reweighting are nontrivial, and are described in chapter 7.

5.4 Data-MC Comparison

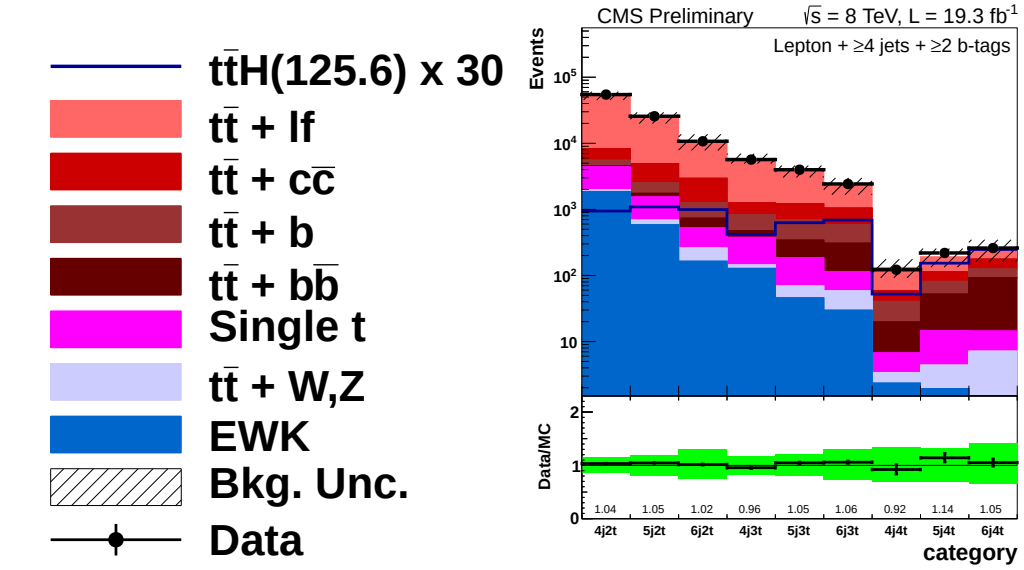
After all the selection criteria and corrections have been applied, we may examine the resulting data, predicted signal and predicted background(s). Table 5.3 shows the predicted event yields for each background in each category, as well as the total background prediction in each category. The number of signal events corresponding to the $m_H = 125.6\,\mathrm{GeV}$ hypothesis is also shown. In addition to the corrections described above, each of the MC samples is normalized by its cross section multiplied by a luminosity of $19.3\,\mathrm{fb}^{-1}$, as in eq. 5.1, corresponding to the total integrated luminosity of the data. Across the categories, the background is dominated by $t\bar{t}+\mathrm{jets}$, with the relative contributions of the different jet flavors varying from category to category. In the $\geq$6 jets, $\geq$4 b-tags category, the ratio of total background events to expected $t\bar{t}H$ events is about 30; this is a significant enhancement in purity compared to the initial ratio of the semi-leptonic $t\bar{t}+\mathrm{jets}$ cross-section to $t\bar{t}H$

	≥6 jets 2 b-tags	4 jets 3 b-tags	5 jets 3 b-tags	≥6 jets 3 b-tags	4 jets 4 b-tags	5 jets ≥4 b-tags	≥6 jets ≥4 b-tags
$t\bar{t}H(125.6)$	31.3 ± 6.4	12.9 ± 2.4	19.4 ± 3.5	21.3 ± 4.4	1.6 ± 0.4	4.8 ± 1.1	7.7 ± 2.0
$t\bar{t}$+lf	7600 ± 1970	4670 ± 790	2590 ± 510	1250 ± 340	73 ± 29	79 ± 33	71 ± 35
$t\bar{t}$+b	520 ± 300	350 ± 190	360 ± 200	280 ± 160	21 ± 12	29 ± 16	33 ± 20
$t\bar{t}+b\bar{b}$	210 ± 120	98 ± 52	157 ± 84	200 ± 110	13.0 ± 7.3	37 ± 21	78 ± 47
$t\bar{t}+c\bar{c}$	1700 ± 1100	430 ± 230	520 ± 280	470 ± 280	19 ± 10	32 ± 18	51 ± 31
$t\bar{t}$+W/Z	98 ± 25	16.1 ± 3.3	23.7 ± 5.0	28.6 ± 6.6	1.1 ± 0.3	2.5 ± 0.7	5.8 ± 1.6
Single t	262 ± 53	233 ± 40	115 ± 21	55 ± 13	3.3 ± 1.6	10.2 ± 5.2	7.3 ± 3.0
W/Z+jets	160 ± 100	122 ± 94	43 ± 38	29 ± 26	2.1 ± 2.4	1.9 ± 1.7	1.2 ± 1.3
Diboson	5.9 ± 1.6	6.3 ± 1.3	2.4 ± 0.7	1.0 ± 0.4	0.3 ± 0.2	0.1 ± 0.1	0.2 ± 0.1
Total bkg	10550 ± 2740	5930 ± 1030	3810 ± 770	2290 ± 600	132 ± 43	192 ± 61	247 ± 88
Data	10724	5667	3983	2426	122	219	260

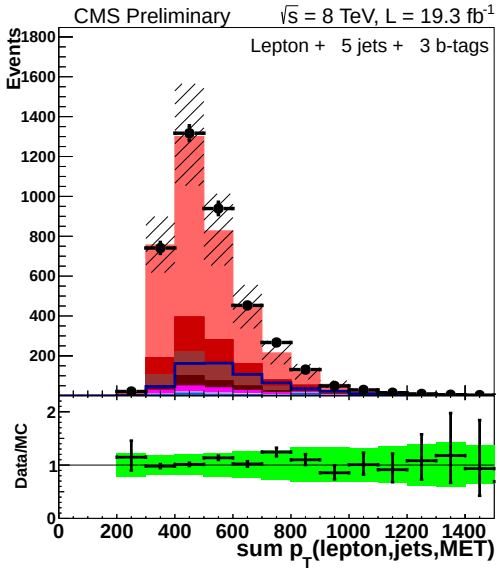
Table 5.3: Predicted signal and background event yields in each of the jet-tag categories, after all event selection criteria and corrections to MC have been applied. The number of observed data events are also shown. The errors on the predicted signal and background are the combined statistical and systematic uncertainties, which will be discussed in chapter 7.

cross-section of about 1000 before event selection and categorization. In each individual category, the total number of background events agrees with the number of observed events, within uncertainty.

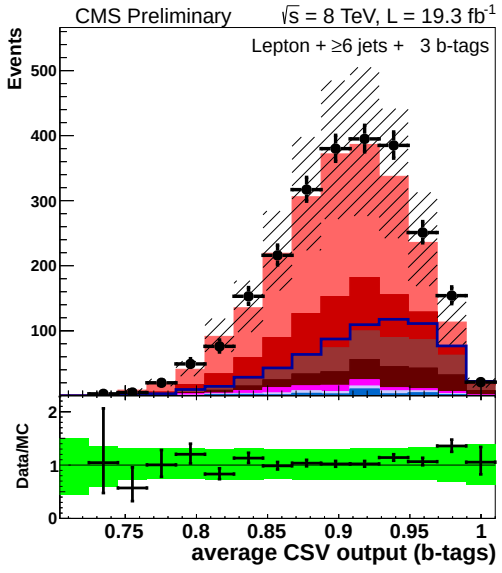
Figure 5.7a shows a graphical representation of the data-to-background MC agreement for each of the categories. Figure 5.7b demonstrates that the CSV reweighting is working as expected, and figure 5.7c shows that there is good agreement between the data and MC p_{T} distributions after the application of the top-p_{T} reweighting and other corrections. As we discuss in the next chapter, we use a variety of event variables to train boosted decision trees to set limits on $t\bar{t}H$ production. Before proceeding with the BDT training, we checked the data/MC agreement of each of the these variables to ensure that they are all well-modeled. Data/MC comparison plots for all the variables used in the analysis are given in appendix A.



(a) Category yields.



(b) Sum p_{T}.



(c) Avg CSV (b-jets).

Figure 5.7: Plots showing overall data/MC agreement per category, data/MC in p_{T} and in the distribution of the CSV output.

Chapter 6
Signal Extraction

Once we have obtained an estimate of the relative contributions of the signal and background events to the different jet-tag categories of the analysis, we may calculate our expected sensitivity to the signal. The simplest way of performing this calculation is by counting events in the different categories, and comparing the number of predicted background events to the number of events predicted by a signal + background hypothesis. As was shown in the previous chapter, the total number of expected background events is quite high compared to the expected number of $t\bar{t}H$ events, even in the most sensitive jet-tag category. If a simple counting experiment were done using only the $\geq$6 jets + $\geq$4 b-tags category, we would not be sensitive to $t\bar{t}H$ production at the 2σ level unless approximately 20 times the expected number of signal events were present in the data, due to the large systematic uncertainty on the background.

In order to obtain competitive results, we take a more sophisticated approach to signal extraction. Sensitivity is first improved by performing a coordinated calculation across all the jet-tag categories. A further improvement is obtained by basing the calculation on a fit to the shapes of discriminating event variables, instead of simply counting events. Variables are identified whose shapes differ between $t\bar{t}H$ and background monte carlo events. Using a given variable, a comparison may be made between the shape of the data, the shape of the background-only prediction, and the shape of the signal + background prediction. The more the variable is able to discriminate between the shapes of the combined background and the $t\bar{t}H$ signal, the more our expected sensitivity to the $t\bar{t}H$ process is improved. To optimize

sensitivity as much as possible, we train boosted decision trees (BDTs)[41] to combine the shape information of multiple event variables into a single overall discriminant for each of the categories. The limits on Higgs production are then calculated using the shapes of these final discriminants.

6.1 Discriminating Variables

The variables used in the analysis are listed in table 6.1. They all individually provide some discriminating power between $t\bar{t}H$ and the dominant $t\bar{t}+$jets background, and are well-modelled (as discussed in chapter 5). Many of these quantities describe event kinematics, and include the p_{T} and energy of different objects, as well as the invariant mass of object combinations. There are a number of variables involving angular separation between specific objects; also, the sphericity and aplanarity variables, together with the Fox-Wolfram moments H_i [34] serve to give an overall picture of angular and p_{T} distributions of objects in the event. We also use the output of the CSV b-tagging discriminant for jets that have been positively identified as b-jets, as well as for jets that fall below the CSV medium working point.

The calculation of the variable "best Higgs boson mass" merits some additional description. It is used in the $\geq$5 jets and $\geq$4 b-tags, and $\geq$6 jets and $\geq$3 b-tags, and $\geq$6 jets and $\geq$4 b-tags categories, where it is possible to fully reconstruct the $t\bar{t}H$ decay. In the case of events with $\geq$5 jets and $\geq$4 b-tags, the highest-p_{T} "loose" jet is added to the rest of the jet collection for the purposes of calculating this variable. This loose jet has the same requirements as the nominal jets as given in chapter 4, except that its p_{T} must fall within 20 GeV $\leq p_{\mathrm{T}} <$ 30 GeV. In the case of events with $\geq$6 jets and 3 b-tags, the non-tagged jet with the CSV discriminant value closest to 0.679 is promoted to a b-tag. We then iterate through all combinations of final state objects and use them to reconstruct two top quark candidates and a Higgs boson candidate. One top quark is constructed from two untagged jets and a b-tagged jet, and the other is made from the lepton, neutrino, and a b-tagged jet. The Higgs is constructed from two b-tagged jets. We require that the lepton and MET together form an on-shell W with a mass of 80 GeV, and we use that constraint to calculate

the neutrino p_z. We iterate over all possible combinations of the various object assignments and calculate a χ^2 value for each combination. The χ^2 value for a given iteration is the following quantity:

$$\chi^2 = \left(\frac{m_t - m_{hadtop}}{\sigma_{hadtop}}\right)^2 + \left(\frac{m_t - m_{leptop}}{\sigma_{leptop}}\right)^2 + \left(\frac{m_W - m_{hadW}}{\sigma_{hadW}}\right)^2 \tag{6.1}$$

where m_{hadtop}, m_{leptop} and m_{hadW} are the reconstructed masses; $m_t = 172.5\,\text{GeV}$ and $m_W = 80.4\,\text{GeV}$; and the widths σ have been estimated by plotting the invariant masses of the correct combinations of objects. The combination that gives the lowest χ^2 is used as the reconstruction of the final state. The "best Higgs boson mass" is the invariant mass of the two highest-p_T b-tagged jets that are *not* assigned to top quarks in the lowest χ^2 combination. The correct combination is found roughly 30% of the time.

6.1.1 Selection of Variables by Category

The variables listed above are the starting point for the formation of the discriminants used in the limit calculation. For each of the BDTs, we identify a subset of the variables to be used as inputs in the training procedure; it is not possible to use the entire list to train each BDT due to the finite amount of available monte carlo statistics. The subset is selected by measuring the "separation" between signal and background for each variable, where the separation $\langle S^2 \rangle$ is defined as [41]:

$$\langle S^2 \rangle = \frac{1}{2} \int \frac{\left(\hat{y}_S(y) - \hat{y}_B(y)\right)^2}{\hat{y}_S(y) + \hat{y}_B(y)} dy, \tag{6.2}$$

where y is the input variable, and $\hat{y}_S$ and $\hat{y}_B$ are the signal and background probability density functions for that input variable in the signal and background samples, respectively. For a given BDT in a given category, the entire list of variables is ranked based on separation, and the most highly-separated variables are chosen to train the BDT. Figure 6.1 shows examples of some of the most highly-separated variables, and tables 6.2 and 6.3 list the variables used as inputs for each of the BDTs.

More information on the BDT training is given below, but it is important to note here that the one-dimensional separation of each of the input variables makes up only part of the

Variable	Description
abs $\Delta\eta$ (leptonic top, bb)	Delta-R between the leptonic top reconstructed by the best Higgs mass algorithm and the b-jet pair chosen by the algorithm
abs $\Delta\eta$ (hadronic top, bb)	Delta-R between the hadronic top reconstructed by the best Higgs mass algorithm and the b-jet pair chosen by the algorithm
aplanarity	Event shape variable equal to $\frac{3}{2}(\lambda_3)$, where λ_3 is the third eigenvalue of the sphericity tensor as described in [6].
ave CSV (tags/non-tags)	Average b-tag discriminant value for b-tagged/non-b-tagged jets
ave ΔR(tag,tag)	Average ΔR between b-tagged jets
ave mass(untag,untag)	Average of the invariant mass of all pairs of jets that are not b-tagged
ave mass(tag,tag)	Average of the invariant mass of all pairs of jets that are b-tagged
best Higgs boson mass	A minimum-chi-squared fit to event kinematics is used to select two b-tagged jets as top-decay products. Of the remaining b-tags, the invariant mass of the two with highest E_t is saved.
best ΔR(b,b)	The ΔR between the two b-jets chosen by the best Higgs boson mass algorithm
closest tagged dijet mass	The invariant mass of the two b-tagged jets that are closest in ΔR
dev from ave CSV (tags)	The square of the difference between the b-tag discriminant value of a given b-tagged jet and the average b-tag discriminant value among b-tagged jets, summed over all b-tagged jets
highest CSV (tags)	Highest b-tag discriminant value among b-tagged jets
H_0, H_1, H_2, H_3	The first few Fox-Wolfram moments [34] (event shape variables)
HT	Scalar sum of transverse momentum for all jets with $p_T > 30$ GeV/c
$\sum p_T$(jets,leptons,MET)	The sum of the p_T of all jets, leptons, and MET
$\sum p_T$(jets,leptons)	The sum of the p_T of all jets, leptons
jet 1, 2, 3, 4 p_T	The transverse momentum of a given jet, where the jet numbers correspond to rank by p_T
lepton p_T	The transverse momentum of the lepton (LJ channel)
lowest CSV (tags)	Lowest b-tag discriminant value among b-tagged jets
mass(lepton,jet,MET)	The invariant mass of the 4-vector sum of all jets, leptons, and MET
mass(lepton,closest tag)	The invariant mass of the lepton and the closest b-tagged jet in ΔR (LJ channel)
max $\Delta\eta$ (jet, ave jet η)	max difference between jet eta and avg deta between jets
max $\Delta\eta$ (tag, ave jet η)	max difference between tag eta and avg deta between jets
max $\Delta\eta$ (tag, ave tag η)	max difference between tag eta and avg deta between tags
median inv. mass (tag pairs)	median invariant mass of all combinations of b-tag pairs
M3	The invariant mass of the 3-jet system with the largest transverse momentum.
MHT	Vector sum of transverse momentum for all jets with $p_T > 30$ GeV/c
MET	Missing transverse energy
min ΔR(lepton,jet)	The ΔR between the lepton and the closest jet (LJ channel)
min ΔR(tag,tag)	The ΔR between the two closest b-tagged jets
min ΔR(jet,jet)	The ΔR between the two closest jets
$\sqrt{\Delta\eta(t^{lep}, bb) \times \Delta\eta(t^{had}, bb)}$	square root of the product of abs $\Delta\eta$ (leptonic top, bb) and abs $\Delta\eta$ (hadronic top, bb)
second-highest CSV (tags)	Second-highest b-tag discriminant value among b-tagged jets
sphericity	Event shape variable equal to $\frac{3}{2}(\lambda_2 + \lambda_3)$, where λ_2 and λ_3 are the second and third eigenvalues of the sphericity tensor as described in [6]
(Σ jet p_T)/(Σ jet E)	The ratio of the sum of the transverse momentum of all jets and the sum of the energy of all jets
tagged dijet mass closest to 125	The invariant mass of the b-tagged pair closest to 125 GeV/c^2
$t\bar{t}b\bar{b}$/$t\bar{t}H$ BDT	BDT used to discriminate between $t\bar{t}b\bar{b}$ and $t\bar{t}H$ in the LJ $\geq$ 6 jets, $\geq$ 4 tags, $\geq$6 jets + 3 tags, and 5 jets + $\geq$4 tags categories. See text for description.

Table 6.1: Event variables used in the boosted decision trees and their descriptions.

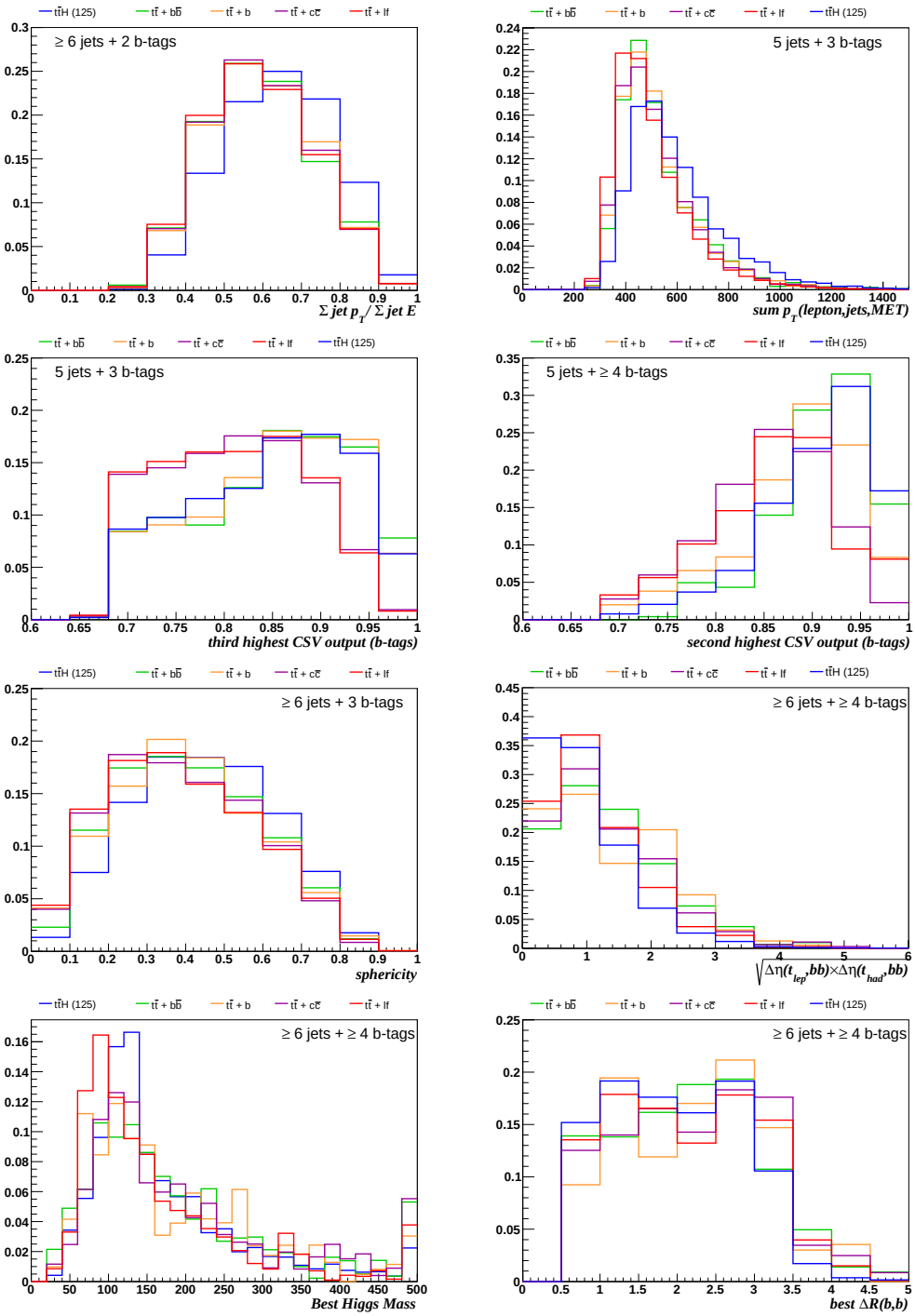


Figure 6.1: Examples of discriminating variables for a variety of categories. The plots are normalized to the number of entries to show the difference in shape between $t\bar{t}H$ and the different $t\bar{t}+$jets components. Several types of variables are shown, including shape, kinematic and CSV variables. Some variables, such as the "third-highest CSV output," are mainly good at separating LF from HF jets, while others, such as "$\Sigma jetp_T/\Sigma jetE$," more uniformly separate the $t\bar{t}+$jets components from $t\bar{t}H$.

information utilized by each BDT. The BDTs also exploit correlations between the variables during the training process. However, it is difficult to incorporate correlations into the criteria for the initial selection of variables. It is not known beforehand precisely how the correlations will be used by the BDT in conjunction with the one-dimensional separating power of each variable. Two-dimensional correlation plots were extensively studied to look for highly-correlated pairs of variables that could provide obvious additional discriminating power; no such variables were found that were not already included by the selection by initial separation. BDTs are sensitive to correlations in more than two dimensions, but these are harder to visualize and evaluate, especially nonlinear multidimensional correlations. In the end, many different configurations of variables were tested. Alternative methods of selecting the variables were investigated, but were found to offer no significant increase in performance. We found that the ranking by separation, although probably not strictly ideal, nevertheless constitutes a well-understood, reproducible procedure for selecting the variables, and results in reliably good BDT performance.

6.2 BDT Configuration

The variables described above are used as inputs to a set of multivariate, machine-learning based classifiers called boosted decision trees (BDTs). The basic structure of a decision tree is illustrated in figure 6.2. The first (or root) node of each tree uses one of the input variables to define an initial cut that splits the training sample; on each side of the cut, another cut is made that attempts to further divide the samples in a way that increases the purity of the signal and/or background. The cuts continue down the tree until a specified depth is reached. The resulting structure is a multidimensional space of rectangular cuts which classifies each event as either signal or background. Typically, a "forest" of such trees is "grown." The growth of the forest is influenced by the boosting algorithm, which reweights events as each successive tree is grown, giving higher priority to events that were misclassified by the previous tree. Finally, the responses of the individual trees are linearly combined to form the output of the entire BDT, where the weights of the linear combination

	4 jets, 3 tags	4 jets, 4 tags
	jet 1 p_T	jet 1 p_T
	jet 2 p_T	jet 2 p_T
	jet 3 p_T	jet 4 p_T
	jet 4 p_T	HT
	M3	$\sum p_T$(jets,lepton,MET)
	$\sum p_T$(jets,lepton,MET)	M3
	HT	ave CSV (tags)
	lowest CSV (tags)	second-highest CSV (tags)
	MHT	third-highest CSV (tags)
	MET	lowest CSV (tags)
	5 jets, 3 tags	5 jets, $\geq$ 4 tags
	jet 1 p_T	max $\Delta\eta$ (tag, ave jet η)
	jet 2 p_T	$\sum p_T$(jets,lepton,MET)
	jet 3 p_T	(Σ jet p_T)/(Σ jet E)
	jet 4 p_T	ave ΔR(tag,tag)
	$\sum p_T$(jets,lepton,MET)	ave CSV (tags)
	(Σ jet p_T)/(Σ jet E)	dev from ave CSV (tags)
	HT	second-highest CSV (tags)
	ave CSV (tags)	third-highest CSV (tags)
	third-highest CSV (tags)	lowest CSV (tags)
	fourth-highest CSV (jets)	ttbb/ttH BDT
$\geq$ 6 jets, 2 tags	$\geq$ 6 jets, 3 tags	$\geq$ 6 jets, $\geq$ 4 tags
$\sum p_T$(jets,lepton,MET)	H_0	(Σ jet p_T)/(Σ jet E)
HT	sphericity	ave ΔR(tag,tag)
mass(lepton,closest tag)	(Σ jet p_T)/(Σ jet E)	$\sqrt{\Delta\eta(t^{lep}, bb) \times \Delta\eta(t^{had}, bb)}$
max $\Delta\eta$ (jet, ave jet η)	max $\Delta\eta$ (jet, ave jet η)	closest tag mass
min ΔR(lepton,jet)	$\sum p_T$(jets,lepton,MET)	max $\Delta\eta$ (tag, ave tag η)
H_2	ave CSV (tags)	ave CSV (tags)
sphericity	second-highest CSV (tags)	third-highest CSV (tags)
(Σ jet p_T)/(Σ jet E)	third-highest CSV (tags)	fourth-highest CSV (tags)
third-highest CSV (jets)	fourth-highest CSV (jets)	best Higgs mass
fourth-highest CSV (jets)	ttbb/ttH BDT	ttbb/ttH BDT

Table 6.2: BDT input variable assignments for the final BDTs in each category. The variables used in each category were selected following the procedure outlined in the text. However, once selected, it is possible to identify trends in the types of variables that offer the best separating power in each of the categories. In general, a mix of different kinematic, shape, and b-tagging variables are used in each BDT, so that a variety of event information is available during training. In the categories that contain greater numbers of jets and tagged jets, more of the CSV variables are used since they offer separating power between events containing a certain number of correctly tagged jets and events containing some jets that were not correctly tagged (either mistagged or incorrectly not tagged). In the categories with lower numbers of jets and tags (such as 4 jets + 3 tags), the best discriminating information is provided by kinematic variables. In categories with the greatest combinatorics, specialized algorithms (such as the "best Higgs mass") and event shape variables are useful in distilling complex event information.

5 jets, $\geq$ 4 tags	$\geq$ 6 jets, 3 tags	$\geq$ 6 jets, $\geq$ 4 tags
ave ΔR(tag,tag)	tagged dijet mass closest to 125	H_3
max $\Delta\eta$ (tag, ave tag η)	(Σ jet p_T)/(Σ jet E)	ave ΔR(tag,tag)
(Σ jet p_T)/(Σ jet E)	$\sqrt{\Delta\eta(t^{lep},bb) \times \Delta\eta(t^{had},bb)}$	closest tagged dijet mass
tagged dijet mass closest to 125	H_1	sphericity
H_1	H_3	max $\Delta\eta$ (tag, ave jet η)
H_3	M3	max $\Delta\eta$ (tag, ave tag η)
$\sum p_T$(jets,lepton,MET)	max $\Delta\eta$ (tag, ave tag η)	mass(lepton,jet,MET)
fourth-highest CSV (tags)	max $\Delta\eta$ (tag, ave jet η)	(Σ jet p_T)/(Σ jet E)
aplanarity	max $\Delta\eta$ (jet, ave jet η)	abs $\Delta\eta$ (leptonic top, bb)
MET	abs $\Delta\eta$ (hadronic top, bb)	abs $\Delta\eta$ (hadronic top, bb)
	abs $\Delta\eta$ (leptonic top, bb)	$\sqrt{\Delta\eta(t^{lep},bb) \times \Delta\eta(t^{had},bb)}$
	sphericity	ave CSV (tags)
	aplanarity	best ΔR(b,b)
	min ΔR(tag,tag)	best Higgs mass
	jet 3 p_T	median inv. mass (tag pairs)

Table 6.3: List of variables used as inputs in each of the $t\bar{t}H/t\bar{t}+b\bar{b}$ BDTs. In contrast to the final BDTs trained in the same categories, b-tagging information does not provide much discriminating power between $t\bar{t}H$ and $t\bar{t}+b\bar{b}$ since both processes nominally contain the same number of real b-jets. Furthermore, $t\bar{t}H$ and $t\bar{t}+b\bar{b}$ are kinematically similar, necessitating the extensive research of event shape variables. In particular, we found differences in the η distributions of objects to be useful, especially the difference in η between the reconstructed tops and the bb-pair assigned to the Higgs by the best Higgs mass algorithm.

are adjusted to optimize performance. For each BDT, we train a boosted forest of 100 trees. The specific BDT method used is the Gradient Boost, and is included as part of the "TMVA" multivariate analysis software package [41]. This algorithm gives similar performance to other boosting algorithms, but is designed to be less susceptible to over-compensation due to noisy events. The boosting procedure works best when the individual trees only weakly classify the events – thus, we limit the trees to a maximum of 5 nodes, and maximum depth of 3 levels (including the root node). We saw that increasing these parameters did not improve performance, and tended to lead to overtraining. The robustness of the boosting procedure was also further enhanced by keeping the learning rate of the algorithm low. The BDTs are trained separately for each of the categories. In the $\geq$ 6 jets + 2 b-tags, 4 jets + 3 b-tags, 5 jets + 3 b-tags, and 4 jets + 4 b-tags categories, a single BDT is trained in each category to discriminate between $t\bar{t}H$ and $t\bar{t}+$jets events. The $\geq$5 jets + $\geq$4 b-tags, $\geq$6 jets + 3 b-tags, and $\geq$6 jets + $\geq$4 b-tags categories each train 2 BDTs: a $t\bar{t}H/t\bar{t}+$jets BDT, as well as an additional BDT specifically trained to discriminate between $t\bar{t}H$ and $t\bar{t}+b\bar{b}$ events.

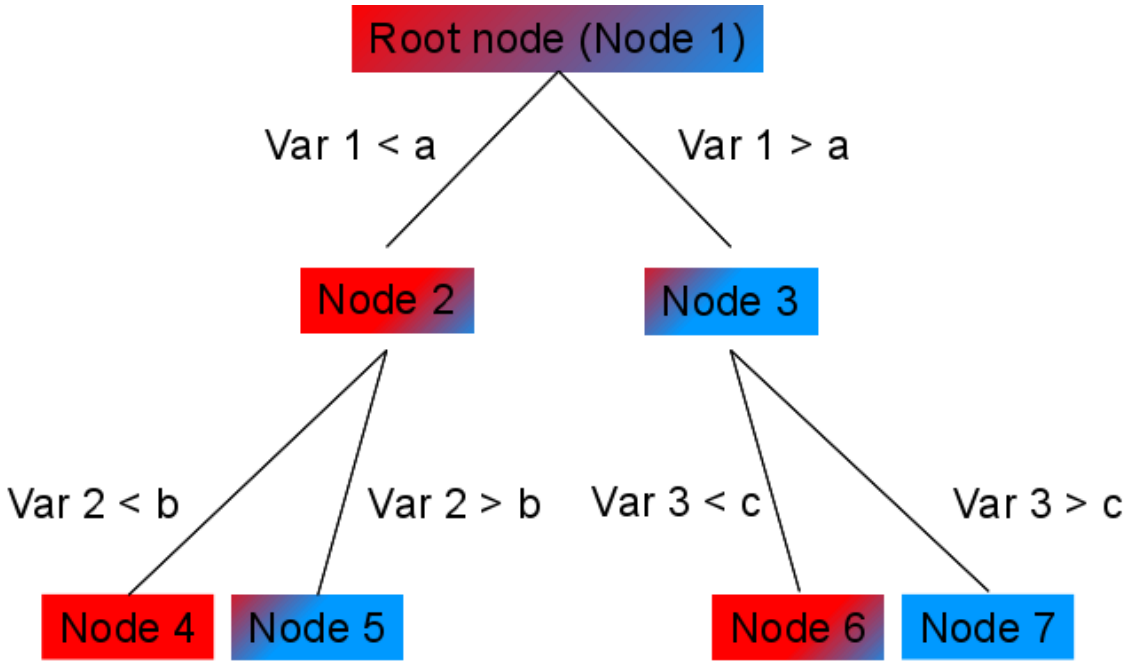


Figure 6.2: Illustration of the structure of a basic decision tree, with 7 nodes in 3 layers. This tree uses a series of cuts on three variables in an attempt to separate the two classes of events, represented by the red and blue shades. The cuts are selected to increase the purity of blue or red events, or both. The initial cut on Variable 1 splits the events into a sample that is more pure for blue events, and another that is more pure for red events. These samples are further divided to increase the purity of the red and blue events, respectively. One can imagine that the cut on Variable 2 was made to optimize the purity of the red events in Node 4, resulting in less blue purity in Node 5. In that case, red events that were misclassified as blue in Node 5 might be "boosted," or given a greater weight when selecting the cuts for the next tree in the forest.

Thus, 10 BDTs in total are used in the analysis: 7 $t\bar{t}H/t\bar{t}+$jets BDTs and 3 $t\bar{t}H/t\bar{t}+b\bar{b}$ BDTs. In the categories that contain both a $t\bar{t}H/t\bar{t}+b\bar{b}$ BDT and a $t\bar{t}H/t\bar{t}+$jets BDT, the output of the $t\bar{t}H/t\bar{t}+b\bar{b}$ BDT is used as one of the input variables of the $t\bar{t}H/t\bar{t}+$jets BDT. In each category (including the 2-BDT categories), only the shape of the final $t\bar{t}H/t\bar{t}+$jets BDT is used in the limit calculation.

6.3 Training Procedure

6.3.1 $t\bar{t}H/t\bar{t}$ BDTs

For each of the final $t\bar{t}H/t\bar{t}+$jets BDTs used in the analysis, the training proceeds as follows. An equal number of $t\bar{t}H$ ($m_H = 125$ GeV) and $t\bar{t}+$jets monte-carlo events are randomly selected, such that the maximum number of available events is used for a given category. These two samples are then each randomly and equally split into one sample that will be used to train the BDT, and another sample that will be used to monitor against overtraining. We do not weight events prior to training. The use of 10 variables for each $t\bar{t}H/t\bar{t}+$jets BDT (as opposed to another number) was optimized to give the best performance with available training statistics. The software trains the BDTs using the boosting procedure described above, and stores them for later use.

At the conclusion of each round of training, we perform a test to determine if a given BDT was overtrained. We use the ROOT implementation of the Kolmogorov-Smirnov test to check for compatibility of the BDT output distributions between the training and test samples; if the test shows a significant disagreement between the samples, then the BDT must be trained again. The Kolmogorov-Smirnov (KS) test is a well-known statistical method for comparing the compatibility of two distributions. It is sensitive to differences in both location and shape of the two distributions being compared. Integrals of the two (normalized) distributions are performed cumulatively from bin to bin, so that the value of the integral at each point of the distribution is determined. The magnitude of the difference between these two integrals at the point where they disagree the most is used to calculate the Kolmogorov test statistic. The ROOT software uses this test statistic to estimate the

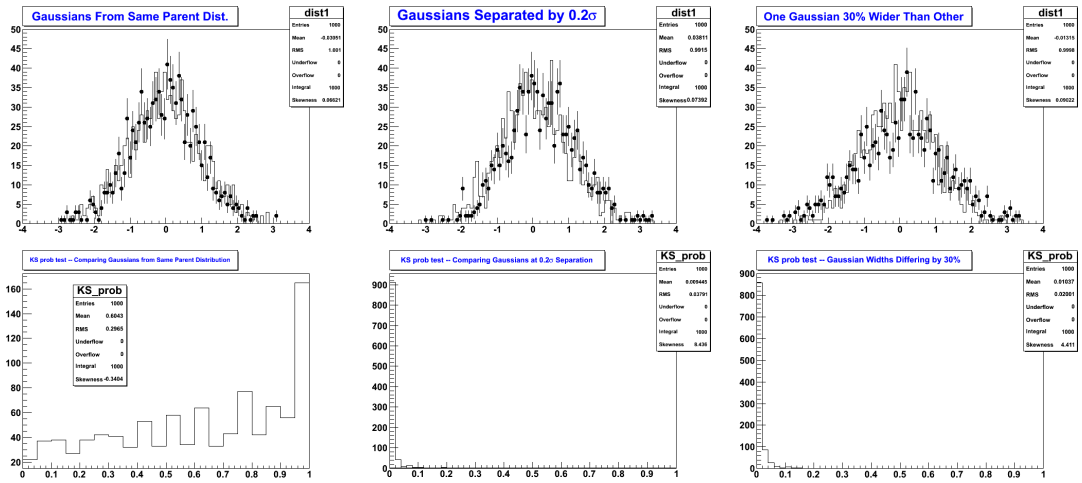


Figure 6.3: Plots showing the range of typical KS-test values in different scenarios. For each scenario, two histograms were each filled with 1000 Gaussian- distributed, random entries (this number was chosen to approximate the statistics used to train the BDTs). A KS test was then performed comparing the two distributions. This was repeated 1000 times, and a histogram was filled with the results of the KS tests. The bottom row shows the histograms filled with the KS-test values, and the top row shows a representative pair of Gaussian distributions from a given trial. The top and bottom plots are grouped together so that each column shows a different scenario. Left: the two Gaussians being compared come from the same parent distribution; center: one Gaussian is displaced by 0.2σ w.r.t. the other; right: one Gaussian is 30% wider than the other. This study demonstrates that the KS test is sensitive to statistically significant differences in both shape and location that may be too small to be visually obvious.

probability that the two samples were derived from the same parent distribution. In the case that the two samples are compatible, the value of the probability is expected to fall uniformly between zero and one[2]. A value close to zero indicates a small probability that the two samples are described by the same distribution. Examples of the different cases are given in figure 6.3. We consider a BDT to have passed the overtraining check if the KS values comparing the training and test samples are above 0.05. This check is done independently for the signal and background distributions.

[2]In reality this is not precisely the case: due to a documented binning effect[30], the KS values are expected to be somewhat no-uniform in the case of identical parent distributions. However, this has no impact on the test's ability to identify incompatible distributions, which is its purpose here.

6.3.2 $t\bar{t}H/t\bar{t}+b\bar{b}$ BDTs

The procedure for training the $t\bar{t}H/t\bar{t}+b\bar{b}$ BDTs is identical to that used to train the $t\bar{t}H/t\bar{t}+$jets BDTs, except that $t\bar{t}+b\bar{b}$ is used as the background instead of inclusive $t\bar{t}+$jets. Equal numbers of $t\bar{t}H$ and $t\bar{t}+b\bar{b}$ events are used in training. The BDT options are the same as described earlier, and the same overtraining checks are carried out. In the $\geq$6 jets + $\geq$4 b-tags and $\geq$6 jets + 3 b-tags categories, 15 variables are used, while in the 5 jets + $\geq$4 b-tags category 10 variables are used due to lower statistics. In categories that use a $t\bar{t}+b\bar{b}$ BDT, the response of the $t\bar{t}+b\bar{b}$ is used as an input variable to the final $t\bar{t}H/t\bar{t}+$jets BDT. It is treated as any other variable in the final BDT, and is selected from the available pool of well-modelled variables using the same ranking-by-separation procedure.

The motivation behind the use of the additional $t\bar{t}H/t\bar{t}+b\bar{b}$ BDTs is illustrated in figure 6.4 and in the left column of figure 6.5. The $t\bar{t}+b\bar{b}$ process is our most difficult background. The $t\bar{t}+b\bar{b}$ final state objects are identical to those expected in $t\bar{t}H$ ($H\rightarrow b\bar{b}$), and the invariant mass distribution of the extra $b\bar{b}$ pair in $t\bar{t}+b\bar{b}$ lies in the same kinematic region as the invariant mass of the $b\bar{b}$ pair from the decay of the Higgs in $t\bar{t}H$. As a result, BDTs that train $t\bar{t}H$ against inclusive $t\bar{t}+$jets tend to focus more on the easier-to-distinguish $t\bar{t}+LF$ and $t\bar{t}+c\bar{c}$ components, at the expense of $t\bar{t}H/t\bar{t}+b\bar{b}$ discrimination. The training of a dedicated $t\bar{t}H/t\bar{t}+b\bar{b}$ BDT has the effect of providing a more level playing field between the different $t\bar{t}+$jets components, and $t\bar{t}H/t\bar{t}+b\bar{b}$ discrimination is improved in the final BDT while maintaining good separation between $t\bar{t}H$ and the lighter $t\bar{t}+$jets components (see the right column of figure 6.5). As will be discussed in the next chapter, the systematic uncertainty on the production of $t\bar{t}$ +HF jets is one of the largest in the analysis. The $\geq$5 jets + $\geq$4 b-tags, $\geq$6 jets + 3 b-tags, and $\geq$6 jets + $\geq$4 b-tags categories are the most sensitive in the analysis, as well as the categories with the highest fraction of $t\bar{t}+b\bar{b}$. The application of the tiered $t\bar{t}H/t\bar{t}+b\bar{b}$ $\rightarrow t\bar{t}H/t\bar{t}+$jets approach in these 3 categories helps to pin down some of the $t\bar{t}$ +HF uncertainty. We found that the use of this system gave us a 20% overall improvement in the expected limit when compared to an analysis that did not use $t\bar{t}H/t\bar{t}+b\bar{b}$ BDTs, but only used a single $t\bar{t}H/t\bar{t}+$jets BDT in each category.

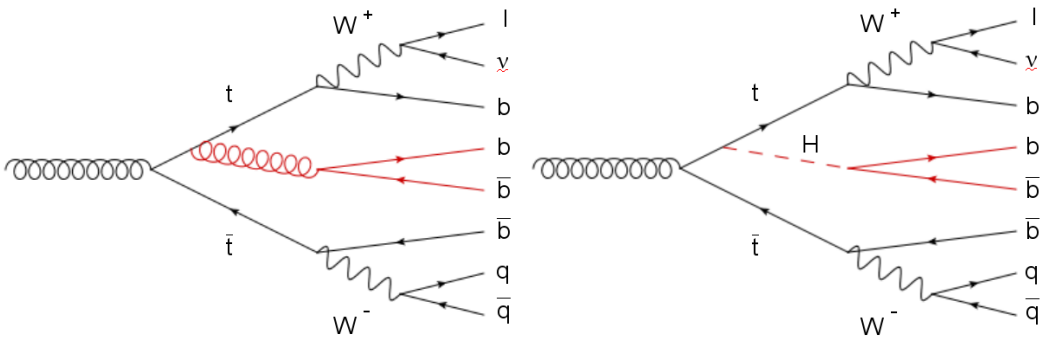


Figure 6.4: Two possible Feynman diagrams for $t\bar{t} + b\bar{b}$ (left) and $t\bar{t}H$ (right), illustrating the similarity between the two processes.

6.4 Validation

As discussed above, overtraining checks are performed for all BDTs. The results are summarized in figures 6.6 and 6.7. We saw no evidence of overtraining in any BDT. In addition, we compared 2-dimensional correlations between the variables in monte-carlo and in the data.

6.5 Data-MC Comparison of BDT Outputs

Figures 6.8 and 6.9 show the data-to-sum-(background) monte-carlo comparison for each of the BDTs.

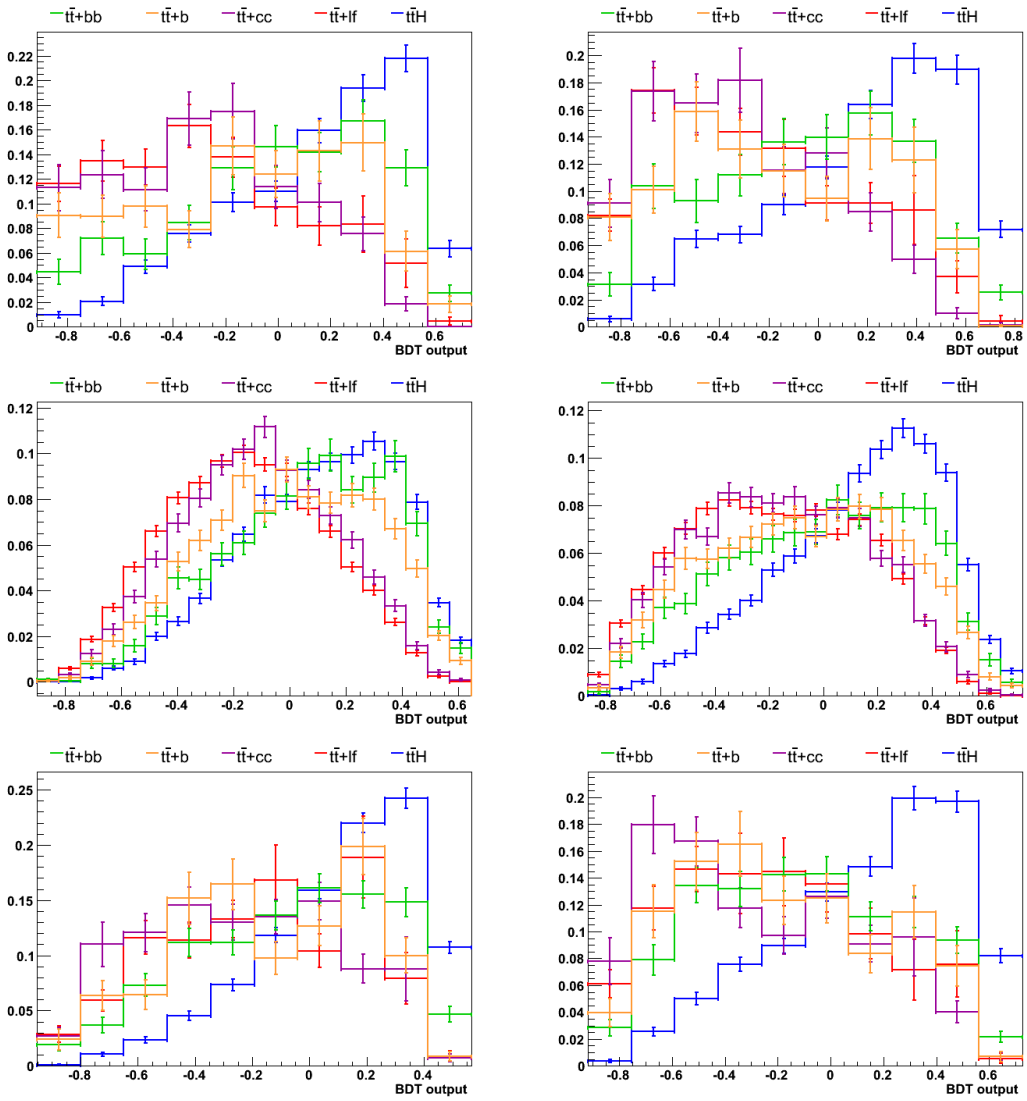


Figure 6.5: Normalized output distributions of the final BDTs for $t\bar{t}H$ and the various flavors of $t\bar{t}$ + jets. Top row: $\geq$5 jets + $\geq$4 b-tags; center row: $\geq$6 jets + 3 b-tags; bottom row: $\geq$6 jets + $\geq$4 b-tags. Left column: BDT trained without use of $t\bar{t}H/t\bar{t}+b\bar{b}$ variable. Right column: BDT trained using $t\bar{t}H/t\bar{t}+b\bar{b}$ variable.

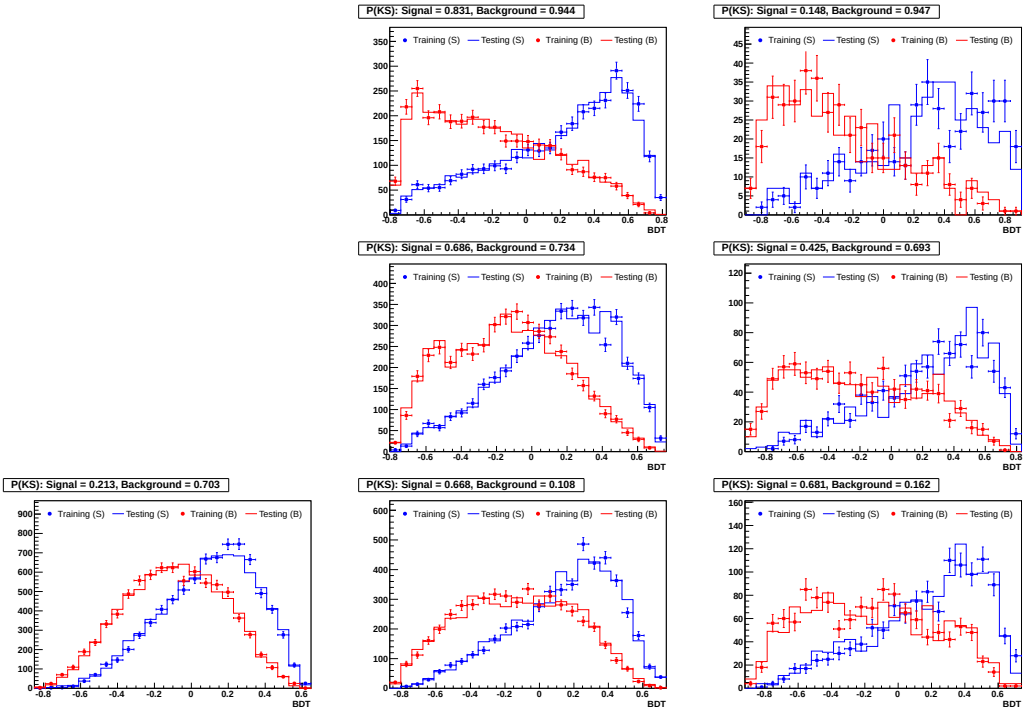


Figure 6.6: Overtraining checks for the final BDTs in each category. The KS test result is shown as a measure of the agreement between training and testing samples. The top, middle and, bottom rows are events with 4, 5, and $\geq$6 jets, respectively, while the left, middle, and right-hand columns are events with 2, 3, and $\geq$4 b-tags, respectively. The background (red) and signal (blue) are shown for the testing (solid line) and training (points) samples.

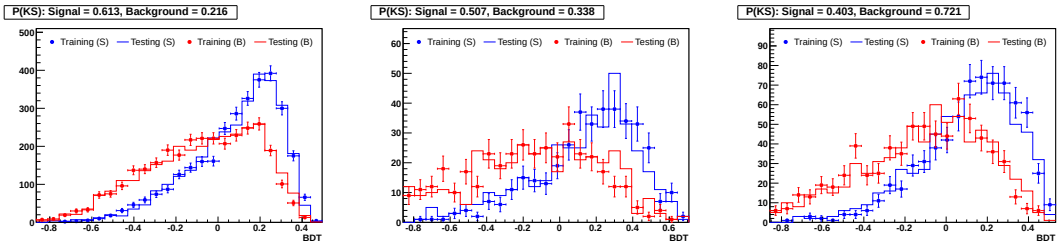


Figure 6.7: Overtraining checks for the $t\bar{t} + b\bar{b}/t\bar{t}H$ BDTs, in the $\geq$6 jets + 3 b-tags (left), 5 jets + $\geq$4 b-tags (center), and $\geq$6 jets + $\geq$4 b-tags (right) categories. The background (red) and signal (blue) are shown for the testing (solid line) and training (points with errors) samples.

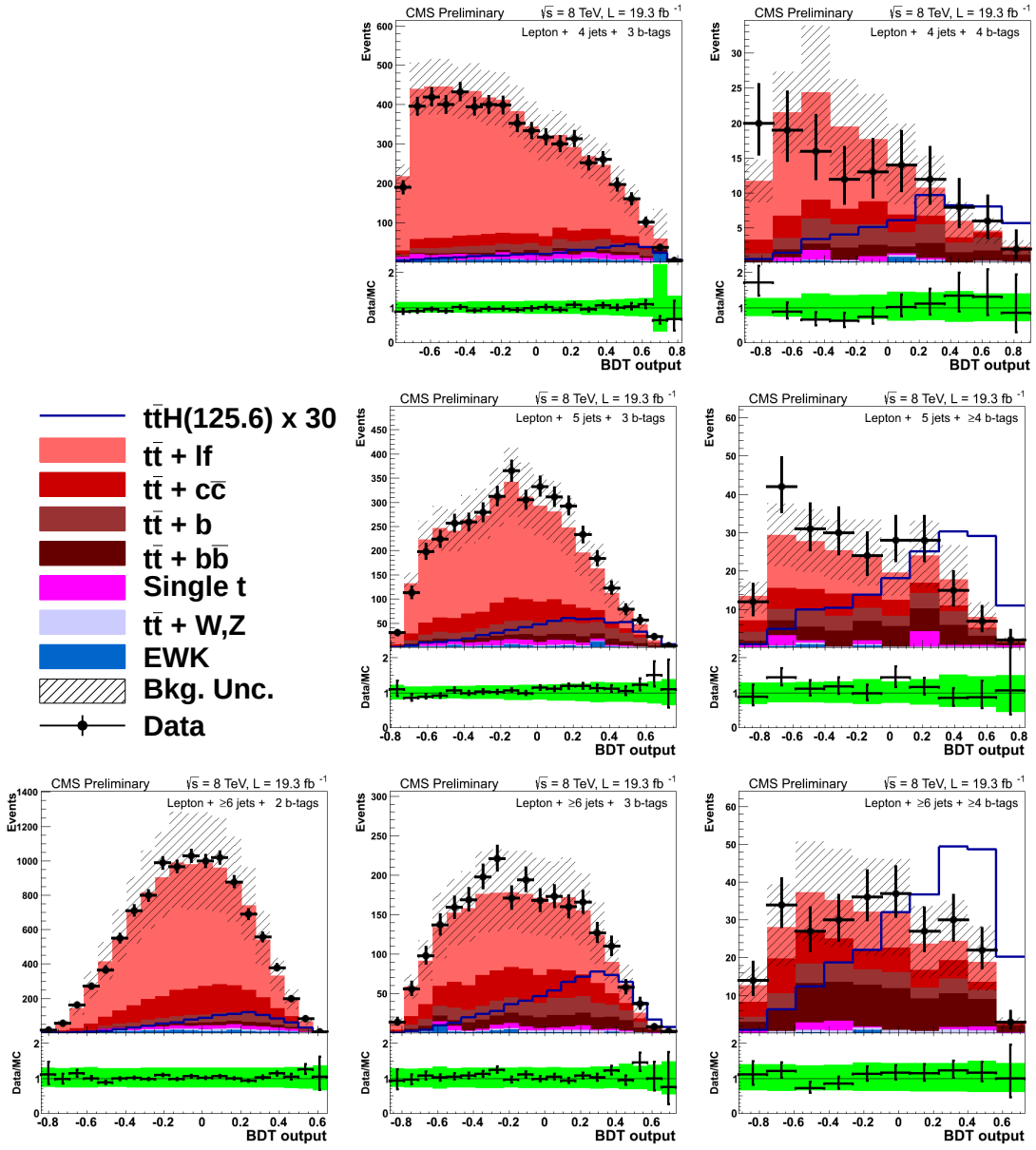


Figure 6.8: Data/MC comparisons for all final BDTs. The top, middle and, bottom rows are events with 4, 5, and $\geq$6 jets, respectively, while the left, middle, and right-hand columns are events with 2, 3, and $\geq$4 b-tags, respectively.

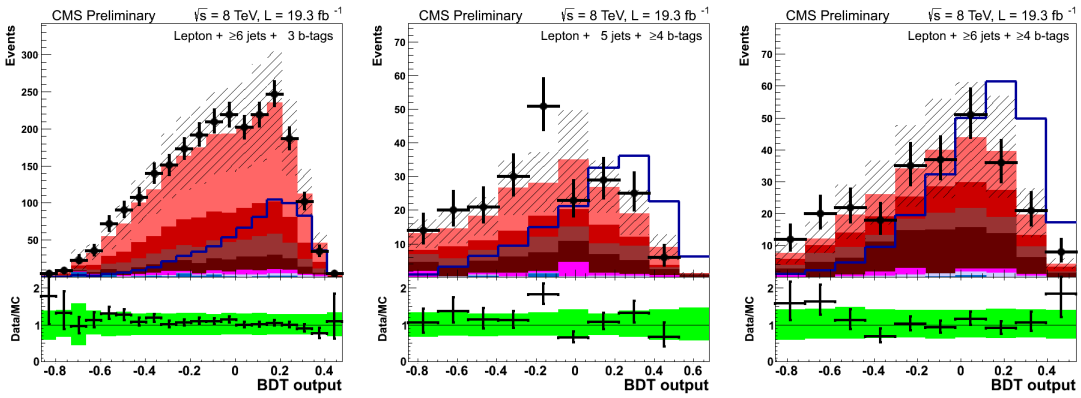


Figure 6.9: Data/MC comparisons for the $t\bar{t}H/t\bar{t}+b\bar{b}$ BDTs in the ≥6 jets + 3 b-tags (left), 5 jets + ≥4 b-tags (center), and ≥6 jets + ≥4 b-tags (right) categories. See figure 6.8 for legend.

Chapter 7
Uncertainties

7.1 Overview

In this analysis, both statistical and systematic uncertainties are considered. The ROOT software calculates and stores the bin-by-bin statistical uncertainty automatically for each histogram object. For a given bin, it is simply the square root of the sum of squares of the weights used to fill the bin. In contrast, the systematic uncertainties must be explicitly specified; the full list is given in table 7.1. Generally speaking, the total uncertainty for a given bin is:

$$\sigma_{tot}^2 = \sigma_{stat}^2 + \sigma_{sys}^2. \tag{7.1}$$

However, due to the complex nature of this analysis, the magnitudes and correlations of both shape and rate uncertainties across the different analysis categories must be considered. The purpose of this chapter is to describe the items listed in table 7.1. The statistical machinery of the limit calculation is discussed in the next chapter.

7.2 Systematics

There are three types of systematic effects considered in this analysis: those that affect only the rates of signal or background processes, those that affect only the shapes of distributions, and those that affect both the rate and the shape. The systematics are applied in the same way to every affected MC sample. Not every systematic affects every sample; however, most do, with the exceptions noted below and in table 7.1. We fluctuate uncertainties

Source	Shape	Remarks
Luminosity	No	Signal and all backgrounds
Lepton ID/Trigger efficiency	No	Signal and all backgrounds
Pileup	No	Signal and all backgrounds
Top p_{T} reweighting	Yes	Only $t\bar{t}$ background
Jet Energy Resolution	No	Signal and all backgrounds
Jet Energy Scale	Yes	Signal and all backgrounds
b-Tag bottom-flavor contamination	Yes	Signal and all backgrounds
b-Tag bottom-flavor statistics (linear)	Yes	Signal and all backgrounds
b-Tag bottom-flavor statistics (quadratic)	Yes	Signal and all backgrounds
b-Tag light-flavor contamination	Yes	Signal and all backgrounds
b-Tag light-flavor statistics (linear)	Yes	Signal and all backgrounds
b-Tag light-flavor statistics (quadratic)	Yes	Signal and all backgrounds
b-Tag Charm uncertainty (linear)	Yes	Signal and all backgrounds
b-Tag Charm uncertainty (quadratic)	Yes	Signal and all backgrounds
QCD Scale ($t\bar{t}H$)	No	Scale uncertainty for NLO $t\bar{t}H$ prediction
QCD Scale ($t\bar{t}$)	No	Scale uncertainty for NLO $t\bar{t}$ and single top predictions
QCD Scale (V)	No	Scale uncertainty for NNLO W and Z prediction
QCD Scale (VV)	No	Scale uncertainty for NLO diboson prediction
PDF (gg)	No	Parton distribution function (PDF) uncertainty for gg initiated processes ($t\bar{t}$, $t\bar{t}Z$, $t\bar{t}H$)
PDF ($q\bar{q}$)	No	PDF uncertainty for $q\bar{q}$ initiated processes ($t\bar{t}W$, W, Z).
PDF (qg)	No	PDF uncertainty for qg initiated processes (single top)
Madgraph Q^2 Scale ($t\bar{t}$+0p,1p,2p)	Yes	Madgraph Q^2 scale uncertainty for $t\bar{t}$+jets split by parton number. There is one nuisance parameter per parton multiplicity and they are uncorrelated.
Madgraph Q^2 Scale ($t\bar{t}$+b/$b\bar{b}$/$c\bar{c}$)	Yes	Madgraph Q^2 scale uncertainty for $t\bar{t}$+b/$b\bar{b}$/$c\bar{c}$.
Madgraph Q^2 Scale (V)	No	Varies by jet bin.
Extra $t\bar{t}$+hf rate uncertainty	No	A 50% uncertainty in the rate of $t\bar{t}$+b, $t\bar{t}+b\bar{b}$, $t\bar{t}+c\bar{c}$.

Table 7.1: Summary of the systematic uncertainties considered in the inputs to the limit calculation. Except where noted, each row in this table is treated as a single, independent nuisance parameter.

independently wherever it is possible to do so; however, some nuisances are correlated, the most notable example being the JES and b-tag systematics. The effect of category-to-category migration of events is taken into account. The rate uncertainties for each sample and each category are summarized in table 5.3, and the combined rate and shape uncertainties for the input variables and BDT outputs and are shown as the shaded regions in figure 6.8, and in the figures in appendix A. Broadly speaking, the largest rate effects include the uncertainty on the amount of $t\bar{t}$ +HF, JES uncertainty and b-tagging uncertainties. Each of the systematic uncertainties considered in this analysis is described below.

7.2.1 Luminosity and Pileup

The overall uncertainty on the integrated luminosity is 4.4%. This is applied uniformly for all samples and all categories.

The pileup reweighting uncertainty is derived by changing the minimum bias cross section used to calculate the pileup reweighting, which is varied by $\pm$ 7% from the default value. The pileup reweighting is recalculated using the shifted cross section, and the uncertainty is determined by applying the new weights and comparing to the nominal reweighting. This uncertainty has a negligible shape effect.

7.2.2 Monte Carlo Cross-Section, Q^2 and Statistical Uncertainties

Several uncertainties affect the normalization and shapes of the individual MC samples. The MC statistical uncertainty is calculated separately for each final BDT bin of each sample, as described earlier; however, to limit the number of nuisance parameters, we neglect bins where the MC statistical uncertainty is negligible compared to the statistical uncertainty of the data, or for which the contribution from signal is negligible.

In addition, there is an uncertainty on the cross section used to normalize each of the individual MC samples. This uncertainty is assessed independently for each sample. Each of the cross sections has been calculated to next-to-leading order (NLO) accuracy or greater. The uncertainty assigned to the cross sections consists of two components: an uncertainty that is due to how the parton density function of the colliding protons is modeled in the

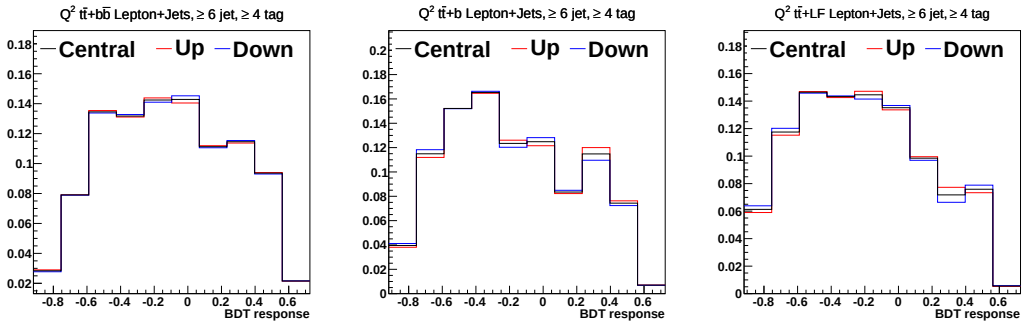


Figure 7.1: Comparison of the BDT output when shifting the Q^2 scale up and down by its uncertainties. Shown are the shift upwards (red) and downwards (blue) relative to the nominal (black) shape for the $t\bar{t}+b\bar{b}$(left) $t\bar{t}+b$ (center) and $t\bar{t}$+LF (right) background samples.. The plots are normalized to unit area.

simulation, and an uncertainty on the QCD scale, which affects the rate of QCD interactions. The dominant ($t\bar{t}+$jets) background has a pdf(QCD scale) uncertainty of 2.6(3)%, and the uncertainty on the pdf(QCD scale) component of the $t\bar{t}H$ cross section is 9(12.5)%. These cross section uncertainties are inclusive uncertainties on the overall normalization of a given process, and are calculated for each MC sample.

We also evaluate an uncertainty on some additional theoretical parameters used by the simulation [49], which are summarized by the Q^2 systematic. The Q^2 uncertainty affects the number of jets produced per event, as well as the jet kinematics; therefore, both rate and shape effects are considered. The uncertainty is applied separately for the $t\bar{t}+LF$, $t\bar{t}+c\bar{c}$, $t\bar{t}+b$ and $t\bar{t}+b\bar{b}$ components of the $t\bar{t}+$jets MC. Figure 7.1 shows Q^2 shape variations for one of the categories. The change in the yields due to Q^2 varies from 14%-23%, depending on the flavor of $t\bar{t}+$jets.

7.2.3 Lepton ID

A Lepton rate uncertainty of 1.4% is applied, and is a single nuisance parameter. This is obtained from adding the separate 1% trigger and 1% lepton isolation/identification uncertainties in quadrature. Although the uncertainties for muons and electrons differ slightly, we treat them identically for simplicity. The trigger and isolation uncertainties are

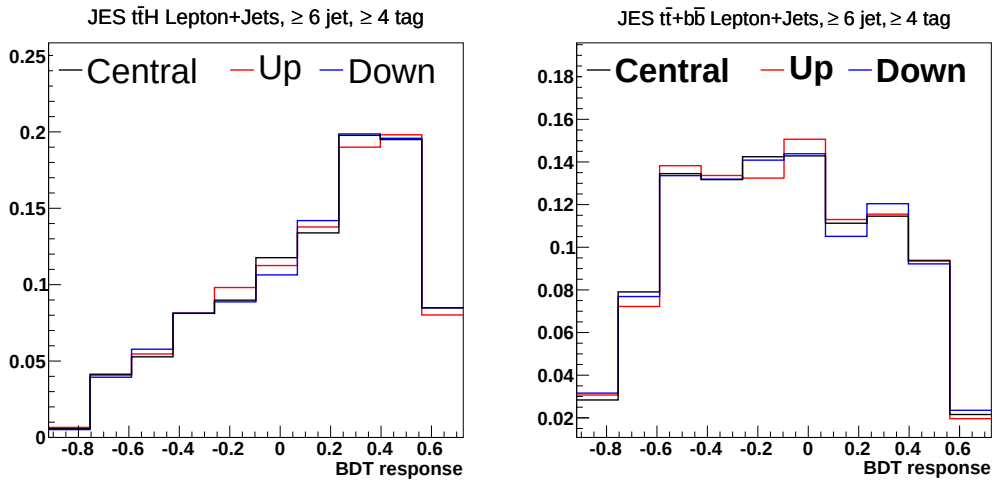


Figure 7.2: Comparison of the final BDT output in ≥ 6 jets ≥ 4 tags, for JES shift upwards (red) and downwards (blue) relative to the nominal (black) shape for the $t\bar{t}H(125)$ signal (left) and the $t\bar{t}+b\bar{b}$ background (right). The plots are normalized to unit area.

conservative estimates, motivated by our own studies, as well as studies by the CMS Muon POG [16].

7.2.4 Jet Energy Corrections

The JES uncertainty has one of the largest effects on analysis sensitivity. The JES affects both rate and shape; events may migrate between categories due to changes in JES, or be added or eliminated completely from the overall selection due to additional or fewer jets passing the p_{T} requirement. The p_{T}- and η-dependant JES correction is adjusted up and down by 1 σ, as described in [14]. This uncertainty has about a 10% effect on rates in the $\geq 6jets$ categories, and a somewhat smaller effect in other categories. The effect on the shapes of $t\bar{t}H$ and $t\bar{t}+b\bar{b}$ events is shown in figure 7.2 for the ≥ 6 jet + ≥ 4 b-tags category.

The uncertainty on the JER correction is assessed by adjusting the correction factor c in equation 5.2 up and down by 1 σ. Although the JER uncertainty affects the jet p_{T} distributions, we found that the shape variation was negligible, so we only consider the uncertainty on the rate in the limit calculation. In our most sensitive category, this has a 1.5% effect on the rate.

7.2.5 Top p_{T} Reweighting

A $\pm 1\ \sigma$ shape uncertainty on the top-p_{T} correction is obtained by not applying the correction, as well as applying twice the correction. This covers the area indicated by the green band in figure 5.4. This correction also has an effect on the rates – it is approximately a 5-7% effect across the different flavors of $t\bar{t}+$ jets in the $\geq 6jet+ \geq 4b-tags$ category. Recall that the top-p_{T} reweighting is only applied to $t\bar{t}+$ jets MC; thus, this uncertainty is only applied to $t\bar{t}+$ jets.

7.2.6 B-tagging

The uncertainty on b-tagged jets has several components that are due to the CSV reweighting procedure, and consist of JES, purity and statistical uncertainties [25]. The JES component is evaluated at the same time the overall JES uncertainty is considered, and is included as an effect on the kinematics of the jets used to derive the CSV SFs. The b-tagging JES uncertainty is thus folded into the other b-tag nuisances, and is not its own separate nuisance parameter. The other two components are evaluated separately for LF and HF jets. The purity uncertainty is the uncertainty on the amount of LF contamination (in the case of calculating the HF SFs) or HF contamination (in the case of calculating the LF SFs). The statistical uncertainties are taken from the MC samples used to derive the CSV SFs. To preserve the normalization during the calculation of the CSV SFs, the statistical uncertainties must be calculated in a way that takes only statistical variations in the shape of the CSV discriminant into account, and does not adjust the overall normalization. Thus, we use two different nuisances for the HF statistics and two nuisances for LF statistics: a nuisance which distorts the CSV distribution by tilting it to one side or the other, and a nuisance which makes a symmetric, parabolic adjustment to the shape, so that the upper and lower ends of the distribution change relative to the center [25]. By changing the shape of the CSV discriminant, the purity and MC-statistics nuisances have an effect on which jets get b-tagged and which do not, so that the event yields in the different categories are affected. In the ≥ 6 jet ≥ 4 category (our most sensitive category, and the category with

sys	shift	$t\bar{t}H(125)$	$t\bar{t}$+LF	$t\bar{t}+b\bar{b}$
Heavy Flavor SF Purity	up	+13.2%	+7.4%	+13.3%
	down	-12.1%	-7.2%	-12.1%
Light Flavor SF Purity	up	-3.4%	-32.2%	-4.4%
	down	+3.4%	+43.9%	+4.4%
Heavy Flavor SF Stat. Err. 1	up	-12.1%	-6.6%	-11.8%
	down	+13.3%	+6.8%	+12.9%
Heavy Flavor SF Stat. Err. 2	up	+8.9%	+5.0%	+9.1%
	down	-8.3%	-4.9%	-8.5%
Light Flavor SF Stat. Err. 1	up	+0.5%	-15.6%	+0.1%
	down	-0.5%	+17.7%	-0.1%
Light Flavor SF Stat. Err. 2	up	+1.8%	+10.1%	+2.1%
	down	-1.7%	-8.9%	-2.0%

Table 7.2: This table summarizes the effect of each of the independent LF and HF b-tag nuisance parameters on the yields of different samples, in the ≥ 6 jet ≥ 4 category. Variations due to JES are not shown. "Stat. Err. 1" and "Stat. Err. 2" refer to the linear and nonlinear components of the respective statistical uncertainties. The light SF purity for $t\bar{t}$+LF events is affected the most; this uncertainty also has the largest effect on the shape of $t\bar{t}$+LF events (see figure 7.3).

the greatest number of jets and b-tagged jets), these nuisances have a rate effect as shown in table 7.2. Figure 7.3 shows examples of the effects on the BDT shape in the ≥ 6 jet + ≥ 4 b-tags category.

Since we do not explicitly calculate SFs for charm-flavor jets, the b-tag uncertainty for these jets is evaluated independently. The c-jet SF is a flat SF of unity, so it is not possible to determine this uncertainty in a similar manner as for the HF and LF SFs. Therefore, we take a conservative approach of applying an uncertainty that is twice the relative HF uncertainty, with the added requirement that this uncertainty must not be less than the HF SF for a given CSV/p_{T} bin. In addition, we assign an extra 50% rate uncertainty on $t\bar{t}+b$, $t\bar{t}+b\bar{b}$ and $t\bar{t}+c\bar{c}$, which is allowed to float independently. This extra uncertainty comes from concerns over differences observed in the cross sections for $t\bar{t}$ +HF processes between LO and NLO predictions. Since $t\bar{t}$ +HF events (especially $t\bar{t}+b\bar{b}$) look most like our signal, the assignment of this extra uncertainty was considered a conservative approach.

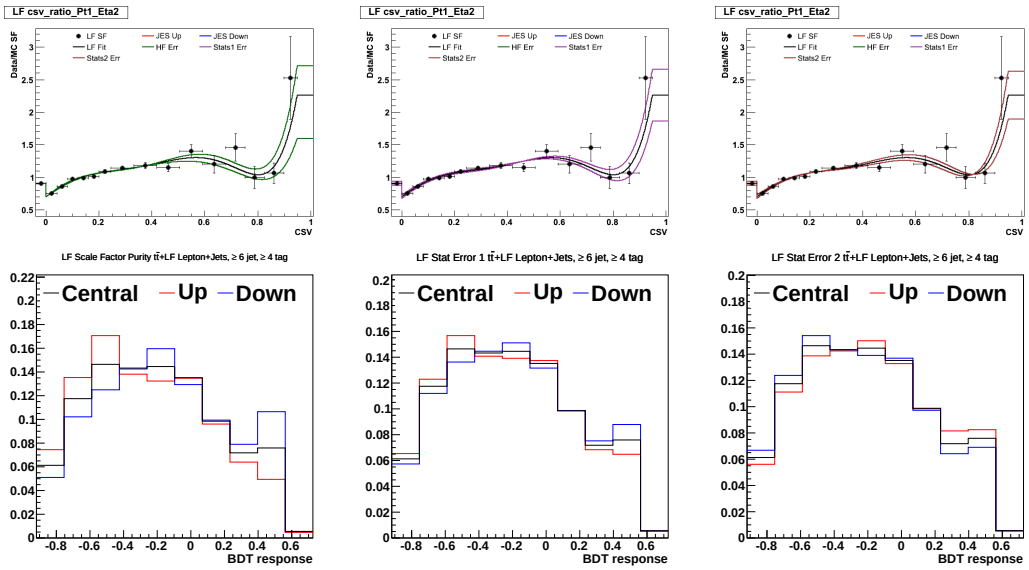


Figure 7.3: Example plots showing the up-and-down variations of selected b-tag shape systematics (top), and the resulting change in shape of the final BDT in the ≥ 6 jet ≥ 4 category (bottom). In a given column, the variations depicted in the top plot are reflected in the change in the shape of the distribution in the bottom plot. The bottom plots all show the change in the $t\bar{t}$ +LF distribution, after varying the HF contamination in the determination of the LF SF (left), and after varying the linear (center) and quadratic (right) distortions that determine the statistical uncertainty of the LF SF extraction. The All plots are normalized to unit area. These are the largest proportional shape variations due to b-tag uncertainties among all the flavors of $t\bar{t} +$ jets in this category.

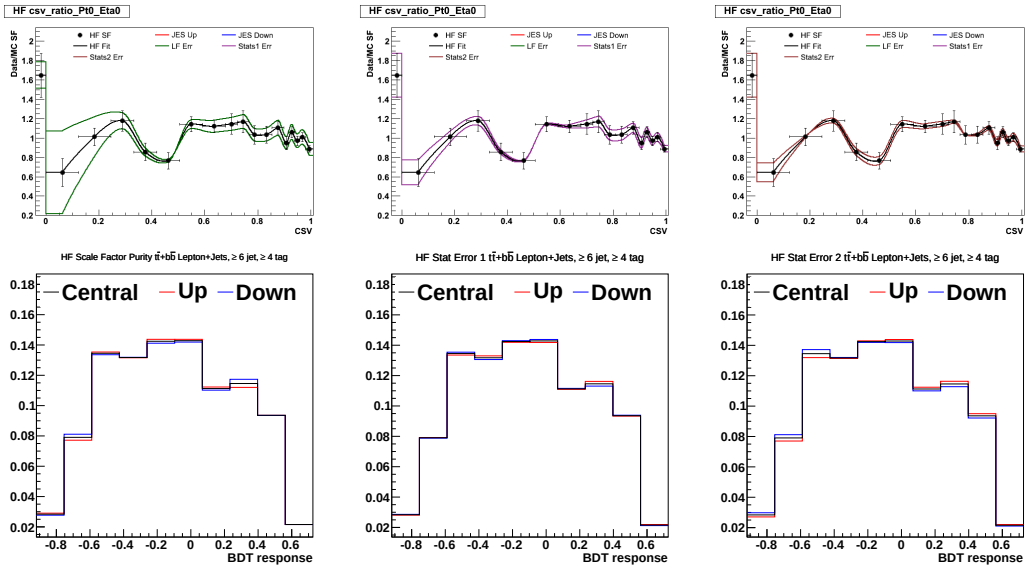


Figure 7.4: Example plots showing the up-and-down variations of selected b-tag shape systematics (top), and the resulting change in shape of the final BDT in the ≥ 6 jet ≥ 4 category (bottom). In a given column, the variations depicted in the top plot are reflected in the change in the shape of the distribution in the bottom plot. The bottom plots all show the change in the $t\bar{t}$ +HF distribution, after varying the LF contamination in the determination of the HF SF (left), and after varying the linear (center) and quadratic (right) distortions that determine the statistical uncertainty of the HF SF extraction. The All plots are normalized to unit area.

Chapter 8
Results

8.1 Statistical Method

We did not observe $t\bar{t}H$. Thus, we present the results of the analysis as an upper limit on the $t\bar{t}H$ production cross-section, given our uncertainties. Since there is no evidence to the contrary, the null (background-only) hypothesis is assumed. While we have not disproven the alternative hypothesis, the limit calculation allows us to be able to reject a portion of its parameter space, quantified as a 95% confidence upper limit on the signal strength modifier, $\mu = \sigma(t\bar{t}H)/\sigma(t\bar{t}H)_{SM}$. The calculation is done in a correlated fashion across the 7 jet-tag categories of the analysis, and takes as inputs the binned distributions of the final $t\bar{t}H/t\bar{t}+$jets BDTs in each of the categories. The BDT distributions for the signal, backgrounds, and data are used, as well as all the different versions of the distributions specified by the nuisance parameter fluctuations.

First, a background-only fit to the data is performed. The fit is done with the sum of the background MC, correlated across all categories, with all rate and shape nuisances allowed to float during the fit (taking correlations in the uncertainties into account). The fit places constraints on the nuisances; going forward, each of the pre-fit uncertainties is replaced by a post-fit uncertainty that is (in general) smaller, depending on the nuisance. Although this is a background-only fit, many background nuisances are shared by the signal, and so the signal uncertainties are also affected by the fit [50]. Figure 8.1 shows the result of this background-only fit in the three most sensitive jet-tag categories.

A modified frequentist approach is used to extract the limits [51, 45]. The limit calculation

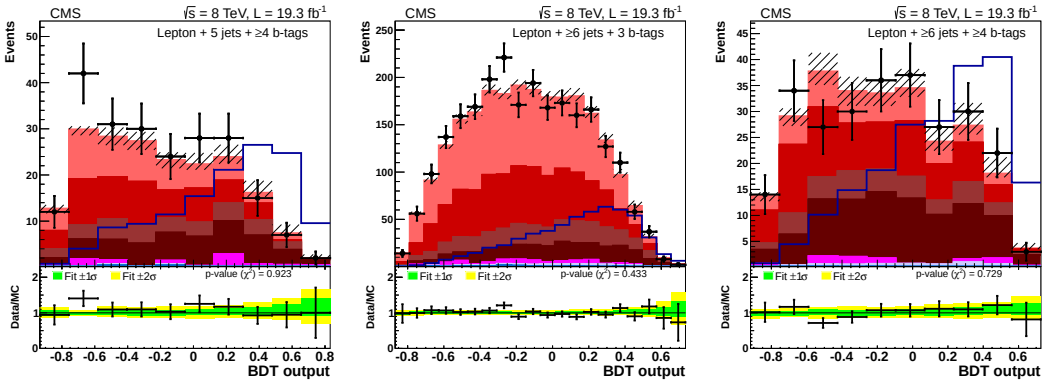


Figure 8.1: Background-only fit to the data in the 5 jets + ≥4 b-tags (left), ≥6 jets + 3 b-tags (center), and ≥6 jets + ≥4 b-tags (right) categories. The post-fit uncertainties are constrained relative to the uncertainties before the fit, as can be seen by comparing to figure 6.8.

uses the test statistic q, where

$$q = -2ln\left(\frac{L(data|\mu s + b, \hat{\theta}_\mu)}{L(data|\hat{\mu} s + b, \hat{\theta})}\right). \tag{8.1}$$

Here, L is a likelihood function that is maximized under certain conditions (explained below). The likelihood function itself is given by:

$$L(data|\mu s + b, \theta) = \prod_i \frac{(\mu s_i + b_i)^{n_i}}{n_i!} e^{-(\mu s_i + b_i)} \times p(\tilde{\theta}, \theta). \tag{8.2}$$

The index i refers to a given bin in the final $t\bar{t}H/t\bar{t}+$jets BDTs, s_i and b_i are the number of predicted signal and background events in that bin, respectively, and μ is the signal strength modifier. The $p(\tilde{\theta}, \theta)$ factor encapsulates the result of the background-only fit to the data: it is a probability distribution that determines the relative probability of some arbitrary values of the nuisances (θ), given the set of post-fit central values of the nuisance parameters, and their corresponding post-fit uncertainties ($\tilde{\theta}$). For the denominator in equation 8.1, the values $\mu = \hat{\mu}$ and $\theta = \hat{\theta}$ are those that maximize the likelihood function L. For the numerator, two cases are considered: the background-only (B) and signal + background (S+B) hypotheses. For the background-only case, the likelihood function in the numerator is maximized after setting $\mu = 0$ (and $\theta_\mu = \theta_0$). For the S+B hypothesis, the likelihood

function is maximized after setting μ to some fixed nonzero value.

A large number of pseudo-experiments are thrown to simulate possible measurements given the B and S+B hypotheses. A single distribution of the test statistic q is constructed using the B pseudo-experiments, and a range of possible distributions of q is obtained for the S+B case by varying the value of μ. Finally, these B and S+B distributions are overlaid, and for a given point on the B distribution, the relative probability of q greater than that point under the B versus the S+B hypothesis is determined:

$$CL_s = \frac{P(q \geq obs|S+B)}{P(q \geq obs|B)}. \tag{8.3}$$

The S+B distribution of q (and corresponding μ) is found such that $1 - CL_s \geq$95%. This μ value is saved, and calculated again using a different point on the B distribution, until the entire B distribution is scanned. The median of the distribution of μ values obtained in this manner is the median expected 95%-confidence upper limit on $t\bar{t}H$ production. In a 1-dimensional plot, the entire histogram of expected 95%-confidence upper limits may be shown, but typically we seek to calculate limits for various values of m_H. In that case, we plot the median expected limit, surrounded by the 1- and 2-σ error bands of the distribution, as a function of m_H. For the actual data, the calculation proceeds in a similar manner, but of course only a single 95%-confidence upper limit is calculated for a given m_H.

8.2 Results of This Analysis

In this section, the results of the search for $t\bar{t}H$ (with $H \to b\bar{b}$) in the lepton+jets (LJ) channel at 8 TeV is presented. The specific software used for the limit calculation is the "combine" package, which is part of the CMS software framework. This is the recommended software; it is the same tool used by each of the CMS Higgs analyses, and in all combinations of Higgs analyses [26].

Figure 8.2 shows the 95% confidence upper limit on the ratio of the $t\bar{t}H$ cross section to the cross section predicted by the Standard Model, as a function of m_H, for the search performed in this analysis channel. The figure shows the median expected limit, and the

$m_H = 125.6\,\text{GeV}$	
Observed Limit:	$\mu < 5.1$
Expected $+2\sigma$	$\mu < 9.5$
Expected $+1\sigma$	$\mu < 7.0$
Expected (Median)	$\mu < 5.0$
Expected -1σ	$\mu < 3.5$
Expected -2σ	$\mu < 2.7$

Table 8.1: The observed and expected 95% confidence upper limits on μ for $t\bar{t}H$ production in the lepton+jets channel, at $m_H = 125.6\,\text{GeV}$.

68% confidence and 95% confidence error bands on the median expected limit. The observed limit is shown as a solid black line. Given the current measured value for the mass of the Higgs boson, the limit was also calculated for the specific case $m_H = 125.6\,\text{GeV}$. This result is shown in table 8.1.

Using the same software, a S+B maximum-likelihood fit to the data was performed to determine the best-fit value of μ in this channel, for $m_H = 125.6\,\text{GeV}$. The result was $\mu = -0.2^{+2.8}_{-2.7}$. This calculation is essentially the same as that which determines the denominator in equation 8.1, except the probability distribution $p(\tilde{\theta}, \theta)$ is determined from a S+B fit of the nuisance parameters (instead of a background-only fit). Although counter-intuitive, one can see how this negative result is possible by examining the Poisson factor in equation 8.2, where a negative contribution from the signal may be used to adjust the normalization of the combined S+B prediction. It should also be pointed out that, due to the large uncertainty, this best-fit result is neither in disagreement with the null hypothesis, nor the $\mu = 1$ hypothesis.

8.3 Combined $t\bar{t}H$ Results

The above result has been combined with the results of other $t\bar{t}H$ searches at CMS. In addition to the current LJ analysis, the search for $t\bar{t}H$ (with $H \to b\bar{b}$) has been performed in the channel where each of the top quarks decays to $l\nu b$ – this is the dilepton channel of $t\bar{t}H$. The LJ and DIL channels have been combined into an overall $t\bar{t}(H \to b\bar{b})$ result in the plots below. The "all hadronic" case, where there are no leptons in the final state, does not

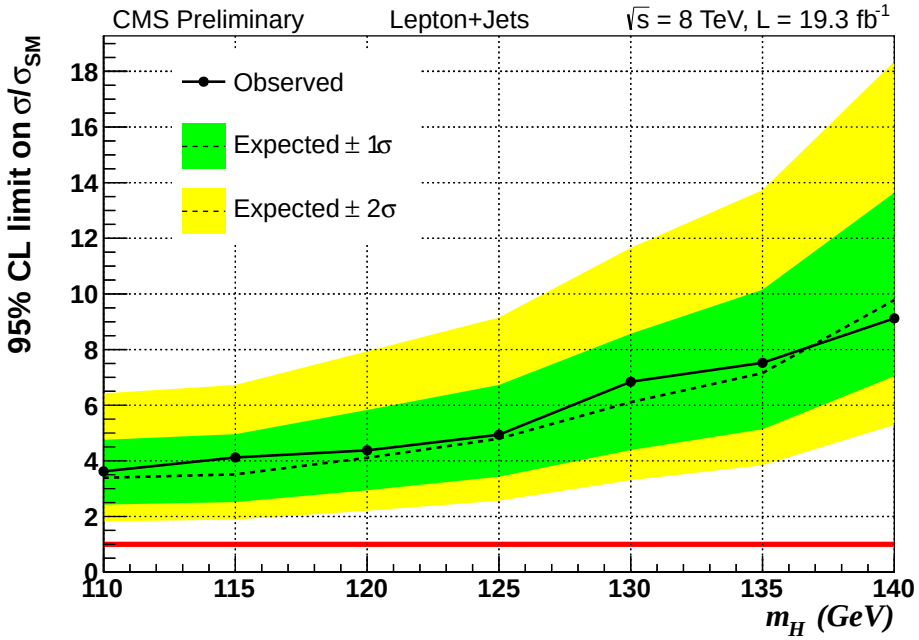


Figure 8.2: The observed and expected 95% confidence upper limits on the signal strength modifier μ, for $t\bar{t}H$ production in the lepton+jets channel, as a function of m_H. The solid black line is the observed limit. The dashed line is the median expected limit, and the green (yellow) regions are the 68% (95%) error bands on the expected limit.

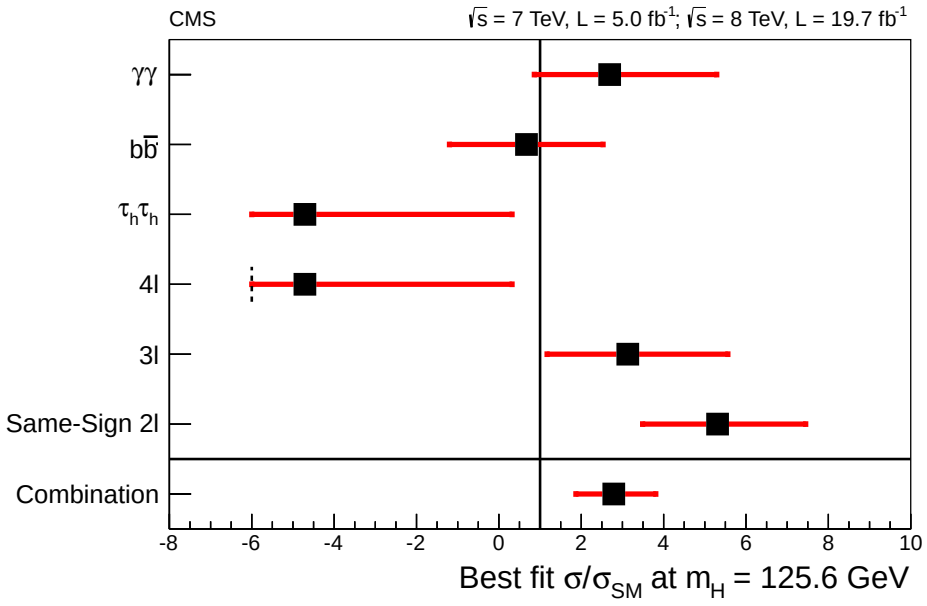


Figure 8.3: The best-fit to the signal strength modifier μ, with $\pm 1\sigma$ error bars. Results are shown separately for the different $t\bar{t}H$ channels, as well as the result for a combined fit using all channels. The Standard Model value is shown as a black vertical line.

currently have a result.

Searches for $t\bar{t}H$ at CMS have also been performed for a variety of other final states. The $t\bar{t}(H \to \tau\bar{\tau})$ analysis searches for a τ pair decaying hadronically in association with the semileptonic mode of $t\bar{t}$, a final state similar to this analysis. Other channels involve multileptonic states, which are designed to look for $t\bar{t}H$ with $H \to W^+W^-$ and $H \to ZZ$. A $t\bar{t}H$ search is also performed in the $H \to \gamma\gamma$ channel. Figure 8.4 shows the limits obtained from combining all current $t\bar{t}H$ results at CMS, and figure 8.3 shows the combined results of the fit to the signal strength μ. It should be noted that special care has been taken to ensure the orthogonality of these results; otherwise such a calculation would not be possible.

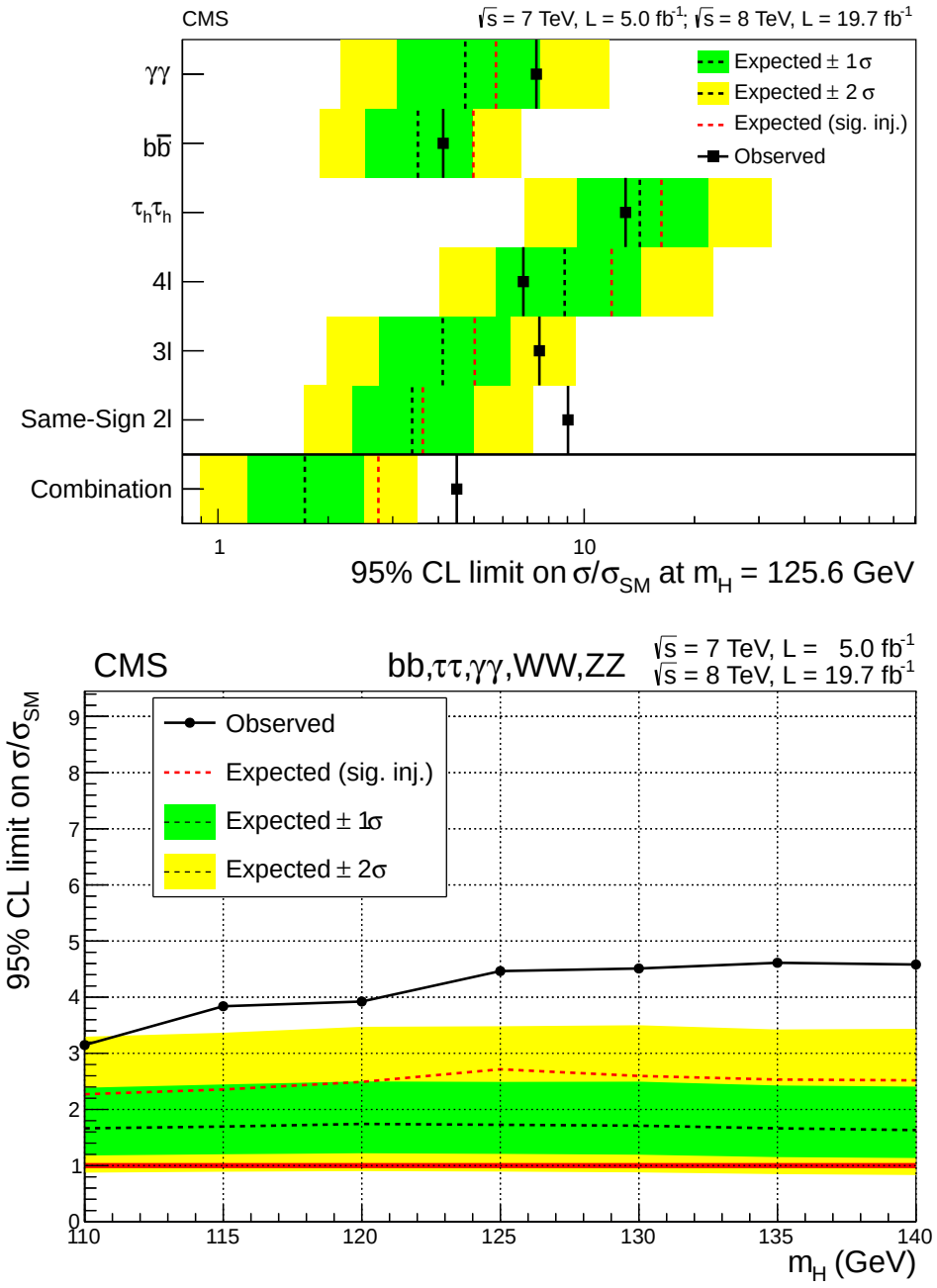


Figure 8.4: The observed and expected 95% confidence upper limits on the signal strength modifier μ, in a combined search for $t\bar{t}H$ production at CMS. The solid black line is the observed limit. The dashed line is the median expected limit, and the green (yellow) regions are the 68% (95%) error bands on the expected limit. The red dashed line represents the median expected limit calculated with signal injected at the Standard Model rate. Top: limits separated by channel, for $m_H = 125.6\,\text{GeV}$. Bottom: the combined limit as a function of m_H.

Chapter 9
Conclusion

In this dissertation, a search for the Standard Model Higgs boson in the $t\bar{t}H$ production mode was performed, using $19.3\,\text{fb}^{-1}$ of data collected at a center-of-mass energy of 8 TeV at CMS. The specific decay channel considered was the semileptonic decay of a pair of top quarks, accompanied by the decay of the Higgs boson to a pair of b-quarks. Data was selected and categorized in an effort to isolate the desired $t\bar{t}H$ events while rejecting background, and a multivariate technique involving tiered BDTs was implemented to further boost analysis sensitivity. After accounting for statistical and systematic uncertainties, this analysis set a 95% upper confidence limit on the $t\bar{t}H$ cross section of 5.1 times the Standard Model expectation, at $m_H = 125.6\,\text{GeV}$. The median expected limit was $\mu < 5.0$, with a 68% CL range of [3.5, 7.0] and a 95% CL range of [2.7, 9.5]. This result was combined with the results of other $t\bar{t}H$ searches at CMS. The combined 95% upper confidence limit on $t\bar{t}H$ was $\mu < 4.5$, with a median expected limit of $\mu < 1.7$. The 68% CL range on the median expected limit of the combination was [1.2, 2.5], and the 95% CL range was [0.9, 3.5].

The Higgs boson is the last piece of the Standard Model puzzle. Although the boson discovered in 2012 has been positively identified as *a* Higgs boson, we have yet to conclude decisively that it is the *Standard Model* Higgs boson. However, this hypothesis seems to be further supported with each passing Higgs physics result. When the LHC resumes pp collisions in 2015, it will be at nearly twice the center-of-mass energy and at half the bunch-spacing as was used previously. Total integrated luminosities in the 100-200 fb^{-1} range can be expected by the end of Run II in 2018. We are about to enter an era where the

$t\bar{t}H$ searches will begin to transition into $t\bar{t}H$ measurements, and where it may be possible to ultimately confirm the SM identity of the newly discovered Higgs boson.

Bibliography

[1] S. Agostinelli et al. GEANT 4: A Simulation Toolkit. *Nucl. Instrum. Meth.*, A503:950–303, 2003.

[2] Ian Aitchison and Anthony Hey. *Gauge Theories in Particle Physics*. Institute of Physics Publishing, Philadelphia, 2004.

[3] Simone Alioli, Paolo Nason, Carlo Oleari, and Emanuele Re. A general framework for implementing NLO calculations in shower Monte Carlo programs: the POWHEG BOX. *JHEP*, 06:043, 2010. arXiv:hep-ph/1002.2581.

[4] ATLAS Collaboration. Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC. *Physics Letters B*, 716, 2012.

[5] F. Beaudette, D. Benedetti, P. Janot, and M. Pioppi. Electron reconstruction within the particle flow algorithm. CMS AN 2010-034, 2010.

[6] J. D. Bjorken and S. J. Brodsky. Statistical model for electron-positron annihilation into hadrons. *Physical Review D*, 1, March 1970. doi:10.1103/PhysRevD.1.1416.

[7] Matteo Cacciari and Gavin P. Salam. The anti-kt jet clustering algorithm. *JHEP*, 2008.

[8] Lea Caminada. *Study of the Inclusive Beauty Production at CMS and Construction and Commissioning of the CMS Pixel Barrel Detector*. Springer, Berlin, 2012.

[9] CERN. CERN accelerator complex web page. home.web.cern.ch/about/accelerators.

[10] CMS. CMS Collaboration Website. cms.web.cern.ch/org/cms-collaboration.

[11] CMS and ATLAS Collaborations. Collider Cross Talk: Higgs spin results from ATLAS and CMS. indico.cern.ch/event/240194/.

[12] CMS b-Tag POG. Methods to apply b-tagging efficiency scale factors. https://-twiki.cern.ch/twiki/bin/viewauth/CMS/BTagSFMethods.

[13] CMS Collaboration. Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC. *Physics Letters B*, 716, 2012.

[14] CMS Collaboration. Energy scale uncertainties. https://twiki.cern.ch/twiki/bin/view/CMSPublic/WorkBookJetEnergyCorrections.

[15] CMS Collaboration. MET optional filters. https://twiki.cern.ch/twiki/bin/view/-CMS/MissingETOptionalFilters.

[16] CMS Collaboration. Muon POG Tag and Probe Recommendations. https://twiki.cern.ch/twiki/bin/viewauth/CMS/MuonTagAndProbe.

[17] CMS Collaboration. CMS Physics Technical Design Report, Vol I: Detector Performance and Software. Technical report, CERN, 2006.

[18] CMS Collaboration. CMS Physics Technical Design Report, Vol II: Physics Performance. Technical report, CERN, 2006.

[19] CMS Collaboration. Particle Flow Event Reconstruction in CMS and Performance for Jets, Taus and MET. CMS PAS PFT-2009-001, April 2009.

[20] CMS Collaboration. Dilepton trigger and lepton identification efficiencies for the top quark pair production cross section measurement at 8 tev in the dilepton decay channel. CMS AN-12-389, 2012.

[21] CMS Collaboration. Measurement of b-tagging efficiency in semi-leptonic decays of $t\bar{t}$ events using the flavor-tag consistency method at 8 tev. CMS Analysis Note AN-12-187, 2012.

[22] CMS Collaboration. Measurement of differential top-quark pair production cross sections in the dilepton channel in pp collisions at 8 tev. CMS-PAS-TOP-12-028, 2012.

[23] CMS Collaboration. Measurement of differential top-quark pair production cross sections in the lepton+jets channel in pp collisions at 8 tev. CMS-PAS-TOP-12-027, 2012.

[24] CMS Collaboration. Measurement of the b-tagging efficiency using μ+jets events at 8 tev. CMS Analysis Note AN-12-432, 2012.

[25] CMS Collaboration. Calibration of the combined secondary vertex b-tagging discriminant using dileptonic ttbar and drell-yan events. CMS Note 2013/130, 2013.

[26] CMS Collaboration. Combination of standard model Higgs boson searches and measurements of the properties of the new boson with a mass near 125 GeV. CMS-PAS-HIG-13-005, 2013.

[27] CMS Collaboration. Search for the standard model higgs boson produced in association with a top-quark pair in pp collisions at the LHC. *JHEP*, 2013. arXiv:1303.0763.

[28] CMS Collaboration. Single muon efficiencies in 2012 data. CMS Performance Note DP-2013-09, March 2013.

[29] CMS Collaboration. Updated measurements of the Higgs Boson at 125 GeV in the two photon decay channel. CMS-PAS-HIG-13-001, 2013.

[30] Jan Conrad and Fred James. Notes on ROOT kolmogorov-smirnov test. http://-root.cern.ch/root/html/TH1.html#TH1:KolmogorovTest, Note 3.

[31] F. Englert and R. Brout. Broken symmetry and the mass of gauge vector mesons. *Phys. Rev. Lett.*, 13, 1964.

[32] J. Alwall et al. Madgraph 5: going beyond. *JHEP*, 06:128, 2011. arXiv:hep-ph/1106.0522.

[33] Fayyazuddin and Riazuddin. *A Modern Introduction to Particle Physics*. World Scientific, New Jersey, 2000.

[34] G. Fox and S. Wolfram. Event shapes in e+e annihilation. *Nuclear Physics B*, 157, 1979.

[35] Paul H. Frampton. *Gauge Field Theories*. Wiley-VCH, Weinheim, 2008.

[36] Gfitter group. Fit results from the current global fit. project-gfitter.web.cern.ch.

[37] David Griffiths. *Introduction to Elementary Particles*. Wiley-VCH, Weinheim, 2008.

[38] Particle Data Group. http://commons.wikimedia.org/wiki/-File:Standard_Model_of_Elementary_Particles.svg.

[39] G. S. Guralnik, C. R. Hagen, and T. W. Kibble. Global conservation laws and massless particles. *Phys. Rev. Lett.*, 13, 1964.

[40] P. W. Higgs. Broken symmetries and the masses of gauge bosons. *Phys. Rev. Lett.*, 13, 1964.

[41] A. Hoecker, P. Speckmayer, J. Stelzer, J. Therhaag, E. von Toerne, and H. Voss. Toolkit for multivariate data analysis with ROOT. arXiv:physics/0703039, 2013.

[42] J. Beringer et al. (Particle Data Group). Phys. Rev. D86, 010001 (2012), with partial update for 2014.

[43] J. Fernandez, for the ATLAS, CMS Collaborations. Higgs searches at cms and atlas, 2009. arXiv:0905.1228.

[44] John David Jackson. *Classical Electrodynamics*. Wiley, New Jersey, 2009.

[45] Thomas Junk. Confidence level computation for combining searches with small statistics. arXiv:hep-ex/9902006v1, 1999.

[46] Mike Lamont. The LHC's first long run. *CERN Courier*, 53(3):25–27, 2013.

[47] LHC Higgs Cross Section Working Group. https://twiki.cern.ch/twiki/bin/view/-LHCPhysics/CrossSections.

[48] Verena Martinez and Benjamin Hooberman. Recent Results on SUSY Searches From CMS. CMS News (cms.web.cern.ch/news), August 2013.

[49] T. Peiffer. q^2 systematic mc samples re-weighting, 2012. https://indico.cern.ch/-getFile.py/access?contribId=5&resId=0&materialId=slides&confId=177717.

[50] Darren Puigh. personal communication.

[51] A. L. Read. Preparation of Search Results: the CLs Technique. *Journal of Physics G: Nuclear Particle Physics*, 28:2693–2704, 2002.

[52] T. Sjostrand, S. Mrenna, and P. Z. Skands. PYTHIA 6.4 physics and manual. *JHEP*, 2006, 2006. arXiv:hep-ph/0603175.

[53] The CMS Collaboration. Determination of Jet Energy Calibration and Transverse Momentum Resolution in CMS. *arXiv*, 6:P11002, 2011.

[54] The CMS Collaboration. Description and performance of track and primary vertex reconstruction with the CMS tracker. CMS Paper TRK-11-001, November 2013.

[55] The CMS Collaboration. Identification of b-quark jets with the CMS experiment. *JINST*, 8, 2013.

[56] The CMS Collaboration et. al. The CMS experiment at the CERN LHC. *JINST*, 3, 2008.

[57] TOTEM. TOTEM experiement web site. totem-experiment.web.cern.ch.

[58] Wikimedia Website. http://commons.wikimedia.org/wiki/-File:Mecanismo_de_Higgs_PH.png.

Appendix A
Data/Monte-Carlo Comparison of Input Variables

This appendix includes Data/Monte-Carlo comparison figures for all the variables used as BDT inputs in the analysis. They are organized by jet-tag category, with the category indicated by the caption.

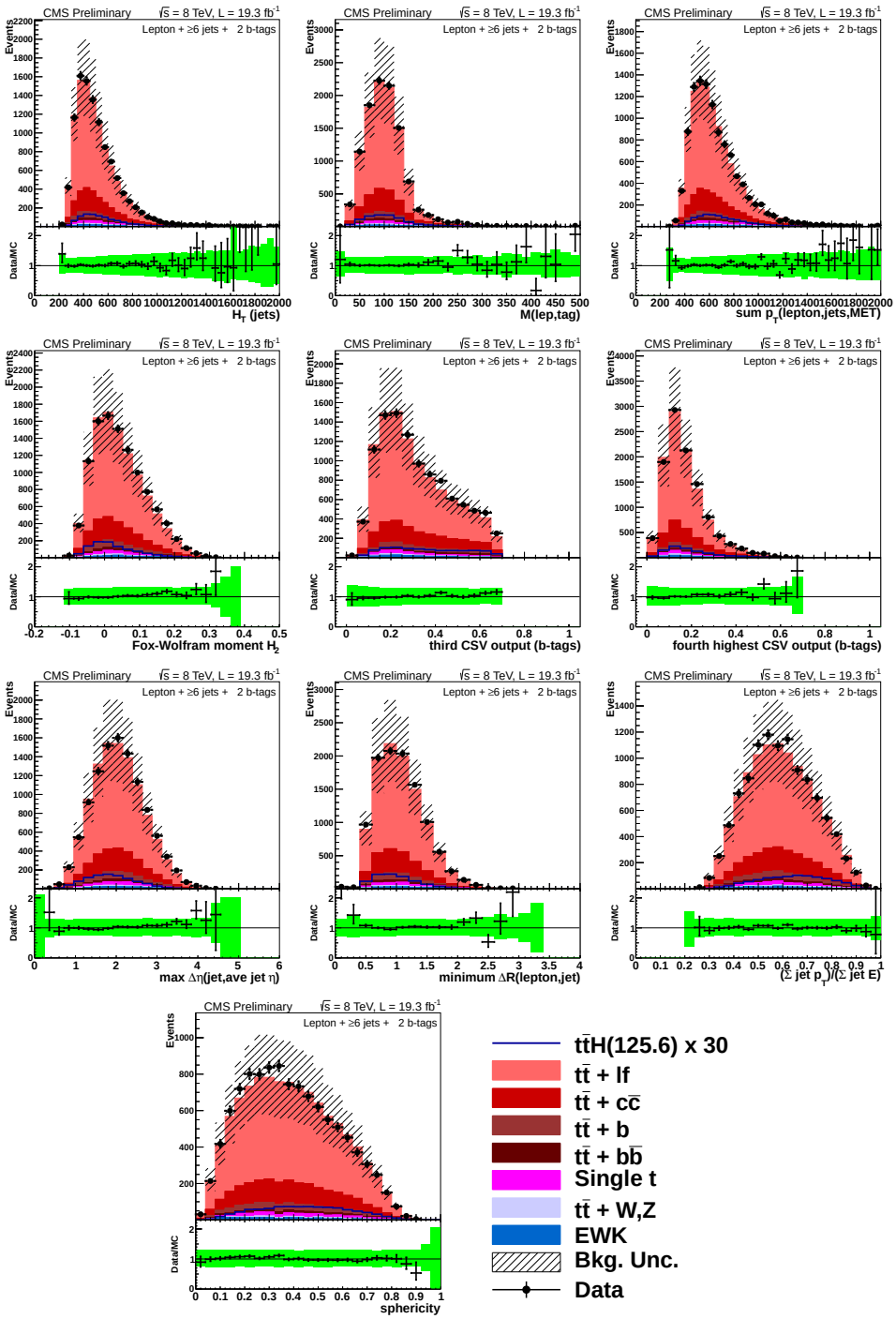


Figure A.1: Data/MC comparisons for events with one lepton and ≥6 jets + 2 b-tags. The uncertainty band includes statistical and systematic uncertainties that affect both the rate and shape of the background distributions.

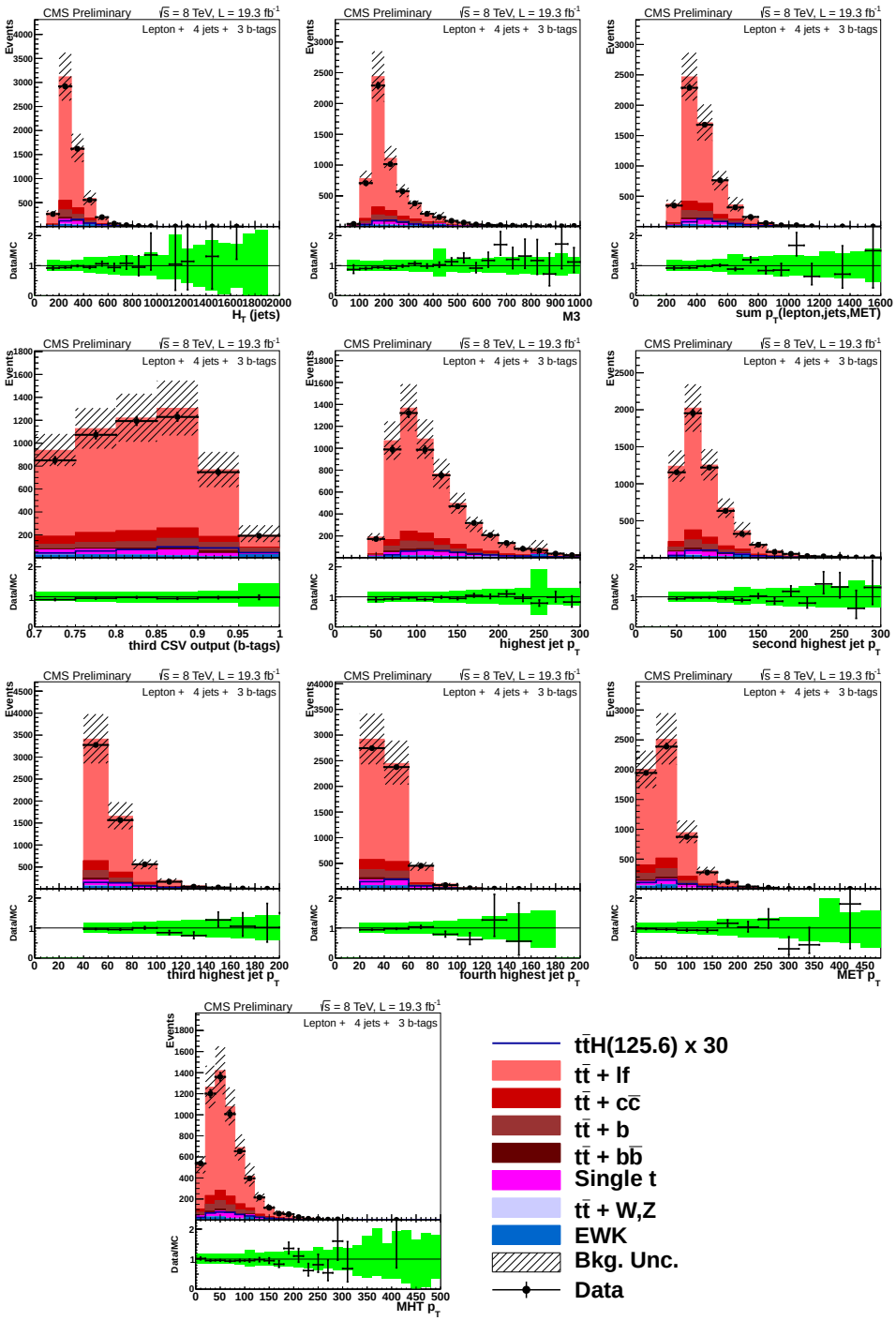


Figure A.2: Data/MC comparisons for events with one lepton and 4 jets + 3 b-tags. The uncertainty band includes statistical and systematic uncertainties that affect both the rate and shape of the background distributions.

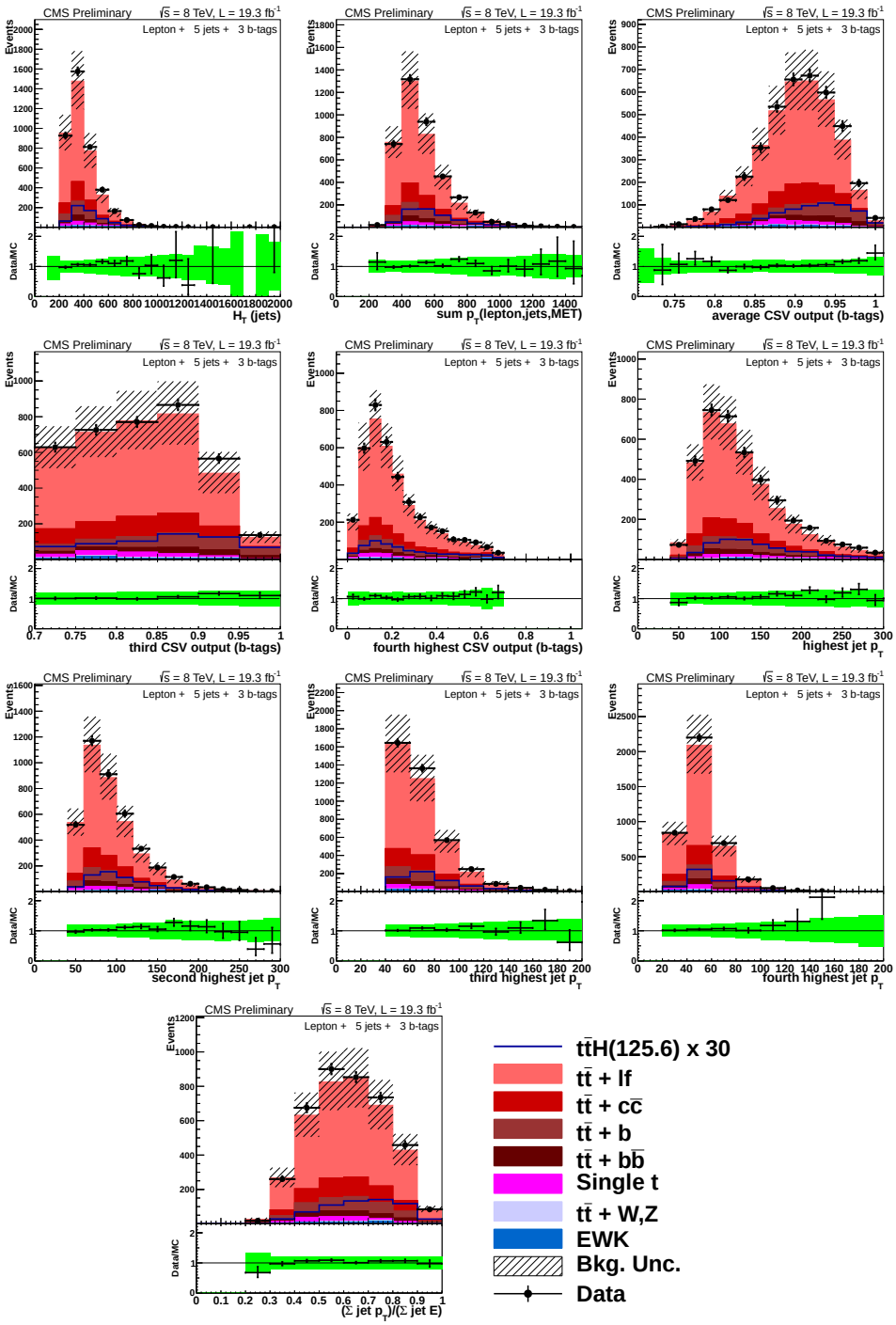


Figure A.3: Data/MC comparisons for events with one lepton and 5 jets + 3 b-tags. The uncertainty band includes statistical and systematic uncertainties that affect both the rate and shape of the background distributions.

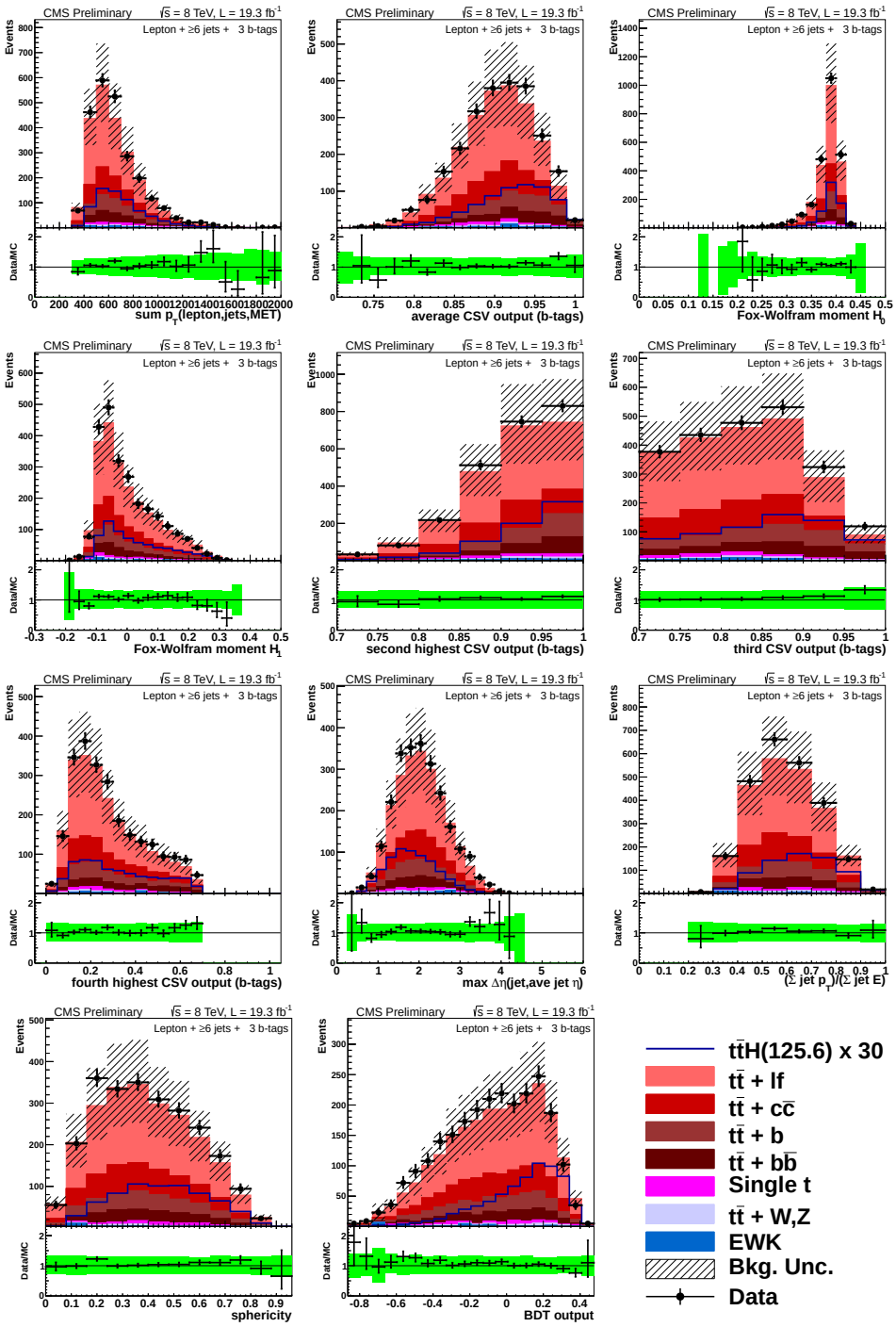


Figure A.4: Data/MC comparisons for events with one lepton and ≥6 jets + 3 b-tags. The uncertainty band includes statistical and systematic uncertainties that affect both the rate and shape of the background distributions.

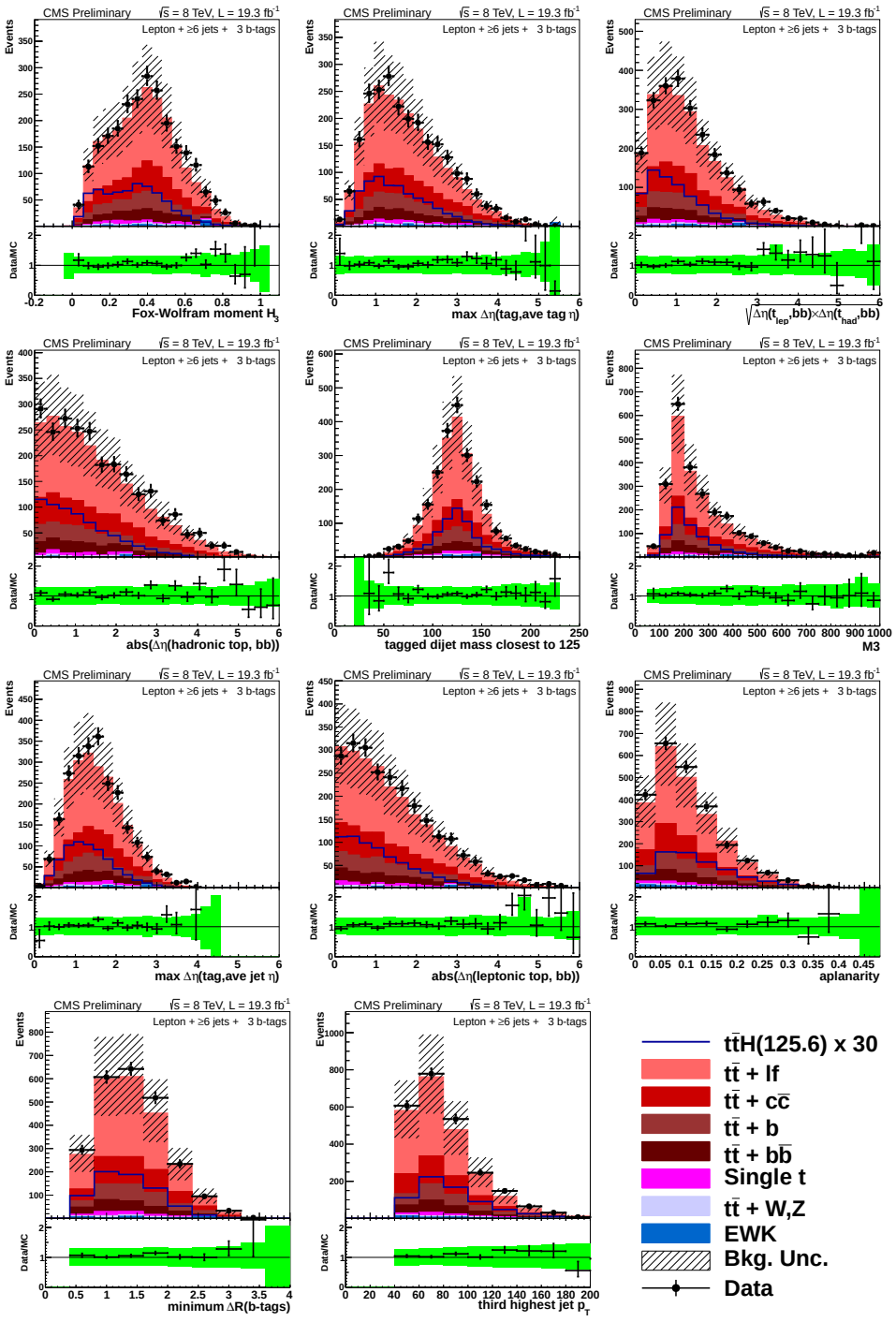


Figure A.5: Data/MC comparisons for events with one lepton and ≥6 jets + 3 b-tags. The uncertainty band includes statistical and systematic uncertainties that affect both the rate and shape of the background distributions.

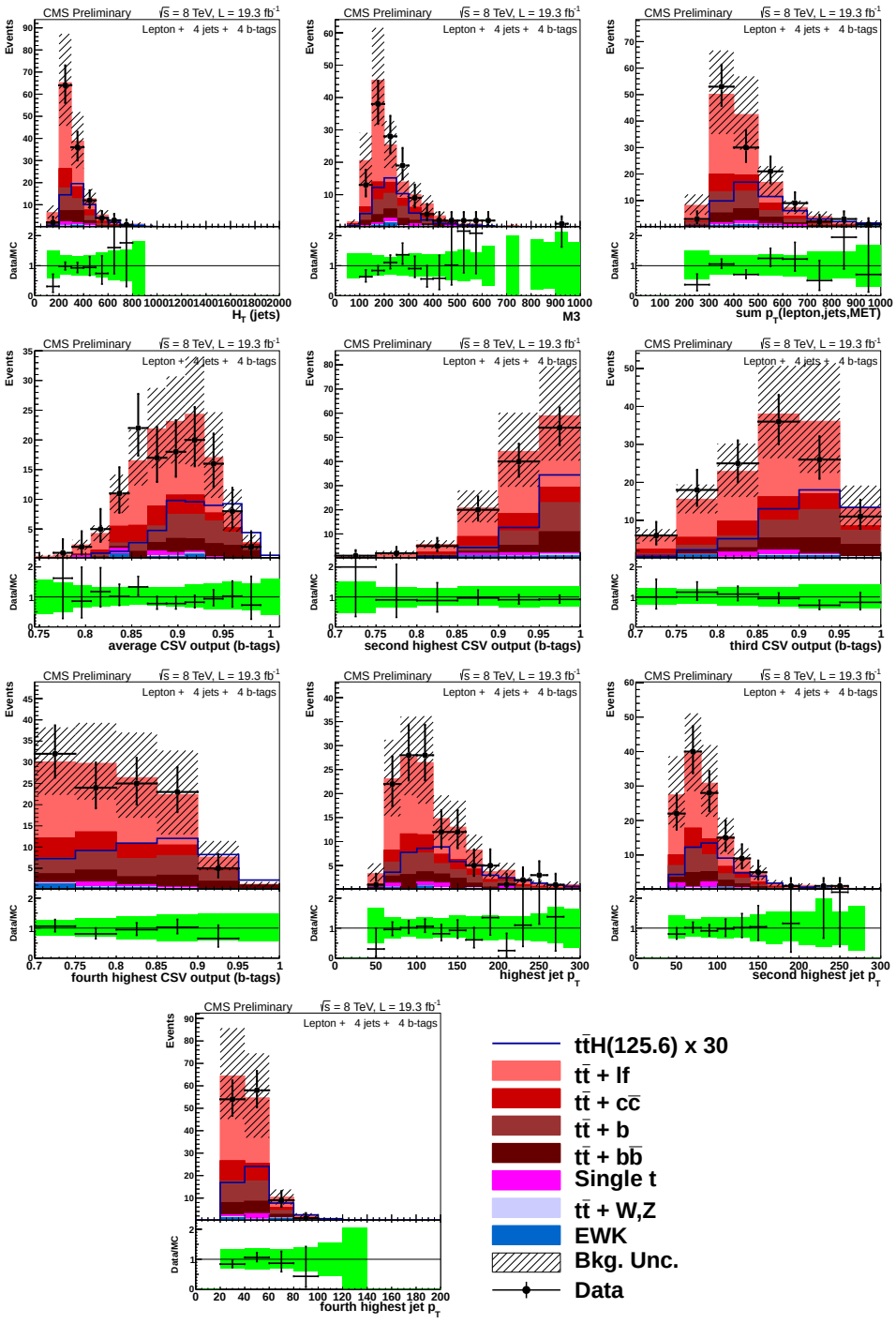


Figure A.6: Data/MC comparisons for events with one lepton and 4 jets + 4 b-tags. The uncertainty band includes statistical and systematic uncertainties that affect both the rate and shape of the background distributions.

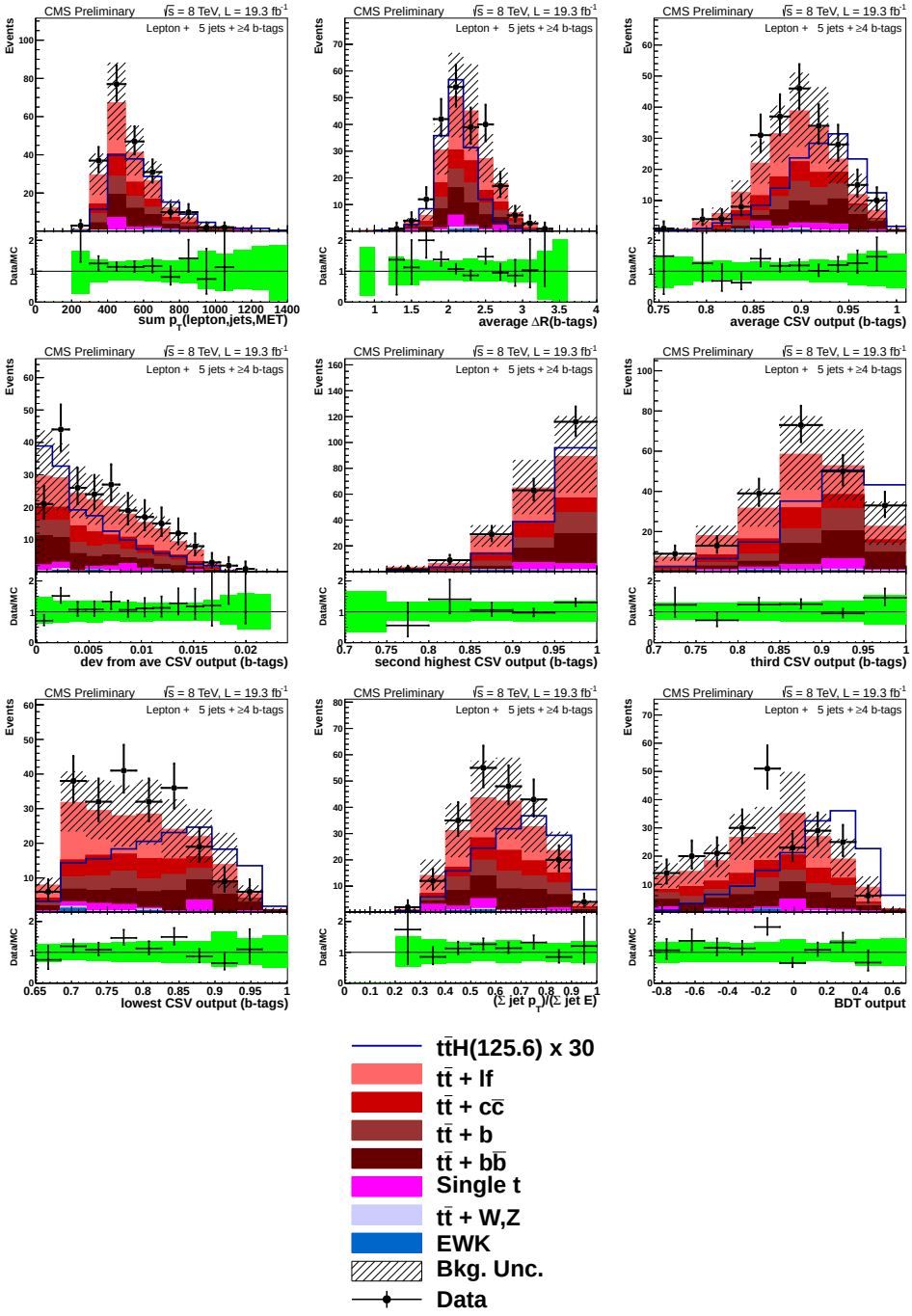


Figure A.7: Data/MC comparisons for events with one lepton and 5 jets +≥4 b-tags. The uncertainty band includes statistical and systematic uncertainties that affect both the rate and shape of the background distributions.

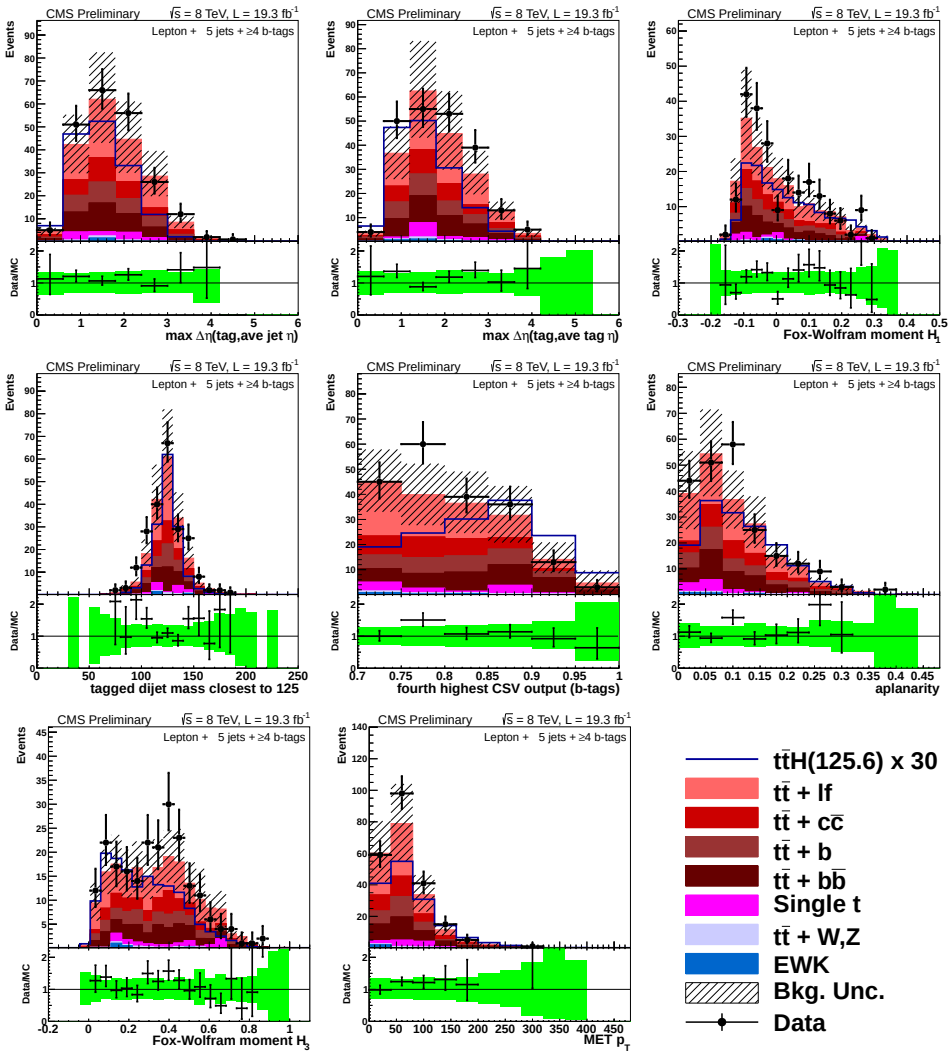


Figure A.8: Data/MC comparisons for events with one lepton and 5 jets $+\geq$4 b-tags. The uncertainty band includes statistical and systematic uncertainties that affect both the rate and shape of the background distributions.

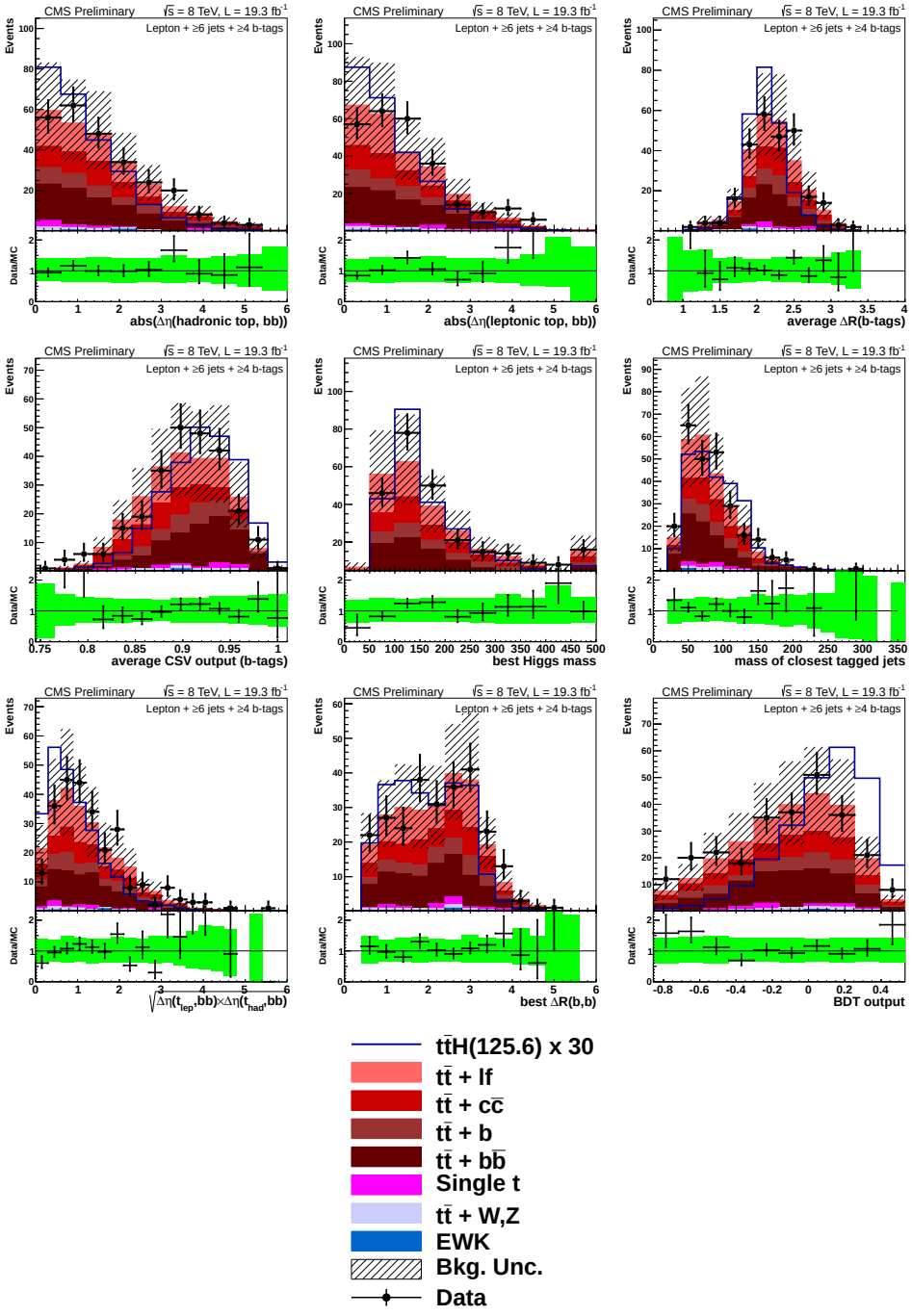


Figure A.9: Data/MC comparisons for events with one lepton and ≥6 jets + ≥4 b-tags. The uncertainty band includes statistical and systematic uncertainties that affect both the rate and shape of the background distributions.

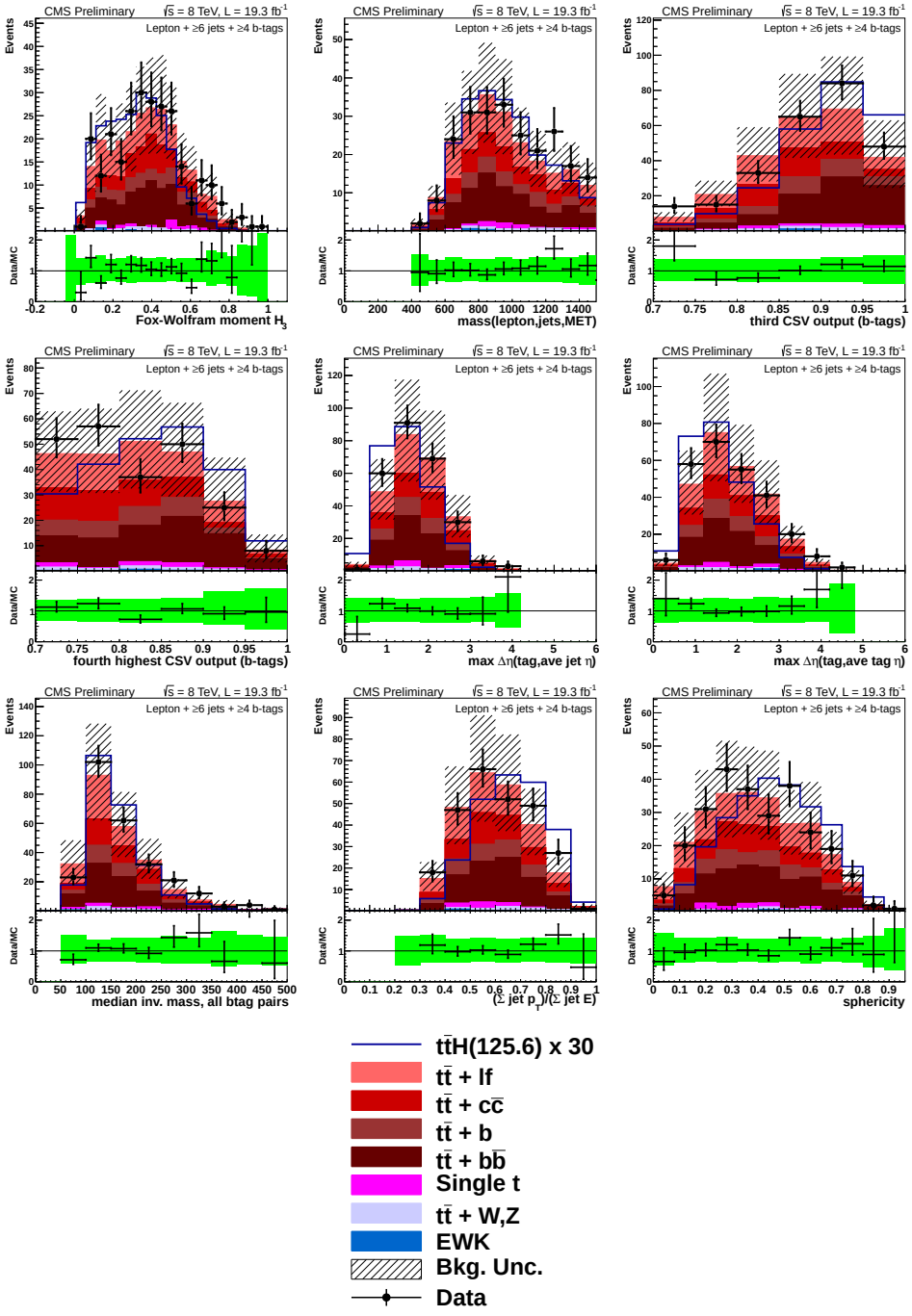


Figure A.10: Data/MC comparisons for events with one lepton and ≥6 jets + ≥4 b-tags. The uncertainty band includes statistical and systematic uncertainties that affect both the rate and shape of the background distributions.