HIGGS SEARCHES WITH BOTTOM QUARKS AND INVISIBLE PARTICLES

By

JIA FU LOW

A DISSERTATION PRESENTED TO THE GRADUATE SCHOOL
OF THE UNIVERSITY OF FLORIDA IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

UNIVERSITY OF FLORIDA

2015

I dedicate this to everyone who contributed to building the CMS experiment.

ACKNOWLEDGMENTS

I would like to thank everyone who has helped me in working on this analysis. I am especially grateful to my Ph.D. advisor Dr. Jacobo Konigsberg and my informal supervisor Dr. Michele de Gruttola at the University of Florida for their teaching and guidance. I am also truly grateful to Dr. Kenneth Bloom, Dr. Roger Kirby (both at the University of Nebraska at Lincoln) and Dr. Joe Incandela (at the University of California at Santa Barbara) for offering me research opportunities under them. I am really appreciative of encouragement and trust from my family and friends. I thank the colleagues involved in the LHC accelerator, the CMS detector, and the Fermilab LHC Physics Center for making this experiment at the highest energy frontier a great success. I acknowledge the support provided by the University of Florida and the U.S. Department of Energy.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Figure page

Abstract of Dissertation Presented to the Graduate School
of the University of Florida in Partial Fulfillment of the
Requirements for the Degree of Doctor of Philosophy

HIGGS SEARCHES WITH BOTTOM QUARKS AND INVISIBLE PARTICLES

By

Jia Fu Low

May 2015

Chair: Jacobo Konigsberg
Major: Physics

An essential search to answer the question of whether the newly discovered boson of mass 125 GeV at the Large Hadron Collider (LHC) is the standard model Higgs boson (H) is the search for decay of the boson into a pair of bottom quarks ($b\bar{b}$), as this is theoretically the dominant decay channel of a low-mass Higgs boson. For best signal-to-background sensitivity, the associated production of the Higgs boson with a vector boson ($V = W$ or Z) is used. The search is carried out in six channels based on the decay of the vector boson: $W(\mu\nu)H(b\bar{b})$, $W(e\nu)H(b\bar{b})$, $W(\tau\nu)H(b\bar{b})$, $Z(\mu\mu)H(b\bar{b})$, $Z(ee)H(b\bar{b})$, and $Z(\nu\bar{\nu})H(b\bar{b})$. This dissertation reports the results of this search, focusing on the details of the analysis for the $Z(\nu\bar{\nu})H(b\bar{b})$ channel. A data sample, recorded by the Compact Muon Solenoid (CMS) experiment at the LHC, that corresponds to 18.9 fb^{-1} integrated luminosity at the center-of-mass energy $\sqrt{s} = 8$ TeV is analyzed. The results of this channel is combined with those of the other channels, and with the earlier $\sqrt{s} = 7$ TeV results, to produce the final CMS Run I results. A mild excess of events is observed above the expected background with a local significance of 2.1 standard deviations. This is compatible with the expectation of 2.1 standard deviations when assuming the production of the standard model Higgs boson signal of mass 125 GeV. The measurements represent the first indication of the $H \to b\bar{b}$ decay at the LHC. These measurements are one of the key components in the global

fit of the Higgs couplings and contributed to the CMS Run I “legacy” publication that describes the properties of the discovered Higgs boson.

CHAPTER 1
INTRODUCTION

1.1 Overview

Thanks to hard work and dedication from countless physicists past and present, we have probed the structure of the universe down to very small scale. Our current understanding at this fundamental scale is described by the standard model (SM) of particle physics. However, it was missing a crucial particle to be a self-consistent model — the Higgs boson H. The Higgs boson is needed in order to explain why the W and Z bosons are massive, unlike the massless photon γ.

On July 4th, 2012, the discovery of a new boson that resembles the SM Higgs boson was announced simultaneously but independently by two experimental collaborations at the Large Hadron Collider (LHC): Compact Muon Solenoid (CMS), and A Toroidal LHC Apparatus (ATLAS). So far, measurements of the properties of this observed boson fit the SM predictions, although there are large uncertainties associated to these measurements. Peter Higgs and François Englert were awarded 2013 Nobel Prize in Physics for proposing the electroweak symmetry breaking mechanism that predicts the existence of the Higgs field and the Higgs boson.

This dissertation describes the search for the standard model Higgs boson, produced in association with a Z boson, in the decay channel of H to two b quarks and Z to two neutrinos, a.k.a. the $Z(\nu\bar{\nu})H(b\bar{b})$ channel. The final state is characterized by a pair of jets induced by b quark hadronization and large missing transverse energy, $E_{\mathrm{T}}^{\mathrm{miss}}$. The misnomer "missing transverse energy" refers to the imbalance of momentum in the transverse plane due to particles that escape the detector undetected (in this case, the neutrinos). A data sample, recorded by the Compact Muon Solenoid (CMS) experiment at the LHC, that corresponds to 18.9 fb^{-1} integrated luminosity at the center-of-mass energy $\sqrt{s} = 8$ TeV is analyzed. The results from this analysis contributed directly to the following CMS publications:

- Search for the VH production with $H \to b\bar{b}$ decay, a.k.a. $VH(b\bar{b})$ [28]. This is the main publication where the results of the $Z(\nu\bar{\nu})H(b\bar{b})$ channel was first published.
- Evidence for the Higgs decay into fermions [30].
- Measurement of WZ and ZZ production with $Z \to b\bar{b}$ decay [31].
- Search for invisible decays of Higgs boson [32].
- Run I legacy results for the properties of the Higgs boson [33].

This dissertation is organized as follows: Chapter 1 introduces the theoretical concepts and gives the motivation for this search; Chapter 2 provides a description of the CMS detector; Chapter 3 gives an account for the simulation tools necessary for making accurate theoretical predictions; Chapter 4 describes the event reconstruction techniques; Chapter 5 presents the details about the analysis strategy; Chapter 6 shows the results from the $Z(\nu\bar{\nu})H(b\bar{b})$ channel alone and from all $VH(b\bar{b})$ channels combined; Finally, Chapter 7 gives the conclusion and future prospects.

1.2 Standard Model of Particle Physics

Particle physics is the study of the elementary particles that make up matter (and antimatter for that matter) and the interactions among them. In the 20th century, particle physicists have come up with a self-consistent model that describes with high precision a vast array of experimental observations, and makes testable predictions. It is called the standard model of particle physics (SM). The elementary particles in the SM is depicted in Fig. 1-1.

Fermions, i.e. quarks [1] and leptons, are spin-$\frac{1}{2}$ particles that make up everyday matter. They are also referred to as "matter particles". For instance, a proton is a bound state of 2 up quarks and 1 down quark; an electron is the lightest charged lepton. There

[1] The name was coined by Murray Gell-Mann and came from the line "Three quarks for Muster Mark", as he postulated correctly that protons and neutrons are composite objects made up of more elementary particles.

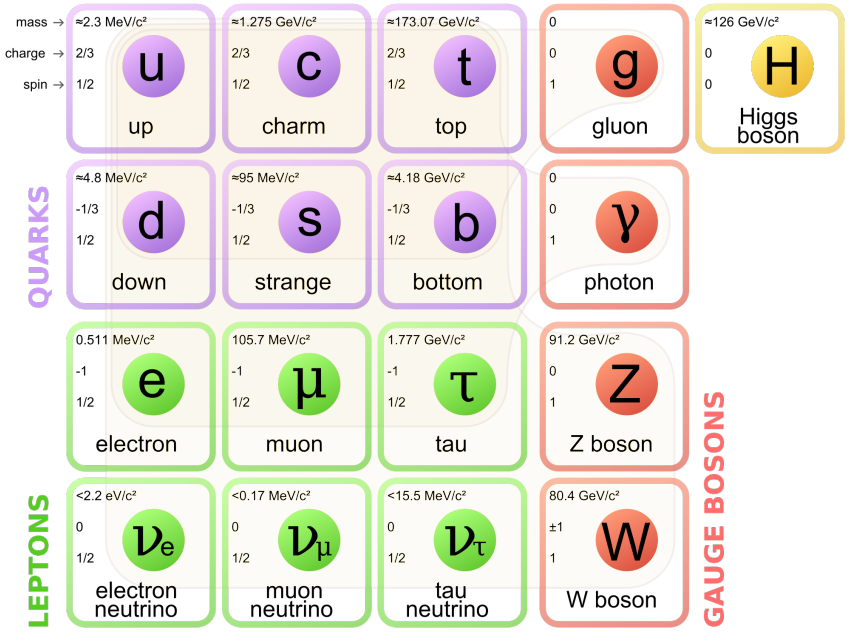

Figure 1-1. Elementary particles in the standard model [1]. There are six quarks, six leptons, four gauge bosons, and one Higgs boson in the current model.

are 3 generations of quarks, 2 flavors in each generation. Up and down quarks belong to the first generation; charm and strange belong to the second; top and bottom belong to the third. Second-generation quarks are heavier than first-generation ones (and thus unstable), but are otherwise identical. Third-generation quarks are in turns heavier than second-generation ones. Similarly, there are also 3 generations of leptons, one charged lepton and one neutrino (neutral lepton) in each generation. Electron and electron neutrino belong to the first generation; muon and muon neutrino belong to the second; tau and tau neutrino belong to the third. Why there are three generations and why there is a mass difference between generations remain a mystery.

Fermions interact with one another by exchanging gauge bosons, i.e. gluon, photon, $W^{\pm}$, and Z^0 bosons, which are spin-1 particles. A gauge boson acts as a carrier of a fundamental force, hence it is also called a “force particle”. We currently know of the existence of four fundamental forces in the universe: the strong (nuclear) force, the weak (nuclear) force, the electromagnetic force, and the gravitational force. Gluons carry the strong force and *glue* quarks together to form hadrons (e.g. protons, neutrons, pions).

Photon, $W^{\pm}$, and Z^0 bosons carry the electroweak force, which is the unification of the electromagnetic force and the weak force. The weak force is responsible for radioactivity and allows heavier quarks to decay via flavor changing processes (e.g. the beta decay), whereas the electromagnetic force, which unifies the electric and magnetic forces, is responsible for forming atoms and molecules. The gravitation force, although prevalent in our daily life, is still outside the SM due to lack of understanding of its quantum nature.

The scalar (spin-0) particle, i.e. the Higgs boson, is introduced into the SM as a way to generate masses of other elementary particles (except neutrinos). To understand why and how it does that, we need to turn to the quantum field theory (QFT) and treat each particle as a vibration of its field. At its heart, the standard model is a series of field equations derived from the gauge symmetries $\mathrm{SU}(3)_c \otimes \mathrm{SU}(2)_L \otimes \mathrm{U}(1)_Y$. The $\mathrm{SU}(3)_c$ symmetry group represents the strong interaction between colored quarks and gluons, described by the quantum chromodynamics (QCD) theory. The $\mathrm{SU}(2)_L \otimes \mathrm{U}(1)_Y$ group, which represents the unification of weak interaction and electromagnetism, is described by the electroweak theory for particles that carry the weak isospin and hypercharge quantum numbers. The group representations of the particle content of the SM are listed in Table 1-1.

Table 1-1. Group representations of the particle content of the standard model.

Field	$\mathrm{SU}(3)_c$	$\mathrm{SU}(2)_L$	$\mathrm{U}(1)_Y$
q_L	3	2	$1/3$
u_R	3	1	$4/3$
d_R	3	1	$-2/3$
l_L	1	2	-1
e_R	1	1	-2
G	8	1	0
W	1	3	0
B	1	1	0
Φ	1	2	1

Under the formulation of the electroweak theory, there are four massless gauge bosons: W^i, $i = 1, 2, 3$ from the $\mathrm{SU}(2)_L$ and B from the $\mathrm{U}(1)_Y$. However, this

contradicts the fact that the 3 vector bosons ($W^{\pm}$, Z^0) we observed are heavy, as W mass $\simeq$ 80.4 GeV and Z mass $\simeq$ 91.2 GeV. [2] It turns out that the electroweak symmetry is broken in our universe with low energy density (low relative to the universe's energy density a tiny fraction of a second after the Big Bang). The theory that explains this spontaneous electroweak symmetry breaking (EWSB), now known as the Brout–Englert–Higgs (BEH) mechanism, posits a scalar field, now known as the Higgs field, that permeates all space. Via interaction with this Higgs field, it is possible to include mass terms into the equations for the elementary particles (or fields), reconciling SM predictions with reality. The concept is based on analogies with the Bardeen-Cooper-Schrieffer theory of superconductivity. The BEH mechanism is described in more detail in the following section.

1.3 Brout–Englert–Higgs Mechanism

In his Nobel prize-winning paper published in 1964, Peter Higgs proposed the mechanism that explains how elementary particles acquire their masses via spontaneous electroweak symmetry breaking [34–36]. The same mechanism was also proposed independently by François Englert and Robert Brout [37], and Gerald Guralnik, C. Richard Hagen, and Tom Kibble [38, 39]. It was incorporated into the successful theory of electroweak unification, developed by Sheldon L. Glashow [40], Steven Weinberg [41] and Abdus Salam [42]. Gerard 't Hooft and Martinus Veltman later proved that the theory is in fact renormalizable [43, 44].

According to the BEH mechanism, the Higgs field has a potential as illustrated in Fig. 1-2. The potential has a rotational symmetry, meaning it is invariant when rotated around the vertical axis. However, the origin ($\Phi = 0$) is not the minimum of the potential.

[2] Particle physicists typically adopt the natural units where the speed of light and the Planck's constant are set to one: $c = 1$ and $\hbar = 1$. Therefore, mass, momentum, and energy are all measured in units of electron volt (eV).

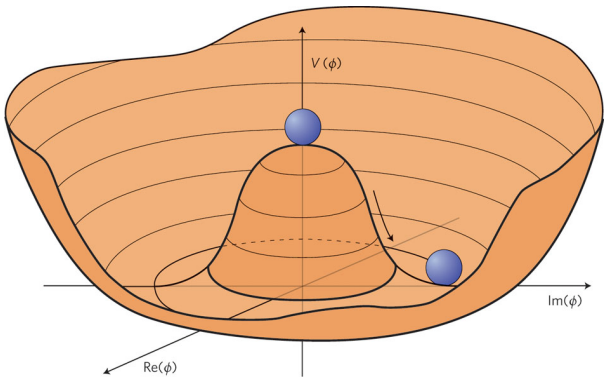

Figure 1-2. Higgs field potential Φ that has a shape resembling a Mexican sombrero [2].

When the ball is at the origin, sitting at the top of the "hill", the ball-and-potential system is still symmetrical under rotation. However, the ball is at an unstable equilibrium. If perturbed slightly, it will roll down the hill until it reaches the "valley", where the potential energy is lowest. As the ball now has a defined direction w.r.t. the potential, the system has lost its rotational symmetry.

Reviews of the mathematical details of the BEH mechanism can be found in the literature [19, 45]. The central idea is to write down the interaction potential of the Higgs field in a renormalizable form:

$$V(\Phi) = \mu^2 \Phi^\dagger \Phi + \lambda \left(\Phi^\dagger \Phi\right)^2 \tag{1–1}$$

where the Higgs field $\Phi \equiv \begin{pmatrix} \phi^+ \\ \phi^0 \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} \phi_1 + i\phi_2 \\ \phi_3 + i\phi_4 \end{pmatrix}$ is a complex doublet that comprises four degrees of freedom (dof), and μ^2, λ are some constants ($\lambda > 0$ by convention). The first term in the potential is the mass term, the second is the self-interaction term. This potential has a global $\mathrm{SU}(2)$ symmetry.

Consider the case $\mu^2 > 0$. $V(\Phi)$ has a "U" shape with ground state at $\Phi = 0$. This is analogous to electrodynamics (QED) plus a charged scalar field. The more interesting case is $\mu^2 < 0$ where $V(\Phi)$ has the shape of a Mexican hat as described previously. The ground state is one of the points on the circle along the valley. We can use a unitary

gauge, choosing ϕ^+ to be zero and ϕ^0 to be real: $\langle\Phi\rangle = \frac{1}{\sqrt{2}}\begin{pmatrix}0\\v\end{pmatrix}$ where $v \equiv \sqrt{-\mu^2/\lambda}$ is the vacuum expectation value (VEV).

The introduction of the Higgs field in the SM yields the Lagrangian term:

$$\mathcal{L}_{\text{H}iggs} = \left(\mathcal{D}^\mu\Phi\right)^\dagger\left(\mathcal{D}_\mu\Phi\right) - V(\Phi) \tag{1–2}$$

The first term is the kinetic energy term with covariant derivative

$$\mathcal{D}_\mu = \partial_\mu + i\frac{g}{2}\sigma^a W_\mu^a - i\frac{g'}{2}YB_\mu \tag{1–3}$$

where $W_\mu^a,\ a = 1,2,3$ and B_μ are the respective $\mathrm{SU}(2)$ and $\mathrm{U}(1)$ gauge fields, g and g' are the gauge couplings, and σ^a are the Pauli matrices. Y is the weak hypercharge quantum number (with the convention $Q = T_3 + Y/2$, Q being the electric charge, and T_3 being the 3rd component of the weak isospin).

The Higgs field can be rewritten in the unitary gauge as $\Phi = \frac{1}{\sqrt{2}}\begin{pmatrix}0\\v+h\end{pmatrix}$ under small perturbations near the minimum of the potential, where h is a real scalar field. Substituting it into the kinetic energy term in Eq. 1–2, assuming $Y = +1$, we get:

$$\frac{1}{2}(0, v)\left(\frac{g}{2}\sigma^a W_\mu^a + \frac{g'}{2}B_\mu\right)^2\begin{pmatrix}0\\v\end{pmatrix} \tag{1–4}$$

If we diagonalize it, the massless gauge fields become the physical gauge fields:

$$W_\mu^\pm = \frac{1}{\sqrt{2}}\left(W_\mu^1 \mp iW_\mu^2\right) \tag{1–5}$$

$$Z_\mu = \frac{gW_\mu^3 - g'B_\mu}{\sqrt{g^2 + g'^2}} \tag{1–6}$$

$$A_\mu = \frac{g'W_\mu^3 + gB_\mu}{\sqrt{g^2 + g'^2}} \tag{1–7}$$

with masses:

$$m_W = \frac{1}{2}gv,\ \ m_Z = \frac{1}{2}\sqrt{g^2 + g'^2}v,\ \ m_A = 0 \tag{1–8}$$

In the SM, v is fixed by the Fermi constant G_F by the equation $v = \left(\sqrt{2}G_F\right)^{-1/2} \approx 246\,\text{GeV}$; g and g' are fixed by the fine structure constant α and the experimentally measurement of m_Z.

In essence, the BEH mechanism states that three degree of freedoms of the Higgs field Φ are used to generate longitudinal polarizations for the massless gauge bosons, turning them into massive gauge bosons. The generator associated to the $\mathrm{U}(1)$ symmetry gives a massless photon.

One remaining degree of freedom of Φ turns into a physical scalar field, which we call the Higgs boson H, with mass:

$$m_H = \sqrt{2\lambda}v \tag{1–9}$$

The origin of m_H is currently unexplained in the SM, so m_H (or equivalently, the Higgs self-coupling parameter λ) remains as a free parameter.

To summarize, in an unbroken electroweak sector, there are massless W^a_μ, B_μ, and complex Φ (3×2, 2, and 4 dof, respectively); after the (minimal case of) spontaneous EWSB, we end up with massive $W^\pm$, Z^0, massless γ, and massive H (2×3, 3, 2, and 1 dof, respectively). The total number of degrees of freedom is always 12.

1.4 Fermion Masses

The BEH mechanism explains the gauge boson masses, but where do the fermion masses come from? The simplest way of including a mass term to a fermion field ψ is to write down a Lagrangian term as such:

$$\mathcal{L} = -m\bar{\psi}\psi = -m\left(\bar{\psi}_L\psi_R + \bar{\psi}_R\psi_L\right) \tag{1–10}$$

where $\psi_{L,R} = \frac{1}{2}\left(1 \mp \gamma_5\right)\psi$ are the left- and right-handed projections. But this term violates gauge invariance, as ψ_L and ψ_R follow different $\mathrm{SU}(2)$ and $\mathrm{U}(1)$ transformations. If the Higgs field exists, it is possible to generate fermion masses through Yukawa

interactions [46, 47]. The general form of such a Lagrangian term is:

$$\mathcal{L} = -m\,(\bar{\psi}_L \Phi \psi_R + \bar{\psi}_R \bar{\Phi} \psi_L) \tag{1–11}$$

It can be shown that this term is invariant under the $\mathrm{SU}(2) \otimes \mathrm{U}(1)$ gauge symmetry.

In the SM, the Lagrangian term for the Yukawa interaction between the Higgs field and the fermionic fields is:

$$\mathcal{L}_{Yukawa} = -\hat{h}_{d_{ij}} \bar{q}_{L_i} \Phi d_{R_j} - \hat{h}_{u_{ij}} \bar{q}_{L_i} \tilde{\Phi} u_{R_j} - \hat{h}_{l_{ij}} \bar{l}_{L_i} \Phi e_{R_j} + h.c. \tag{1–12}$$

where q_L and l_L are the left-handed quark and lepton doublets, d_R, u_R, e_R are the right-handed quark and lepton singlets, and $\tilde{\Phi} = i\sigma_2 \Phi^*$. For the d-type quark, after spontaneous EWSB and diagonalization of the terms, $\hat{h}_{d_{ij}} \rightarrow \lambda_d \mathbf{1_{3\times 3}}$, it receives an effective coupling:

$$-\frac{1}{\sqrt{2}} \lambda_d \bar{q}_{L_i} \begin{pmatrix} 0 \\ v + h \end{pmatrix} d_R + h.c. \tag{1–13}$$

We identify as the mass term:

$$m_d = \frac{1}{\sqrt{2}} \lambda_d v \tag{1–14}$$

The Yukawa coupling constant λ_d is a free parameter which has to be determined experimentally. Note that the mass is directly proportional to the coupling constant.

Similarly, u-type quark and charged lepton e also acquire their masses from the Higgs field through Yukawa interactions. Since there is no right-handed neutrino in the SM, neutrinos should be massless. The observations of neutrino oscillations [48–51], which evidently show that neutrinos have non-zero masses, is outside of the SM.

1.5 Higgs Production and Decay Mechanisms

The SM is a fully calculable effective theory once all its 19 free parameters are fixed. Prior to the turn on of the LHC, 18 of them had been experimentally measured. The remaining free parameter is the Higgs boson mass m_H. Although there exists indirect constraints as to what the value of m_H can be, a direct observation of the Higgs

boson is necessary to ensure that the SM is the "correct" theory in our universe with low energy density. [3]

The SM Higgs boson is predicted to have zero spin, even parity, zero electric charge, and zero color charge. Its couplings to other elementary particles are summarized in Fig. 1-3. At a proton-proton or proton-antiproton collider, the four most important Higgs boson production modes are: gluon fusion (ggH), vector boson fusion ($q\bar{q}H$), associated production with a vector boson (VH), and associated production with top quarks ($t\bar{t}H$). In ggH, the Higgs boson is produced via a virtual top-quark loop; in $q\bar{q}H$, the Higgs boson is produced in association with a quark-antiquark pair; in VH and $t\bar{t}H$, the Higgs boson is produced either in association with a W/Z boson or with a top-antitop quark pair. Fig. 1-4 shows the Feynman diagrams of these production modes.

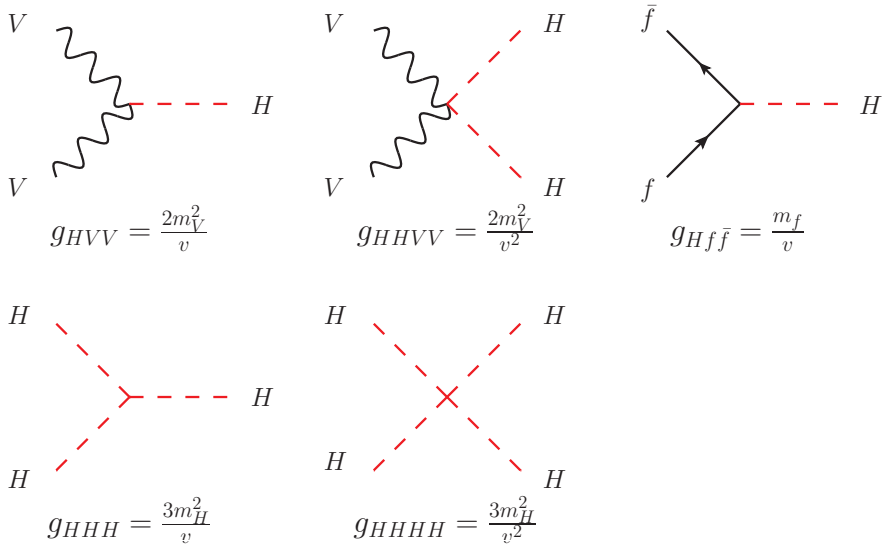

Figure 1-3. Feynman diagrams of the interactions between the Higgs boson and other standard model particles.

[3] Note that we already know that the SM is incomplete, as it fails to explain the absence of antimatter in our observable universe, the dark matter, the gravitational force, the mass of the neutrinos, etc. In fact, there is another problem in the SM, often referred to as the hierarchy problem, as theoretically the Higgs boson mass ought to be close to the Planck mass (about 10^{18} GeV) due to quantum corrections, instead of at the electroweak scale (about 10^2 GeV).

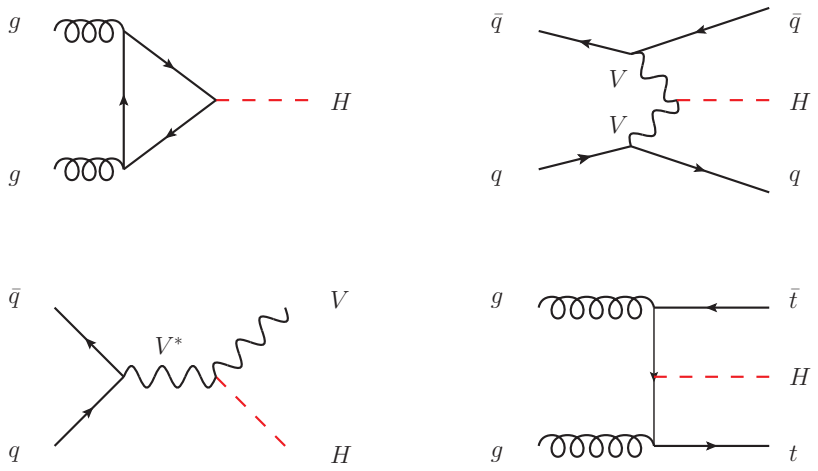

Figure 1-4. Feynman diagrams of the Higgs production modes at the LHC. From left to right, top to bottom: gluon-gluon fusion, vector boson fusion, associated production with a vector boson, associated production with top quarks.

Theorists have calculated the Higgs production cross sections at $\sqrt{s} = 8\,\text{TeV}$ and decay branching fractions at the LHC by scanning a range of m_H values [3, 4]. These are plotted in Fig. 1-5.

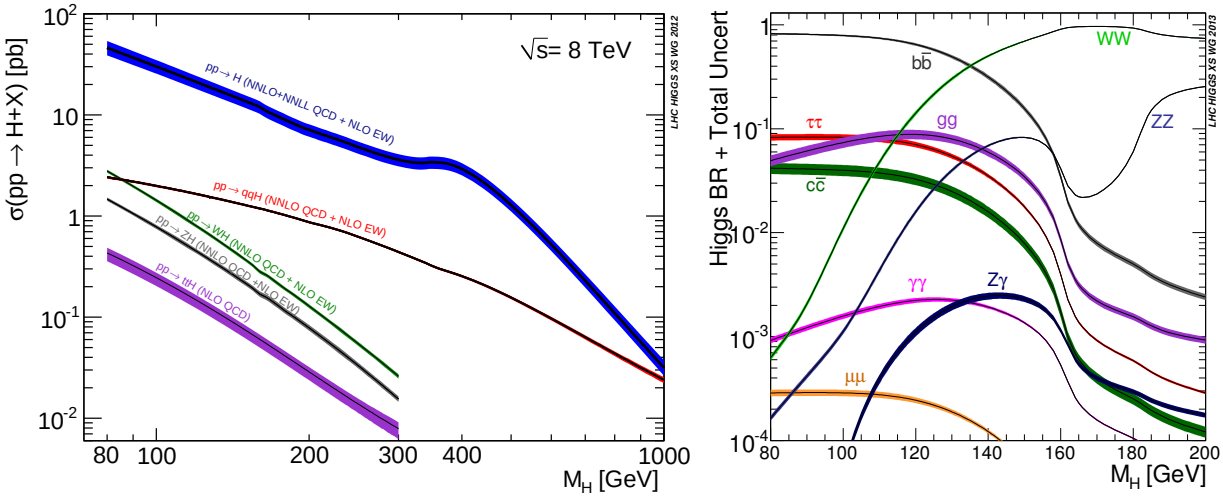

Figure 1-5. Left: Higgs production cross sections as a function of the Higgs boson mass at the LHC running at center-of-mass energy $\sqrt{s} = 8\,\text{TeV}$. Right: Higgs decay branching fractions as a function of the Higgs boson mass [3, 4].

For m_H =125 GeV, ggH mode has the highest production cross section, followed by $q\bar{q}H$, WH, ZH, and $t\bar{t}H$:

$$\begin{aligned}
\sigma(ggH) &= 19.27^{+7.2\%+7.5\%}_{-7.8\%-6.9\%}\,\text{pb},\\
\sigma(q\bar{q}H) &= 1.578^{+0.2\%+2.6\%}_{-0.2\%-2.8\%}\,\text{pb},\\
\sigma(WH) &= 0.7046^{+1.0\%+2.3\%}_{-1.0\%-2.3\%}\,\text{pb},\\
\sigma(ZH) &= 0.4153^{+3.1\%+2.5\%}_{-3.1\%-2.5\%}\,\text{pb},\\
\sigma(t\bar{t}H) &= 0.1293^{+3.8\%+8.1\%}_{-9.3\%-8.1\%}\,\text{pb}
\end{aligned}$$

where the associated uncertainties are due to variations in QCD scale, and in parton density functions (PDFs) and strong coupling constant α_s (to be described in Chapter 3.3). Note that a barn is a unit of cross-sectional area ($1\,\text{b} = 10^{-24}\,\text{cm}^2$).

The predicted branching fractions for the Higgs boson of 125 GeV mass are (from highest to lowest):

$$\begin{aligned}
\mathcal{B}(H\to b\bar{b}) &= 0.577^{+3.21\%}_{-3.27\%}, & \mathcal{B}(H\to W^+W^-) &= 0.215^{+4.26\%}_{-4.20\%},\\
\mathcal{B}(H\to gg) &= 0.0857^{+10.22\%}_{-9.98\%}, & \mathcal{B}(H\to \tau^+\tau^-) &= 0.0632^{+5.71\%}_{-5.67\%},\\
\mathcal{B}(H\to c\bar{c}) &= 0.0291^{+12.17\%}_{-12.21\%}, & \mathcal{B}(H\to ZZ) &= 0.0264^{+4.28\%}_{-4.21\%},\\
\mathcal{B}(H\to \gamma\gamma) &= 0.00228^{+4.98\%}_{-4.89\%}, & \mathcal{B}(H\to Z\gamma) &= 0.00154^{+9.01\%}_{-8.83\%},\\
\mathcal{B}(H\to \mu^+\mu^-) &= 0.000219^{+6.01\%}_{-5.86\%}
\end{aligned}$$

Direct searches for the Higgs boson at particle colliders began with the Large Electron Positron (LEP) experiments at CERN. In 2003, they established an exclusion limit on the Higgs boson mass that excludes $m_H < 114.4$ GeV at the 95% confidence level (C.L.) [52].

Tevatron experiments, CDF and D0, then excluded at the 95% C.L. two mass ranges: $90 < m_H < 109$ GeV and $149 < m_H < 182$ GeV using about 10 fb^{-1} of $\sqrt{s} = 1.96$ TeV $p\bar{p}$ collision data [53]. In the mass range between 115 and 140 GeV,

they reported an excess of events corresponding to a local significance of 3.0 standard deviations (denoted by 3σ) at $m_H = 125\,\text{GeV}$. Assuming the SM Higgs boson is present with 125 GeV mass, the expected local significance is 1.9σ. The best-fit signal strength, defined as the ratio of the observed signal yield to the SM expectation, is $1.44^{+0.59}_{-0.56}$, which is within 1 standard deviation. In that search, the sensitivity for the Higgs boson mass below 130 GeV is strongly driven by the $H \to b\bar{b}$ decay channels [54].

On July 4th 2012, the discovery of a Higgs boson-like particle with $m_H \approx 125\,\text{GeV}$ was announced simultaneously but independently by the ATLAS and CMS collaborations at the LHC [55, 56]. In the early data analyses, signals were observed most strongly in the decay modes of $H \to ZZ$ and $H \to \gamma\gamma$. ATLAS observed a combined significance of 5.9σ including all the high-priority decay channels, while CMS observed a combined significance of 4.9σ. With the full 7 TeV and 8 TeV Run I data, ATLAS and CMS have performed comprehensive measurements of the properties of the observed Higgs boson including its mass, coupling constants, spin-parity (J^P) quantum numbers, etc. All the measurements are so far consistent with the SM predictions. In particular, the combined signal strength is $1.00 \pm 0.09\,(\text{stat})^{+0.08}_{-0.07}\,(\text{theo}) \pm 0.07\,(\text{syst})$ for CMS [33] and $1.18 \pm 0.10(\text{stat})^{+0.08}_{-0.07}(\text{theo}) \pm 0.07(\text{syst})$ for ATLAS [57]. Using the high-resolution ZZ and $\gamma\gamma$ channels, the Higgs boson mass m_H is measured to be $125.02^{+0.26}_{-0.27}\,(\text{stat})^{+0.14}_{-0.15}\,(\text{syst})\,\text{GeV}$ for CMS [33] and $125.36 \pm 0.37\,(\text{stat}) \pm 0.18\,(\text{syst})\,\text{GeV}$ for ATLAS [58].

Consistency of the $H \to W^+W^-$ and $H \to ZZ$ couplings and the Higgs spin-parity quantum numbers with the SM expectation suggests that the observed boson very likely plays a role in electroweak symmetry breaking. Furthermore, the direct measurements of the Higgs couplings to fermions indicate that the observed boson also very likely serves as the source of fermion mass generation through Yukawa interaction. CMS observed an excess significance of 3.8σ (expected 4.4σ) when combining the results of the $b\bar{b}$ and $\tau^+\tau^-$ channels [30]. ATLAS observed a 4.5σ significance (expected 3.4σ)

using only the $\tau^+\tau^-$ channel [59]. In the $b\bar{b}$ channel alone, the observed significance is 2.3σ (expected 2.1σ) for CMS [28], and 1.4σ (expected 2.6σ) for ATLAS [60].

1.6 $Z(\nu\bar{\nu})H(b\bar{b})$ Channel

$H \to b\bar{b}$ is the dominant Higgs decay channel for $m_H \lesssim 2m_W$ (as shown in Fig. 1-5). Along with $H \to \tau^+\tau^-$, they probe the Yukawa interaction between the Higgs boson and the down-type fermions. The study of these decays are integral to our understanding of the origin of the fermion masses.

Due to color confinement, quarks and gluons cannot exist in isolation but very quickly form a spray of highly collimated color-neutral hadrons via a process called "hadronization". The kinematics of a quark can be recovered by clustering the hadron products as a "jet", and measuring the jet kinematics. Hadronization of b quark ("b jet") is distinct as it can have displaced tracks and/or a secondary vertex within (to be described in Chapter 4.8). Even so, in the final state with 2 b tagged jets, QCD multijet process, $pp \to b\bar{b}$, presents an overwhelming background that has 10^7 times higher cross section.

One way to drastically improve the signal-to-background ratio is to select the VH production mode instead, in which the Higgs boson is produced in association with a vector boson V that decays into leptons and/or neutrinos. The final state consists of 2 b jets plus the decay products of V. Although this production mode has $\sim$ 100 times smaller cross section than the dominant production mode ggH, the QCD background can be reduced to negligible levels. However, significant background still exists, primarily arising from production of W and Z bosons in association with jets, singly and pair-produced top quarks, and dibosons. Representative leading-order Feynman diagrams of some of these background processes that can mimic the VH signature are displayed in Fig. 1-6. Except for single top and dibosons, these processes have production cross sections that are several orders of magnitude larger than VH

production. Cross sections of various SM processes have been measured by CMS, and they are summarized in Fig. 1-7.

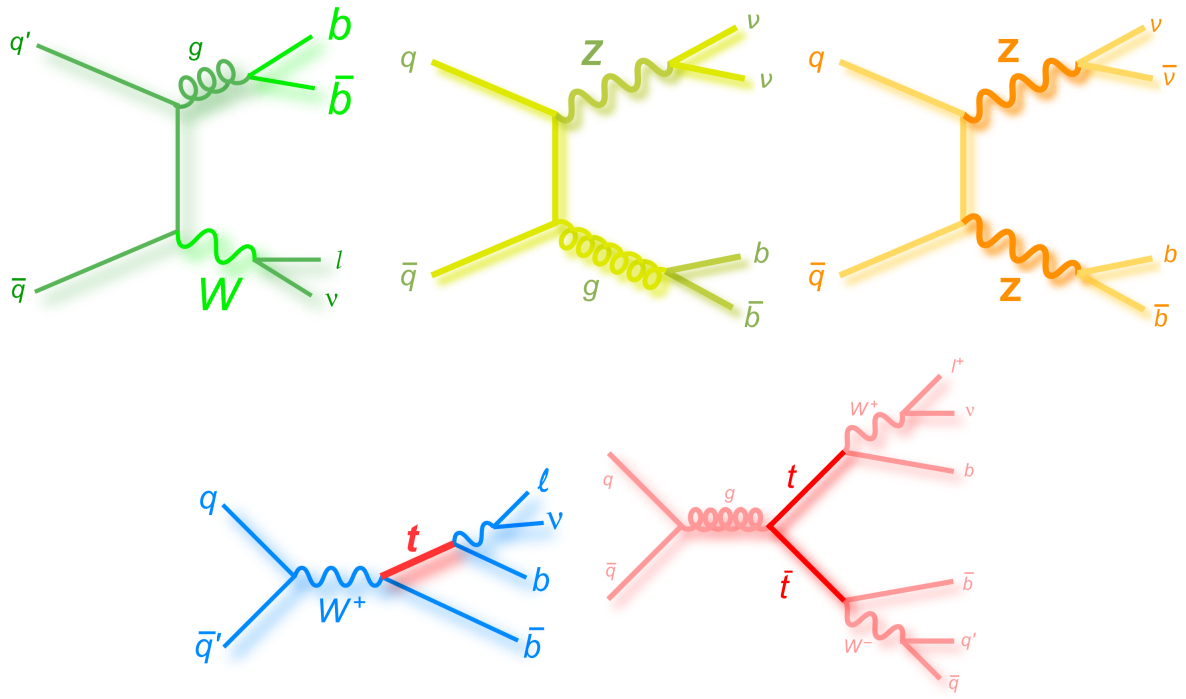

Figure 1-6. Representative leading-order Feynman diagrams of various background production processes: W+jets (top left), Z+jets (top center), ZZ (top right), s-channel single top (bottom left), top pairs (bottom right) [5].

In the CMS $VH(b\bar{b})$ analysis, 6 different channels are studied: $W(\mu\nu)H(b\bar{b})$, $W(e\nu)H(b\bar{b})$, $W(\tau\nu)H(b\bar{b})$, $Z(\mu\mu)H(b\bar{b})$, $Z(ee)H(b\bar{b})$, and $Z(\nu\bar{\nu})H(b\bar{b})$. As neutrinos can escape the detector without a trace, its presence can only be inferred from the missing transverse energy $E_{\mathrm{T}}^{\mathrm{miss}}$. $E_{\mathrm{T}}^{\mathrm{miss}}$ is the magnitude of the missing transverse momentum vector $\vec{p}_{\mathrm{T}}^{\,\mathrm{miss}}$, which is defined as the negative of the vectorial sum of the transverse momenta of all reconstructed particles in a given event. This dissertation describes the analysis in the particular $Z(\nu\bar{\nu})H(b\bar{b})$ channel, using the 2012 $\sqrt{s} = 8$ TeV data. To enhance the sensitivity, a multivariate discriminator is used, combining various discriminating variables that help separating signal events from background. A binned maximum likelihood fit is performed on the output distribution of the discriminator in real data, using the signal and background templates from the simulation as input

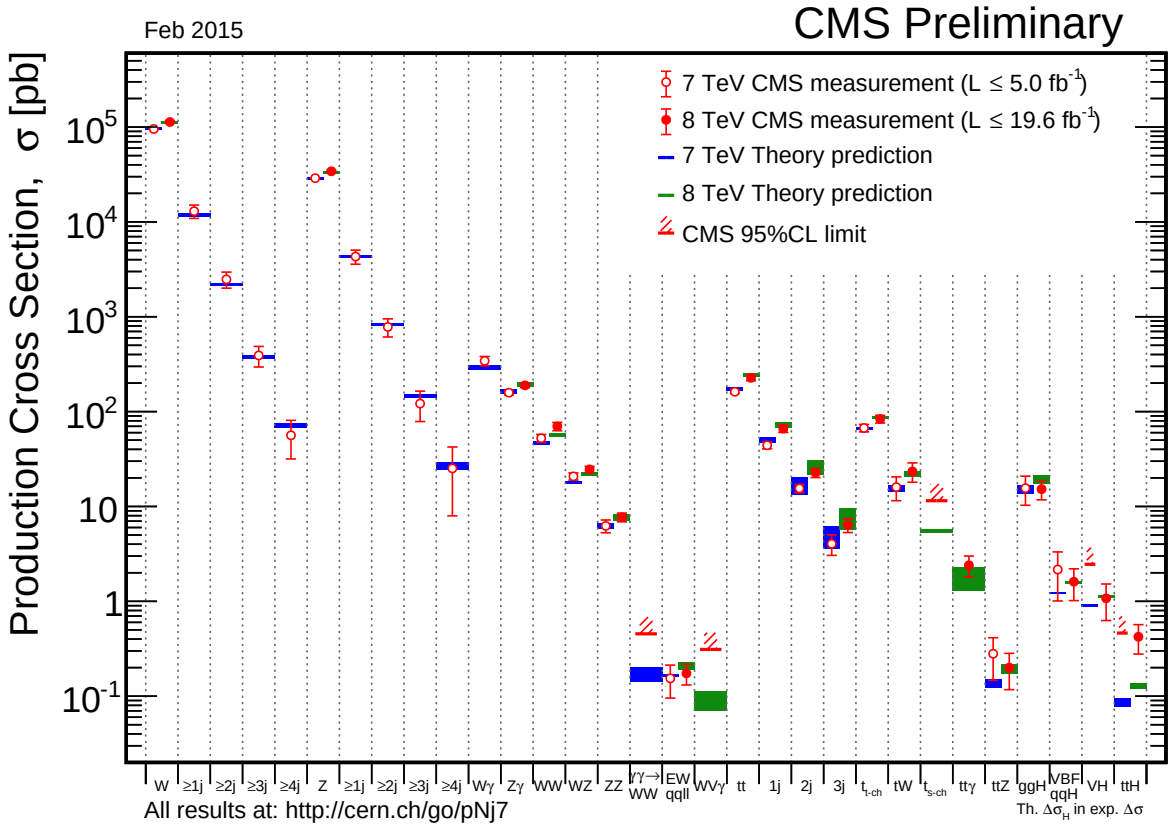

Figure 1-7. Summary plot of CMS cross section measurements of various standard model processes [6].

and incorporating their associated uncertainties. Upper limits at the 95% C.L. on the *VH* production cross section times the $H \to b\bar{b}$ branching fraction are computed for m_H in the 110–135 GeV range, using the $Z(\nu\bar{\nu})H(b\bar{b})$ channel alone and using all the channels combined. The significance of any excess of events above the expectation from the background processes, as well as the excess compatibility with the SM Higgs boson signal, are also evaluated. Previous analysis carried out on the $\sqrt{s} = 7$ TeV data that was published in Ref. [62] is included into the final results.

CHAPTER 2
EXPERIMENTAL APPARATUS

2.1 Large Hadron Collider

To find something as elusive as the Higgs boson, physicists need a powerful machine. The Large Hadron Collider is a proton-proton (*pp*) collider that can accelerate two proton beams up to 4 TeV per beam and collide them at a center-of-mass energy of $\sqrt{s} = 8$ TeV during "Run I" (2010-2013). It has achieved a peak instantaneous luminosity, i.e. rate of collision per cross-sectional unit, of $7.7 \times 10^{33}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$. Currently, it is the particle collider with the highest energy and highest instantaneous luminosity in the world. With future upgrades, it will reach the design *pp* collision energy of 14 TeV and likely exceed the design instantaneous luminosity of $1 \times 10^{34}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$.

The LHC is the latest addition in the accelerator complex at the European Organization for Nuclear Research, a.k.a. CERN, in Geneva, Switzerland (see Fig. 2-1). Its accelerator ring is 27 km in circumference, [1] which is *large*, in a tunnel about 100 m underground. [2] The protons are sent in bunches, guided around the accelerator ring by thousands of superconducting electromagnets that are cooled to -271.3° C or 1.9 K. One of these superconducting electromagnets is displayed in Fig. 2-2. Besides *pp* collisions, the LHC also made lead-lead collisions at $\sqrt{s_{\mathrm{NN}}} = 2.76$ TeV and lead-proton collisions at $\sqrt{s_{\mathrm{NN}}} = 5.02$ TeV.

The LHC was approved for construction in 1994. Four experiments were originally conceived: A Large Ion Collider Experiment (ALICE), A Toroidal LHC ApparatuS (ATLAS), Compact Muon Solenoid (CMS), and Large Hadron Collider beauty (LHCb). ALICE studies heavy ion physics including quark-gluon plasma formation. LHCb studies

[1] Due to gravity of the Moon on the Earth's crust, the circumference can vary by 1 mm. This change must be taken into account in calculating the beam energy [67].

[2] The same tunnel used to house the LEP experiment.

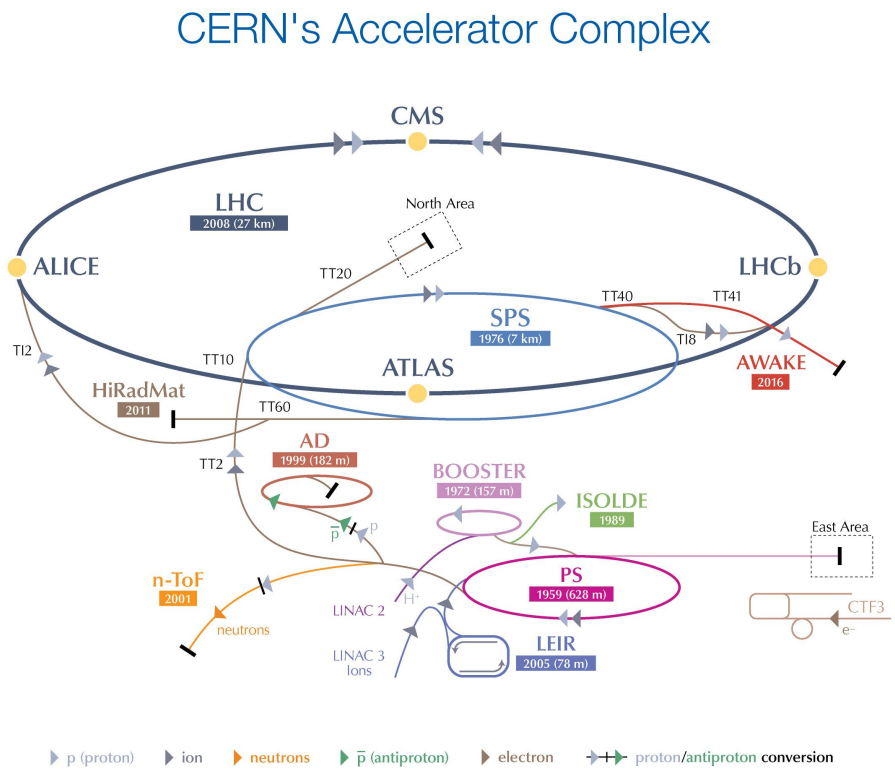

Figure 2-1. CERN accelerator complex and the LHC experiments [7].

Figure 2-2. 3D-rendered cut-out section of a LHC superconducting dipole magnet [8].

B meson physics and investigates matter-antimatter asymmetry. ATLAS and CMS are two general-purpose detectors that study a wide range of SM processes and new physics processes, including but not limited to: Higgs search, supersymmetry, extra dimensions, and dark matter. ATLAS and CMS are the largest experiments, each involving more than 3000 physicists and engineers from many different countries.

The LHC saw its first beam circulation in 2008. However, an electrical problem with the magnet connections damaged some of its superconducting magnets and caused the LHC to shut down. It restarted in late 2009 and made the first *pp* collisions at $\sqrt{s} = 0.9$ TeV. In 2010, the collision energy was ramped up to 7 TeV; in 2012, it was raised again to 8 TeV. It concluded Run I on Feb 14th, 2013, entering the first long shutdown for maintenance and upgrade work.

The number of events N produced by a collider machine can be written as:

$$N = \sigma \cdot \int \mathcal{L}(t)\, \mathrm{d}t \tag{2–1}$$

where σ is the *pp* interaction cross section, and $\mathcal{L}$ is the instantaneous luminosity. Interaction cross sections of different processes at a *pp* or $p\bar{p}$ collider as a function of $\sqrt{s}$ are shown in Fig. 2-3. The total cross section at $\sqrt{s} = 8$ TeV at the LHC was measured to be 101.7 ± 2.9 mb, with a value for the inelastic cross section of 74.7 ± 1.7 mb [68].

$\mathcal{L}$ is the most important parameter for a collider after $\sqrt{s}$. Assuming both beams have an identical Gaussian profile in the transverse plane, it can be expressed as [69]:

$$\mathcal{L} = \frac{N_p^2 k_b f_{\mathrm{rev}}}{4\pi\sigma_x^*\sigma_y^*} F \tag{2–2}$$

$$= \frac{N_p^2 k_b f_{\mathrm{rev}} \gamma}{4\pi\epsilon_n \beta^*} F \tag{2–3}$$

where N_p is the number of protons per punch, k_b is the number of bunches, f_{rev} is the revolution frequency, γ is the relativistic Lorentz boost factor, σ_x^* and σ_y^* are the horizontal and vertical beam sizes at the interaction point (IP), ϵ_n is the normalized

transverse beam emittance, β^* is the beta function at the IP, and F is the geometrical reduction factor due to the crossing angle between proton beams. β^* and ϵ_n control the transverse beam size at the IP.

In year 2012, the bunch spacing of 50 ns was used, which allows the maximum number of bunches of 1380. The design bunch spacing is 25 ns, which allows more bunches but introduces potential difficulties in electronic readout and noise suppression. The integrated luminosities delivered by the LHC to the four experiments in 2012, as well as the amount recorded successfully by CMS, are shown in Fig. 2-4. CMS recorded 21.79 fb^{-1} out of 23.30 fb^{-1} (93.5%). The mean number of *pp* interactions per bunch crossing was approximately 21 and reached up to 40, as shown in Fig. 2-5. The additional interactions overlapping with the collision event of interest are labeled as "pileup". There are two classes of pileup interactions: in-time pileup refers to those in the same bunch crossing as the collision of interest, and out-of-time (OOT) pileup refers to those in the bunch crossings just before and after the collision of interest. Pileup effects can deteriorate energy measurements and particle identifications. Full description of the LHC technical design can be found in Ref. [70].

2.2 Compact Muon Solenoid Detector

The Compact Muon Solenoid apparatus is located at 'Point 5' (one of the 8 interaction points) of the LHC ring at Cessy, France. It looks like a giant cylindrical onion, with multiple detector layers built around and inside a huge superconducting solenoid, as shown in Fig. 2-6. The solenoid is a niobium-titanium coil of 6 m in diameter that is used to provide a uniform axial magnetic field of 3.8 T inside along the z direction. [3] This strong magnetic field is needed to bend the trajectories of charged particles — a larger curvature of trajectory provides a more precise momentum measurement (see Sec. 2.2.1). Inside the solenoid, there are a silicon pixel and strip tracker, a lead

[3] This is about 100,000 times stronger than the Earth's magnetic field (25–65 μT).

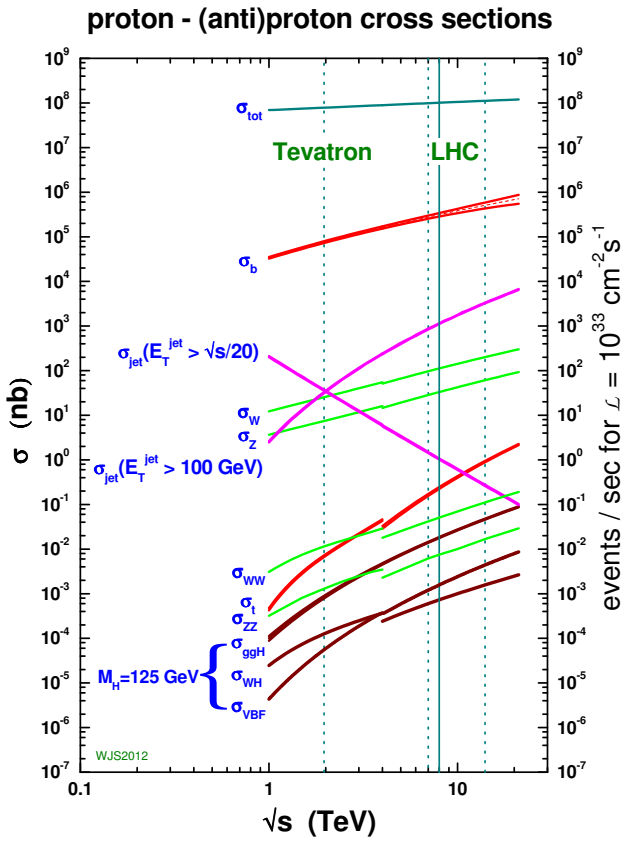

Figure 2-3. Standard Model cross sections of different processes as a function of collider energy, assuming m_H =125 GeV. The discontinuity is due to the Tevatron being a $p\bar{p}$ collider whereas the LHC is a pp collider [9].

tungstate crystal electromagnetic calorimeter (ECAL), and a brass and scintillator hadron calorimeter (HCAL). Outside the solenoid, there is a flux-return steel yoke that weighs 12,500 tonnes. The muon system consists of gas-ionization detectors embedded in the return yoke. Extensive forward calorimeter detectors cover the very forward region near the beam pipe. The tracker, ECAL, HCAL and muon system all consist of a cylindrical barrel section and two endcap sections. The whole CMS apparatus is 14.6 m in diameter, 21.6 m in length, and 14,500 tonnes in weight, hence *compact*. In total, there are $\sim 10^8$ electronic channels being read out during each collision.

Note that CMS uses a right-handed coordinate system that assumes the origin at the nominal interaction point, x axis pointing radially inwards to the center of the LHC

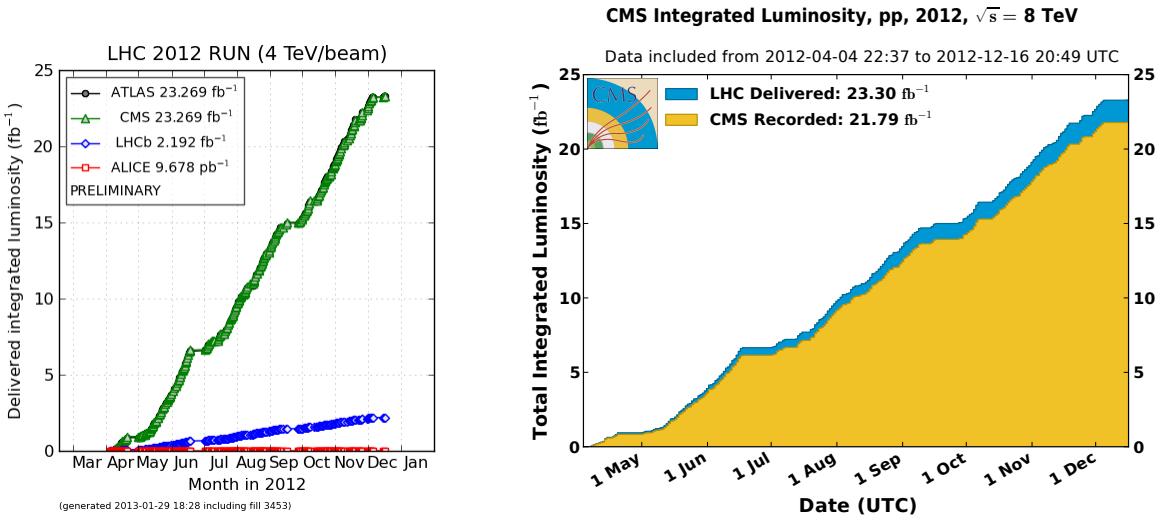

Figure 2-4. Left: Integrated luminosities delivered to the four experiments by the LHC in 2012 [10]. Right: Integrated luminosity delivered to (blue) and recorded by (yellow) CMS in 2012 [11].

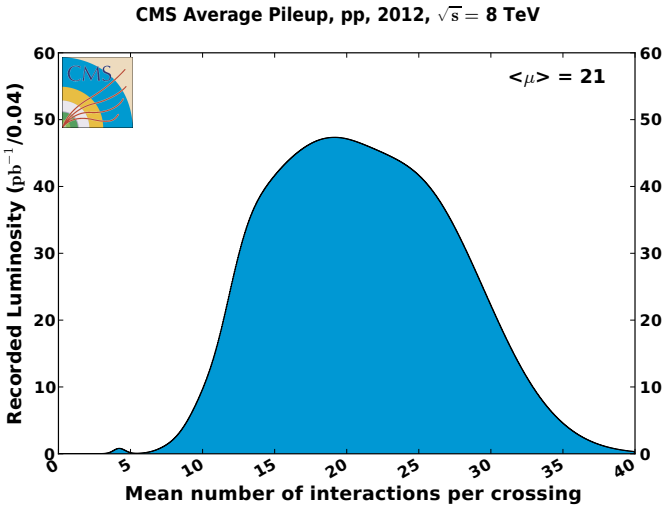

Figure 2-5. Number of *pp* interactions per bunch crossing during 2012 data taking [11].

ring, *y* axis pointing vertically upwards, and *z* axis along the counterclockwise beam direction. The polar angle θ is measured from the positive *z* axis and the azimuthal angle ϕ is measured from the *x* axis in the *x*-*y* plane. In collider experiments, rapidity $Y \equiv \frac{1}{2}\ln\frac{E+p_z}{E-p_z}$ and pseudorapidity $\eta \equiv -\ln[\tan(\theta/2)]$ are sometimes preferred, as difference in rapidity or pseudorapidity is invariant under *z* boost. Pseudorapidity invariance is only valid for ultra-relativistic ($p \gg m$) particles, but is more often used

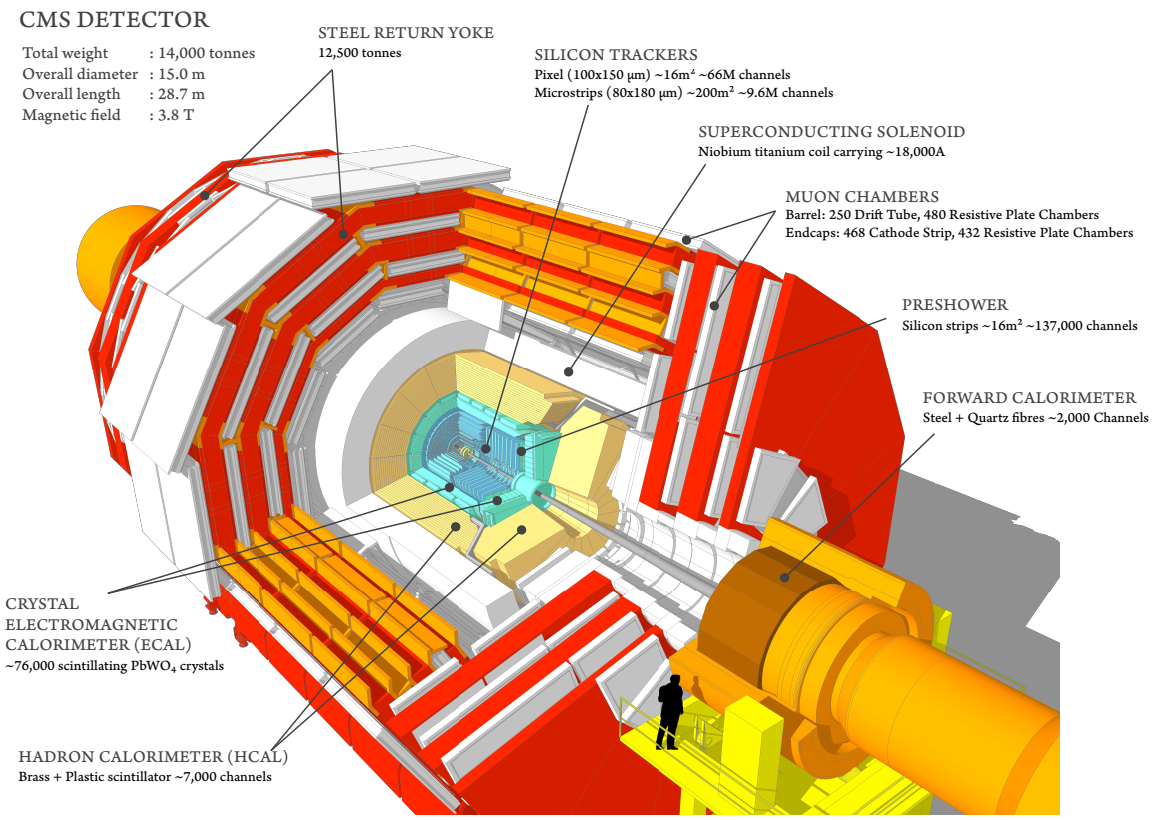

Figure 2-6. Schematic layout of the CMS detector [12].

because it is experimentally difficult to measure particle mass. The variable $\Delta R \equiv \sqrt{(\Delta\eta)^2 + (\Delta\phi)^2}$ is used to measure the separation of particles.

The main design goals of the CMS detector are [14, 71]:

- Muon detection with good muon identification and momentum resolution, good dimuon mass resolution and good muon charge assignment at high momentum (up to 1 TeV). Muon is so important that it becomes the experiment's middle name;
- Charge-particle tracking with good momentum resolution and reconstruction efficiency, and efficient tagging of τ and b jets;
- Electromagnetic calorimeter with high granularity, good energy resolution for electrons and photons, and efficient rejection of pions;
- Hadron calorimeter with hermetic geometric coverage and good energy resolution to reconstruct jets and missing transverse energy.

Full description of the CMS technical design can be found in Ref. [16].

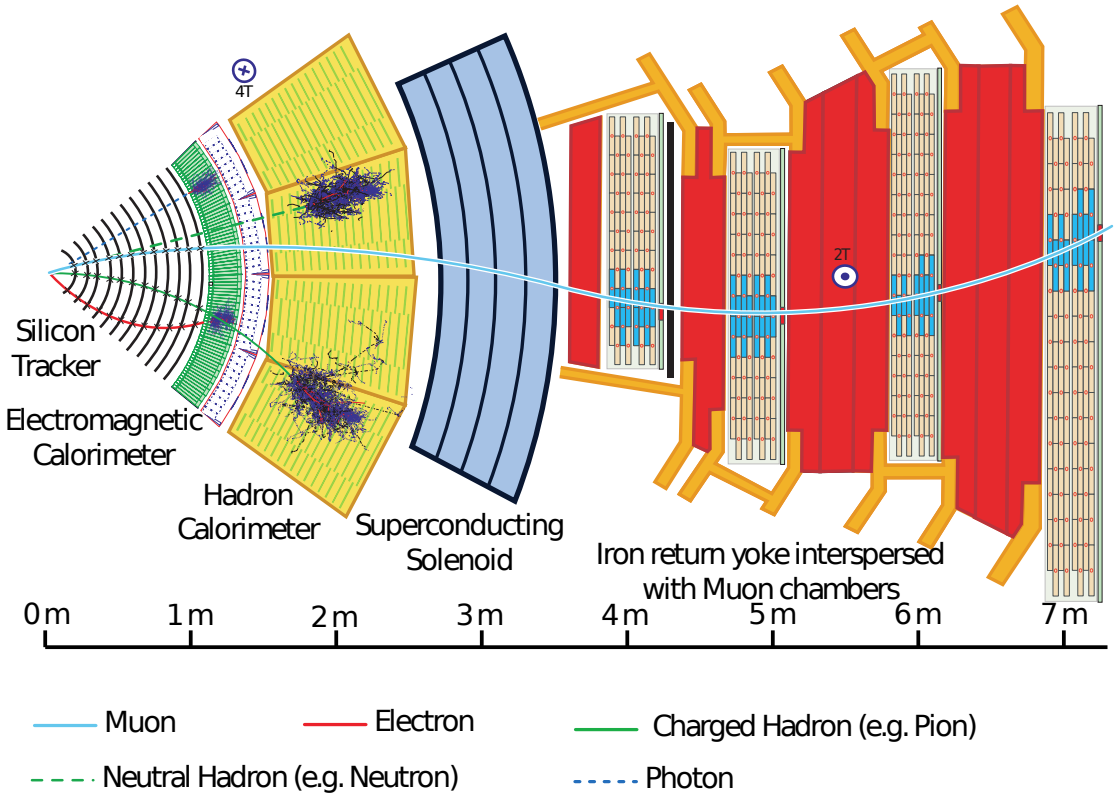

Figure 2-7. Identification of five types of particles by different subdetectors as they traverse through the CMS detector.

The detector layers (or subdetectors) of CMS exploit the distinct properties of particles to identify them and measure their energy or momentum with good resolution as stated in the goals. CMS can identify five types of particles: muon, electron, charged hadron (e.g. charged pions, charged kaons, protons), neutral hadron (e.g. K_L^0 meson, neutrons), and photons. The interactions of these five types of particles with different subdetectors and their signatures are illustrated in Fig. 2-7. They are described in detail in the following subsections. In addition, there are “invisible” particles that can go through the detector undetected, e.g. neutrinos. Neutrinos are neutral and very rarely interact with material. Their presence can only be inferred from $E_{\mathrm{T}}^{\mathrm{miss}}$.

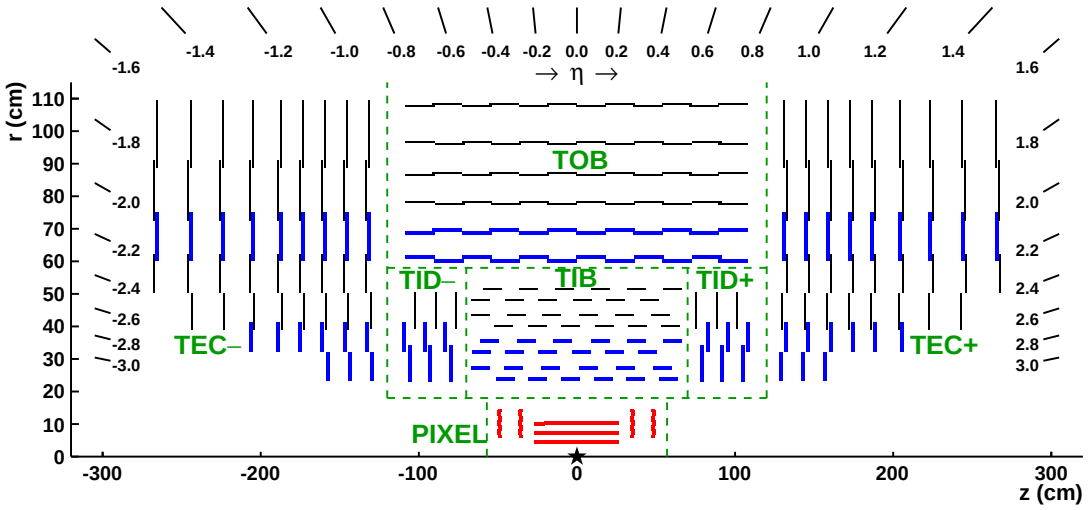

Figure 2-8. Schematic layout of the CMS tracker in the *r*-*z* plane. Each line represents a detector module. Blue lines represent stereo modules [13].

2.2.1 Tracker

The tracker is the subsystem closest to the interaction point. It measures the positions of charged particles as they pass through the magnetic field, and reconstructs them as tracks. Ideally, in a uniform magnetic field $\vec{B} = B\hat{z}$, the radius of track curvature R and transverse momentum p_{T} are related by [19]:

$$p_{\mathrm{T}} \cos\lambda = 0.3\, B\, R \tag{2–4}$$

where λ is the dip angle, B is in tesla and R is in meters. The factor of 0.3 comes from the speed of light c after appropriate unit conversion. What is experimentally measured is the sagitta of the track:

$$s \approx \frac{L^2}{8\,R} = \frac{0.3\, B\, L^2}{8\, p_{\mathrm{T}}} \tag{2–5}$$

for $L \ll R$ where L is the chord length or lever arm. The relative uncertainty on the p_{T} measurement is proportional to the sagitta uncertainty, δs, times p_{T}:

$$\frac{\delta p_{\mathrm{T}}}{p_{\mathrm{T}}} = \frac{8}{0.3\, B\, L^2} \delta s\, p_{\mathrm{T}} \tag{2–6}$$

The tracker consists of 1,440 silicon pixel and 15,148 silicon strip detector modules, covering active areas of 1 m^2 and 198 m^2 respectively. In total, there are 66 million pixel and 9.3 million strip channels. The fine segmentation is needed to deal with large flux of charged particles and to provide high momentum resolution. The pixels also provide high spatial resolution for the measurement of the track impact parameters and the reconstruction of secondary vertices.

The tracker covers the pseudorapidity range $|\eta| < 2.5$. The pixel detector is divided into barrel (BPIX) layers and endcap (FPIX) disks. The silicon strip detector is divided into four regions: inner barrel (TIB) layers, outer barrel (TOB) layers, endcap (TEC) disks, and inner disks (TID). The schematic layout of the tracker is shown in Fig. 2-8. There are 3 BPIX layers, 4 TIB layers, 6 TOB layers, (on each endcap side) 2 FPIX disks, 3 TID disks, and 9 TEC disks. The pixel detector provides 2–3 points of measurement, while the silicon strip detector provides 10–14 points.

The tracker is designed to be as light as possible to minimally interfere with particles in order to minimize effects on later measurements. However, due to the material in electrical cables, cooling tubes, support structures, and so on, several interactions can happen as particles traverse the tracker, e.g. multiple scattering, bremsstrahlung, photon conversion and nuclear interactions. The characteristic distance in a material that a particle can travel before it undergoes an EM interaction is known as the radiation length X_0. It is defined as the mean free path length over which the energy of a high-energy electron reduces by the factor of $1/e$, predominantly due to bremsstrahlung. An analogous distance is known as the (nuclear) interaction length λ_I, which is defined as the mean free path length over which the number of colored particles reduces by the factor of $1/e$ due to inelastic nuclear interactions. The material budget of the tracker in units of X_0 of the material and in units of λ_I are shown in Fig. 2-9.

In the barrel region, the spatial resolution of the pixels is about 10 μm for the r-ϕ measurement and about 20 μm for the z measurement. [4] The strip detector modules typically give only r-ϕ measurement, with a resolution of 23-53 μm (depending on layers). Certain layers use "stereo" modules that also give a coarse z measurement with a resolution of 230-530 μm. The occupancy is $\sim$ 0.01% per pixel per bunch crossing, and $\sim$ 1–3% per strip per bunch crossing.

In the barrel region, isolated charged particles of p_T = 100 GeV have resolutions of approximately 2.8% in p_T and 10 μm and 30 μm in the transverse and longitudinal impact parameters. For non-isolated charged particles of $1 < p_T < 10$ GeV, the track resolutions are typically 1.5% in p_T and 25–90 μm and 45–150 μm in the transverse and longitudinal impact parameters. The position resolution of the reconstructed primary vertex is 10–12 μm in each of the three spatial coordinates [13].

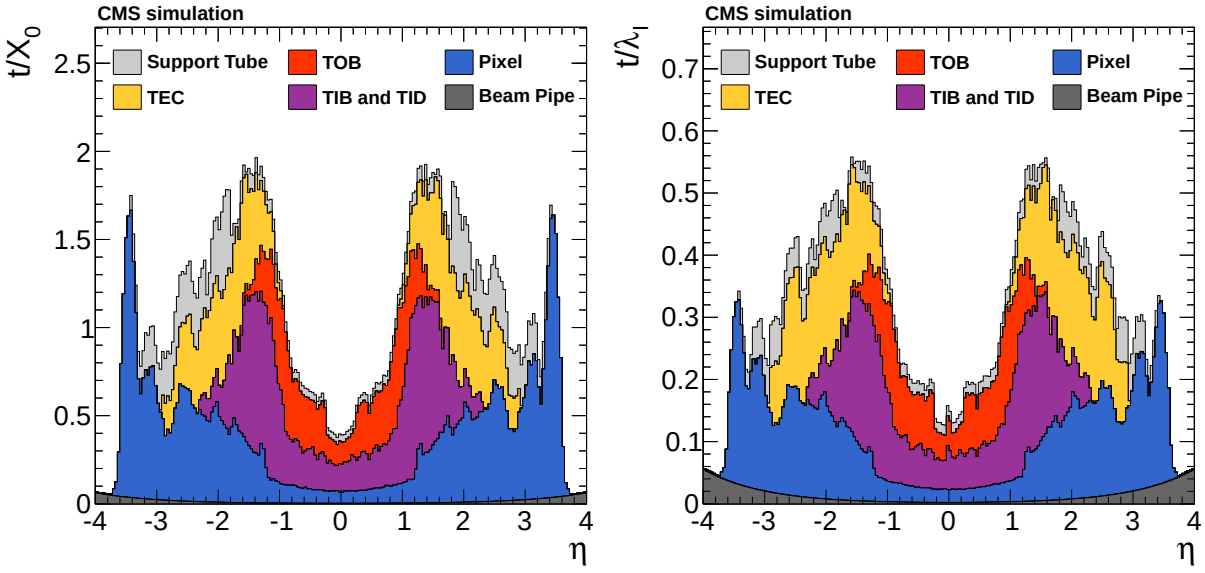

Figure 2-9. Material budget in units of radiation length X_0 (left) and interaction length λ_T (right) as a function of pseudorapidity η for the different parts of the tracker [13].

[4] For comparison, the diameter of human hair is about 100 μm.

2.2.2 Electromagnetic Calorimeter

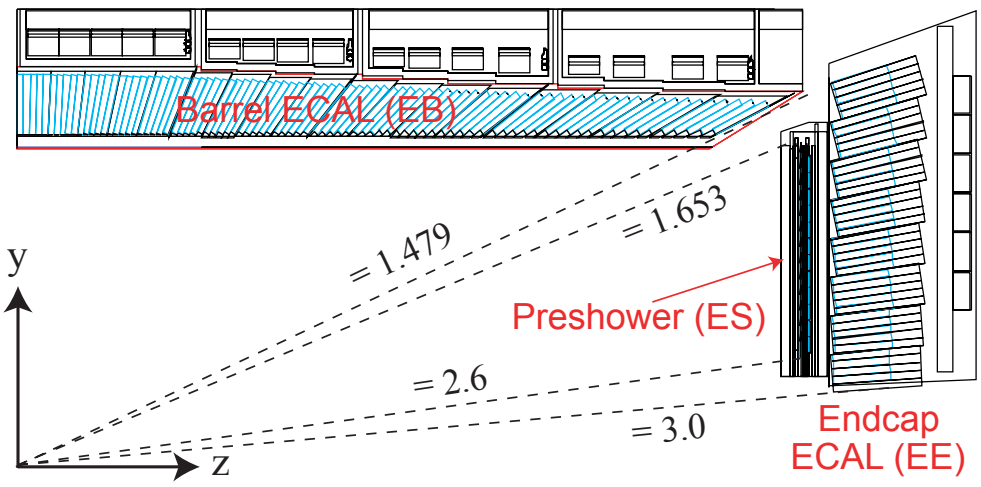

Figure 2-10. Schematic layout of a quadrant of the CMS ECAL in the *y*-*z* plane [14].

The ECAL consists of 75,848 lead tungstate ($PbWO_4$) scintillating crystals that provide precise energy measurement of electrons and photons. High-energy electrons and photons passing through matter with high atomic number Z initiate an electromagnetic (EM) "shower", i.e. generation of electrons and photons in cascade via e^+e^- pair production and bremsstrahlung, losing their energies in the process. As their energies fall below a critical value, they stop generating more particles and instead lose their energies primarily by ionization. (For photons, ionization is preceded by photon conversion.) The scintillating crystals absorb the energy of low-energy electrons and re-emit a proportional fraction of the energy as light, which is then detected by photodetectors. The peak of the light spectrum is at 425 nm (blue light). The ECAL is homogeneous, meaning the entire volume simultaneously serves as absorber as well as active medium. The crystals are chosen because they provide high granularity, are fast and are radiation resistant.

The ECAL is divided into a barrel (EB) region, which covers $|\eta| < 1.479$, and two endcap (EE) regions, which cover $1.479 < |\eta| < 3.0$. In the barrel region, the crystal cross section is approximately 0.0174×0.0174 in η-ϕ, or $22 \times 22\,\text{mm}^2$ at the front face and $26 \times 26\,\text{mm}^2$ at the rear face. The crystals have a radiation length of

25.8 X_0. The scintillation light is detected by silicon avalanche photodiodes (APDs). In the endcap regions, the crystal cross section is $28.62 \times 28.62\,\text{mm}^2$ at the front face and $30 \times 30\,\text{mm}^2$ at the rear face. The crystal length is 24.7 X_0. The photodetectors are vacuum phototriodes (VPTs).

In addition, there is a preshower (ES) detector which covers $1.653 < |\eta| < 2.6$, in front of EE. It consists of two planes of silicon strip detectors which lie behind disks of lead absorber at depths of 2 X_0 and 3 X_0. Its main job is to identify neutral pions. It also helps electron identification against minimum ionizing particles (MIPs) and improves the position measurement of electrons and photons. The schematic layout of the ECAL is shown in Fig. 2-10.

The ECAL energy resolution, $\sigma(E)/E$, for electrons is measured in beam tests and fit as a function of energy, E:

$$\left(\frac{\sigma(E)}{E}\right)^2 = \left(\frac{S}{\sqrt{E}}\right)^2 + \left(\frac{N}{E}\right)^2 + C^2 \qquad (2\text{–}7)$$

assuming Gaussian energy distribution. S is the stochastic term, N is the noise term and C is the constant term. Fig. 2-11 shows a representative result for a barrel array of 3×3 crystals. Since the S term is usually more important, the energy resolution is roughly $\sigma(E)/E \approx 3\%/\sqrt{E}$ for barrel and $\sigma(E)/E \approx 6\%/\sqrt{E}$ for endcap.

2.2.3 Hadron Calorimeter

The HCAL interacts with charged and neutral hadrons via strong interaction. It consists of brass absorber plates that cause the hadrons to initiate hadronic showers via nuclear interactions. Brass is chosen to maximize amount of material in order to contain the showers, because it has a short interaction length λ_I. Brass is also non-magnetic, hence it is suitable for the HCAL which is immersed in the strong solenoid magnetic field. Plastic scintillator tiles embedded with wavelength-shifting (WLS) fibers are interleaved in the absorber plates. The scintillation light is collected by WLS

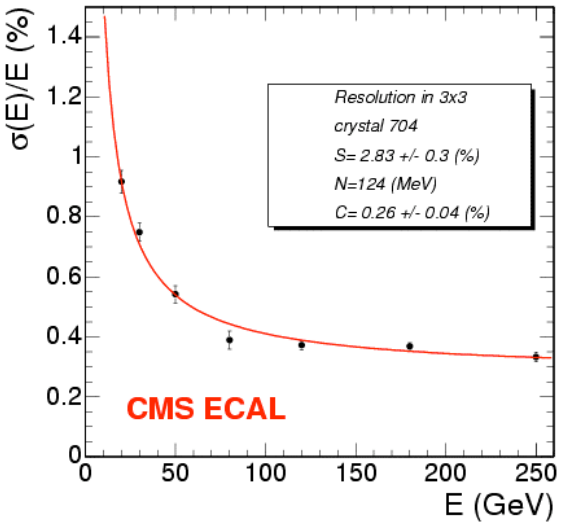

Figure 2-11. ECAL energy resolution, $\sigma(E)/E$, as a function of electron energy in a representative barrel array of 3×3 crystals [15].

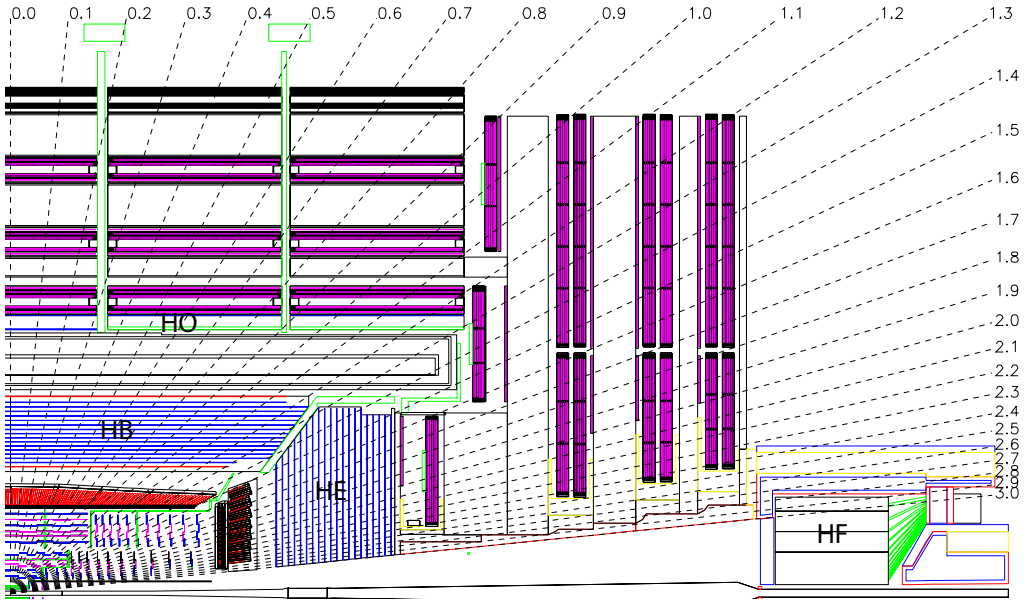

Figure 2-12. Schematic layout of a quadrant of the CMS HCAL in the *y*-*z* plane [16].

fibers and channeled to photodetectors. Hybrid photodiodes (HPDs) are used as the photodetectors because they can provide gain while operating in strong magnetic fields.

The HCAL completely surrounds the ECAL and is divided into a barrel (HB) region, which covers $|\eta| < 1.4$, and two endcap (HE) regions, which cover $1.3 < |\eta| < 3$ with some overlap. Compared to the ECAL, the HCAL has larger depth and coarser segmentation. In the barrel region, the size of each individual scintillator tile is 0.087×0.087 in η-ϕ. The HCAL cells map on to 5×5 ECAL crystals arrays to form calorimeter towers projecting radially outwards from the nominal interaction point. In the endcap regions, the size in ϕ varies from 0.087–0.174 and the size in η varies from 0.087–0.35, depending on η. The endcap calorimeter towers have a larger size but the matching ECAL arrays contain fewer crystals. HB is complemented by a hadron outer (HO) calorimeter, which consists of an additional layer of scintillators, outside the solenoid. It acts as a tail-catcher to ensure that hadronic showers are sampled with nearly 11 interaction lengths. It has a similar segmentation as HB.

Extended coverage for $3 < |\eta| < 5$ is provided by a steel/quartz-fiber hadron forward (HF) calorimeter, located 11 m away from the interaction point. The Cherenkov light emitted in the quartz fibers is detected by photomultipliers (PMTs). HF experiences high hadron rates, so the active medium (quartz) must be radiation hard. HF is needed to ensure hermetic geometric coverage that is necessary for good $E_{\mathrm{T}}^{\mathrm{miss}}$ measurement. The schematic layout of the HCAL is shown in Fig. 2-12.

The HCAL material thickness varies in the range 7–11 λ_I without HO, or 10–15 λ_I with HO, depending on η. The combined ECAL+HCAL energy resolution in the barrel region is approximately $\sigma(E)/E \approx 84.7\%/\sqrt{E}$, after correcting for non-linearity energy response [72].

2.2.4 Muon System

A muon is not stopped by the calorimeters as it is a minimum ionizing particle (MIP), meaning its energy loss by ionization as it passes through matter is close

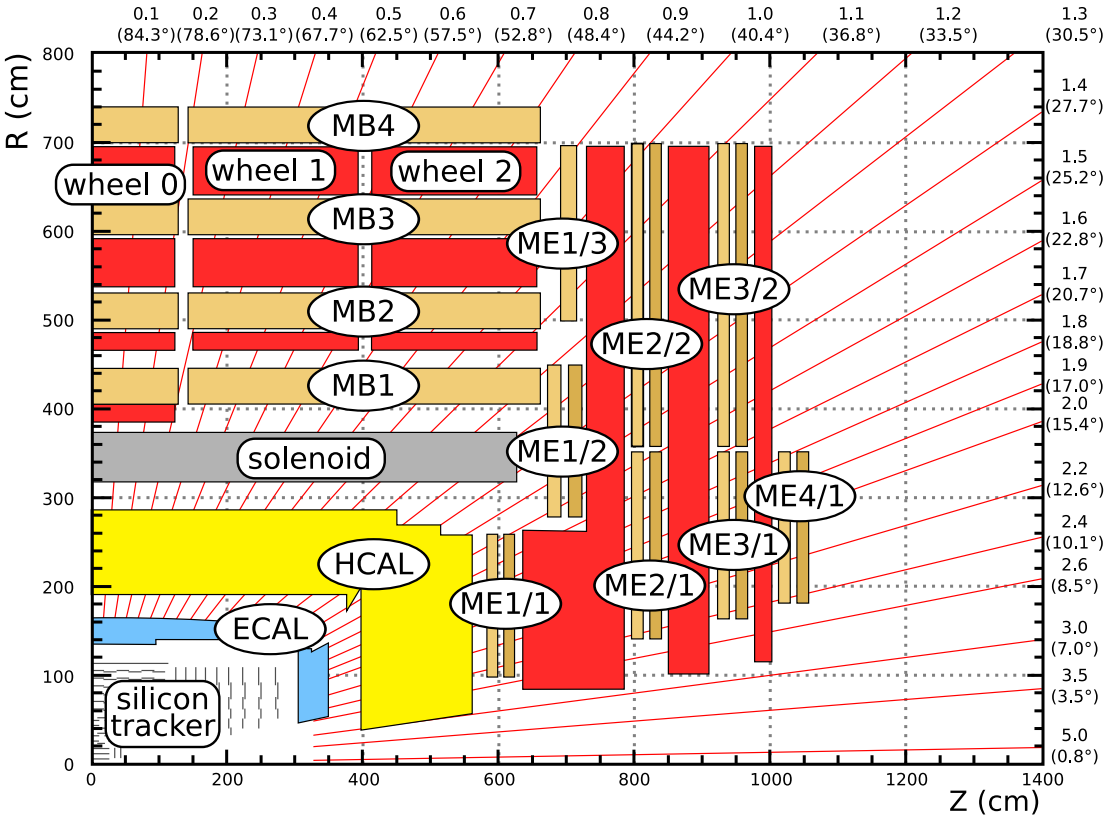

Figure 2-13. Schematic layout of a quadrant of the CMS muon system in the *r*-*z* plane. The red blocks between the muon stations represent the iron yoke [17].

to the minimum. To provide efficient muon identification and improved momentum measurement, dedicated muon chambers are built. They are placed in the return yoke and exploit the return magnetic field. The muon system consists of 4 muon stations, and is divided into a barrel region ($|\eta| < 1.2$) and two endcap regions ($0.9 < |\eta| < 2.4$ with some overlap). 3 different types of gaseous detectors are used. Each station consists of several layers of aluminum drift tubes (DTs) in the barrel region and cathode strip chambers (CSCs) in the endcap region. Resistive plate chambers (RPCs) are present in barrel and endcap regions up to $|\eta| < 1.6$. The schematic layout of the muon system is shown in Fig. 2-13.

DT technology is used in the barrel region because the magnetic field is low and almost-uniform, the muon rate is relative low, and neutron-induced background is negligible. CSC technology is used in the endcaps where the magnetic field is strong

and uneven, high muon rate and neutron-induced background rate. CSCs also have a higher radiation resistance. RPCs have fast response and very high time resolution that helps to identify the correct bunch crossing of observed muons. However, their position resolution is coarser than that of the DTs or CSCs.

Muons found in the muon system are matched to tracks found in the silicon tracker. A global momentum fit using hit information from both systems improves the momentum resolution by an order of magnitude at low momenta, compared to using information from the muon system only. This can be seen in Fig. 2-14 for two different pseudorapidity ranges. The global muon p_T resolution is in the range of 1–3% at p_T of 10–100 GeV, and $<$10% at $p_T \sim 1$ TeV.

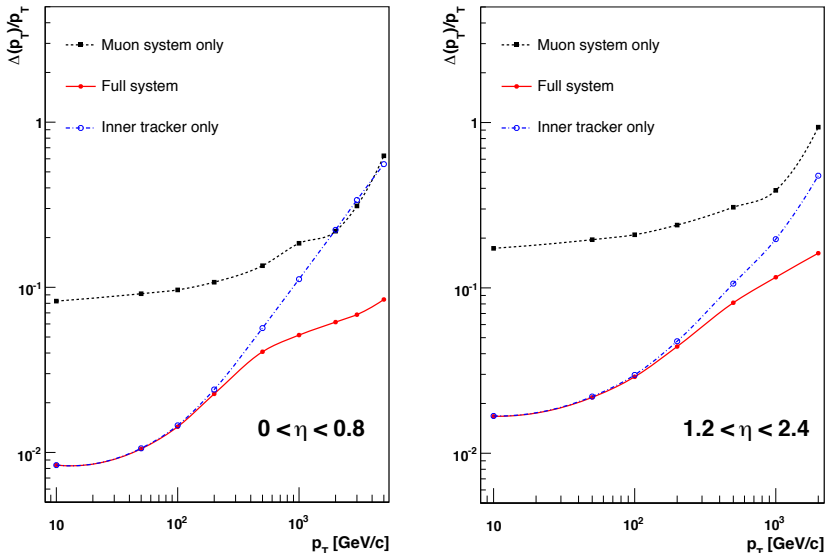

Figure 2-14. Muon transverse momentum resolution as a function of the transverse momentum using the muon system only, the inner tracker only, and both, for two pseudorapidity ranges: $|\eta| < 0.8$ (left) and $1.2 < |\eta| < 2.4$ (right) [16].

2.2.5 Trigger and Data Acquistion

The LHC is designed to produce order of 10^9 collisions per second. To write out and analyze this amount of data far exceeds current technological capabilities. Fortunately, most of these events are uninteresting and can be discarded. CMS implements a

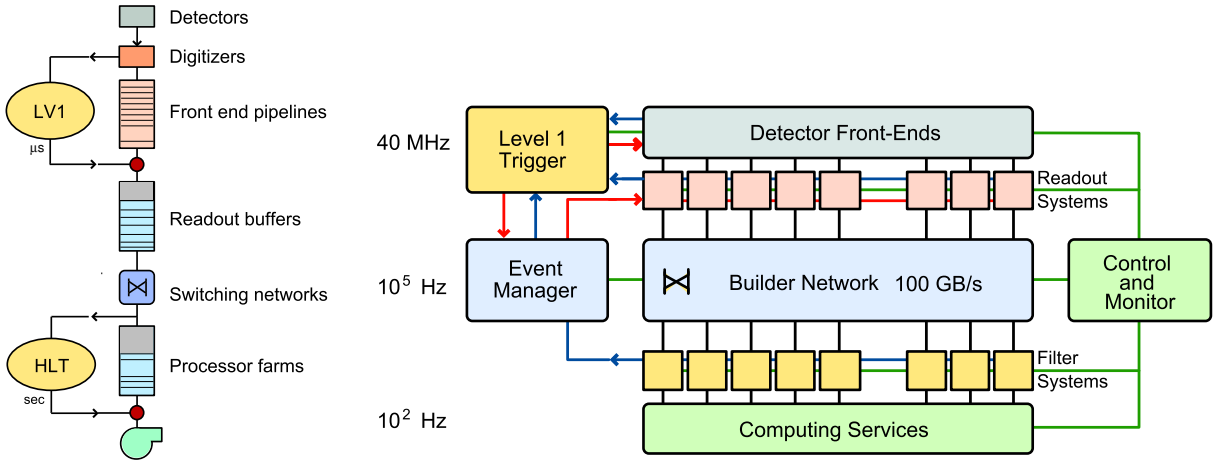

Figure 2-15. Left: Data flow in the two-level trigger system with Level-1 Trigger and High Level Trigger [18]. Right: Architecture of the data acquisition system [16].

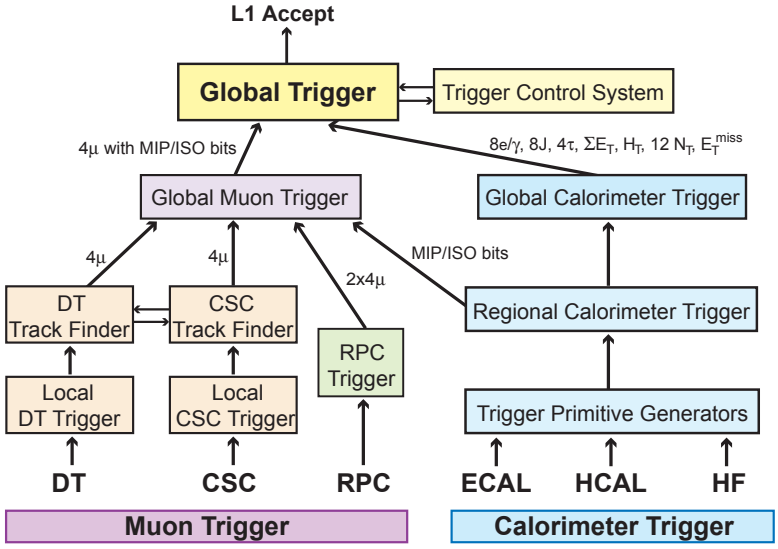

Figure 2-16. Architecture of the Level-1 Trigger [16].

two-level trigger system that processes events online and select the most interesting events for permanent storage and offline analysis [18, 73, 74].

The first level (L1) of the CMS trigger system is composed of custom-designed, largely programmable hardware, such as field-programmable gate arrays (FPGAs), application-specific integrated circuits (ASICs), and programmable memory lookup tables (LUTs). It uses information from the calorimeters and the muon system to make trigger decisions in a fixed time interval of less than 4 μs. The trigger processing time

is called the latency. The very short latency is forced by the large input rate and the finite amount of data that can be retained in the "pipeline" while waiting for the trigger decisions to be made for the previous events. The L1 Trigger reduces the event rate from 40 MHz (20 MHz during Run I) to around 100 kHz.

The high level trigger (HLT) software algorithms, executed on a farm of about 13,000 CPUs (short for central processing units), further decrease the event rate to around 400 Hz before data storage. Unlike the L1 Trigger, the HLT uses information from all subdetectors including the tracker and performs complex calculations similar to those used in the offline software, but modified to comply with the strict constraint in the online processing time. For instance, the HLT tracking is done using less iterations and tighter seed requirements. The mean processing time per event at the HLT is about 150 ms. (In contrast, the offline reconstruction takes O(10) seconds per event.) The data flow and architecture of the trigger and data acquisition (DAQ) system are displayed in Figs. 2-15 & 2-16.

There are about 400 paths (or algorithms) in the HLT "menu" that select various interesting physics signatures. When any trigger path transmits the decision to keep a given event ("fires"), detector raw data for the event are read out, reformatted, and stored in one or more primary datasets (PDs). The PDs are datasets organized according to trigger signatures or analysis use-cases with the goal of having more or less the same event rate across different PDs. A few example PDs are: `SingleMuon`, `SingleElectron`, `DoubleElectron`, and `MET`. Note that an event may be stored in more than one PD if it happens to be accepted by trigger paths that fall into different PDs.

The recorded raw data are transferred to computing centers that are distributed worldwide at various collaborating institutes, interconnected by and managed through the Worldwide LHC Computing Grid (WLCG) project [75]. These computing centers provide several functions: storage of the recorded data for the lifetime of the experiment,

production and storage of the simulated data, data transfer and distribution, offline event reconstruction, data-intensive analysis, etc. The challenges of managing the very large data samples have been addressed through construction of a modular system of loosely coupled components with well-defined interfaces, with emphasis on scalability [76].

CMS makes use of three major data tiers: RAW, RECO, and AOD. The RAW events contain the full detector raw data after online formatting and a record of the trigger decisions made at the L1 and HLT. RAW data are permanently archived. The event size is O(1.5) MB/event. The RECO (short for reconstructed) data tier contains the reconstructed high-level physics objects (tracks, vertices, leptons, jets, etc) and all the associated hits and clusters. The event size is O(250) kB/event. The AOD (short for analysis object data) data tier is a subset of RECO intended for use in a wide range of physics analyses. Only a limited amount of low-level information is kept in AOD to minimize disk usage. The event size is O(50) kB/event.

CHAPTER 3
EVENT SIMULATION

3.1 Overview of Event Simulation

In order to understand how a theoretical particle physics process would manifest itself in the detector and how to design an experiment to look for such a process, it is necessary to generate events from the process and simulate what they would look like when reconstructed in the CMS detector. An ensemble of such events is needed in order to populate all points in the phase space of the process according to the quantum mechanical probabilities. Given the dimension of phase space is 3 per outgoing particle, an event generation typically involves integration over a dimension of $3n - 4$ for n outgoing particles. (4 dimensions are subtracted due to overall conservation of four-momentum.) The large dimensionality renders analytical integration practically impossible. Instead, numerical integration is needed. One simple and robust numerical integration technique is the Monte Carlo (MC) method based on repeated random sampling, described in Sec. 3.2.

Furthermore, the simulation of particle interactions with the detector, the reconstruction of physics objects, and other stages of event simulation also involve many instances that are probabilistic, or having many degrees of freedoms, or a combination of both. It turns out that the MC method is also a natural fit to all these problems. Thus, the Monte Carlo method is used extensively in particle physics, so much that simulation is often known simply as “MC”.

3.2 Monte Carlo Method

The Monte Carlo (MC) method was introduced by Stanislaw Ulam and John von Neumann in the 1940s. based on random sampling. It is widely used to solve physical and mathematical problems, especially when the analytic form is unknown, too complex, or involves large number of degrees of freedom.

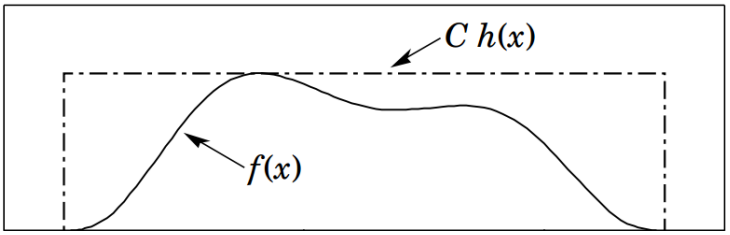

Figure 3-1. Illustration of the acceptance-rejection method [19].

A simple example is known as the acceptance-rejection method. Suppose, for a variable x, its probability density function $f(x)$ is known and can be enclosed entirely inside a shape which is C times an easily generated distribution $h(x)$, such as a "box". Hence, $C \cdot h(x) \geq f(x)$ for all x, as illustrated in Fig. 3-1. Typically, both $f(x)$ and $h(x)$ are normalized to unit area, hence $C \geq 1$. Then, the integral of $f(x)$ is found as follows:

1. Generate a trial x according to $h(x)$.
2. Evaluate $f(x)$ and $C \cdot h(x)$.
3. Generate a random number $u \in [0,1]$ and test if $u \cdot C \cdot h(x) \leq f(x)$. If yes, the trial is accepted; otherwise it is rejected.
4. Repeat from step 1.

Given a well-behaved pseudorandom number generator and a number of trials N, the sampled distribution converges to actual integral with an error estimate $\propto 1/\sqrt{N}$ regardless of the number of dimensions. More advanced MC techniques are used when non-uniform integration phase space is involved to improve convergence rate (see Ref. [19]).

3.3 Event Generation

An MC event generator turns the abstract quantum mechanical rules that govern a pp collision into experimental observables such as momentum, mass, charge, flavor, and time of flight. A broad range of physics is involved — from hard physics at very short distances to soft physics at long distances. The typical energy scale of QCD, Λ_{QCD}, is a few hundred MeV and the distance scale is $1/\Lambda_{QCD} \sim$ a few femtometers. The QCD factorization theorem [77] states that calculations can be factorized into short-distance

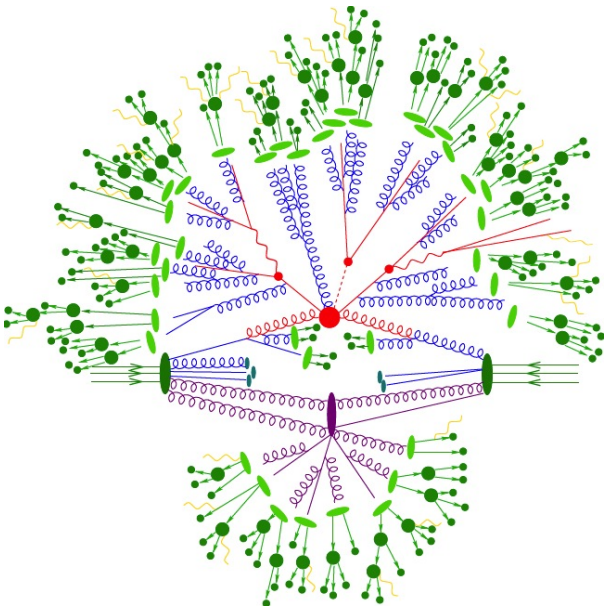

Figure 3-2. Illustration of a *pp* collision event [20]. See text for details.

behavior (below $1/\Lambda_{QCD}$), which can be reliably calculated by perturbation theory of QCD, and long-distance physics, which is not (currently) calculable but is universal and can be described by phenomenological models that are tuned using empirical data.

An event generation is split into several steps [20, 78]:

- **Hard scattering**. Hard scattering is the interaction between two incoming partons that involves the highest momentum exchange. Kinematics of hard scattering is determined by the matrix element (ME) of the scattering process and the parton density functions (PDFs). ME describes the transition from an initial state to a final state, as derived from Feynman rules. Decays of short-lived resonances such as top, W and Z are usually included in ME so that spin correlations are preserved. PDFs describe the densities of different types of partons in a hadron as a function of the momentum fraction. They are obtained from fits to data.

- **Final state radiation**. Final state radiation (FSR) is the gluon emission, a.k.a. QCD bremsstrahlung, that is associated with outgoing colored partons. It is implemented using parton shower (PS) algorithms that can deal with soft singularities, where one of the parton energies vanishes, and collinear singularities, where two partons become collinear. PS is an approximation to perturbation theory. It repeats $1 \to 2$ branching (e.g. $q \to qg$, $g \to gg$, $g \to q\overline{q}$) in cascade to evolve a parton into more partons ("shower").

- **Initial state radiation**. Initial state radiation (ISR) is the gluon emission from incoming partons that participate in the hard scattering. ISR is implemented like FSR, but it evolves backwards in time.

- **Hadronization and hadron decay**. Hadronization is the formation of color-singlet hadrons from a colored parton due to color confinement. This happens in the non-perturbative regime, and is modeled by phenomenological fragmentation models. Stable hadrons are hadrons that have a long enough lifetime to hit the detector sensors. Many of the hadrons, as well as τ-lepton, are unstable and decay further. Hadron decays are usually treated using matrix elements with additional corrections. Photon radiations are also included.

- **Underlying event**. Underlying event (UE) includes any hadronic activity not attributed to the particles participating in the hard scattering or to the hadronization of ISR and FSR. It is mainly due to the hadronization of remainder partons not participating in the hard scattering and to the hadronization of beam remnants not involved in other scatterings. UE is correlated to the hard scattering in flavor, color and momentum space. It is described by a phenomenological model with parameters that are tuned to match data. Note that additional collisions that involve different protons in the same bunch crossing are referred to as pileup interactions, and each pileup interaction has its own underlying event.

- **Photon radiation**. Photon radiation is the photon emission from outgoing EM particles.

An illustration of a *pp* collision event is shown in Fig. 3-2. Hard scattering between two incoming partons, as well as gluon radiations from those partons, are shown in red. Gluon radiations from outgoing partons are shown in blue. Underlying event activity due to secondary interactions from incoming partons is shown in purple. Hadronization of outgoing partons and hadron decays are shown in light green and dark green. Photon radiations are shown in yellow.

Leading-order (LO) matrix elements, from the tree-level diagrams, are typically used as the starting point for a process. Leading-log (LL) parton shower is then applied to develop high-multiplicity final states. For instance, ME of $2 \rightarrow 2$ process is used along with PS to produce $2 \rightarrow$ many. Care must be taken to match ME and PS without gap or double counting. To get more accurate predictions, next-to-leading-logarithmic (NLL) or next-to-leading-order (NLO) corrections are sometimes needed, with more complicated

matching schemes. The ratio of NLO (or next higher orders like NNLO, ...) cross section prediction over LO is known as K-factor.

Due to truncation of higher order α_s terms, there are two unphysical scales in MC event generation: factorization scale μ_F and renormalization scale μ_R. μ_F arises from the inclusion of PDFs which are resummed to all orders; μ_R is used to cut off ultraviolet divergence due to loop diagrams. Variations due to scale dependency are taken as theoretical uncertainties. However, they can be underestimated, especially when new Feynman diagrams that only appear at higher order lead to dramatic change in certain region of phase space.

3.4 Detector Simulation

GEANT4 [79] is used to model the CMS detector response to the passage of particles through it. A wide range of known particle interactions with matter and external electromagnetic fields are included in the software. Detailed description of detector is simulated, including geometry, alignment, densities and types of material, and subsystem conditions. Then, it can accurately simulate propagation of particles from event generator through different parts of the detector, including particle trajectories, energy loss, response or hits in sensitive detector components, secondary interactions, signal digitization and readout, etc. The simulated signals are stored and processed in the same way as real detector signals in subsequent event reconstruction.

CHAPTER 4
EVENT RECONSTRUCTION

4.1 From Detector Signals to Physics Objects

The physics objects that make up an event include tracks, primary vertices, leptons, photons, jets, and $E_{\mathrm{T}}^{\mathrm{miss}}$. From these reconstructed objects, one can characterize a final state for each event. This is the first step in the attempt to assign an event to a specific physics process achieved by further event selection criteria on these objects, as described in the next sections.

Event reconstruction is carried out using the central CMS software framework (CMSSW) that also provides interfaces to various MC event generators, GEANT4, etc. CMSSW is written in C++ in large part, following an object-oriented design. This analysis is done using the software version `CMSSW_5_3_3_patch2`.

4.2 Particle-Flow Algorithm

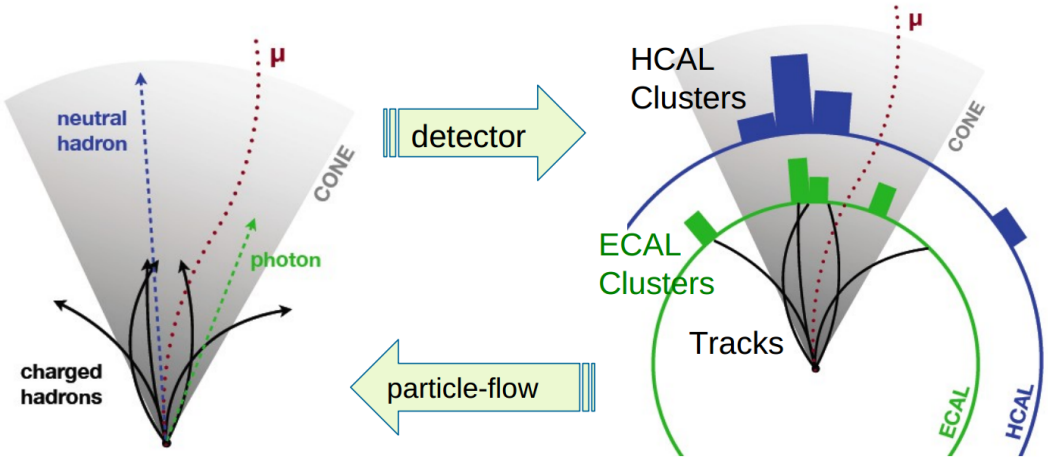

Figure 4-1. Illustration of particle-flow algorithm.

CMS runs a global event reconstruction algorithm, called the particle-flow (PF) algorithm [80, 81], that reconstructs and identifies each individual particle with an optimized combination of all subdetector information. PF starts with more primitive elements created locally in the subdetectors, i.e. silicon tracks, muon tracks, ECAL

clusters, and HCAL clusters. The elements are topologically linked to form blocks, and the content of each block is analyzed and interpreted as PF particles. The link algorithm gets rid of possible double counting of information about the same particle from different subdetectors. Five particle types are identified — photon, electron, muon, charged hadron and neutral hadron. Each PF particle is then assigned momentum and energy based on its type:

- **Electron**. Electrons are identified as a silicon track and potentially many ECAL clusters corresponding to this track extrapolation to the ECAL and to possible bremsstrahlung photons emitted along the way through the tracker material. The energy of electrons is determined from a combination of the track momentum at the primary vertex, the energy of the corresponding ECAL cluster, and the energy sum of all compatible bremsstrahlung photons.
- **Photon**. Photons are identified as ECAL clusters not linked to the extrapolation of any silicon track to the ECAL. The energy of photons is directly obtained from the ECAL measurement, corrected for zero-suppression effects.
- **Muon**. Muons are identified as a silicon track consistent with either a muon track or several hits in the muon system, associated with a MIP signature in the calorimeters. The energy of muons is obtained from the track momentum at the primary vertex.
- **Charged hadron**. Charged hadrons are identified as silicon tracks that are not identified as electrons or muons. The energy of charged hadrons is determined from a combination of the track momentum and the corresponding ECAL and HCAL energy deposits, corrected for zero-suppression effects and for the response function of the calorimeters to hadronic showers.
- **Neutral hadron**. Neutral hadrons are identified as HCAL clusters not linked to any silicon track, or as ECAL and HCAL energy excesses w.r.t. the expected charged hadron energy deposit. The energy of neutral hadrons is obtained from the corresponding ECAL and HCAL energy deposits after corrections.

An illustration of the PF algorithm is shown in Fig. 4-1. The reconstructed PF particles are treated as if they are MC generated particles in the subsequent reconstructions of jets and $E_{\mathrm{T}}^{\mathrm{miss}}$, as well as to identify hadronic τ decays and to quantify lepton isolation.

4.3 Track Reconstruction

The trajectory of a charged particle in an axial magnetic field is reconstructed as a track parametrized by 5 parameters: signed transverse curvature ρ (which is proportional to particle charge times p_T), azimuthal angle ϕ, polar angle $\cot\theta$, longitudinal impact parameter d_z, signed transverse impact parameter d_0, all defined at the point of closest approach (d_x, d_y, d_z) of the track to the nominal beam axis, where $d_0 = -d_y \cos\phi + d_x \sin\phi$. Promptly produced tracks, or prompt tracks, are tracks assumed to be originating close to the interaction point ($d_0 \lesssim 2\,\text{cm}$); displaced tracks refer to tracks with large d_0.

To achieve both high track-finding efficiency and low rate of fake tracks in a high-occupancy environment, CMS adopts an iterative tracking strategy [13]. The Combinatorial Track Finder (CTF) algorithm, an extension of the Kalman filter [82], is repeated multiple times, starting with tight criteria and followed by progressively looser criteria. The first iterations search for tracks that are relatively easy to find, e.g. prompt, high-p_T tracks. Once a track is found, hits associated with the track are excluded from consideration (masked), so that hit combinatorial complexity is reduced in the subsequent iterations that look for more difficult classes of tracks, e.g. low-p_T tracks, displaced tracks from decays of long-lived particles (e.g. $K^0_S \to \pi^+\pi^-$, $\Lambda^0 \to p\pi^-$), and tracks from secondary interactions (e.g. photon conversions, nuclear interactions).

The CMS implementation uses a series of 7 iterations. Each iteration involves the following four steps:

- **Seed generation**. A seed is an initial track candidate that is either a triplet of hits or a pair of hits with an additional constraint from the beamspot or a vertex. (The beamspot is the luminous region over which pp interactions occur.) It defines the initial estimate of the track parameters and their uncertainties. Seeds are typically generated from the inner pixel layers because of low channel occupancy and precise & unambiguous 3D hit position measurement. Only seeds that have acceptable track parameters, such as minimum p_T and maximum d_0 and $|d_z|$, are accepted (see Table 4-1).

- **Track finding**. This is based on the Kalman filter technique. The filter makes an inside-out extrapolation of the trajectory using the parameters provided by the seed. Hits compatible with the trajectory in the successive layers are added. The estimated track parameters and uncertainties are updated. Missing hits are allowed by adding ghost hit at the positions where the trajectory is expected to produce a hit. This search continues until either the end of the tracker is reached or the number of ghost hits exceeds a threshold.

- **Track fitting**. The collection of hits that are assigned to the trajectory is refitted using a Kalman filter and smoother to obtain the best possible estimate of the track parameters. The amount of material crossed is taken into account to estimate the effects of multiple scattering and energy loss. Constraints added during seed generation are removed.

- **Track selection**. A fraction of the reconstructed tracks are fake, i.e. not associated with a genuine charged particle. A selection is applied to set the quality of the tracks and remove fake tracks based on the number of layers that have associated hits, the goodness of fit (χ^2/ndof, ndof $=$ number of degrees of freedom), and the compatibility with originating from a vertex. The "high-purity" quality flag is used in many physics analyses.

Table 4-1. Configurations of seed generation for each of the 7 iterations used in the track reconstruction [83]. d_z is measured w.r.t. the beamspot, except for iteration 2 (denoted by *) where it is w.r.t. a pixel vertex.

Step	Seed type	Seed layers	min p_T[GeV]	max d_0	max $\lvert d_z \rvert$
0	triplet	pixel	0.6	0.02 cm	4.0σ
1	triplet	pixel	0.2	0.02 cm	4.0σ
2	pair	pixel	0.6	0.0015 cm	0.09^* cm
3	triplet	pixel	0.3	1.5 cm	15 cm
4	triplet	pixel/TIB/TEC	0.4–0.6	1.5 cm	10 cm
5	pair	TIB/TID/TEC	0.7	2.0 cm	10 cm
6	pair	TOB/TEC	0.6	6.0 cm	30 cm

At the very end, the tracks from all six iterations are put into a single track collection. Sometimes, a given seed may yield more than one tracks, or different seeds may yield the same track. When the fraction of shared hits between any pair of tracks exceeds a threshold, one of them is deemed as a duplicate and is removed, based on the number of hits and the goodness of fit. The CTF software is also used at the HLT. To cope with trigger constraints, it is tuned to run much faster by using less iterations and imposing tighter requirements on the seeds.

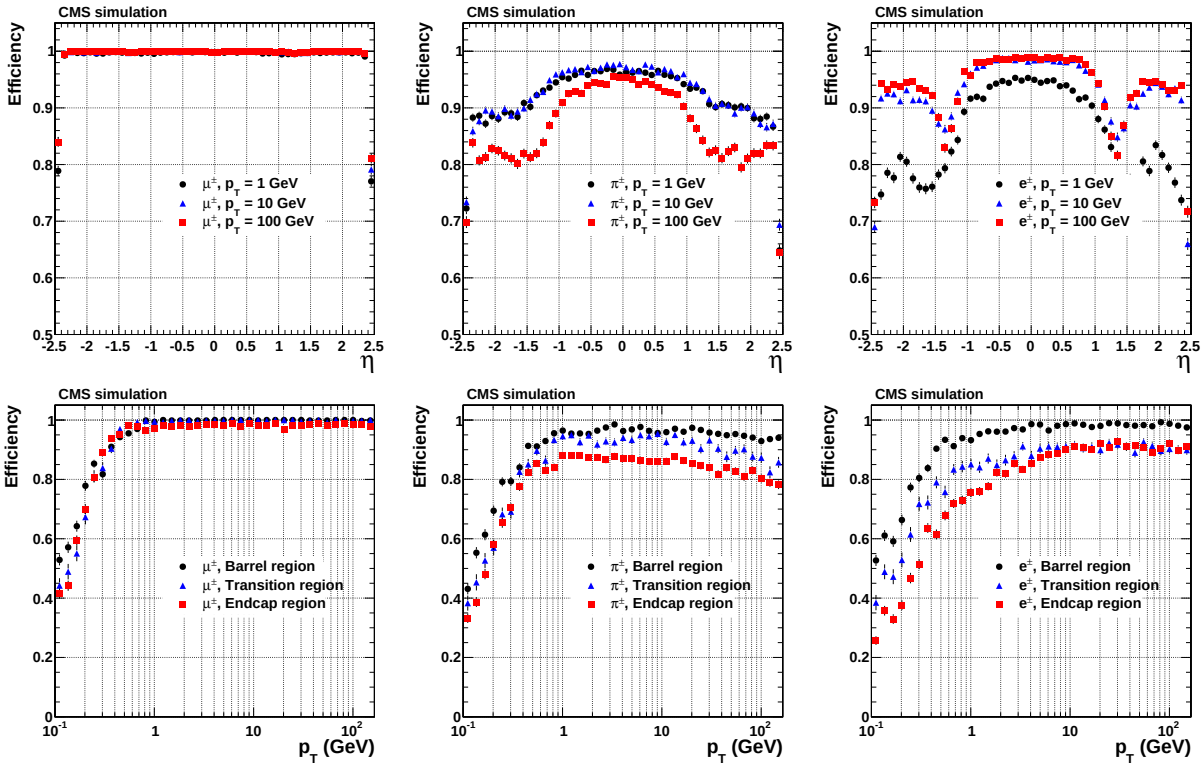

Figure 4-2. Top: Track reconstruction efficiencies as a function of η for p_T = 1, 10, and 100 GeV respectively. Bottom: Track reconstruction efficiencies as a function of p_T for different η intervals (0–0.9, 0.9–1.4 and 1.4–2.5) [13].

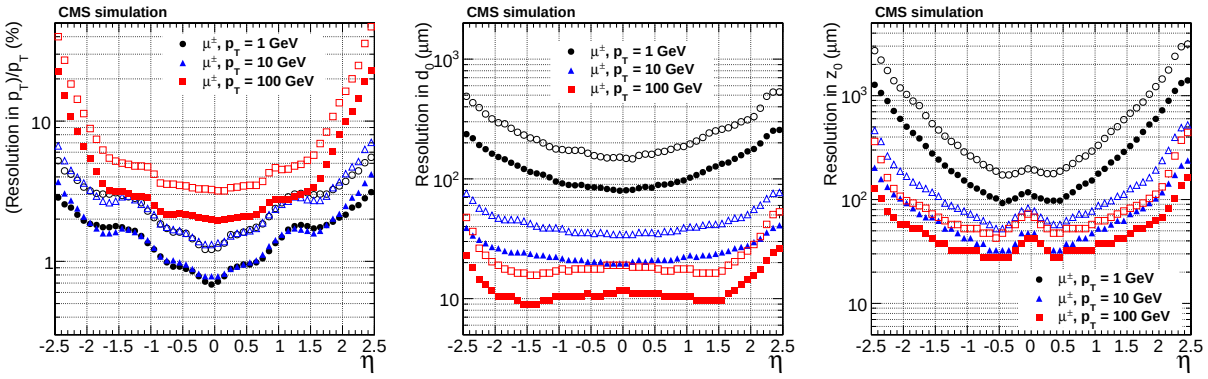

Figure 4-3. Resolution as a function of η in the muon transverse momentum p_T (left), transverse impact parameter d_0 (middle), and longitudinal impact parameter d_z (right). In each plot, resolution is shown for p_T = 1, 10, and 100 GeV respectively [13].

Fig. 4-2 shows the simulated track reconstruction efficiencies for three different kinds of charged particles: muons, charged pions and electrons. For isolated muons with $p_T > 0.9$ GeV, the efficiency is >99% over the full tracker acceptance range, independent of p_T. Pions are affected by inelastic nuclear interactions, thus have up to 20% inefficiency, worse for higher p_T and for transition and endcap regions. Electrons lose a large fraction of energy via bremsstrahlung, which causes worse efficiency at low p_T. The tracker is capable of reconstructing tracks with p_T as low as 0.1 GeV.

Fig. 4-3 shows the track parameters for muons, sampled from 68% and 90% of the entries in the distribution of the track residuals. The track residuals are the differences between reconstructed and generated track parameters. For 100 GeV muons in the central region, the p_T resolution is approximately 2–3%, the d_0 resolution is 10 μm and the d_z resolution is 30 μm. At high transverse momentum, the impact parameter resolutions are dominated by the position resolution of the innermost pixel hit; at lower momenta, the p_T, d_0, d_z resolutions are degraded due to multiple scattering.

4.4 Primary Vertex Reconstruction

A primary vertex (PV) is the position of a *pp* interaction vertex from where tracks originate. A deterministic annealing (DA) [13, 84] algorithm is used to perform track clustering to find the signal production vertex and all additional pileup vertices. The algorithm solves for the global minimum in an optimization problem that involves many degrees of freedom in a way that is analogous to how a physical system approaches the lowest energy state through a series of gradual temperature reductions.

In the CMS implementation, a track selection is first applied to choose prompt tracks. The DA process is initiated at a very high "temperature" state corresponding to having only one vertex in the event. As temperature cools down, a vertex can be split into two vertices, and the tracks are assigned to the closest vertices based on the track d_z's. The process continues until it reaches a minimum temperature where the possibility of incorrectly splitting genuine vertices becomes more important than

the efficiency of resolving nearby vertices. Then, all the candidate vertices and their associated tracks are returned.

An adaptive vertex fit [85] is done on every candidate vertex with more than one track to get the best estimate of vertex parameters, including the x, y and z position and covariance matrix, the vertex ndof, and the weights of the tracks (a weight represents the likelihood of correct vertex assignment). Reconstructed primary vertices are required to have a z position within 24 cm of the nominal origin of the detector, a radial position within 2 cm of the beamspot and the vertex fit must include at least 4 tracks.

After the selection, the reconstructed vertex with the largest value of $\sum_{\text{all tracks}} p_{\text{T track}}^2$ is identified as the primary event vertex, i.e. the production vertex of the hard-scattering process. It is used as the reference vertex for all relevant physics objects that are reconstructed with the PF algorithm.

An independent vertex reconstruction using only pixel hits is also performed. The very fast reconstruction speed that can be achieved is very valuable for many HLT applications and for seed generation used in tracking.

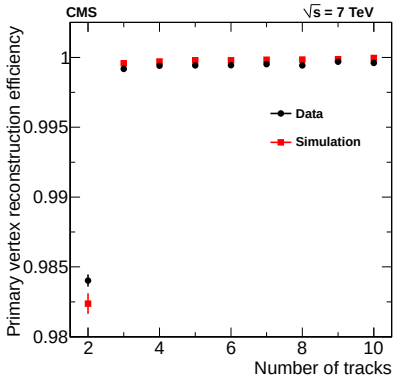

Figure 4-4. Primary vertex reconstruction efficiency as a function of the number of tracks in a cluster, measured in data and in simulation [13].

Fig. 4-4 shows the primary vertex reconstruction efficiency, which is close to 100% when more than two tracks are used to reconstruct the vertex. There is a good agreement between simulation and data. Fig. 4-5 shows the resolutions in x and z using both minimum-bias and jet-enriched data samples. As the number of tracks increase,

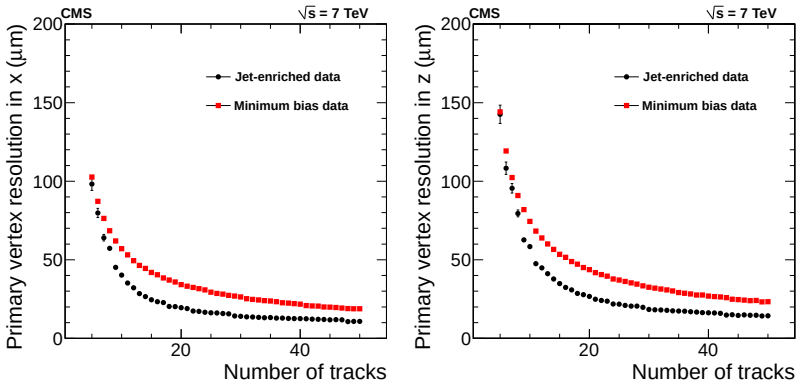

Figure 4-5. Primary vertex resolution in x (left) and z (right) as a function of the number of tracks in a cluster, measured in data selecting two kinds of events with different average track p_T values [13].

the resolutions get better. For minimum-bias events, the resolutions in x and z are less than 20 μm and 25 μm respectively; for jet-enriched events which are characteristic of interesting events, the resolutions are better, about 10 μm and 12 μm respectively.

4.5 Lepton and Tau Reconstruction

Besides the standard PF event reconstruction, dedicated algorithms for muons and electrons have been used.

Muons are reconstructed using two algorithms [86]: one in which a track in the silicon tracker is matched to signals in the muon chambers, and another in which a global track fit is performed, seeded by signals in the muon chambers. Muon candidates are required to be successfully reconstructed by both algorithms in the $|\eta| < 2.4$ range. Further identification criteria are imposed on these muon candidates to reduce the fraction of tracks misidentified as muons. These include the number of hits in the tracker and in the muon system, the quality of the global muon track fit and the consistency with the primary vertex.

Electrons are reconstructed by associating an ECAL energy cluster to a track reconstructed by a Gaussian-sum filter (GSF) algorithm in the silicon tracker [87, 88]. Electron identification relies on a multivariate technique that combines observables sensitive to the amount of bremsstrahlung emitted along the electron trajectory, the

geometrical and momentum matching between the electron trajectory and associated cluster, as well as shower shape observables in the cluster. To reject electrons produced by photon conversion, the electrons associated to a GSF track with a missing hit in any of the innermost three layers of the pixel tracker are rejected. Electrons are considered in the ECAL fiducial volume defined by $|\eta| < 1.44$ and $1.57 < |\eta| < 2.5$, excluding the transition region between the ECAL barrel and endcaps where electron reconstruction is suboptimal.

Prompt charged leptons from W or Z decays are expected to be isolated from other activity in the event. (Throughout this analysis, lepton ℓ refers to either muon or electron, but not to tau). For each lepton candidate, a cone is constructed around the track direction and the lepton isolation is quantified as:

$$R_{\text{iso}}^{\ell} = \frac{\sum_{\text{HS-charged}} p_{\text{T}} + \max\left[0, \sum_{\text{neutrals}} p_{\text{T}} + \sum_{\text{photons}} p_{\text{T}} - p_{\text{T}}(\text{PU})\right]}{p_{\text{T}}{}^{\ell}} \qquad (4\text{–}1)$$

where $\sum_{\text{HS-charged}} p_{\text{T}}$, $\sum_{\text{neutrals}} p_{\text{T}}$, and $\sum_{\text{photons}} p_{\text{T}}$ are respectively the scalar sums of the p_{T} of charged hadrons from the hard-scatter vertex, of neutral hadrons, and of photons within $\Delta R < 0.4$ around the lepton. The $p_{\text{T}}(\text{PU})$ term is a subtraction of additional energy coming from the pileup particles. The effective area method is used for electrons, $p_{\text{T}}(\text{PU}) \equiv \rho \times A_{\text{eff}}$, where A_{eff} is the geometric area of the isolation cone corrected for the residual η-dependence of the average pileup energy deposition, and ρ is the estimated per-event average energy density arising from the neutral particles. The $\Delta\beta$ method is used for muons, $p_{\text{T}}(\text{PU}) \equiv 0.5 \times \sum_{\text{PU-charged}} p_{\text{T}}$, where $\sum_{\text{PU-charged}} p_{\text{T}}$ is the scalar sum of the p_{T} of charged hadrons associated to pileup vertices. The factor of 0.5 corrects for the different fraction of charged and neutral particles in the isolation cone. If this isolation quantity exceeds approximately 10%, the lepton is rejected; the exact requirement depends on the lepton η, p_{T}, and flavor. Including the isolation requirement, the total efficiency to reconstruct muons is in the 87–91% range, depending on p_{T} and η. The corresponding efficiency for electrons is in the 81–98% range.

The hadronically-decaying taus are reconstructed using the hadron plus strips (HPS) algorithm [89] which uses charged hadrons and photons to reconstruct tau decays. (Throughout this analysis, tau τ refers to tau with 1-prong hadronic decay). Reconstructed taus are required to be in the $|\eta| < 2.1$ range. In the first step of reconstruction, charged hadrons are reconstructed using the PF algorithm. Since neutral pions are often produced in hadronic tau decays, the HPS algorithm is optimized to reconstruct neutral pions in the ECAL as objects called "strips". The strip reconstruction starts by centering one strip on the most energetic electromagnetic particle and then looking for other particles in a window of 0.05 in η and 0.20 in ϕ. Strips satisfying $p_{\mathrm{T}}(\mathrm{strip}) > 1$ GeV are combined with the charged hadrons to reconstruct the hadronic tau candidate. In the final step of reconstruction, all charged hadrons and strips are required to be contained within a shrinking cone size of $\Delta R = 2.8/p_{\mathrm{T}}(\tau)$, where $p_{\mathrm{T}}(\tau)$ is measured from the reconstructed tau candidate. Further identification criteria are imposed on the tau candidate to reduce the fraction of electron and muons misidentified as taus. These include the tau candidate passing an anti-electron discriminator and an anti-muon discriminator. The isolation requirement for taus is that the sum of transverse momenta of particle-flow charged hadron and photon candidates, with $p_{\mathrm{T}} > 0.5$ GeV and within a cone of $\Delta R < 0.5$, be less than 2 GeV. The tau reconstruction efficiency is approximately 50% while the misidentification rate from jets is about 1%.

4.6 Jet Reconstruction

Hadronic jets are the experimental signature of quarks and gluons. In the offline analysis, jets are clustered from reconstructed PF particles by the anti-k_{t} algorithm [90] with a size parameter R of 0.5. The anti-k_{t} algorithm uses a sequential recombination

scheme that is infrared- and collinear-safe, [1] as provided by the FASTJET [91] package. The algorithm clusters the pair of particles that are closest (smallest in distance measure d_{ij}), then repeatedly clusters the next closest pair, until some stopping criteria is met. d_{ij} is defined between the ith particle and jth particle as:

$$d_{ij} = \min\left(p_{\mathrm{T}_i}^{-2}, p_{\mathrm{T}_j}^{-2}\right)\frac{\Delta R_{ij}^2}{R^2} \tag{4–2}$$

where $\Delta R_{ij}^2 = \left(\eta_i - \eta_j\right)^2 + \left(\phi_i - \phi_j\right)^2$. As a result, the hardest jets in an event are usually perfectly circular in the η-ϕ space.

Jet momentum is determined as the vectorial sum of all particle momenta in the jet. Jet energy corrections (JECs) [92] are applied to calibrate on average the measured (raw) jet energy to its particle-level energy which would have been obtained if all particles inside the jet cone at the vertex were measured perfectly. CMS uses factorized corrections with the following components:

- **Offset**. Offset correction removes the contribution due to pileup and electronic noise. It is parametrized as a function of the median pileup energy density that is calculated using the jet area method [93] that takes into account the η dependence.
- **Relative**. Relative correction corrects for the non-uniformity of calorimeter response in η.
- **Absolute**. Absolute correction corrects for the non-linearity of calorimeter response in p_{T}.
- **Residual**. Residual correction accounts for the remaining small differences between data and simulation. It is only applied on real data events.

The JECs are derived from MC, and are confirmed with in situ measurements with the energy balance of dijet and photon+jet events. To mitigate pileup effects, charged

[1] Infrared safety means that the addition of a soft emission does not lead to an additional jet; collinear safety means that the collinear splitting of a particle does not split a jet into two.

hadrons that are identified as originating from pileup vertices are excluded from jet clustering. This is referred to as the "charged hadron subtraction" (CHS) approach.

For a typical jet, roughly 65% of its energy is expected to be carried by charged hadrons, 25% by photons and 10% by neutral hadrons. It is checked that the expected PF jet composition agrees very well with real data, down to 1–2% in barrel, as studied in $Z \rightarrow \mu^+\mu^-$ events [21]. The energy fractions carried by different types of PF jet constituents are shown in Fig. 4-6.

Because of the use of tracker and ECAL information that provide very precise momentum and energy measurements of charged hadrons and photons, PF jet energy resolution (JER) is vastly improved compared to calo jets, i.e. jets that are clustered from only calorimeter towers. The PF jet energy resolution typically amounts to approximately 15% at 10 GeV, 8% at 100 GeV, and 4% at 1 TeV, to be compared to 40%, 12%, and 5% when using calo jets [92]. Nonetheless, calo jets are very valuable at the HLT and continue to be used because of the much faster reconstruction speed.

Jet energy scale (JES) uncertainties arise from several sources, including:

- Physics modeling in MC such as showering, underlying event, etc.
- Detector modeling in MC such as noise, zero suppression, detector response, etc.
- Potential biases in the methodologies used to derive the jet energy corrections.

In the central region ($|\eta| < 2.5$), the total jet energy scale uncertainty is less than 3% for $p_T > 50$ GeV jets, mainly due to pileup, jet flavor and extrapolation. The JEC uncertainty increases in the forward region ($|\eta| > 2.5$) due to out-of-time pileup and time dependence. Dependence of JEC uncertainty on η and p_T are depicted in Fig. 4-7.

The core of the distribution of jet p_T resolution is found to be broader in data than in simulation by approximately 10% in the central region and up to 30% in the forward region [94]. Thus, as correction, the jet p_T in simulation is smeared by the discrepancy in width, parametrized by η, and the size of the correction is taken as the JER uncertainty.

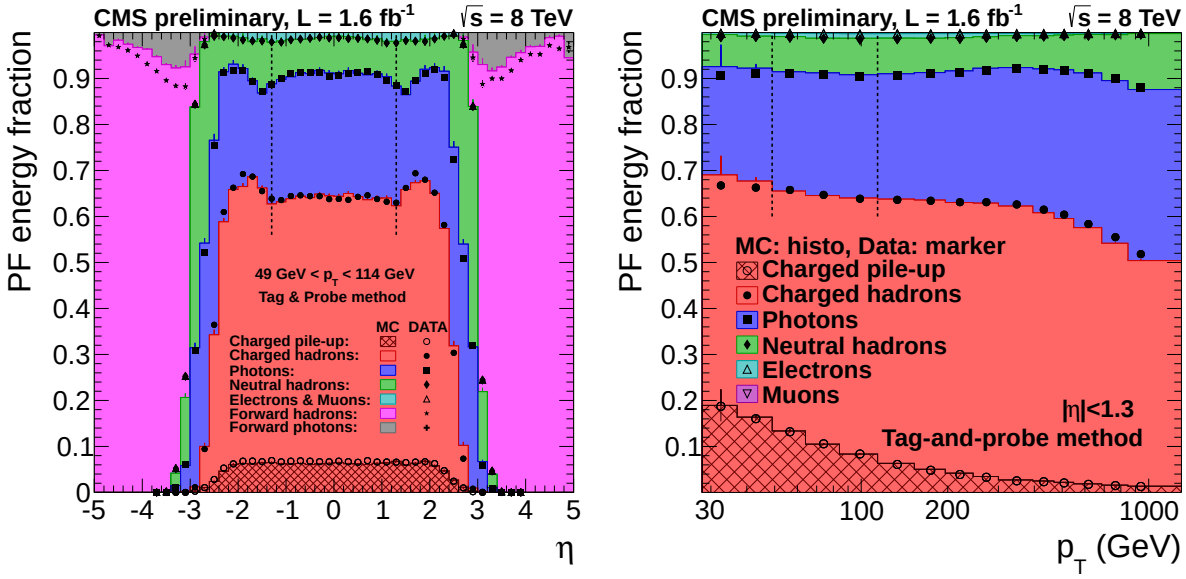

Figure 4-6. Energy fractions of different types of PF jet constituents as a function of η (left) and as a function of p_{T} (right). Jet energy corrections are applied. Very good agreement between data and simulation is found [21].

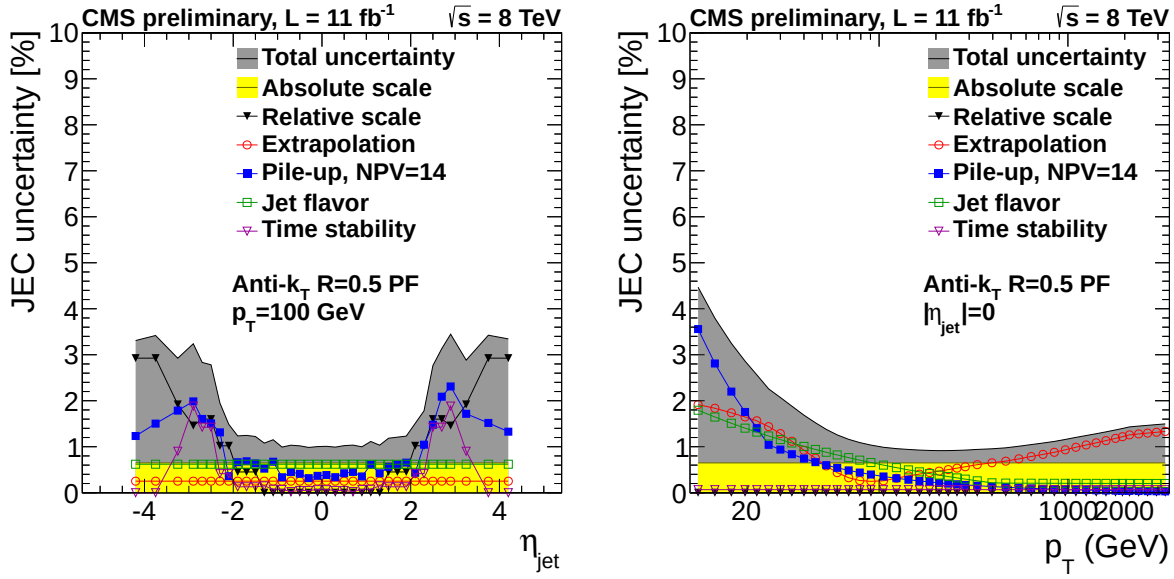

Figure 4-7. Jet energy uncertainty (combined and factorized to different sources) as a function of p_{T} for PF jets of $|\eta| = 0$ (left) and as a function of η for PF jets of $p_{\mathrm{T}} = 100\,\mathrm{GeV}$ (right) [22].

To remove spurious jets that are likely originating from instrumental effects, additional selection criteria based on the electromagnetic & hadronic energy fractions, and the number of constituents are applied. These are commonly referred to as the PF jet identification cuts [95]. Moreover, pileup collisions can produce many low p_T QCD jets. At large incidence rate, some low p_T jets can overlap and be clustered into a single high p_T jet. Such a jet is considered as a pileup jet. A multivariate pileup jet identication algorithm based on vertex information and jet shape information has been developed [96]. In this analysis, jets are required to pass the loose working points of the PF jet ID and of the pileup jet ID in order to be considered in the Higgs boson reconstruction.

4.7 Missing Transverse Energy Reconstruction

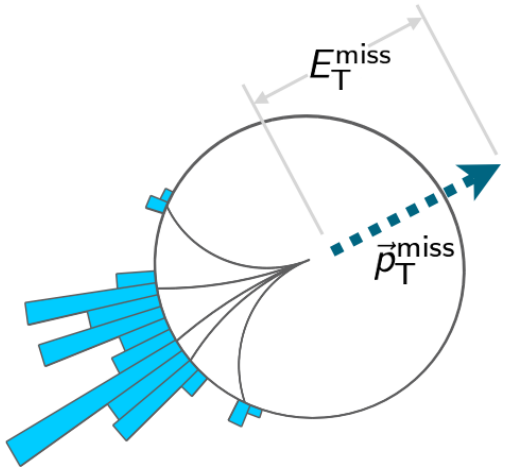

Figure 4-8. Sketch of how $\vec{p}_T^{miss}$ and E_T^{miss} are defined.

PF E_T^{miss} is constructed using PF particles and is used exclusively in the offline analysis. As a reminder, E_T^{miss} is the magnitude of the missing transverse momentum vector $\vec{p}_T^{miss}$, which is defined as the negative of the vectorial sum of the $\vec{p}_{T,i}$ of all reconstructed particles in a given event ($\vec{p}_T^{miss} \equiv -\sum_i \vec{p}_{T,i}$). A sketch of how $\vec{p}_T^{miss}$ and E_T^{miss} are defined is shown in Fig. 4-8. It can be thought of as the signature of invisible particles that either interact very minimally or do not interact at all with the detector. On

the other hand, it is also very sensitive to particle momentum mismeasurements, particle misidentification, detector malfunctions, particles impinging on poorly instrumented regions, particles from pileup interactions etc. In addition, $\sum E_T$ is defined as the associated scalar sum of the particle transverse energies ($E_T = E \sin\theta$).

Corrections are applied to the raw PF E_T^{miss} scale to make it a better estimate of true E_T^{miss} in the event [23, 97]. The so-called "Type-I" correction is a propagation of the jet energy corrections to $\vec{p}_T^{miss}$. The correction is applied according to the following:

$$\text{corr } \vec{p}_T^{miss} = \vec{p}_T^{miss} - \sum_{\text{jets}} \left(\text{corr } \vec{p}_{T,\text{jet}} - \vec{p}_{T,\text{jet}}\right) \tag{4–3}$$

considering all jets that have less than 0.9 of their energy in the ECAL and corrected p_T > 10 GeV.

The so-called "Type-0" correction aims to reduce the dependency on the number of pileup interactions. In the minimum bias pp interactions, E_T^{miss} is expected to be close to zero, so the vectorial $\vec{p}_T$ sum of charged particles is expected to be equal to that of neutral particles. However, due to the non-linearity response and minimum energy thresholds of the calorimeters, $\vec{p}_T^{miss}$ tend to be aligned with the direction of the vectorial $\vec{p}_T$ sum of neutral particles. The correction is derived from simulation, parametrized as a function of the direction and magnitude of the vectorial $\vec{p}_T$ sum of pileup charged particles.

In the 2012 data, several sources of events with anomalously large E_T^{miss} are found, including detector noise in ECAL & HCAL, non-functioning ECAL channels, non-instrumented regions of the detector, non-collision particles from cosmic rays and beam halo, etc. Specific cleaning algorithms have been developed to identify and remove these anomalous-E_T^{miss} events [98]. After the event cleaning, the agreement between data and simulation improves significantly, as depicted in Fig. 4-9.

E_T^{miss} performance is typically studied in events where a well-measured Z boson or isolated γ is identified. In such an event, the vector boson transverse momentum is

denoted by $\vec{q}_T$; the hadronic recoil, defined as the vectorial $\vec{p}_T$ sum of all reconstructed particles except the vector boson, is denoted by $\vec{u}_T$. By the momentum conservation in the transverse plane, $\vec{q}_T + \vec{u}_T + \vec{p}_T^{\,miss} = 0$. $\vec{u}_T$ provides a measure of the induced E_T^{miss} in an event. Resolution of E_T^{miss} reconstruction is assessed by looking at the spread of $\vec{u}_T$. This is shown in Fig. 4-10 separately for the parallel ($u_{||}$) and perpendicular ($u_\perp$) components of $\vec{u}_T$ w.r.t. the direction of $\vec{q}_T$. One can see that each additional pileup interaction degrades the PF E_T^{miss} resolution by 3.3–3.7 GeV in quadrature. Though, the PF E_T^{miss} performs significantly better than the calo E_T^{miss}, which is constructed using only calorimeter towers.

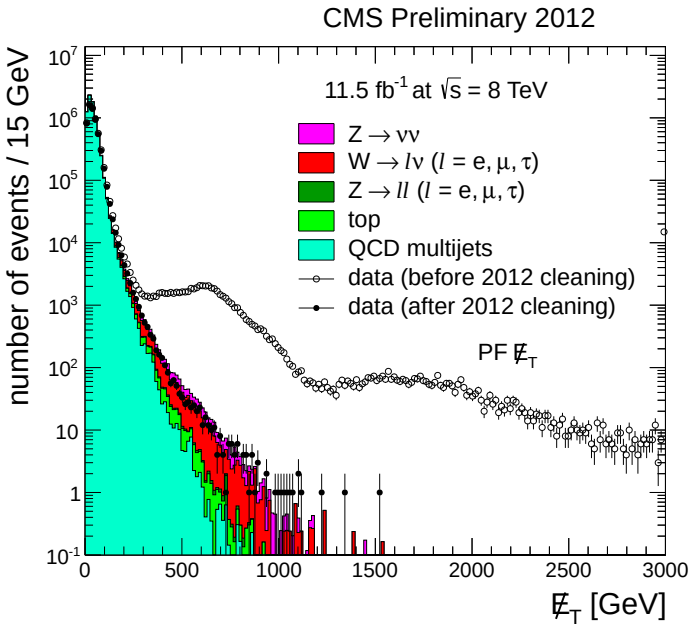

Figure 4-9. PF E_T^{miss} distributions for dijet events without 2012 cleaning algorithms applied (open markers), with 2012 cleaning algorithms applied (filled markers), and events from MC (filled histograms) [23].

4.8 *b* Jet Identification

b tagging refers to the identification of jets containing *B* hadrons (e.g. $B^{\pm}$, B^0, B_s, B_c, Λ_b) which result from the hadronization of *b* quark. *B* hadrons decay via the weak interaction with a relatively long lifetime $\tau \sim 1.572 \pm 0.009$ ps [99], corresponding to

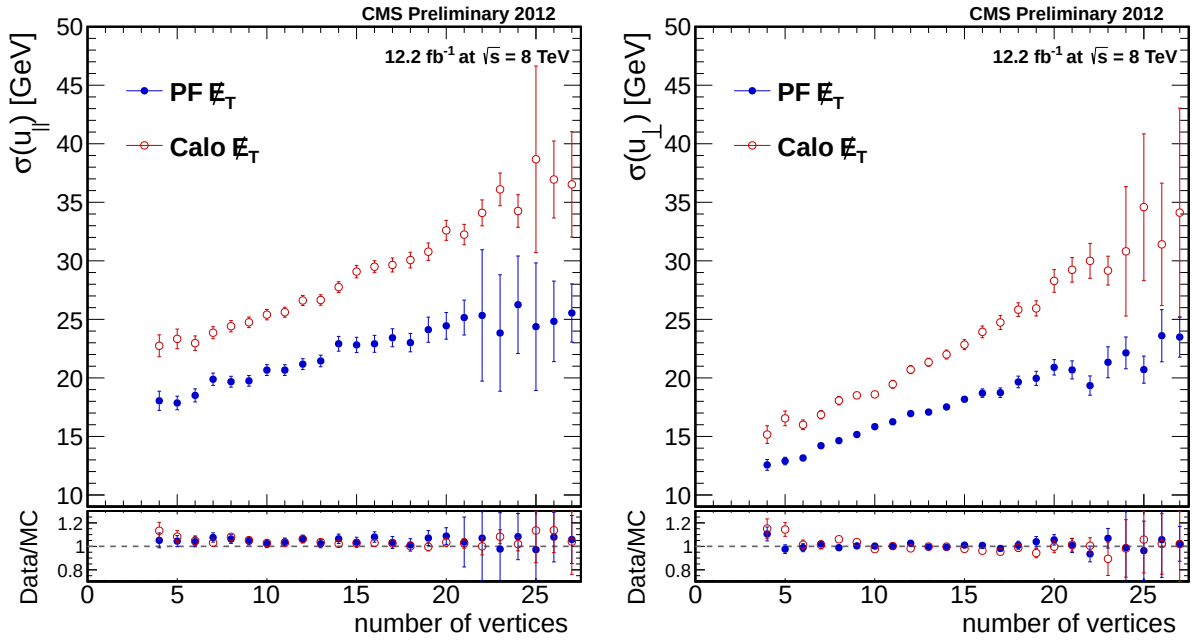

Figure 4-10. Resolutions of the parallel (left) and perpendicular (right) components of the hadronic recoil for PF $E_{\mathrm{T}}^{\mathrm{miss}}$ and Calo $E_{\mathrm{T}}^{\mathrm{miss}}$ *vs.* the number of vertices, measured in $Z \rightarrow \mu^{+}\mu^{-}$ events. The bottom panel shows the data/MC ratio [23].

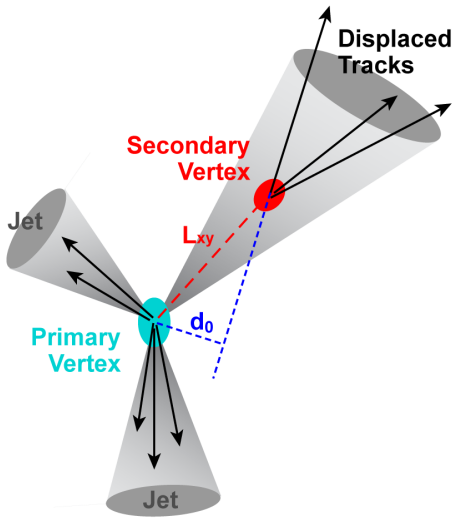

Figure 4-11. Illustration of a *b* jet with displaced tracks and a secondary vertex.

a decay length $c\tau \sim 450\,\mu$m. For a 30 GeV *b* quark, the Lorentz boost factor $\beta\gamma$ is about 6 and the decay length is boosted to $\beta\gamma c\tau \sim 2.7$ mm. Consequently, *b* jets can have highly displaced tracks, i.e. tracks with significant impact parameter d_0, and/or a secondary vertex (SV), i.e. the decay vertex of *B* hadron which is away from its production vertex. An illustration of a *b* jet is shown in Fig. 4-11.

B hadrons and *b* jets have other distinctive properties. A *b* quark has a mass of 4.78 ± 0.06 GeV (in the 1S scheme) [19]. This large mass leads to large invariant mass of the system of tracks at the secondary vertex. The semileptonic branching fraction of *B* hadrons is large; with $\mathcal{B}(b \to \ell^-) = 0.1071 \pm 0.0022$ & cascade $\mathcal{B}(b \to c \to \ell^+) = 0.0801 \pm 0.0018$ for each flavor of $\ell = e, \mu$ [100]. Moreover, a *b* quark has a hard fragmentation function, meaning that the *B* hadron carries a large fraction of the original *b* quark momentum. Also, the multiplicity of charged particles per *B* hadron decay is high: about 5 on average.

Several *b* tagging algorithms have been designed to discriminate *b* (denoted by heavy flavor (HF)) jets from *u*,*d*,*s*,*c*, or *g* (denoted by light flavor (LF)) jets. The best performing ones include Track Counting High Purity (TCHP), Jet Probability (JP), and Combined Secondary Vertex (CSV) [24, 25]. Each tagger has 3 operating points — loose (L), medium (M), and tight (T) — corresponding to misidentification (misid) probability for *udsg* jets of 10%, 1%, and 0.1%, respectively. In the simulation, a reconstructed jet is matched with a generated parton and assigned the flavor of the parton if the parton is within $\Delta R < 0.3$ of the jet. Should more than one parton be matched to a given jet, the flavor assigned is that of the heaviest parton. In particular, jets originating from $g \to b\bar{b}$ are classified as *b* jets.

In this analysis, the CSV tagger is chosen because of its better discriminating power against *c* jets, which constitute a challenging background. CSV is a multivariate likelihood-based discriminator that combines the information about track impact parameters and secondary vertices within jets. It outputs a discriminant value between 0

(least b-like) and 1 (most b-like). For b tagging purpose, d_0 is redefined as the distance to the primary vertex at the point of closest approach (PCA) of the trajectory in the transverse plane. The sign of d_0 is the sign of the scalar product of the jet axis with the vector pointing from PV to PCA. The resolution of d_0 depends strongly on the p_T and η of the track, thus the d_0 significance, defined as the ratio of the d_0 to its estimated uncertainty, is used as an observable. Typical d_0 uncertainty is 10–100 μm. Only tracks fulfilling certain criteria on e.g. p_T, number of hits, normalized χ^2, impact parameters, and angular and spatial distance between track and jet are used in the tagger.

Secondary vertices within jets are reconstructed by using the adaptive vertex reconstructor (AVR) algorithm, which is an iterative application of adaptive vertex fit [25, 101]. The flight distance significance, defined as the distance between PV and SV divided by the uncertainty of the distance, is used as an observable. Other properties of secondary vertices include the invariant mass and vectorial $\vec{p_T}$ sum of the associated tracks at SV, the track multiplicity at SV, etc. Only secondary vertices fulfilling certain criteria on e.g. number of shared tracks with PV, SV flight direction, SV mas, and compatibility with K^0 hypothesis are considered. When no secondary vertex is found, CSV combines tracks with d_0 significance > 2 to make a "pseudo vertex", so that a subset of SV-based quantities can still be computed without performing an actual vertex fit. When even this is not possible, CSV reverts to using only the track impact parameters.

The variables used in CSV include [25]:

- Type of secondary vertex: real, pseudo or nothing;
- 2D flight distance significance (in the transverse plane) of SV;
- Secondary vertex mass;
- Number of tracks at SV;
- Number of tracks in the jet;

- Ratio of energy carried by tracks at SV w.r.t. all tracks in the jet;
- Pseudorapidity of each track at SV w.r.t. the jet axis;
- 2D d_0 significance of the first track that raises the invariant mass above the charm quark mass of 1.5 GeV when subsequently summing up tracks ordered by decreasing d_0 significance;
- 3D d_0 significance of each track in the jet.

The discriminator is trained separately for every type of secondary vertex, separately for *b* vs *c* and *b* vs *udsg*, in bins of $p_T \otimes \eta$. The distributions of 3D d_0 significance, 3D flight distance significance of SV, secondary vertex mass, and CSV discriminant are shown in Fig. 4-12.

At the loose operating point (10% misid probability), CSV achieves a *b* tagging efficiency of about 85%. For higher *b* jet purity, approximately 70% *b* tagging efficiency is achieved with a misid probability of only 1.5%. Efficiency for *b* jets and misid probabilities for light flavor jets using the JP tagger with loose operating point (JPL) and using the CSV tagger with medium operating point (CSVM) [2] are presented in Fig. 4-13. These performance measurements are obtained directly from multijet events with soft muon (muon from semileptonic *B* hadron decay) and $t\bar{t}$ events. The CSVM tagger has slightly lower efficiency but much lower misid probability. The *b* jet identification efficiency and the *c* jet misidentification probability increase with p_T up to $p_T < 100$–200 GeV, and decrease after that. This dependence is due to a convolution of the track impact parameter resolution (which is worse at low p_T), of the hadron decay lengths (which scale with p_T), of the secondary vertex reconstruction efficiency (worse with collimated tracks at high p_T), and of the track selection criteria The misidentification probability for *udsg* jets rises continuously with p_T due to the logarithmic increase of the

[2] CSVL corresponds to CSV > 0.244, CSVM corresponds to CSV > 0.679, and CSVT corresponds to 0.898.

number of particles in jets and the higher fraction of merged hits in the innermost layers of the tracking system.

Measurements of the b tagging efficiency are performed with a number of methods (PtRel, System8, IP3D, LT, FTM, FTC, ...) that cross check one another [24]. When there is a discrepancy between data and simulation, corrections can be applied to simulated events using a scale factor SF_b, defined as the ratio of the efficiency in data to the efficiency in MC. The efficiency measured in data and predicted in simulation using the FTC method in $t\bar{t}$ events, as well as the scale factor SF_b are presented in Fig. 4-14. At a reference p_T of 100 GeV, the SF_b for the L, M, T operating points are 0.980 ± 0.016, 0.965 ± 0.024, and 0.935 ± 0.032, respectively [102].

Similarly, the scale factor for the misid probability, SF_{light} has been measured in and is also shown in Fig. 4-14. In the p_T range of 80–120 GeV, the SF_{light} for the 3 operating points with statistical and systematic uncertainties are $1.10 \pm 0.01 \pm 0.05$, $1.17 \pm 0.02 \pm 0.15$, and $1.26 \pm 0.07 \pm 0.28$, respectively. The scale factors SF_b and SF_{light}, including their uncertainties, are applied in the analysis via a "CSV reshaping" method [103]. The method derives the differential scale factor for all CSV discriminant values, as a function of jet p_T, η, and flavor, and adjusts the jet CSV values in the simulated events such that the b tagging efficiency and the misid probability reproduce the performance in data.

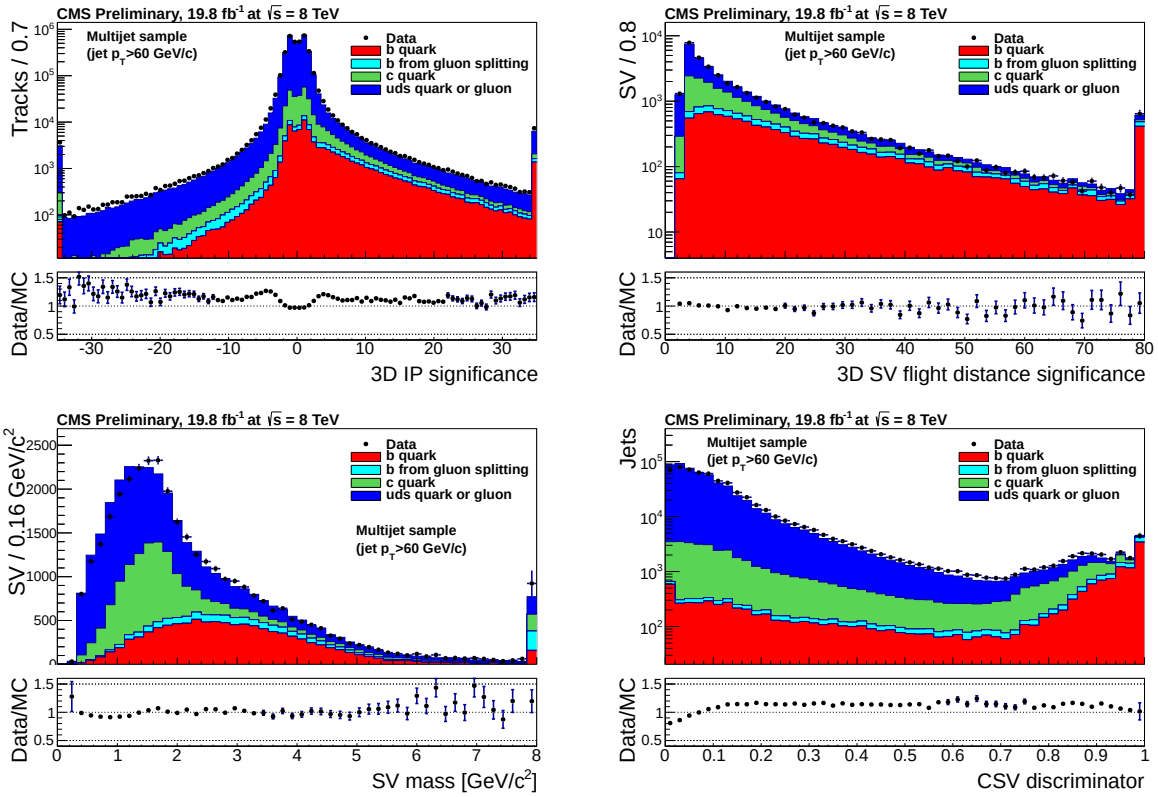

Figure 4-12. Distributions of 3D impact parameter significance (top left), 3D flight distance significance of secondary vertex (top right), secondary vertex mass (bottom left), CSV discriminant (bottom right) in QCD multijet events [24].

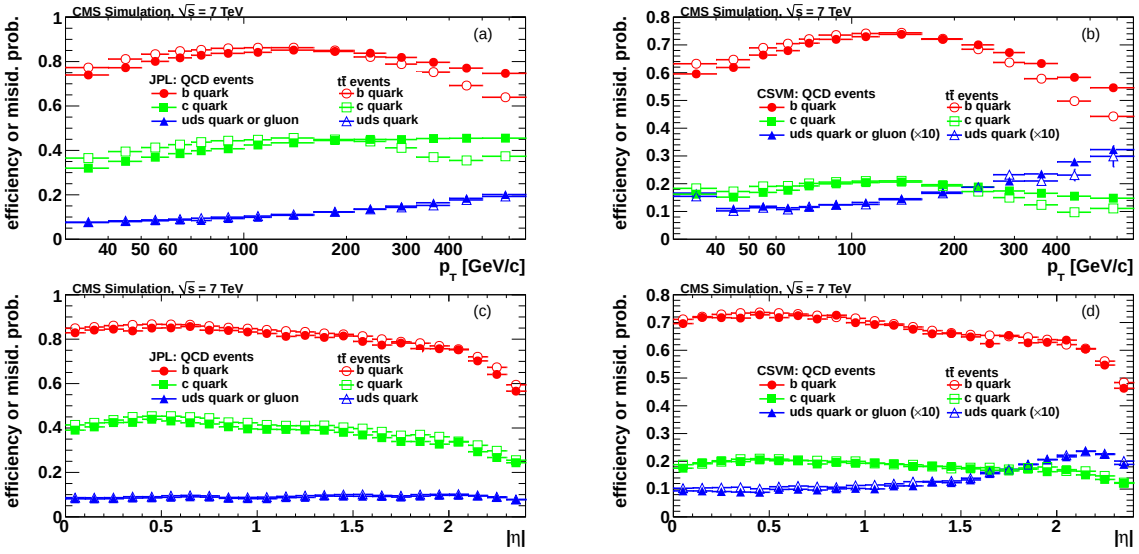

Figure 4-13. Efficiency for *b* jets and misidentification probabilities for light flavor jets using the (a, c) JPL tagger and (b, d) CSVM tagger as a function of (a, b) jet p_T and (c, d) $|\eta|$ in multijet events (filled symbols) and $t\bar{t}$ events (open symbols) [25].

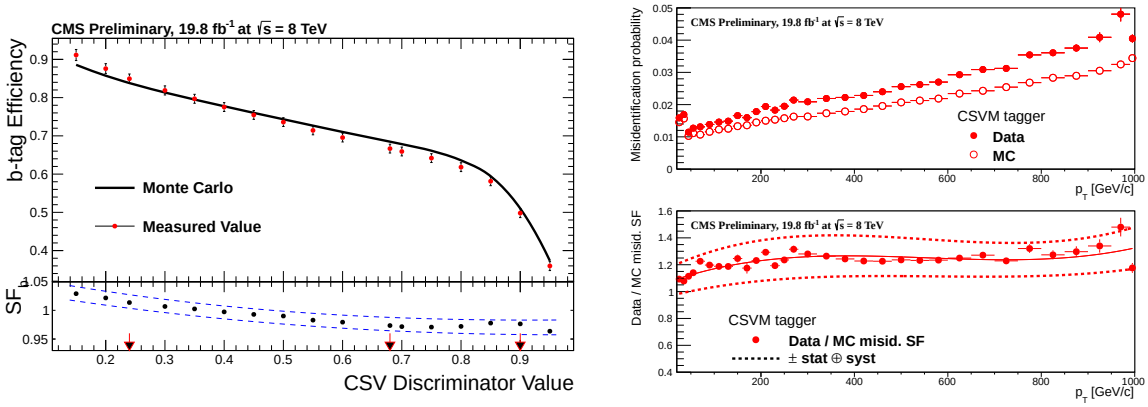

Figure 4-14. Left: *b* tagging efficiency in data and simulation and the scale factor SF_b in $t\bar{t}$ events. Right: Misidentification probability in data and simulation and the scale factor SF_{light} for the CSVM operating point in multijet events with a muon [24].

CHAPTER 5
ANALYSIS STRATEGY

5.1 Overview of Analysis Strategy

The analysis strategy is based on the reconstruction and identification of the $Z \to \nu\bar{\nu}$ decay and $H \to b\bar{b}$ decay in the kinematic region where the Z and H point to opposite directions in the transverse plane. The Higgs boson candidate is made up of the pair of jets (dijets) in the event, each with $|\eta| < 2.5$ and $p_{\mathrm{T}} > 30\,\mathrm{GeV}$, for which the the vectorial sum of their transverse momenta, $p_{\mathrm{T}}(\mathrm{jj})$, is the highest. This is motivated by the fact that the two daughters carry the original p_{T} of the Higgs boson, so the $p_{\mathrm{T}}(\mathrm{jj})$ of the two daughters is usually higher than the $p_{\mathrm{T}}(\mathrm{jj})$ of a random combination of jets. These jets are then required to be tagged by the CSV algorithm, with the value of the CSV discriminator above a minimum threshold. The Z boson candidate is identified by the presence of non-negligible amount of $E_{\mathrm{T}}^{\mathrm{miss}}$ in the event. Under the hypothesis of VH production, the Z and H candidates typically recoil away from each other with a large azimuthal opening angle, $\Delta\phi(\mathrm{jj})$, between them, resulting in the back-to-back topology. Furthermore, the dijets are expected to be central due to the large invariant mass of the VH system. The final state should be clean with minimal additional jet activity due to initial state radiation, as VH is initiated by $q\bar{q}$ annihilation at the leading order.

Background events arise from production of W and Z bosons in association with jets (V+jets), single top production ($t/\bar{t}$), top pair production ($t\bar{t}$), dibosons (VV), and QCD multijet processes:

- V+2 jets production has about 10^3 larger cross section compared to VH production. Its kinematics is similar to that of the VH. This background process can be reduced by requiring two b tagged jets. After b tagging, $Z(\nu\bar{\nu})$+$b\bar{b}$ and $W(\ell\nu)$+$b\bar{b}$ (when ℓ is not reconstructed or is out of acceptance) constitute an

Text and materials in this Chapter were adapted from the CMS publication Phys. Rev. D 89, 012003 (2014), American Physical Society. The author's work contributed to the publication.

irreducible background. Compared to VH, the p_{T} spectrum of V+jets production is softer and the dijet invariant mass spectrum is a sharply falling one.

- VV, especially $Z(\nu\bar{\nu})Z(b\bar{b})$, has only a few times larger cross section compared to VH production. It is also an irreducible background, with Z mass being very close to the search region for Higgs mass.

- $t/\bar{t}$ and $t\bar{t}$ production with lost lepton in $W \to \ell\nu$ decays also makes final state with $E_{\mathrm{T}}^{\mathrm{miss}}$ and $b\bar{b}$. Top background usually has different event topology such as higher jet multiplicity.

- QCD multijet production can have fake $E_{\mathrm{T}}^{\mathrm{miss}}$ due to jet energy mismeasurements. The $E_{\mathrm{T}}^{\mathrm{miss}}$ direction in QCD dijet events tend to align with one of the jets.

The background processes are substantially reduced by applying b tagging requirements and kinematic requirements. Large $E_{\mathrm{T}}^{\mathrm{miss}}$ requirement is particularly effective in reducing the QCD multijet background. Events with identified prompt leptons and hadronic taus from W and Z decays are already included in the search regions of other $VH(b\bar{b})$ channels. To keep the channels orthogonal to one another, lepton veto and hadronic tau veto are applied in this channel.

The search region is categorized into 3 boost regions defined by: $100 < E_{\mathrm{T}}^{\mathrm{miss}} < 130, 130 < E_{\mathrm{T}}^{\mathrm{miss}} < 170, E_{\mathrm{T}}^{\mathrm{miss}} > 170$ GeV. They are referred to as the low-, intermediate-, and high-boost regions, respectively. Because of different signal and background content, each boost region has different sensitivity. The analysis strategy is optimized and performed individually in each region. The results from all regions are then combined.

In order to improve the b jet energy resolution, which in turn improves the dijet invariant mass resolution, a multivariate regression technique using the boosted decision tree (BDT) algorithm is applied to further correct b jet energies. Finally, a multivariate signal classifier, again using BDT, is trained to give the best separation between signal and various background hypotheses. It is the final BDT discriminant, i.e. the output of the signal classifier, on which a binned maximum likelihood fit is performed,

using the signal and background templates from the simulation as input, to extract the fraction of signal events (if they exist).

This Chapter is organized as follows: the samples and the triggers used in this analysis are listed in Secs. 5.2 and 5.3. The multivariate b jet energy regression is explained in Sec. 5.4. The event selection and the use of BDT signal classifier are described in detail in Sec. 5.5. Corrections to Monte Carlo simulation are necessary to reduce the discrepancy between simulation and real data. These corrections are derived from background-enriched control regions in real data and they are described in Sec. 5.6. Finally, various systematic uncertainties associated to the background estimation and signal prediction are given in Sec. 5.7. Complete documentation about this analysis can be found in Ref. [103].

5.2 Data and Simulation

Simulated samples of signal and background events are produced using various MC event generators, with the CMS detector response modeled with GEANT4 [79]. They provide guidance in the optimization of the analysis and the initial estimate of various background contributions.

The Higgs boson signal samples are produced using the NLO POWHEG [104] event generator. Production cross sections for WH and ZH and $H \to b\bar{b}$ branching fractions as a function of m_H are listed in Table 5-1. [1] The MADGRAPH 5.1 [106] generator is used for the W+jets, Z+jets, VV, and $t\bar{t}$ samples. The single top samples, including the tW-, t-, and s-channel processes, are produced with POWHEG. The QCD multijet samples are generated by PYTHIA 6.4 [107]. The production cross sections for the VV and $t\bar{t}$ samples are rescaled to the cross sections calculated using the NLO MCFM generator [108], while the cross sections for the W+jets and Z+jets samples are rescaled to NNLO cross sections calculated using the FEWZ program [109–111].

[1] Note that the CERN Report 2 [105] numbers are used at the time of the analysis.

The default set of parton density functions used to produce the NLO POWHEG samples is the NLO MSTW2008 set [112], while the LO CTEQ6L1 set [113] is used for the other samples. For parton showering and hadronization the POWHEG and MADGRAPH samples are interfaced with HERWIG++ [114] and PYTHIA, respectively. The PYTHIA parameters for the underlying event description are set to the Z2* tune [115]. The TAUOLA [116] library is used to simulate tau decays.

This analysis is performed in the boosted regime, so differences in the p_T spectrum of the V and H bosons between data and MC may introduce systematic effects in the signal acceptance and efficiency estimates. Two calculations are available that evaluate the NLO electroweak (EWK) [117–119] and NNLO QCD [120] corrections to VH production in the boosted regime. The relative NLO EWK corrections are calculated using HAWK as a function of the generated p_T of the vector boson separately for the $W(\ell\nu)H(b\bar{b})$, $Z(\ell\ell)H(b\bar{b})$, and $Z(\nu\bar{\nu})H(b\bar{b})$ modes. For the NNLO QCD correction, the relative efficiency of the veto on additional jet activity is calculated as a function of the generated p_T of the Higgs boson. Both the EWK and QCD corrections are shown in Fig. 5-1 and are applied to the signal samples.

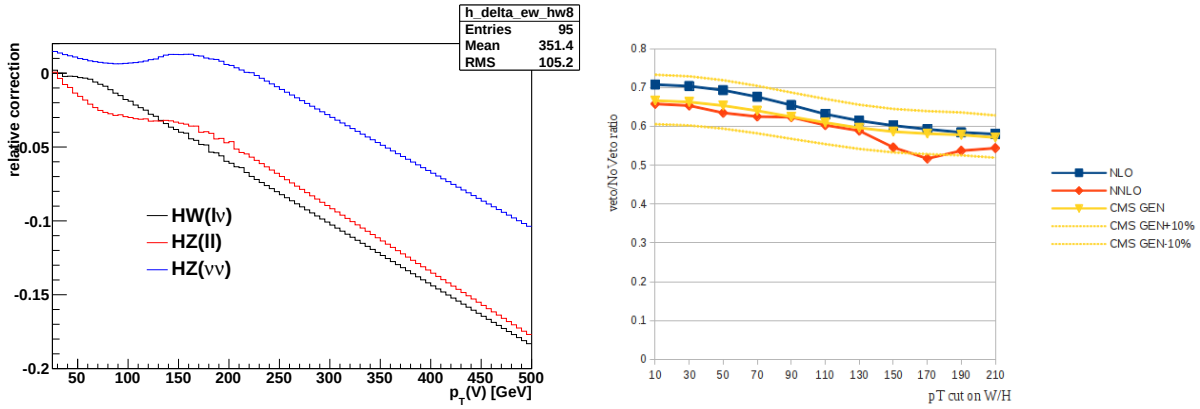

Figure 5-1. Left: Relative NLO EWK corrections as a function of the vector boson p_T for different modes. Right: Relative efficiency of the veto on additional jet activity as a function of the Higgs boson p_T.

The full 2012 MET primary dataset recorded at $\sqrt{s} = 8$ TeV are used, excluding runs 207883–208307. Those runs, containing about 0.6 fb^{-1} of integrated luminosity,

are affected by a pixel misalignment issue that impacts b tagging. Table 5-2 reports the real data samples used in this analysis and their approximate integrated luminosities. Table 5-3 reports the simulation samples (from `Summer12` MC production campaign) and their cross sections (multiplied by branching fractions where applicable). Appropriate pileup reweighting is applied to the simulation samples [121, 122].An example of the effect of reweighting as validated in the $t\bar{t}$-enriched control region is shown in Fig. 5-2. The reweighted distribution of the number of reconstructed primary vertices in simulation is in agreement with data.

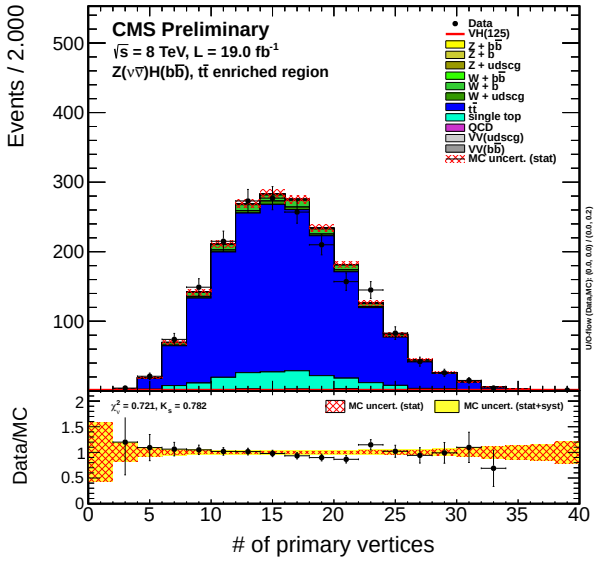

Figure 5-2. Distribution of the number of reconstructed primary vertices in data compared to simulation in the $t\bar{t}$ control region.

5.3 Trigger

Events that are consistent with the signal hypothesis are collected by dedicated trigger paths, all requiring $E_{\mathrm{T}}^{\mathrm{miss}}$ to be above a given threshold. Extra requirements are added to keep the trigger rates manageable as the instantaneous luminosity increases and to reduce the $E_{\mathrm{T}}^{\mathrm{miss}}$ thresholds in order to increase signal acceptance. In particular, at $E_{\mathrm{T}}^{\mathrm{miss}} < 130\,\mathrm{GeV}$, triggers that require b tagging are utilized. As instantaneous luminosity exceeded $3 \times 10^{33}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$, the triggers from the 2012A period were revised

Table 5-1. *WH* and *ZH* production cross sections at $\sqrt{s} = 8\,\text{TeV}$ and $H \to b\bar{b}$ branching fractions for $105 \leq m_H \leq 150\,\text{GeV}$, as well as the associated theoretical uncertainties.

m_H[GeV]	$\sigma(WH)$[pb]	$\pm\sigma(WH)$[%]	$\sigma(ZH)$[pb]	$\pm\sigma(ZH)$[%]	$\mathcal{B}(H \to b\bar{b})$	$\pm\mathcal{B}(H \to b\bar{b})$[%]
105	1.2290	+3.6, −4.1	0.6750	+4.9, −4.9	0.771	+1.9, −2.0
110	1.0600	+3.9, −4.4	0.5869	+5.4, −5.4	0.745	+2.1, −2.2
115	0.9165	+4.0, −4.5	0.5117	+5.6, −5.5	0.704	+2.4, −2.5
120	0.7966	+3.5, −4.0	0.4483	+5.0, −4.9	0.648	+2.8, −2.8
125	0.6966	+3.7, −4.1	0.3943	+5.1, −5.0	0.577	+3.2, −3.2
130	0.6095	+3.7, −4.1	0.3473	+5.4, −5.3	0.493	+3.7, −3.8
135	0.5351	+3.5, −4.1	0.3074	+5.4, −5.2	0.403	+4.2, −4.3
140	0.4713	+3.6, −4.2	0.2728	+5.6, −5.4	0.315	+3.4, −3.4
145	0.4164	+3.9, −4.5	0.2424	+6.0, −5.8	0.232	+3.7, −3.7
150	0.3681	+3.4, −4.0	0.2159	+5.7, −5.4	0.157	+4.0, −4.0

Table 5-2. List of 2012 CMS data samples from `MET` primary dataset used in this analysis and their approximate integrated luminosities.

Period	Dataset	$\mathcal{L}[\text{fb}^{-1}]$
2012A	/MET/Run2012A-13Jul2012-v1 (runs 207883–208307 excluded)	0.796
	/MET/Run2012A-recover-06Aug2012-v1	0.081
2012B	/MET/Run2012B-13Jul2012-v1	4.412
2012C	/MET/Run2012C-24Aug2012-v1	0.474
	/MET/Run2012C-EcalRecover_11Dec2012-v1	0.133
	/MET/Run2012C-PromptReco-v2	6.330
2012D	/MET/Run2012D-PromptReco-v1	6.712
	Total	18.938

to overcome harsher pileup condition. These triggers then remained stable during the 2012B,C,D periods.

The triggers used in this analysis are:

- `HLT_PFMET150` (2012ABCD). This trigger collects any event with online $E_{\text{T}}^{\text{miss}} > 150\,\text{GeV}$ throughout the 2012 data taking. This trigger attains a plateau efficiency of 99% at roughly offline $E_{\text{T}}^{\text{miss}}$ of 190 GeV.

- `HLT_DiCentralPFJet30_PFMHT80` (2012A), `HLT_DiCentralJetSumpT100_dPhi05_DiCentralPFJet60_25_PFMET100_HBHENoiseCleaned` (2012BCD). The 2012A version requires the presence of two central (i.e. $|\eta| < 2.6$ at HLT) jets with $p_{\text{T}} > 30\,\text{GeV}$ and $E_{\text{T}}^{\text{miss}}$ threshold of 80 GeV; the 2012BCD version requires $E_{\text{T}}^{\text{miss}} > 100\,\text{GeV}$, at least two central jets with a dijet vectorial sum $p_{\text{T}} > 100\,\text{GeV}$ and individual jet p_{T} above 60 and 25 GeV respectively. Any event with a jet of $p_{\text{T}} > 40\,\text{GeV}$ closer than 0.5 radians in azimuthal angle to the $E_{\text{T}}^{\text{miss}}$

Table 5-3. List of 8 TeV Monte Carlo samples from CMS `Summer12` campaign used in this analysis. Different generator binnings and different generators have been tested to maximize the available statistics and to study systematic uncertainties.

Process	Dataset	$\sigma \times \mathcal{B}$ [pb]
$VH(m_H$ =125)	ZH_ZToNuNu_HToBB_M-125_8TeV-powheg-herwigpp	0.4153 × 0.200 × 0.577
	WH_WToLNu_HToBB_M-125_8TeV-powheg-herwigpp	0.7046 × 0.326 × 0.577
$Z(\nu\bar{\nu})$+jets	ZJetsToNuNu_PtZ-70To100_8TeV-madgraph	127.61
	ZJetsToNuNu_PtZ-100_8TeV-madgraph	83.005
	ZJetsToNuNu_50_HT_100_TuneZ2Star_8TeV_madgraph	495.56
	ZJetsToNuNu_100_HT_200_TuneZ2Star_8TeV_madgraph	208.39
	ZJetsToNuNu_200_HT_400_TuneZ2Star_8TeV_madgraph	51.240
	ZJetsToNuNu_400_HT_inf_TuneZ2Star_8TeV_madgraph	6.8562
	ZJetsToNuNu_Pt-100_8TeV-herwigpp (for syst.)	83.005
$W(\ell\nu)$+jets	WJetsToLNu_PtW-70To100_TuneZ2star_8TeV-madgraph	557.57
	WJetsToLNu_PtW-100_TuneZ2star_8TeV-madgraph	297.57
	WJetsToLNu_PtW-100_TuneZ2star_8TeV_ext-madgraph-tarball	297.57
	WJetsToLNu_PtW-180_TuneZ2star_8TeV-madgraph-tarball	34.293
	WJetsToLNu_PtW-100_8TeV-herwigpp (for syst.)	297.57
$t\bar{t}$	TTJets_FullLeptMGDecays_8TeV-madgraph	26
	TTJets_SemiLeptMGDecays_8TeV-madgraph	104
	TTJets_HadronicMGDecays_8TeV-madgraph	104
	TT_CT10_TuneZ2star_8TeV-powheg-tauola (for syst.)	234
$t/\bar{t}$	T_tW-channel-DR_TuneZ2star_8TeV-powheg-tauola	11.1
	Tbar_tW-channel-DR_TuneZ2star_8TeV-powheg-tauola	11.1
	T_s-channel-DR_TuneZ2star_8TeV-powheg-tauola	3.79
	Tbar_s-channel-DR_TuneZ2star_8TeV-powheg-tauola	1.76
	T_t-channel_TuneZ2star_8TeV-powheg-tauola	56.4
	Tbar_t-channel_TuneZ2star_8TeV-powheg-tauola	30.7
VV	WW_TuneZ2star_8TeV_pythia6_tauola	56.75
	WZ_TuneZ2star_8TeV_pythia6_tauola	33.85
	ZZ_TuneZ2star_8TeV_pythia6_tauola	8.297

direction is vetoed. The 2012BCD trigger achieves a plateau efficiency of 95% at roughly offline $E_{\mathrm{T}}^{\mathrm{miss}}$ of 160 GeV.

- `HLT_DiCentralJet20_CaloMET65_BTagCSV07_PFMHT80` (2012A), `HLT_DiCentralPFJet30_PFMET80_BTagCSV07` (2012BCD). The 2012A version requires two central jets with $p_{\mathrm{T}} > 20$ GeV and that at least one central jet with $p_{\mathrm{T}} > 20$ GeV be tagged by the online CSV algorithm with a threshold of 0.7 on the output of the CSV discriminant. The 2012BCD version requires two central jets with $p_{\mathrm{T}} > 30$ GeV instead and the same *b* tagging requirement. This online *b*tagging requirement has an efficiency that is equivalent to that of the tight offline requirement, CSV > 0.898. $E_{\mathrm{T}}^{\mathrm{miss}}$ is required to be greater than 80 GeV for both versions. The triggers achieve an efficiency of roughly 90% at roughly offline $E_{\mathrm{T}}^{\mathrm{miss}}$ of 130 GeV when considering at least one tight *b* tagged jet offline.

A logical OR combination of three triggers is used in the analysis (meaning the set of events are the union of the sets of events collected by the three triggers). The exact combination varies for 2012A *vs.* 2012BCD period. These triggers are all seeded by the OR combination of L1 $E_{\mathrm{T}}^{\mathrm{miss}} > 36\,\mathrm{GeV}$ and $E_{\mathrm{T}}^{\mathrm{miss}} > 40\,\mathrm{GeV}$ triggers (`L1_ETM36` OR `L1_ETM40`).

The efficiencies of the simulated triggers are parametrized and corrected as a function of $E_{\mathrm{T}}^{\mathrm{miss}}$ and $\mathrm{CSV}_{\mathrm{max}}$ to match the efficiencies measured in data. $\mathrm{CSV}_{\mathrm{max}}$ is the highest CSV value of the pair of jets that constitute the Higgs boson candidate. This approach takes into account the non-negligible correlations among the various trigger paths. It also characterizes the online *b* tagging efficiency and its dependency on jet p_{T} and η, as the geometry and trigger algorithm are simulated in a way that are as close as possible to the actual trigger environment. The trigger efficiencies in simulation and in data, as well as the data/MC correction scale factors, parametrized as a function of $E_{\mathrm{T}}^{\mathrm{miss}}$ for each individual trigger and the combination of all triggers are shown in Fig. 5-3. The trigger efficiencies in real data are evaluated in an orthogonal sample collected by an unbiased single isolated muon trigger with muon $p_{\mathrm{T}} > 24\,\mathrm{GeV}$ requirement: `HLT_IsoMu24_eta2p1`. The correction scale factors are taken from the ratio of the efficiency in data over the efficiency in MC. They are always very close to unity, except for low $E_{\mathrm{T}}^{\mathrm{miss}}$ values, for which they never exceed 5%.

For events passing exclusively the trigger with *b* tagging requirement, an additional correction as a function of $\mathrm{CSV}_{\mathrm{max}}$ is evaluated. A very similar trigger is used as the unbiased trigger: `HLT_DiCentralPFJet30_PFMET80`. It requires two central jets with $p_{\mathrm{T}} > 30\,\mathrm{GeV}$ and $E_{\mathrm{T}}^{\mathrm{miss}} > 80\,\mathrm{GeV}$, but it is prescaled and does not require *b* tagging. The data/MC correction parametrized as a function of $E_{\mathrm{T}}^{\mathrm{miss}}$ is applied before the evaluation. The derived correction is shown in the Fig. 5-4. These corrections are very close to unity for high values of $\mathrm{CSV}_{\mathrm{max}}$, but 10–20% higher for low values, suggesting the *b* tag fake rate is higher in data than in MC.

For $E_{\mathrm{T}}^{\mathrm{miss}} > 130\,\mathrm{GeV}$, the combined trigger efficiency for $Z(\nu\bar{\nu})H(b\bar{b})$ signal events is near 100% with respect to the offline event reconstruction and selection, described in Sec. 5.5. For events with $E_{\mathrm{T}}^{\mathrm{miss}}$ between 100 and 130 GeV the efficiency is about 88%.

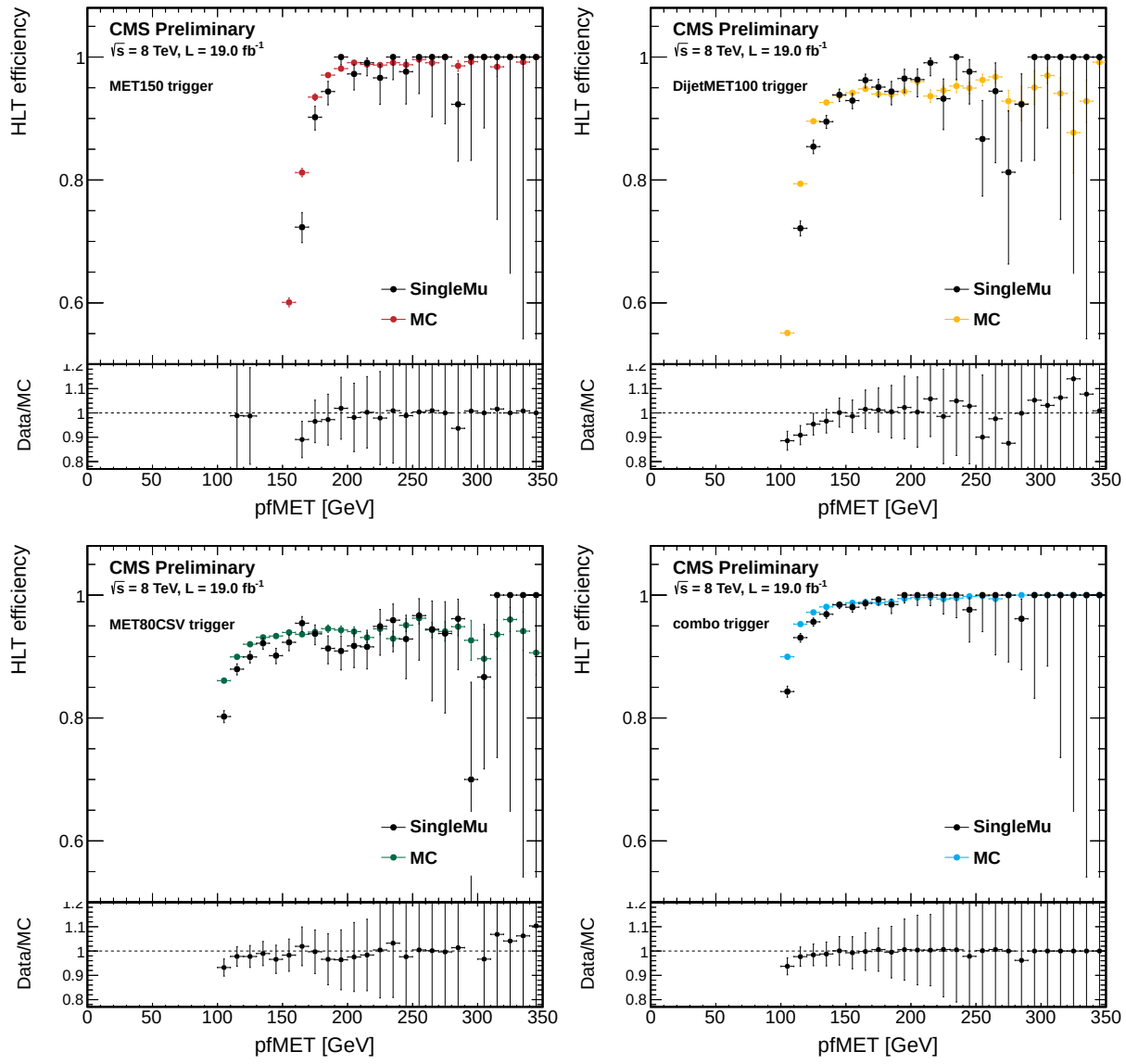

Figure 5-3. Trigger efficiencies of the three trigger paths and the logical OR combination of them plotted as a function of $E_{\mathrm{T}}^{\mathrm{miss}}$. The data/MC trigger efficiency scale factors are shown in the bottom panel.

5.4 *b* Jet Energy Regression

Using the standard jet reconstruction in CMS, the dijet invariant mass resolution of the two *b* jets from the Higgs decay is approximately 10%, with a few percent bias on

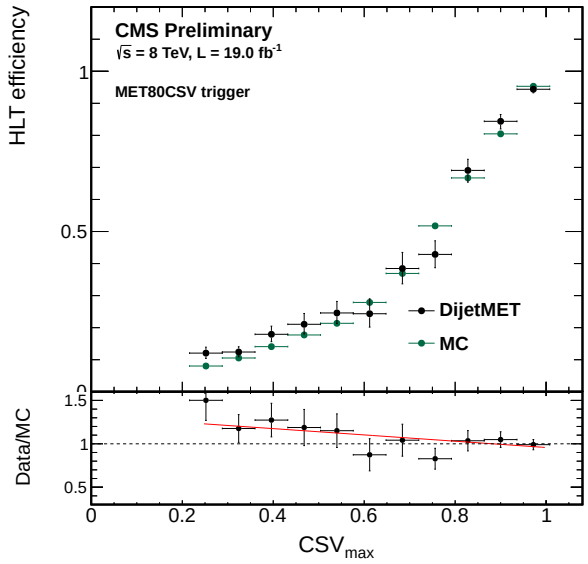

Figure 5-4. Exclusive efficiency of the trigger with *b* tagging requirement as a function of CSV_{max}, after applying the first trigger efficiency scale factors. The residual data/MC trigger efficiency scale factors are shown in the bottom panel.

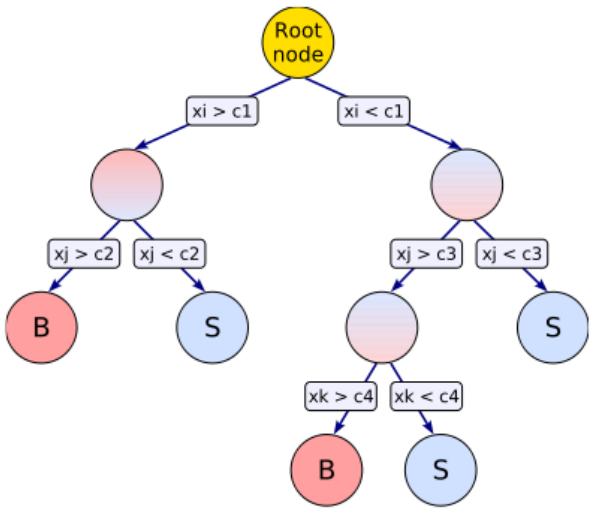

Figure 5-5. Schematic view of a decision tree [26].

the value of the mass peak. The mass resolution generally improves for larger p_T of the reconstructed Higgs boson.

To further improve the dijet invariant mass resolution, a dedicated multivariate regression using the boosted decision tree (BDT) algorithm has been developed. BDT is a supervised machine learning algorithm that is widely used to solve classification and regression problems. A simple decision tree is sketched in Fig. 5-5. Starting from the root node, a training sample is split by a sequence of binary decisions into subsamples (a.k.a. nodes) until one of the termination criteria is met. Each binary decision is made by a cut on the discriminating variable that maximizes the separation between signal and background. The terminal nodes are classified as either signal or background depending on the majority of events in the nodes (purity or impurity). For regression trees, the splitting decisions are made to minimize the average squared error in the output value of the target variable, and each terminal node represents a specific value of the target variable. The termination criteria may depend on the purity of the node, the number of remaining events in the node, or the depth of the tree.

Decision tree is appealing because it is robust against outliers and inclusion of weak variables. It is also easy to interpret, as it is essentially a partitioning of the feature space spanned by all the input variables. However, a single decision tree is very sensitive to overfitting, i.e. fitting to statistical fluctuation of the sample used in training. Thus, a boosted decision tree is frequently used to avoid overfitting. Boosting means generating an ensemble of decision trees which are trained using reweighted events. For each successive tree, events that are more often misclassified by previous trees are supplied larger weights. The final classifier (regressor) is a (weighted) average of the individual decision (regression) trees. The BDT implementation from the TMVA library [26] is used in this analysis.

The BDT regression technique for *b* jet energy correction was first used by the CDF experiment [123]. An additional correction, beyond the standard CMS jet energy

corrections, is computed for individual jets that make up the Higgs boson candidate in an attempt to recalibrate to the true b quark energy. To this end, a specialized BDT is trained on an ensemble of simulated $H \rightarrow b\bar{b}$ events with different m_H hypotheses (110–150 GeV) to avoid bias on m_H, using 14 discriminating variables as input and the p_T of the particle-level jet as target variable. The inputs to the BDT include variables related to several properties of the secondary vertex (when reconstructed), information about tracks and jet constituents, and other variables related to the energy reconstruction of the jet. Because of semileptonic B hadron decays, b jets contain, on average, more non-isolated leptons (referred to as soft leptons) and a larger fraction of missing energy due to neutrinos than jets from other quarks or gluons. Therefore, in cases where a soft lepton is found within $\Delta R < 0.5$ around the jet, the following variables are also included in the regression: p_T of the lepton, ΔR between the lepton and the jet direction, and p_T of the lepton relative to the jet direction. The full list of variables are given in Table 5-4.

Table 5-4. Variables used in the training of the BDT jet energy regression.

Variable	Description
raw p_T	Transverse momentum of the jet before jet energy corrections
p_T	Transverse momentum of the jet after jet energy corrections
E_T	Transverse energy of the jet after jet energy corrections
m_T	Transverse mass of the jet after jet energy corrections
JEC uncertainty	Uncertainty of the jet energy corrections applied on the jet
$N_{\text{constits.}}$	Number of PF candidates in the jet
p_T(lead. track)	Transverse momentum of the leading-p_T track in the jet
SV L_{3D}	3D flight length of the secondary vertex in the jet (if any)
SV L_{3D} uncertainty	Resolution of SV L_{3D} (if any)
SV mass	Invariant mass of the system of associated tracks at the secondary vertex (if any)
SV p_T	Magnitude of the vectorial p_T sum of associated tracks at the secondary vertex (if any)
SL p_T	Transverse momentum of soft lepton in the jet (if any)
SL ΔR	Distance in η-ϕ between the soft lepton (if any) and the jet direction
SL p_T^{rel}	Transverse momentum of soft lepton (if any) relative to the jet direction

The average improvement on the mass resolution, measured on simulated signal samples, when the regression technique is applied is approximately 15%, resulting in an increase in the analysis sensitivity of 10–20%. An example of the improvement is shown in Fig. 5-6 for simulated $Z(\ell\ell)H(b\bar{b})$ events where the improvement in resolution

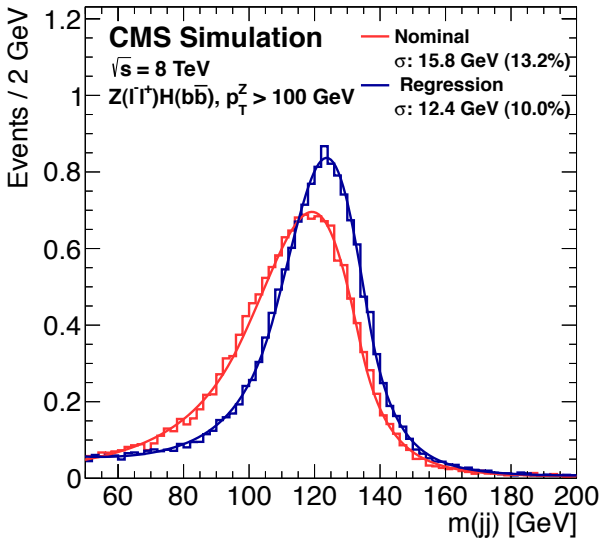

Figure 5-6. Dijet invariant mass distribution for simulated samples of $Z(\ell\ell)H(b\bar{b})$ events (m_H = 125 GeV), before (red) and after (blue) the regression correction is applied. A Bukin function [27] is fit to the distribution [28].

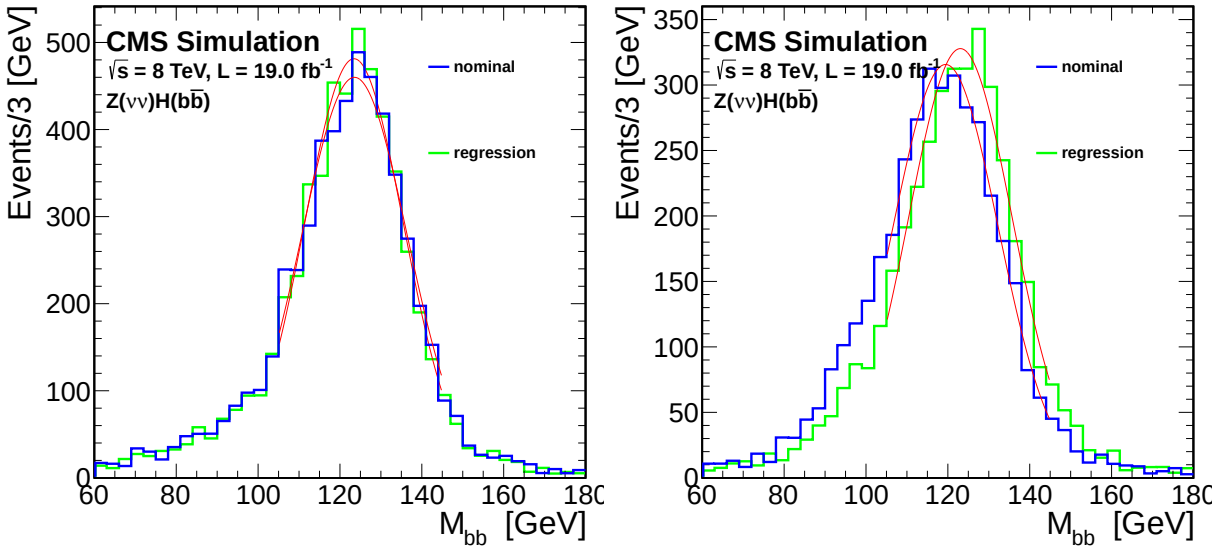

Figure 5-7. Comparison of the reconstructed dijet invariant mass for Higgs boson candidates in $Z(\nu\bar{\nu})H(b\bar{b})$ events before and after the regression for the case where the b jet does not (left) or does (right) contain a soft lepton.

is approximately 25%. [2] For the $Z(\nu\bar{\nu})H(b\bar{b})$ channel, the majority of the improvement happens in the cases where semileptonic B decays occur and a fraction of jet energy is lost due to neutrinos. Some of the loss is recovered by the regression when a soft lepton is detected. The performance is shown for cases where the b jet does or does not contain a soft lepton in Fig. 5-7.

As a result of the improved dijet invariant mass resolution, the separation between the ZH and ZZ resonance positions also increases. An example in the $Z(\ell\ell)H(b\bar{b})$ channel is shown in Fig. 5-8.

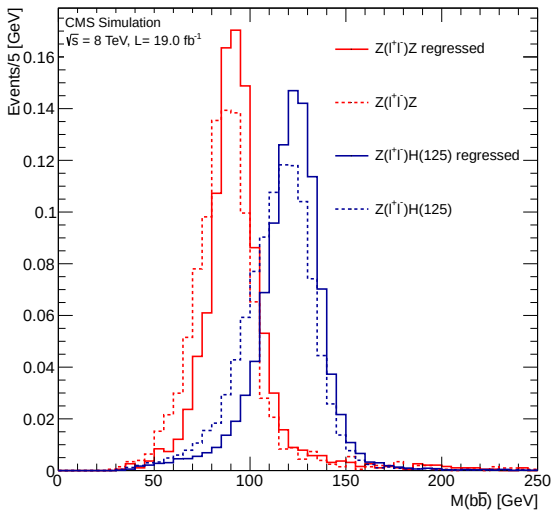

Figure 5-8. $ZZ(b\bar{b})$ and $ZH(b\bar{b})$ resonances for the $Z(\ell\ell)H(b\bar{b})$ channel before and after the regression is applied in the simulation. Regression helps to increase the separation.

The validation of the regression technique in data is done with samples of $Z \to \ell^+\ell^-$ events with two b tagged jets and in top-enriched samples in the lepton+jets final state. In the $Z \to \ell^+\ell^-$ case, when the jets are corrected by the regression procedure,

[2] For the $Z(\ell\ell)H(b\bar{b})$ channel, the $E_{\mathrm{T}}^{\mathrm{miss}}$ in the event and the azimuthal angle between $E_{\mathrm{T}}^{\mathrm{miss}}$ and the jet are also considered in the regression. It is the channel that sees the best improvement.

the distribution of the p_T balance between the Z boson candidate (reconstructed from an oppositely-charged pair of leptons) and the Higgs boson candidate is improved to be better centered at unity and narrower than when the regression correction is not applied. p_T balance is defined as the ratio between the p_T of the dijet system and the p_T of the dilepton system:

$$p_T \text{ balance} = \frac{p_T(jj)}{p_T(\ell\ell)} \tag{5–1}$$

The p_T balance distribution in data *vs.* MC before and after the regression is shown in Fig. 5-9. In the top-enriched case, the reconstructed top-quark mass distribution is closer to the nominal top-quark mass and also narrower than when the correction is not applied. In both cases, the distributions for data and the simulated samples are in very good agreement after the regression correction is applied.

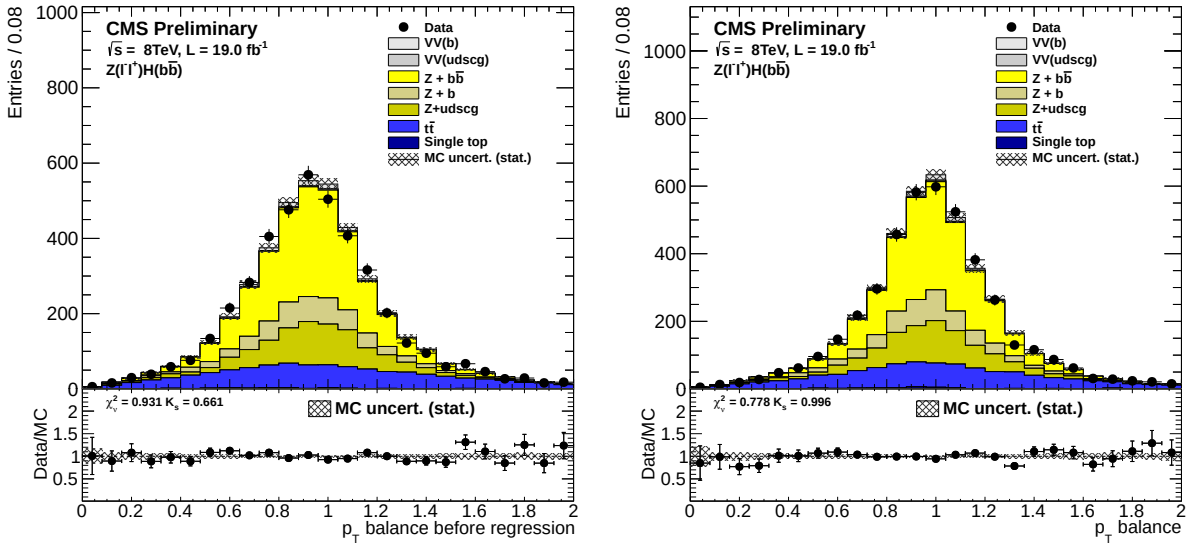

Figure 5-9. Distribution of the ratio between the $p_T(jj)$ and the p_T of the dilepton system on data *vs.* MC before (left) and after (right) regression in the Z+$b\bar{b}$ control region for the $Z(\ell\ell)H(b\bar{b})$ channel.

5.5 Event Selection

Event selection for the signal region in this channel starts by requiring the Z boson and Higgs boson reconstruction criteria described previously are satisfied. In particular, the identification of $Z \to \nu\bar{\nu}$ decays requires the E_T^{miss} in the event to be within the low-,

intermediate-, and high-boost regions: $100 < E_{\mathrm{T}}^{\mathrm{miss}} < 130, 130 < E_{\mathrm{T}}^{\mathrm{miss}} < 170, E_{\mathrm{T}}^{\mathrm{miss}} > 170$ GeV. For this channel, $E_{\mathrm{T}}^{\mathrm{miss}}$ is also sometimes referred to as the vector boson transverse momentum, $p_{\mathrm{T}}(V)$.

QCD multijet background can mimic the signature of $Z(\nu\bar{\nu})H(b\bar{b})$ events when one or more jets are poorly measured, leading to potentially large apparent missing energy. To ensure that $E_{\mathrm{T}}^{\mathrm{miss}}$ does not originate from mismeasured jets or some instrumental effects, three event requirements are made. First, for the high-boost region, a requirement of $\min \Delta\phi(E_{\mathrm{T}}^{\mathrm{miss}}, \mathrm{jet}) > 0.5$ radians is applied on the azimuthal angle between the $E_{\mathrm{T}}^{\mathrm{miss}}$ direction and the closest jet with $|\eta| < 2.5$ and $p_{\mathrm{T}} > 25$ GeV. For the low- and intermediate-boost regions the requirement is tightened to $\min \Delta\phi(E_{\mathrm{T}}^{\mathrm{miss}}, \mathrm{jet}) > 0.7$ radians. As shown in Fig. 5-10, the distribution of this variable peaks sharply at small angles for QCD events, while it is more randomly distributed for processes where $E_{\mathrm{T}}^{\mathrm{miss}}$ arises from neutrinos. The second requirement is that the azimuthal angle between the missing transverse energy direction as calculated from reconstructed charged tracks only (with $p_{\mathrm{T}} > 0.5$ GeV) and the $E_{\mathrm{T}}^{\mathrm{miss}}$ direction, $\Delta\phi(E_{\mathrm{T}}^{\mathrm{miss}}, E_{\mathrm{T}}^{\mathrm{miss}}{}_{\mathrm{trk}})$, should be smaller than 0.5 radians. This variable is uncorrelated to the $\min \Delta\phi(E_{\mathrm{T}}^{\mathrm{miss}}, \mathrm{jet})$ variable and further reduces QCD background. The third requirement is made for the low-boost region where the $E_{\mathrm{T}}^{\mathrm{miss}}$ significance, defined as the ratio between the $E_{\mathrm{T}}^{\mathrm{miss}}$ and the square root of the total transverse energy of the PF particles, should be greater than 3. The QCD multijet background is reduced to negligible levels by these three event requirements.

For the Higgs candidate dijet pair, the kinematic requirements are: the highest-p_{T} jet must have $p_{\mathrm{T}} > 60$ GeV, the second highest-p_{T} jet must have $p_{\mathrm{T}} > 30$ GeV, and the vectorial sum of their transverse momenta must be above 100 GeV for the low-boost region or 130 GeV for the intermediate- and high-boost regions. The azimuthal opening angle between the V and H candidates must be larger than 2 radians. Events with dijet invariant mass greater than 250 GeV are rejected. These requirements

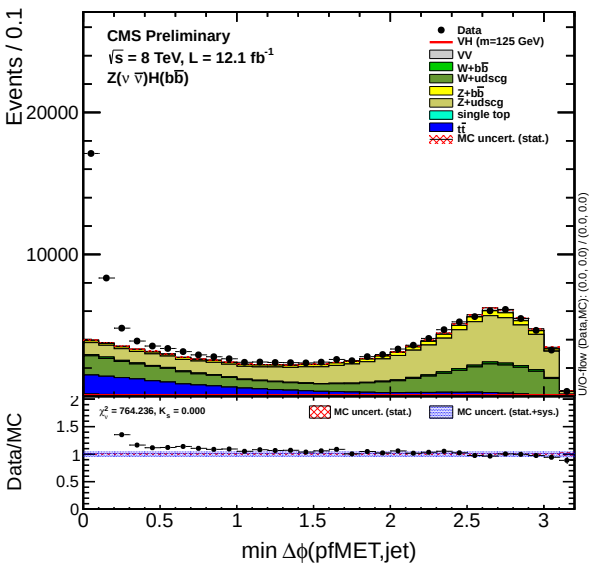

Figure 5-10. Distribution of min $\Delta\phi(E_{\mathrm{T}}^{\mathrm{miss}}, \mathrm{jet})$ in data (points with errors) and the sum of all backgrounds from simulation (histogram). The QCD contribution is clearly visible at small angles.

are set to be loose enough to have as high signal acceptance as possible, yet tight enough to be safe from effects due to kinematic thresholds imposed at trigger level. The b tagging requirements are: the highest-CSV jet of the pair must pass the CSVM working point (> 0.679), and the second highest-CSV jet must pass the CSVL working point (> 0.244). The V+jets background is reduced significantly by the b tagging requirements and subprocesses where the two jets originate from genuine b quarks dominate the signal region.

To reduce background events from $t\bar{t}$ and VV production, events with any number of isolated leptons of $p_{\mathrm{T}} > 20\,\mathrm{GeV}$, $N_{\mathrm{a}\ell} > 0$ are rejected. To further reduce $t\bar{t}$ background, events with 2 or more additional jets, $N_{\mathrm{aj}} \geq 2$ are also not accepted. Jets are counted as additional jets if they satisfy $p_{\mathrm{T}} > 25\,\mathrm{GeV}$ and $|\eta| < 4.5$. The full selection for the signal regions for all three boost regions are listed in Table 5-5. Including the intermediate- and low-boost regions improve the analysis sensitivity by roughly 15% and 5% respectively, compared to using only the high-boost region.

Table 5-5. Selection criteria that define the signal region. Entries marked with "–" indicate that the variable is not used. If different, the entries in square brackets indicate the selection for the different boost regions as defined in the first row of the table. Kinematic variables are in units of GeV, and angular variables are in radians.

Variable	Selection
$E_{\mathrm{T}}^{\mathrm{miss}}$	$[100-130]$ $[130-170]$ $[>170]$
$p_{\mathrm{T}}(\mathrm{j}_1)$	>60
$p_{\mathrm{T}}(\mathrm{j}_2)$	>30
$p_{\mathrm{T}}(\mathrm{jj})$	$[>100]$ $[>130]$ $[>130]$
$m(\mathrm{jj})$	<250
$\mathrm{CSV}_{\mathrm{max}}$	>0.679
$\mathrm{CSV}_{\mathrm{min}}$	>0.244
N_{aj}	$[<2]$ [–] [–]
$N_{\mathrm{a}\ell}$	$=0$
$\Delta\phi(V,H)$	>2.0
$\min\Delta\phi(E_{\mathrm{T}}^{\mathrm{miss}},\mathrm{jet})$	$[>0.7]$ $[>0.7]$ $[>0.5]$
$\Delta\phi(E_{\mathrm{T}}^{\mathrm{miss}},E_{\mathrm{T}}^{\mathrm{miss}}{}_{\mathrm{trk}})$	<0.5
$E_{\mathrm{T}}^{\mathrm{miss}}$ significance	$[>3]$ [–] [–]

In the final stage of the analysis, to better discriminate signal against background under different Higgs boson mass hypotheses, an event BDT classifier is trained separately at each mass value using simulated samples for signal and all background processes. The training of this BDT is performed with events that pass the signal selection. The set of event input variables used, listed in Table 5-6, is chosen by iterative optimization from a larger number of potentially discriminating variables. All the variables that enter the BDT training in the high-boost region are displayed in Fig. 5-11. Among the most discriminant variables are the dijet invariant mass distribution ($m(\mathrm{jj})$), the number of additional jets (N_{aj}), the value of CSV for the Higgs boson daughter with the second largest CSV value ($\mathrm{CSV}_{\mathrm{min}}$), and the distance between Higgs boson daughters ($\Delta R(\mathrm{jj})$). The variable rankings in terms of importance in the event BDT classifier for the high-boost region are listed in Table 5-7. It has been suggested that variables related to techniques that study in more detail the substructure of jets could improve the sensitivity of the $H \to b\bar{b}$ searches [29]. In this analysis, several combinations of such variables

were considered as additional inputs to the BDT discriminant. However they did not yield significant gains in sensitivity and are not included in the final training used.

Table 5-6. Variables used in the training of the event BDT discriminant.

Variable	Description
$m(\mathrm{jj})$	Dijet invariant mass
$p_{\mathrm{T}}(\mathrm{j}_1)$, $p_{\mathrm{T}}(\mathrm{j}_2)$	Transverse momentum of each H boson daughter
$p_{\mathrm{T}}(\mathrm{jj})$	Dijet transverse momentum
$\Delta\eta(\mathrm{jj})$	Difference in η between H boson daughters
$\Delta R(\mathrm{jj})$	Distance in η-ϕ between H boson daughters
$E_{\mathrm{T}}^{\mathrm{miss}}$	Missing transverse energy
$\mathrm{CSV}_{\mathrm{max}}$	Value of CSV for the H boson daughter with largest CSV value
$\mathrm{CSV}_{\mathrm{min}}$	Value of CSV for the H boson daughter with second largest CSV value
N_{aj}	Number of additional jets
$\Delta\phi(V,H)$	Azimuthal angle between $E_{\mathrm{T}}^{\mathrm{miss}}$ and dijets
$\min\Delta\phi(E_{\mathrm{T}}^{\mathrm{miss}},\mathrm{jet})$	Azimuthal angle between $E_{\mathrm{T}}^{\mathrm{miss}}$ and the closest jet
$\Delta\theta_{\mathrm{pull}}$	Color pull angle [124]
$\mathrm{maxCSV}_{\mathrm{aj}}$	Maximum CSV of the additional jets in an event
$\min\Delta R(H,\mathrm{aj})$	Minimum distance between an additional jet and the H boson candidate

Table 5-7. Variable rankings in terms of importance in the event BDT classifier for the high-boost region.

Rank	Variable
1	$m(\mathrm{jj})$
2	$\min\Delta\phi(E_{\mathrm{T}}^{\mathrm{miss}},\mathrm{jet})$
3	$\mathrm{CSV}_{\mathrm{min}}$
4	$\mathrm{maxCSV}_{\mathrm{aj}}$
5	N_{aj}
6	$\Delta R(\mathrm{jj})$
7	$\mathrm{CSV}_{\mathrm{max}}$
8	$E_{\mathrm{T}}^{\mathrm{miss}}$
9	$\min\Delta R(H,\mathrm{aj})$
10	$\Delta\eta(\mathrm{jj})$
11	$p_{\mathrm{T}}(\mathrm{j}_2)$
12	$\Delta\theta_{\mathrm{pull}}$
13	$\Delta\phi(V,H)$
14	$p_{\mathrm{T}}(\mathrm{jj})$
15	$p_{\mathrm{T}}(\mathrm{j}_1)$

A fit is performed to the shape of the output distribution of the event BDT discriminant to search for events compatible with the Higgs boson hypothesis. Before testing all events through this final discriminant, events are classified based on where they fall in

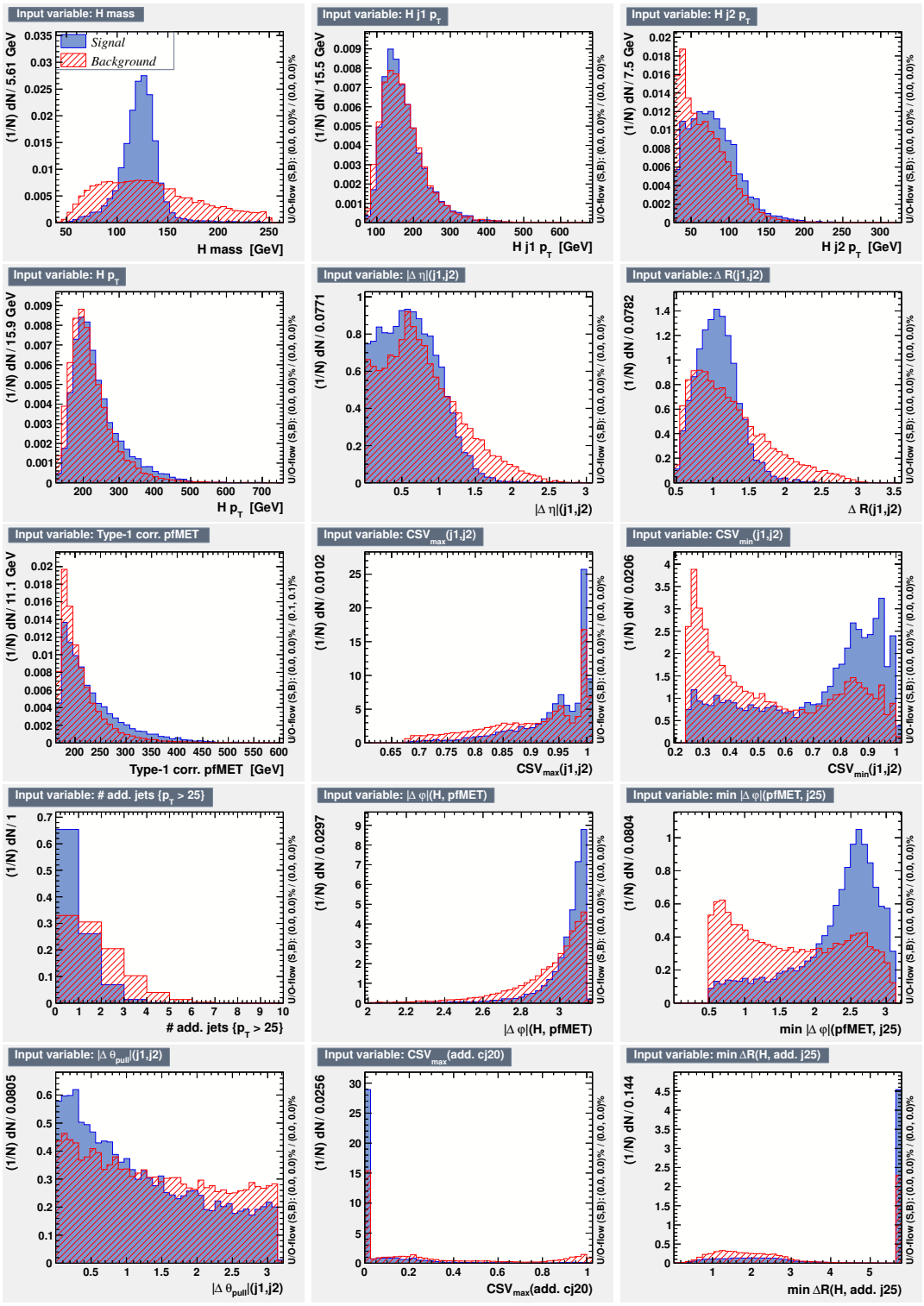

Figure 5-11. Variables used in the training of the event BDT discriminant in the high-boost region. Signal is shown in blue and the sum of backgrounds is shown in red. The normalizations of the histograms are arbitrary.

the output distributions of several other background-specific BDT discriminants that are trained to discern signal from individual background processes. This technique, similar to the one used by the CDF collaboration [125], divides the samples into four distinct subsets that are enriched in $t\bar{t}$, V+jets, VV, and VH. The increase in the analysis sensitivity from using this technique in the $Z(\nu\bar{\nu})H(b\bar{b})$ channel is 5–10%.

The first background-specific BDT discriminant is trained to separate $t\bar{t}$ from VH, the second one is trained to separate V+jets from VH, and the third one separates VV from VH. The output distributions of the background-specific BDTs are used to separate events into four subsets: those that fail a requirement on the $t\bar{t}$ BDT are classified as $t\bar{t}$-like events; those that pass the $t\bar{t}$ BDT requirement but fail a requirement on the V+jets BDT are classified as V+jets-like events; those that pass the V+jets BDT requirement but fail the requirement on the VV BDT are classified as VV-like events; finally, those that pass all BDT requirements are considered VH-enriched events. The events in each subset are then run through the final event BDT discriminant and the resulting distribution, now composed of four distinct subsets of events, is used as input to the fitting procedure.

As a validation of the multivariate approach to this analysis, these BDT discriminants are also trained to find diboson signals (ZZ and WZ, with $Z \to b\bar{b}$) rather than the VH signal. The event selection used in this case is identical to that used for the VH search.

As a cross-check to the BDT-based analysis, a simpler analysis is done by performing a fit to the shape of the dijet invariant mass distribution of the two jets associated with the reconstructed Higgs boson, $m(\mathrm{jj})$. The event selection for this analysis is more restrictive than the one used in the BDT analysis and is optimized for sensitivity in this single variable. Table 5-8 lists the event selection of the $m(\mathrm{jj})$ analysis. Since the diboson background also exhibits a peak in the $m(\mathrm{jj})$ distribution from $Z \to b\bar{b}$ decays, the distribution is also used to measure the consistency of the diboson rate with

the expectation from the standard model. A consistent rate measurement would support the validity of the estimate of the background processes in the *VH* search.

Table 5-8. Selection criteria that define the signal region for the $m(\mathrm{jj})$ analysis. Entries marked with "–" indicate that the variable is not used. If different, the entries in square brackets indicate the selection for the different boost regions as defined in the first row of the table. Kinematic variables are in units of GeV, and angular variables are in radians.

Variable	Selection
$E_{\mathrm{T}}^{\mathrm{miss}}$	$[100-130]$ $[130-170]$ $[>170]$
$p_{\mathrm{T}}(\mathrm{j}_1)$	$[>60]$ $[>60]$ $[>80]$
$p_{\mathrm{T}}(\mathrm{j}_2)$	>30
$p_{\mathrm{T}}(\mathrm{jj})$	$[>110]$ $[>140]$ $[>190]$
$\mathrm{CSV}_{\mathrm{max}}$	>0.898
$\mathrm{CSV}_{\mathrm{min}}$	>0.5
N_{aj}	$=0$
$N_{\mathrm{a}\ell}$	$=0$
$\Delta\phi(V,H)$	>2.95
$\min\Delta\phi(E_{\mathrm{T}}^{\mathrm{miss}},\mathrm{jet})$	$[>0.7]$ $[>0.7]$ $[>0.5]$
$\Delta\phi(E_{\mathrm{T}}^{\mathrm{miss}},E_{\mathrm{T}}^{\mathrm{miss}}{}_{\mathrm{trk}})$	<0.5
$E_{\mathrm{T}}^{\mathrm{miss}}$ significance	$[>3]$ [–] [–]

5.6 Background Estimation

Appropriate control regions are identified in data and used to validate the simulation modeling of the distributions used as input to the BDT discriminants, and to obtain scale factors used to adjust the simulated event yield estimates for the major background processes: production of *W* and *Z* bosons in association with jets and $t\bar{t}$ production. For the *W* and *Z* background processes, the control regions are defined such that they are enriched in either heavy-flavor (HF) or light-flavor (LF) jets. Furthermore, these processes are split according to how many of the two jets selected as the Higgs boson candidate originate from *b* quarks. Separate scale factors are obtained for each case. The notation used is: $V+$ *udscg* for the case where none of the jets originate from a *b* quark, $V+b$ for the case where only one of the jets is from a *b* quark, and $V+b\bar{b}$ for the case where both jets originate from *b* quarks. The control regions are depleted of signal events by relaxing *b* tagging requirements, using the sideband of dijet invariant

mass ($m(\mathrm{jj}) < 100$ or $m(\mathrm{jj}) > 140$), or requiring additional jets or in the event. For W and $t\bar{t}$ background, an isolated lepton is required. Table 5-9 lists the selection criteria used to define the control regions.

To get the scale factors by which the simulated event yields are adjusted, a set of binned likelihood fits is simultaneously performed to CSV distributions of jets for events in the control regions. These fits are done separately for each channel. Several other distributions of variables are also tried and consistent scale factors are obtained. These scale factors account not only for event yield discrepancies, but also for potential residual differences in physics object reconstruction and selection. Therefore, separate scale factors are used for each background process in the different channels. The uncertainties in the scale factor determination include two components: the statistical uncertainty due to the finite size of the samples and the systematic uncertainty. The latter is obtained by subtracting, in quadrature, the statistical component from the full uncertainty which includes the effect of various sources of systematic uncertainty such as b tagging, jet energy scale, and jet energy resolution.

Table 5-10 shows the scale factors and uncertainties for the three boost regions. Table 5-11 shows the correlation matrix from the fit for the high-boost region. The scale factors are found to be close to unity for all processes except for $V + b$, for which the scale factors are consistently found to be closer to two. In this case, most of the excess occurs in the region of low $\mathrm{CSV}_{\mathrm{min}}$ values in which events with two displaced vertices are found relatively close to each other, within a distance $\Delta R < 0.5$ defined by the directions of their displacement trajectories with respect to the primary vertex. This discrepancy is interpreted as arising mainly from mismodeling in the generator parton shower of the process of gluon splitting to $b\bar{b}$. In this process the dominant contribution typically contains a low-p_{T} b quark that can end up not being reconstructed as a jet above the p_{T} threshold used in the analysis, or that is merged with the jet from the more energetic b quark. These discrepancies are consistent with similar observations in other

studies of the production of vector bosons in association with heavy-flavor quarks by the ATLAS and CMS experiments [126–128].

Table 5-9. Definition of the background-enriched control regions. Entries marked with "–" indicate that the variable is not used. If different, the entries in square brackets indicate the selection for the different boost regions as defined in the first row of the table. Kinematic variables are in units of GeV, and angular variables are in radians.

Variable	$Z+$LF	$Z+$HF	$t\bar{t}$	$W+$LF	$W+$HF
$E_{\mathrm{T}}^{\mathrm{miss}}$	[100 – 130] [130 – 170] [> 170]	[100 – 130] [130 – 170] [> 170]	[100 – 130] [130 – 170] [> 170]	[100 – 130] [130 – 170] [> 170]	[100 – 130] [130 – 170] [> 170]
$p_{\mathrm{T}}(\mathrm{j}_1)$	> 60	> 60	> 60	> 60	> 60
$p_{\mathrm{T}}(\mathrm{j}_2)$	> 30	> 30	> 30	> 30	> 30
$p_{\mathrm{T}}(\mathrm{jj})$	[> 100] [> 130] [> 130]	[> 100] [> 130] [> 130]	[> 100] [> 130] [> 130]	[> 100] [> 130] [> 130]	[> 100] [> 130] [> 130]
$m(\mathrm{jj})$	< 250	< 250, $\notin$ [100–140]	< 250, $\notin$ [100–140]	< 250	< 250, $\notin$ [100–140]
$\mathrm{CSV}_{\mathrm{max}}$	[0.244–0.898]	> 0.679	> 0.898	[0.244–0.898]	> 0.679
$\mathrm{CSV}_{\mathrm{min}}$	–	> 0.244	–	–	> 0.244
N_{aj}	[< 2] [–] [–]	[< 2] [–] [–]	≥ 1	= 0	= 0
$N_{\mathrm{a}\ell}$	= 0	= 0	= 1	= 1	= 1
$\Delta\phi(V,H)$	–	> 2.0	–	–	> 2.0
$\min\Delta\phi(E_{\mathrm{T}}^{\mathrm{miss}},\mathrm{jet})$	[> 0.7] [> 0.7] [> 0.5]	[> 0.7] [> 0.7] [> 0.5]	[> 0.7] [> 0.7] [> 0.5]	[> 0.7] [> 0.7] [> 0.5]	[> 0.7] [> 0.7] [> 0.5]
$\Delta\phi(E_{\mathrm{T}}^{\mathrm{miss}},E_{\mathrm{T}}^{\mathrm{miss}}{}_{\mathrm{trk}})$	< 0.5	< 0.5	–	–	–
$E_{\mathrm{T}}^{\mathrm{miss}}$ significance	[> 3] [–] [–]	[> 3] [–] [–]	[> 3] [–] [–]	[> 3] [–] [–]	[> 3] [–] [–]

Table 5-10. Data/MC scale factors for the three boost regions derived from the control regions. The quoted uncertainties have two components: the statistical uncertainty from the fit (first set of ± values) and a systematic uncertainty that accounts for possible data/MC shape differences in the discriminating variables (second set of ± values).

Process	Low $p_{\mathrm{T}}(V)$	Intermediate $p_{\mathrm{T}}(V)$	High $p_{\mathrm{T}}(V)$
$W+udscg$	$0.83\pm0.02\pm0.04$	$0.93\pm0.02\pm0.04$	$0.93\pm0.02\pm0.03$
$W+b$	$2.30\pm0.21\pm0.11$	$2.08\pm0.20\pm0.12$	$2.12\pm0.22\pm0.10$
$W+b\bar{b}$	$0.85\pm0.24\pm0.14$	$0.75\pm0.26\pm0.11$	$0.71\pm0.25\pm0.15$
$Z+udscg$	$1.24\pm0.03\pm0.09$	$1.19\pm0.03\pm0.07$	$1.17\pm0.02\pm0.08$
$Z+b$	$2.06\pm0.06\pm0.09$	$2.30\pm0.07\pm0.08$	$2.13\pm0.05\pm0.07$
$Z+b\bar{b}$	$1.25\pm0.05\pm0.11$	$1.11\pm0.06\pm0.12$	$1.12\pm0.04\pm0.10$
$t\bar{t}$	$1.01\pm0.02\pm0.04$	$0.99\pm0.02\pm0.03$	$0.99\pm0.02\pm0.03$

Table 5-11. Correlation matrix from the scale factor fit for the high-boost region.

	$W+udscg$	$W+b$	$W+b\bar{b}$	$Z+udscg$	$Z+b$	$Z+b\bar{b}$	$t\bar{t}$
$W+udscg$	1.000	–	–	–	–	–	–
$W+b$	−0.276	1.000	–	–	–	–	–
$W+b\bar{b}$	0.153	−0.076	1.000	–	–	–	–
$Z+udscg$	−0.305	0.476	0.057	1.000	–	–	–
$Z+b$	0.397	−0.304	0.135	−0.426	1.000	–	–
$Z+b\bar{b}$	0.289	0.177	−0.302	0.272	−0.185	1.000	–
$t\bar{t}$	0.241	−0.407	−0.008	−0.179	−0.052	0.018	1.000

Figs. 5-12 and 5-13 show examples of distributions for variables in the simulated samples and in data for different control regions. Many other distributions of the input

variables are plotted in Appendix. The scale factors described above have been applied to the corresponding simulated samples.

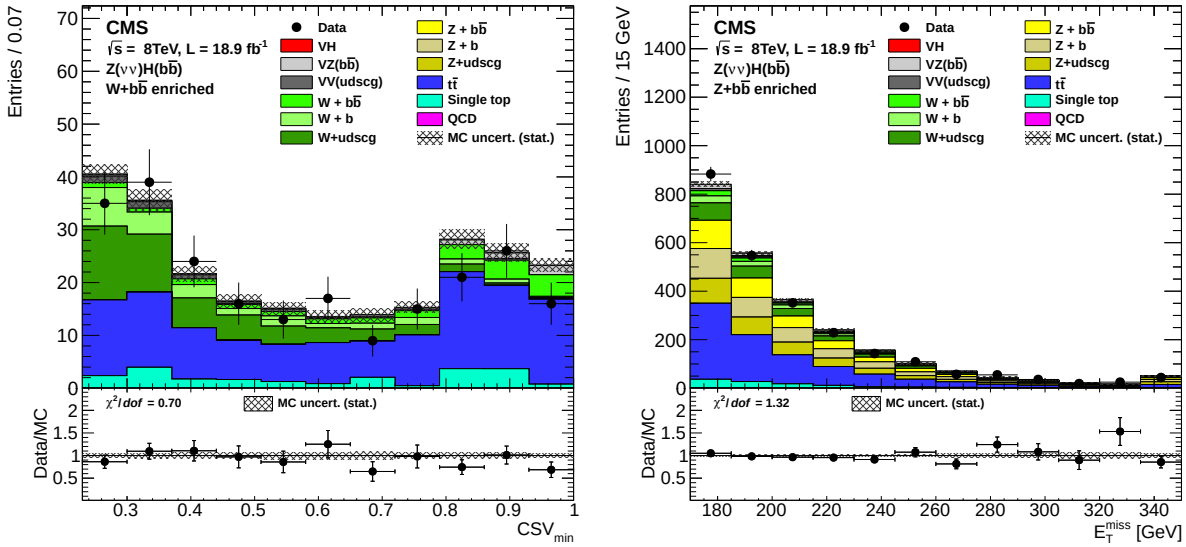

Figure 5-12. Left: CSV_{min} distribution for the $W+$HF high-boost control region. Right: E_T^{miss} distribution for the $Z+$HF high-boost control region. Simulation samples are shown after applying the data/MC scale factors.

Existing Monte Carlo QCD samples currently limit the ability to accurately predict the event yields of QCD multijet background due to insufficient effective luminosity in the relevant kinematic regime. For the high-boost and intermediate-boost regions, samples that have been checked generally predict negligible QCD backgrounds after applying all the selection criteria, but the level of statistical accuracy is not sufficient to make a definitive statement. For the low-boost region, there can be a few QCD events that survive the "anti-QCD" event requirements described in Sec. 5.5. Due to large QCD cross section, these events carry large weights, corresponding to a large number of expected events. Therefore, a procedure, based on the standard ABCD method, has been developed to obtain data-driven estimates for QCD background in this analysis.

In order to build reasonable shapes for the QCD background for use in the BDT analysis, the anti-QCD cuts are specifically not applied on the MC QCD samples. For each boost region, the distribution of the BDT discriminant output of the QCD events

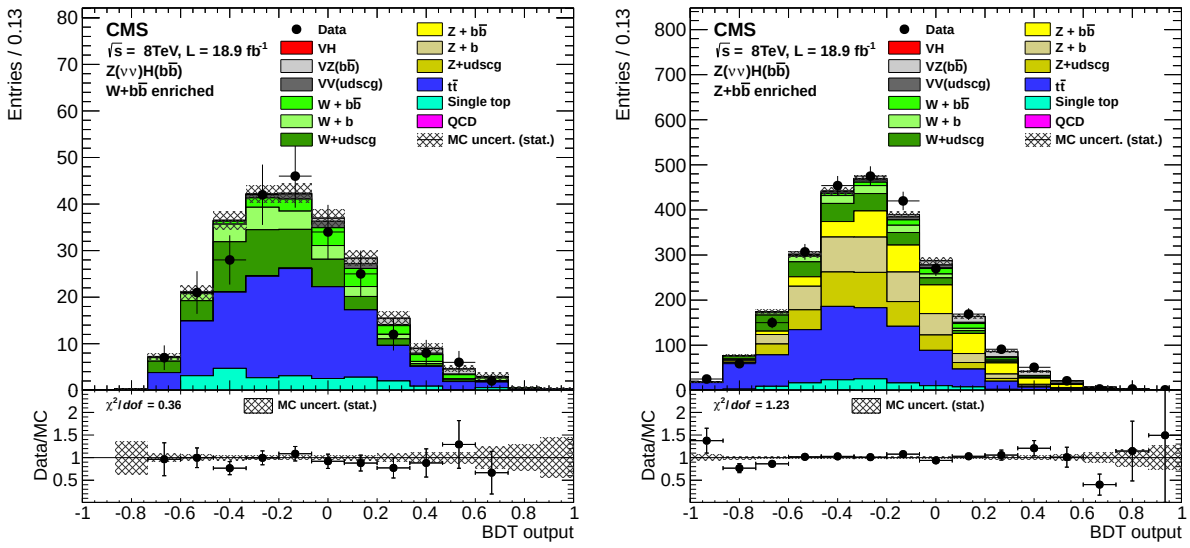

Figure 5-13. Left: Event BDT discriminant output for the $W+$HF high-boost control region. Right: Event BDT discriminant output for the $Z+$HF high-boost control region. Simulation samples are shown after applying the data/MC scale factors.

are scaled by the ratio of number of events after and before those cuts. In addition, a data/MC scale factor is obtained by inverting the anti-QCD cuts to select a region dominated purely by QCD. The data/MC scale factor is found to be close to 2 in all the boost regions. With this method, the QCD shapes are included in the final fit with event yields that are predicted to be $<$5% in low-boost region, $<$1% in the intermediate-boost, and $<$0.1% in the high-boost.

In the V+jets control regions, a shape difference in the vector boson transverse momentum of the simulated samples with respect to real data has been observed. The observed data has a softer spectrum, so a negative correction with increasing $p_{\mathrm{T}}(V)$ is necessary to correct for the effect. This negative correction is expected to stem from higher order electroweak corrections to the vector boson production, similar to the correction that are applied to the simulated signal samples as discussed in Sec. 5.2.

An ad-hoc correction is derived by fitting the data/MC ratio in the $W+$LF and $Z+$LF control regions. The correction that is applied in $Z(\nu\bar{\nu})H(b\bar{b})$ is given by:

$$c = 1 - 0.0025 \cdot (\max(p_{\mathrm{T}}(Z), 130) - 130) \qquad \text{for } Z\text{+jets} \qquad (5\text{–}2)$$

$$c = 1 - 0.0010 \cdot (\max(p_{\mathrm{T}}(W), 150) - 150) \qquad \text{for } W\text{+jets} \qquad (5\text{–}3)$$

It is applied to the $p_{\mathrm{T}}(V)$ spectrum at the generator level.

The impact of the correction on the reconstructed $E_{\mathrm{T}}^{\mathrm{miss}}$ spectrum is shown in Fig. 5-14. The effect of the $p_{\mathrm{T}}(V)$ reweighting on data in the $Z+$HF control region is shown in Fig. 5-15 and an improvement in the agreement between data and simulation is found. The reweighting is applied prior to the fit to obtain the scale factors.

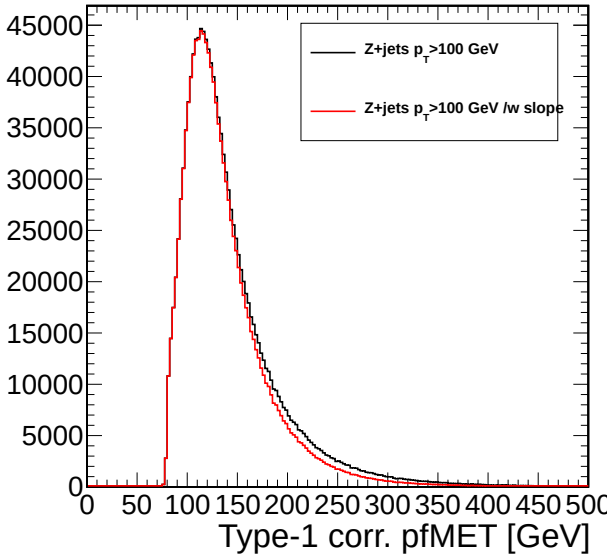

Figure 5-14. Impact of the $p_{\mathrm{T}}(V)$ reweighting on the reconstructed $E_{\mathrm{T}}^{\mathrm{miss}}$ spectrum for events with $p_{\mathrm{T}}(Z)$>100 GeV at the generator level. The nominal MC is shown in black and the $p_{\mathrm{T}}(V)$-reweighted MC is shown in red.

5.7 Systematic Uncertainties

The systematic uncertainties that affect the results presented in this analysis are listed in Table 5-12. Information about each source of systematic uncertainty is given, including whether it affects the shape or normalization of the BDT output, the uncertainty in signal or background event yields, and the relative contribution to the expected

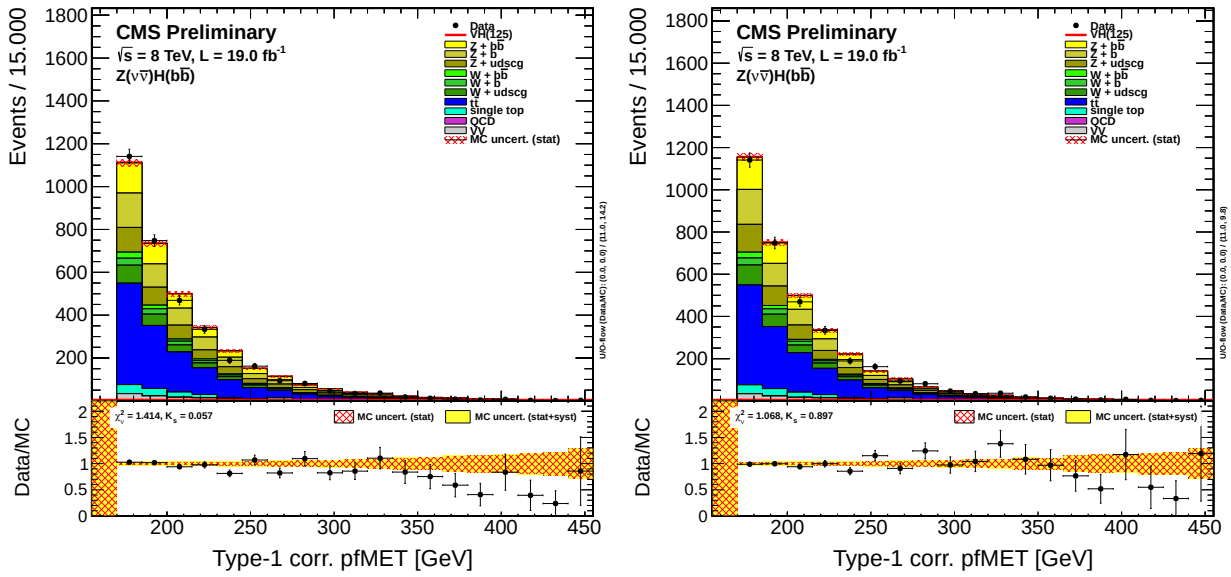

Figure 5-15. Distributions of the $E_{\mathrm{T}}^{\mathrm{miss}}$ in the ZjHF control region using the nominal MC (left) and $p_{\mathrm{T}}(V)$-reweighted MC (right).

uncertainty in the signal strength, μ (defined as the ratio of the best-fit value for the production cross section for a 125 GeV Higgs boson, relative to the standard model cross section). The systematic uncertainties are described in more detail below.

The uncertainty in the CMS luminosity measurement is estimated to be 2.6% for the 2012 $\sqrt{s} = 8$ TeV data [129]. The parameters describing the $Z(\nu\overline{\nu})H(b\overline{b})$ trigger efficiency turn-on curve have been varied within their statistical uncertainties and for different assumptions on the methods used to derive the efficiency. This results in an event yield uncertainty of about 3%.

The jet energy scale is varied within its uncertainty as a function of jet p_{T} and η. The efficiency of the analysis selection is recomputed to assess the variation in event yields. Depending on the process, a 2–3% yield variation is found. The effect of the uncertainty on the jet energy resolution is evaluated by smearing the jet energies according to the measured uncertainty. Depending on the process, a 3–6% variation in event yields is obtained. The uncertainties in the jet energy scale and resolution also have an effect on the shape of the BDT output distribution. The impact of the jet energy

Table 5-12. Information about each source of systematic uncertainty, including whether it affects the shape or normalization of the BDT output, the uncertainty in signal or background event yields, and the relative contribution to the expected uncertainty in the signal strength, μ. Due to correlations, the total systematic uncertainty is less than the sum in quadrature of the individual uncertainties. The last column shows the percentage decrease in the total signal strength uncertainty, including statistical, when removing that specific source of uncertainty. The ranges quoted are due to the differences for different background processes and the different Higgs boson mass hypotheses.

Source	Type	Event yield uncertainty range (%)	Individual contribution to μ uncertainty (%)	Effect of removal on μ uncertainty (%)
Luminosity	norm.	2.6	< 2	< 0.1
Trigger	shape	3	< 2	< 0.1
Jet energy scale	shape	2–3	5.0	0.5
Jet energy resolution	shape	3–6	5.9	0.7
Missing transverse energy	shape	3	3.2	0.2
b tagging	shape	3–15	10.2	2.1
Signal cross section (scale and PDF)	norm.	4	3.9	0.3
Signal cross section (p_T boost, EWK/QCD)	norm.	2/5	3.9	0.3
Monte Carlo statistics	shape	1–5	13.3	3.6
Backgrounds (data estimate)	norm.	10	15.9	5.2
Single top (simulation estimate)	norm.	15	5.0	0.5
Dibosons (simulation estimate)	norm.	15	5.0	0.5
MC modeling (V+jets and $t\bar{t}$)	shape	10	7.4	1.1

scale uncertainty is determined by recomputing the BDT output distribution after shifting the energy scale up and down by its uncertainty. Similarly, the impact of the jet energy resolution is determined by recomputing the BDT output distribution after increasing or decreasing the jet energy resolution. An uncertainty of 3% is assigned to the event yields of all processes in the $Z(\nu\bar{\nu})H(b\bar{b})$ channels due to the uncertainty related to the missing transverse energy estimate.

Data/MC *b* tagging scale factors are measured in heavy-flavor enhanced samples of jets that contain muons and are applied consistently to jets in signal and background events. The measured uncertainties for the *b* tagging scale factors are: 3% per *b* quark tag, 6% per *c* quark tag, and 15% per mistagged jet (originating from gluons and other light-flavor quarks) [25]. These translate into yield uncertainties in the 3–15% range, depending on the channel and the specific process. The shape of the BDT output

distribution is also affected by the shape of the CSV distributions and an uncertainty is assigned according to a range of variations of the CSV distributions.

The total VH signal cross section has been calculated to NNLO accuracy, and the total theoretical uncertainty is $\approx$ 4% [4], including the effect of scale variations and PDF uncertainties [112, 130–133]. The estimated uncertainties of the NLO electroweak corrections are 2% for both the ZH and WH processes. The estimate for the NNLO QCD correction results in an uncertainty of 5% for both the ZH and WH processes.

The uncertainty in the background event yields estimated from data is approximately 10%. For V+jets, the differences in the shape of the BDT output distribution between events generated with the MADGRAPH and the HERWIG++ MC generators are considered as a shape systematic uncertainty. For $t\bar{t}$ the differences in the shape of the BDT output distribution between the one obtained from the nominal MADGRAPH samples and those obtained from the POWHEG and MC@NLO [134] generators are considered as shape systematic uncertainties.

An uncertainty of 15% is assigned to the event yields obtained from simulation for single top production in the tW- and t-channels. For the diboson backgrounds, a 15% cross section uncertainty is assumed. These uncertainties are consistent with the CMS measurements of these processes [135, 136]. The limited number of MC simulated events is also taken into account as a source of systematic uncertainty.

The combined effect of the systematic uncertainties results in an increase of about 15% on the expected upper limit on the Higgs boson production cross section and in a reduction of 15% on the expected significance of an observation when the Higgs boson is present in the data at the predicted standard model rate.

CHAPTER 6
RESULTS

6.1 $VH(b\bar{b})$ Results

6.1.1 BDT analysis

Results are obtained from combined signal and background binned likelihood fits to the shape of the output distribution of the BDT discriminants. These discriminants are trained separately for each boost region and for each Higgs boson mass hypothesis in the 110–135 GeV range. In the simultaneous fit to all boost regions, the BDT shape and normalization for signal and for each background component are allowed to vary within the systematic and statistical uncertainties described in Sec. 5.7. These uncertainties are treated as independent nuisance parameters in the fit. All nuisance parameters, including the scale factors described in Sec. 5.6, are adjusted by the fit. The combined $VH(b\bar{b})$ results, obtained from the simultaneous fit to all the six $VH(b\bar{b})$ channels, are presented as the main results. The possible correlations among the channels are taken into account in the fit. Results pertaining to the $Z(\nu\bar{\nu})H(b\bar{b})$ channel alone are presented when available. Information about analyses in the other channels can be found in Ref. [28]. A desciption of the statistical methodology applied in this analysis can be found in Appendix.

In total, 14 BDT distributions are considered, one from each boost region in each channel. Fig. 6-1 shows the BDT output distributions after the fit for the three boost regions of the $Z(\nu\bar{\nu})H(b\bar{b})$ channel, for the $m_H = 125$ GeV mass hypothesis. The four partitions in the left panel correspond to the subsets enriched in $t\bar{t}$, V+jets, VV, and VH production, after all selection criteria as described in Sec. 5.5 have been applied. The bottom right panel shows the right-most, VH-enriched, partition in more detail. Every

Text and materials in this Chapter were adapted from the CMS publication Phys. Rev. D 89, 012003 (2014), American Physical Society. The author's work contributed to the publication.

distribution is accompanied by the ratio of the number of events observed in data to that of the Monte Carlo prediction for signal and backgrounds. For completeness, all 14 BDT distributions used in the fit are shown in Appendix. Table 6-1 lists, for partial combinations of channels, the total number of events in the four highest bins of their corresponding BDT for the expected backgrounds, for the 125 GeV SM Higgs boson signal, and for data. A mild excess compatible with the presence of the SM Higgs boson is observed. Fig. 6-2 combines the BDT outputs of all channels where the events are gathered in bins of similar expected signal-to-background ratio, as given by the value of the output of their corresponding BDT discriminant (trained with a Higgs boson mass hypothesis of 125 GeV). The ratio of the data to the background-only prediction and to the predicted sum of background and SM Higgs boson signal with a mass of 125 GeV is also shown. The observed excess of events in the bins with the largest signal-to-background ratio is consistent with what is expected from the production of the 125 GeV SM Higgs boson.

Table 6-1. Observed total number of events for partial combinations of channels in the four highest bins of their corresponding BDT for the expected backgrounds (B), for the 125 GeV SM Higgs boson signal (S), and for 8 TeV data. Also shown is the signal-to-background ratio (S/B).

	$W(\ell\nu)H(b\bar{b})$			$W(\tau\nu)H(b\bar{b})$	$Z(\ell\ell)H(b\bar{b})$		$Z(\nu\bar{\nu})H(b\bar{b})$		
Process	Low $p_{\mathrm{T}}(V)$	Int. $p_{\mathrm{T}}(V)$	High $p_{\mathrm{T}}(V)$		Low $p_{\mathrm{T}}(V)$	High $p_{\mathrm{T}}(V)$	Low $p_{\mathrm{T}}(V)$	Int. $p_{\mathrm{T}}(V)$	High $p_{\mathrm{T}}(V)$
$V+b\bar{b}$	25.2	22.4	15.9	4.3	158.6	36.2	177.3	98.3	68.2
$V+b$	3.1	2.9	9.6	1.2	95.8	14.6	84.7	58.3	27.6
$V+udscg$	4.5	8.5	10.0	2.5	62.3	8.7	57.6	31.0	21.6
$t\bar{t}$	113.2	106.5	50.3	22.6	107.0	6.9	153.8	87.4	39.2
Single top	24.1	20.3	14.7	7.4	2.9	0.4	54.5	20.1	11.7
$VV(udscg)$	0.3	1.3	1.2	0.2	2.4	0.4	2.3	1.5	1.4
$VZ(b\bar{b})$	1.1	1.4	2.3	1.1	11.0	2.7	9.5	6.9	7.7
Total background	171.7	163.4	104.1	39.4	439.8	69.8	539.7	303.5	177.4
VH	3.0	6.0	8.3	1.4	5.5	6.3	8.5	8.5	11.5
Data	185	182	128	35	425	77	529	322	188
S/B (%)	1.7	3.7	8.0	3.4	1.3	9.0	1.6	2.8	6.5

The 95% confidence level (C.L.) upper limits on the product of the VH production cross section times the $H \to b\bar{b}$ branching fraction, with respect to the expectations for a standard model Higgs boson ($\sigma/\sigma_{\mathrm{SM}}$) are obtained for the 8 TeV $Z(\nu\bar{\nu})H(b\bar{b})$ analysis. At each mass point the observed limit, the median expected limit, and the 1

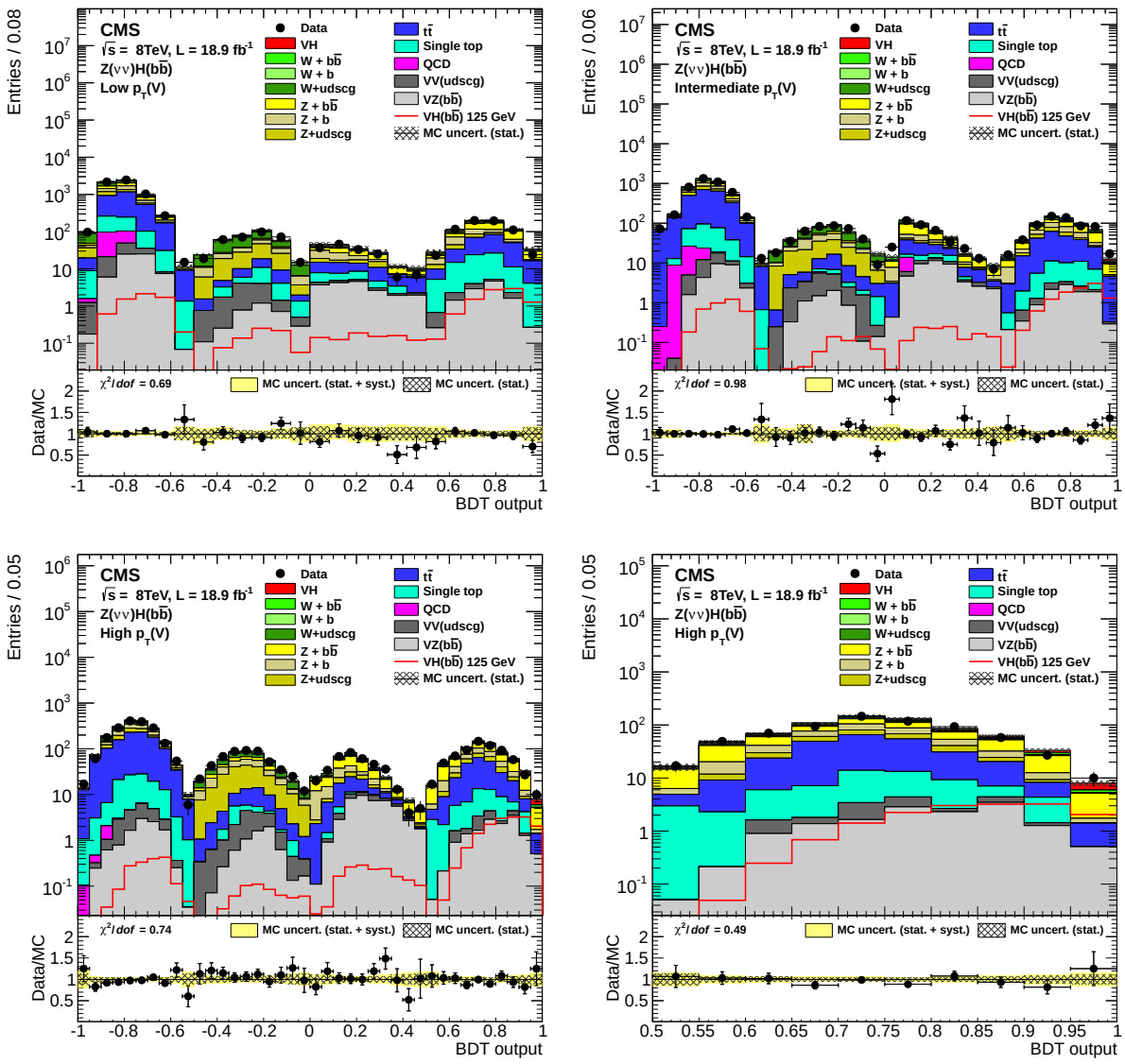

Figure 6-1. Post-fit BDT output distributions for $Z(\nu\bar{\nu})H(b\bar{b})$ in the low-boost region (top left), the intermediate-boost (top right), and the high-boost (bottom left). Bottom right: *VH*-enriched partition of the high-boost region is shown in more detail.

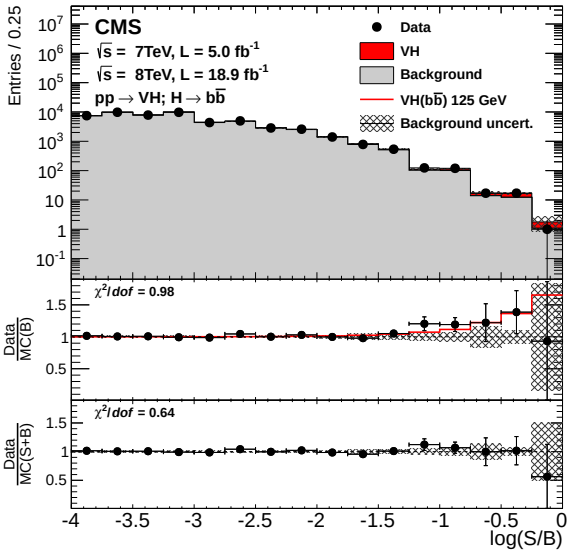

Figure 6-2. Combination of all channels into a single distribution. Events are sorted in bins of expected signal-to-background ratio. The ratios of the data to the background-only prediction and to the signal+background prediction are also shown.

and 2 standard deviation bands are calculated using the modified frequentist method CL_s [137–139]. Fig. 6-3 displays the limits. The limits are also obtained combining the results of all channels, for all boost regions, and including the previous 7 TeV results [62]. The left panel of Fig. 6-4 displays the final results.

For a Higgs boson mass of 125 GeV, using the 8 TeV $Z(\nu\bar{\nu})H(b\bar{b})$ results alone, the expected limit is 1.6 and the observed limit is 2.6. For the full 7 and 8 TeV combination of $VH(b\bar{b})$ channels, the corresponding expected limit is 0.95 and the observed limit is 1.89. Given that the resolution for the reconstructed Higgs boson mass is $\approx$ 10%, these results are compatible with a Higgs mass of 125 GeV. This is demonstrated by the red dashed line in the left panel of Fig. 6-4, which represents the expected limit obtained from the sum of expected background and the signal of a SM Higgs boson with a mass of 125 GeV.

For all channels an excess of events over the expected background contributions is indicated by the fits of the BDT output distributions. The probability (p-value) to

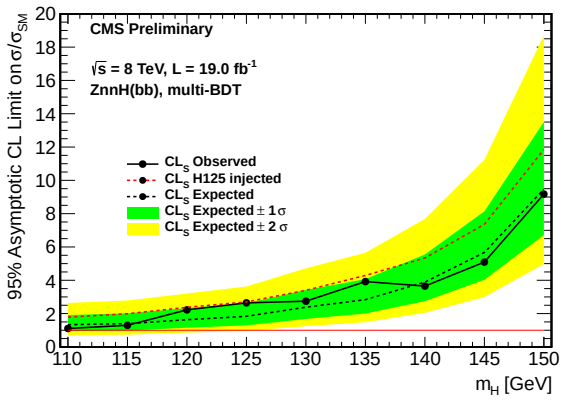

Figure 6-3. Expected and observed 95% C.L. upper limits on the product of the *VH* production cross section times $\mathcal{B}$ ($H \to b\bar{b}$) w.r.t. SM expectations, for the $Z(\nu\bar{\nu})H(b\bar{b})$ channel only.

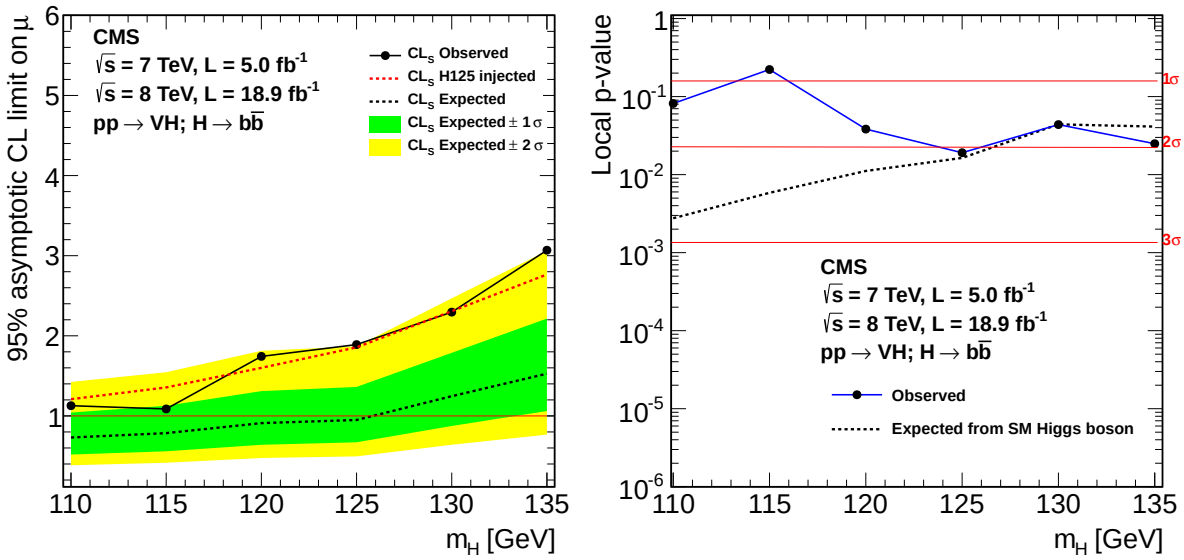

Figure 6-4. Left: Expected and observed 95% C.L. upper limits on the product of the *VH* production cross section times $\mathcal{B}$ ($H \to b\bar{b}$) w.r.t. SM expectations. Right: Local *p*-values of the observed excess for the background-only hypothesis.

observe data as discrepant as observed under the background-only hypothesis is shown in the right panel of Fig. 6-4 as a function of the assumed Higgs boson mass. For $m_H = 125\,\text{GeV}$, the excess of observed events corresponds to a local significance of 2.1 standard deviations away from the background-only hypothesis. This is consistent with the 2.1 standard deviations expected when assuming the standard model prediction for Higgs boson production.

The relative sensitivity of the channels that are topologically distinct is demonstrated in Table 6-2 for $m_H = 125$ GeV. The table lists the expected and observed limits and local significances for the $W(\ell\nu)H(b\bar{b})$ and $W(\tau\nu)H(b\bar{b})$ channels combined, for the $Z(\ell\ell)H(b\bar{b})$ channels combined, and for the $Z(\nu\bar{\nu})H(b\bar{b})$ channel.

Table 6-2. Expected and observed 95% C.L. upper limits on the product of the VH production cross section times the $H \to b\bar{b}$ branching fraction, w.r.t. the expectations for the 125 GeV SM Higgs boson, for partial combination of channels. Also shown are the expected and observed local significances.

	$\sigma/\sigma_{\text{SM}}$ (95% C.L.) median expected	$\sigma/\sigma_{\text{SM}}$ (95% C.L.) observed	Significance expected	Significance observed
$W(\ell\nu,\tau\nu)H(b\bar{b})$	1.6	2.3	1.3	1.4
$Z(\ell\ell)H(b\bar{b})$	1.9	2.8	1.1	0.8
$Z(\nu\bar{\nu})H(b\bar{b})$	1.6	2.6	1.3	1.3
All channels	0.95	1.89	2.1	2.1

The best-fit values of the production cross section for a 125 GeV Higgs boson, relative to the standard model cross section (signal strength μ), are shown in the left panel of Fig. 6-5 for the $W(\ell\nu)H(b\bar{b})$ and $W(\tau\nu)H(b\bar{b})$ channels combined, for the $Z(\ell\ell)H(b\bar{b})$ channels combined, and for the $Z(\nu\bar{\nu})H(b\bar{b})$ channel. The observed signal strengths are consistent with each other, and the value for the signal strength for the combination of all channels is 1.0 ± 0.5. In the right panel of Fig. 6-5, the correlation between the signal strengths for the separate production processes, WH and ZH is shown. The two production modes are consistent within uncertainties. This figure contains slightly different information than the one on the left panel as some final states contain signal events that originate from both WH and ZH production processes. The WH process contributes approximately 20% of the Higgs boson signal event yields in the $Z(\nu\bar{\nu})H(b\bar{b})$ channel, resulting from events in which the lepton is outside the detector acceptance, and the $Z(\ell\ell)H(b\bar{b})$ process contributes less than 5% to the $W(\ell\nu)H(b\bar{b})$ channel when one of the leptons is outside the detector acceptance. The

dependency of the combined signal strength on the value assumed for the Higgs boson mass is shown in the left panel of Fig. 6-6.

In the right panel of Fig. 6-6 the best-fit values for the κ_V and κ_b parameters are shown. The parameter κ_V quantifies the ratio of the measured Higgs boson couplings to vector bosons relative to the SM value. The parameter κ_b quantifies the ratio of the measured Higgs boson partial width into $b\bar{b}$ relative to the SM value. They are defined as: ${\kappa_V}^2 = \sigma_{VH} \Big/ \sigma_{VH}^{\mathrm{SM}}$ and ${\kappa_b}^2 = \Gamma_{b\bar{b}} \Big/ \Gamma_{b\bar{b}}^{\mathrm{SM}}$, with the SM scaling of the total width [140]. By definition, $(\kappa_V, \kappa_b) = (1, 1)$ in the SM. The measured couplings are consistent with the expectations from the standard model, within uncertainties.

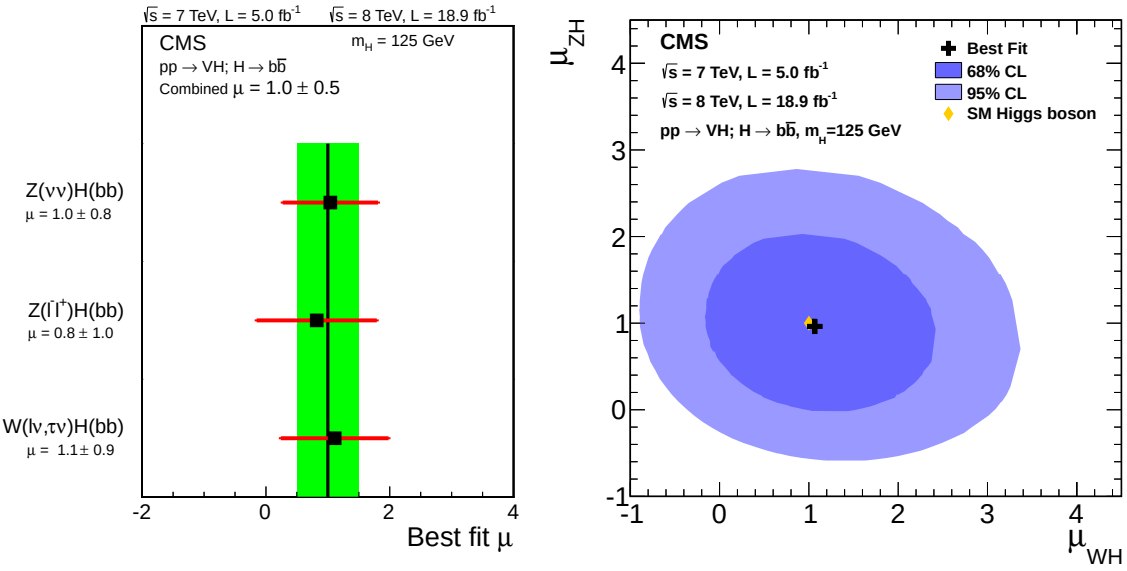

Figure 6-5. Left: Best-fit value of the signal strength μ for a 125 GeV Higgs boson, for partial combinations of channels and for all channels combined (band). Right: Best-fit values for the μ_{ZH}, μ_{WH} signal strength parameters for a 125 GeV Higgs boson.

6.1.2 $m(\mathrm{jj})$ cross-check analysis

The dijet invariant mass distributions for the $Z(\nu\bar{\nu})H(b\bar{b})$ channel in the more selective $m(\mathrm{jj})$ cross-check analysis are plotted in Fig. 6-7.

The left panel of Fig. 6-8 shows a weighted dijet invariant mass distribution for the combination of all $VH(b\bar{b})$ channels, in all boost regions, in the combined 7 and 8 TeV

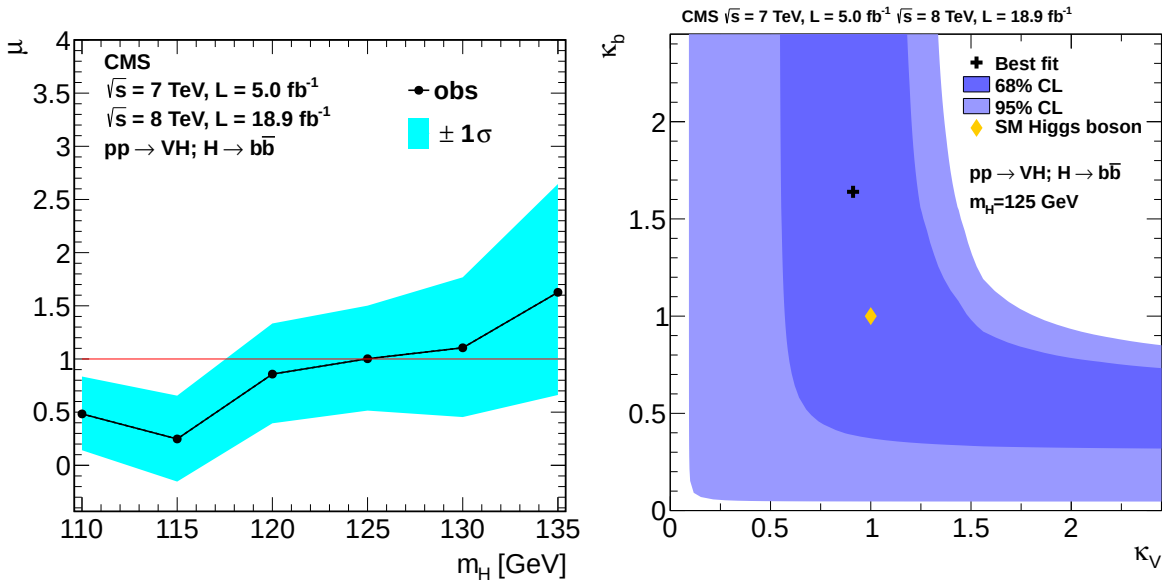

Figure 6-6. Left: Signal strength for all channels combined as a function of the value assumed for the Higgs boson mass. Right: Best-fit values for the κ_V and κ_b parameters. The cross indicates the best-fit values and the diamond shows the SM point.

data, using the event selection for the $m(\mathrm{jj})$ analysis described in Sec. 5.5. For each channel, the relative event weight for each boost region is obtained from the ratio of the expected number of signal events to the sum of expected signal and background events in a window of $m(\mathrm{jj})$ values between 105 and 150 GeV. The expected signal used corresponds to the production of the SM Higgs boson with a mass of 125 GeV. The weight for the highest-boost region is set to 1.0 and all other weights are adjusted proportionally. Also shown in the right panel of Fig. 6-8 is the same weighted dijet invariant mass distribution with all backgrounds subtracted except diboson production. In addition, Fig. 6-9 presents the weighted $m(\mathrm{jj})$ distributions for partial combinations of channels. The data are consistent with the presence of a diboson signal from ZZ and WZ channels, with $Z \to b\bar{b}$, with a rate consistent with the standard model prediction from the MADGRAPH generator, together with a small excess consistent with the production of the standard model Higgs boson with a mass of 125 GeV. For the $m(\mathrm{jj})$ analysis, a fit to the dijet invariant mass distribution results in a measured Higgs

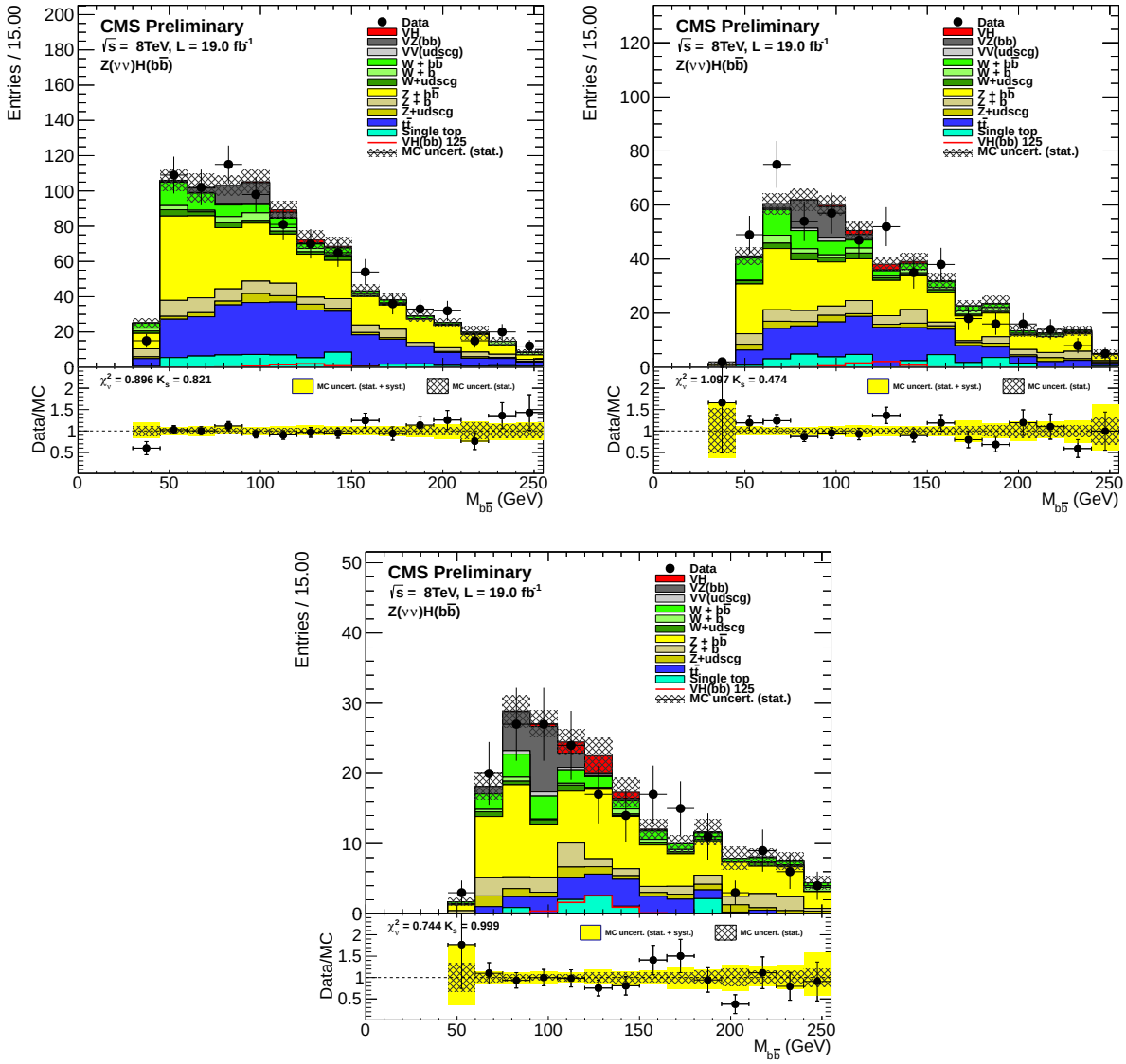

Figure 6-7. Distributions of the dijet invariant mass for the $Z(\nu\bar{\nu})H(b\bar{b})$ channel in the more selective $m(\mathrm{jj})$ analysis in the low-boost region (top left), the intermediate-boost (top right), and the high-boost (bottom).

boson signal strength, relative to that predicted by the standard model, of $\mu = 0.8 \pm 0.7$, with a local significance of 1.1 standard deviations with respect to the background-only hypothesis. For a Higgs boson of mass 125 GeV, the expected and observed 95% C.L. upper limits on the production cross section, with respect to the standard model prediction, are 1.4 and 2.0, respectively.

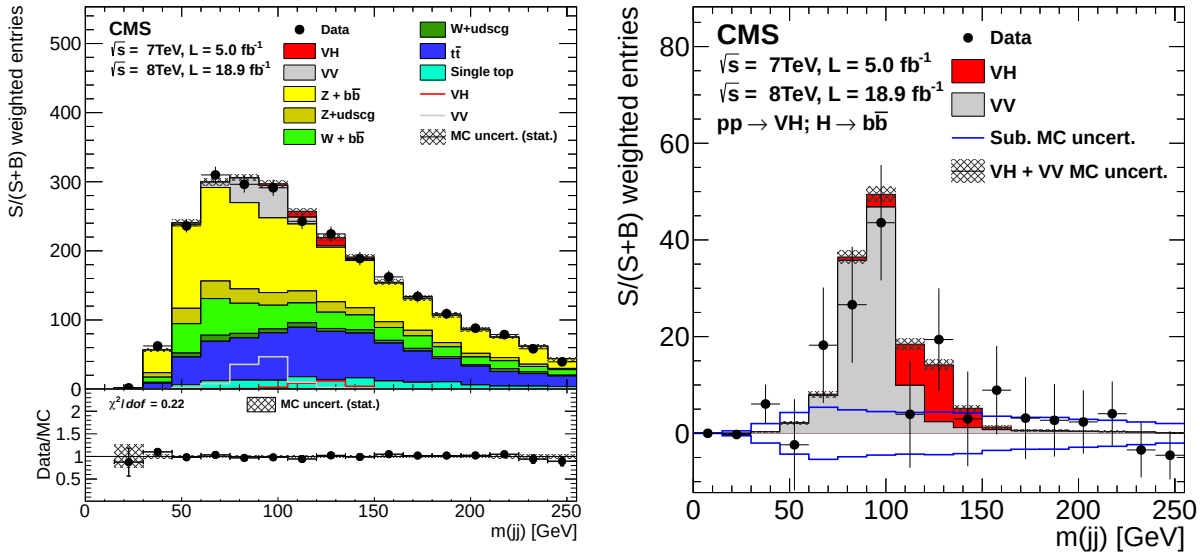

Figure 6-8. Left: Weighted dijet invariant mass distribution, combined for all channels. See text for details. The bottom inset shows the data/MC ratio. Right: Same distribution with all backgrounds, except *VV*, subtracted.

6.2 Diboson Signal Extraction

As a validation of the multivariate technique, BDT discriminants are trained targeting production of *ZZ* and *WZ* with $Z \to b\bar{b}$ decays as signal. All other processes, including *VH* production (at the predicted standard model rate for a 125 GeV Higgs mass), are treated as background. This is done for the 8 TeV dataset only. The BDT output distributions for the $Z(\nu\bar{\nu})H(b\bar{b})$ channel are plotted in Fig.6-10.

The observed excess of events for the combined *WZ* and *ZZ* processes with $Z \to b\bar{b}$ differs by 7.5 standard deviations from the event yield as predicted from the background-only hypothesis. The expected significance is 6.3 from the signal+background hypothesis. The corresponding signal strength, relative to the prediction from the

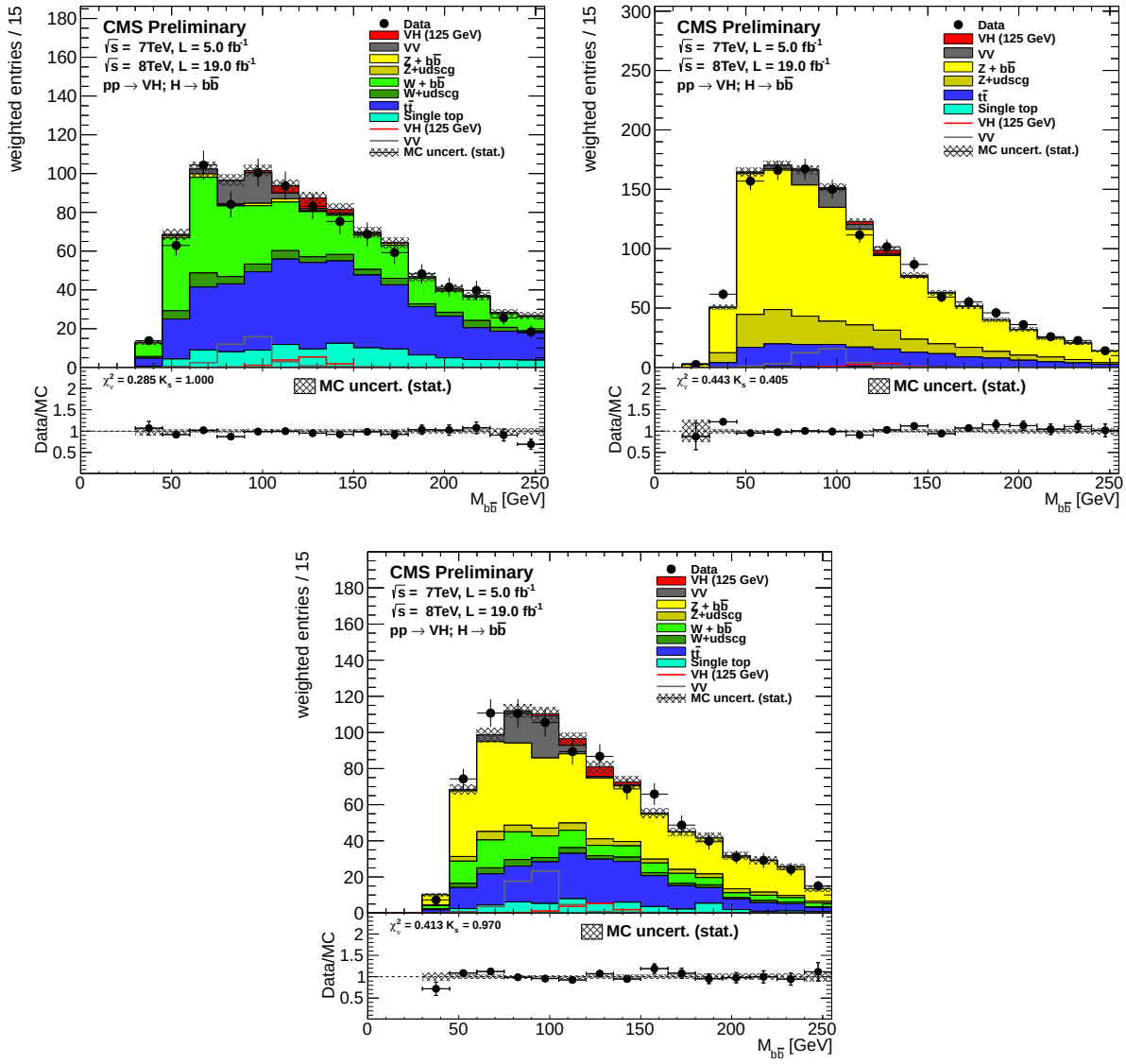

Figure 6-9. Weighted dijet invariant mass distributions for partial combinations of channels: $W(\ell\nu, \tau\nu)H(b\bar{b})$ (top left), $Z(\ell\ell)H(b\bar{b})$ (top right), and $Z(\nu\bar{\nu})H(b\bar{b})$ (bottom). See text for details. The bottom inset shows the data/MC ratio.

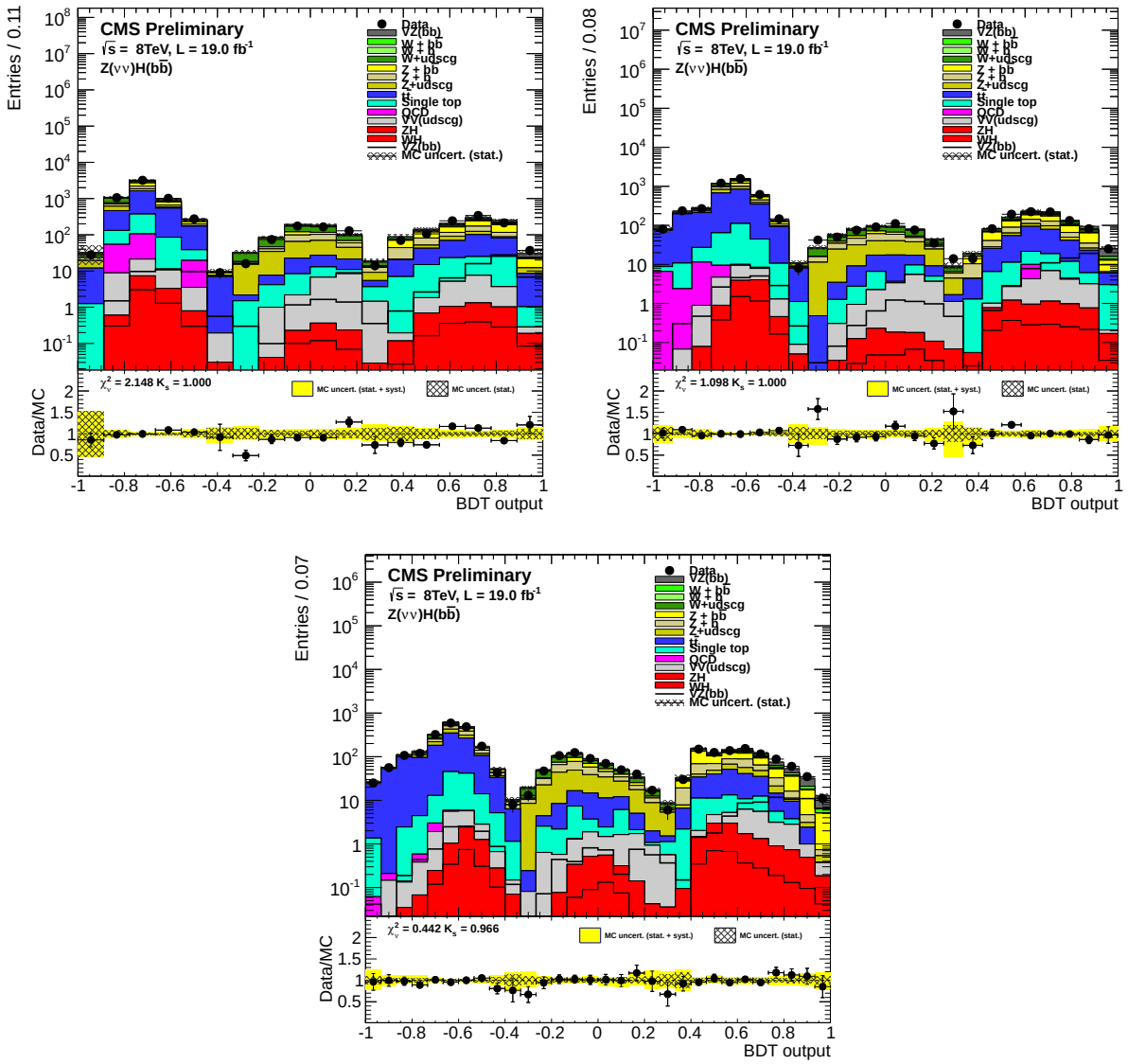

Figure 6-10. Post-fit BDT output distributions trained to find the production of ZZ and WZ with $Z \to b\bar{b}$ decays for the $Z(\nu\bar{\nu})H(b\bar{b})$ channel in the low-boost region (top left), the intermediate-boost (top right), and the high-boost (bottom).

diboson MADGRAPH generator mentioned in Sec. 5.2, and rescaled to the cross section from the NLO MCFM generator, is measured to be $\mu_{WZ,ZZ} = 1.19^{+0.28}_{-0.23}$.

The observation of the well-known diboson production using the same techniques as applied in the $H \to b\bar{b}$ search lends a lot of confidence to the $H \to b\bar{b}$ search results. Furthermore, the diboson process can be used as a "standard candle" for calibration purposes when larger amount of data is accumulated in the future.

6.3 Run I Legacy Higgs Combination Results

CMS has published the final "Run I legacy" measurements of the properties of the Higgs boson in Ref. [33], including results from a comprehensive set of Higgs decay channels: $\gamma\gamma$, ZZ, W^+W^-, $\tau^+\tau^-$, $b\bar{b}$, and $\mu^+\mu^-$. The global Higgs combination results are obtained from simultaneous likelihood fits of all channels, with all of the systematic and theoretical uncertainties profiled in the fits. Since the publication of $VH(b\bar{b})$ results, a more accurate prediction of the $p_T(Z)$ spectrum in the ZH production mode has become available [141–144], taking into account the contribution of the gluon-gluon initiated process, $gg \to ZH$. This process arises from NLO calculations and has a sizeable contribution in the high-p_T kinematic regime, which happens to be the most sensitive categories of the analysis. Therefore, the $gg \to ZH$ contribution has been incorporated into the legacy Higgs combination. [1] This leads to an increase of the expected signal yields by 10–30% for $p_T(Z) > 150$ GeV in the $Z(\nu\bar{\nu})H(b\bar{b})$ and $Z(\ell\ell)H(b\bar{b})$ channels. Overall, the expected excess significance for $VH(b\bar{b})$ increases from 2.1 standard deviations to 2.5; while the observed significance remains unchanged at 2.1 standard deviations. On the other hand, the signal strength decreases slightly to $\mu = 0.890^{+0.469}_{-0.441}$.

The best-fit value for the combined signal strength of all Higgs production and decay channels, assuming m_H = 125 GeV, is $\mu = 1.00^{+0.14}_{-0.13}$, consistent with the expectation

[1] The WH and ZH cross section numbers from the CERN Report 3 [145] are used.

for the SM Higgs boson. Fig. 6-11 displays the signal strengths obtained in different independent combinations of channels, grouped by predominant decay modes and tags targeting specific production mechanisms. For $H \to b\bar{b}$ decays, two production tags are used: VH and $t\bar{t}H$. The plot provides the compatibility test of all the channels included in the combination.

Several other tests of coupling strengths have been carried out, and no statistically significant deviations from the SM expectation are found. In the SM, the Higgs boson is directly responsible for the particle masses: the Yukawa coupling between the Higgs boson and the fermion is proportional to the mass of that fermion; and the gauge coupling to the massive vector boson is proportional to the square of the mass of that vector boson. A graphical representation of the fit result for coupling strength deviations as a function of the particle mass is shown in Fig. 6-12. In the plot, the ordinates are different for fermions and massive vector bosons: $\lambda_f = \kappa_f m_f / v$ for fermions and $\sqrt{g_V/(2v)} = \sqrt{\kappa_V} m_V / v$ for vector bosons, where v is the vacuum expectation value. The linear relationship from the fit result agrees with the SM expectation, demonstrating that the Higgs boson is very likely fundamental to mass generation of both fermions and vector bosons.

6.4 Comparison with ATLAS Results

The ATLAS experiment has also performed a search for the SM VH production with $H \to b\bar{b}$ decay in five decay channels: $W(\mu\nu)H(b\bar{b})$, $W(e\nu)H(b\bar{b})$, $Z(\mu\mu)H(b\bar{b})$, $Z(ee)H(b\bar{b})$, and $Z(\nu\bar{\nu})H(b\bar{b})$. Their data samples correspond to integrated luminosities of 4.7 fb^{-1} at $\sqrt{s} = 7$ TeV and 20.3 fb^{-1} at $\sqrt{s} = 8$ TeV. The left panel of Fig. 6-13 shows the event yields as a function of $\log_{10}(S/B)$ for data, background and Higgs boson signal with m_H =125 GeV from the ATLAS multivariate analysis on the 8 TeV data (analogous to Fig. 6-2); the right panel shows the distribution of $m(\mathrm{jj})$ in data after subtraction of all background except for VV, as obtained with the 8 TeV ATLAS cut-based analysis (analogous to Fig. 6-8).

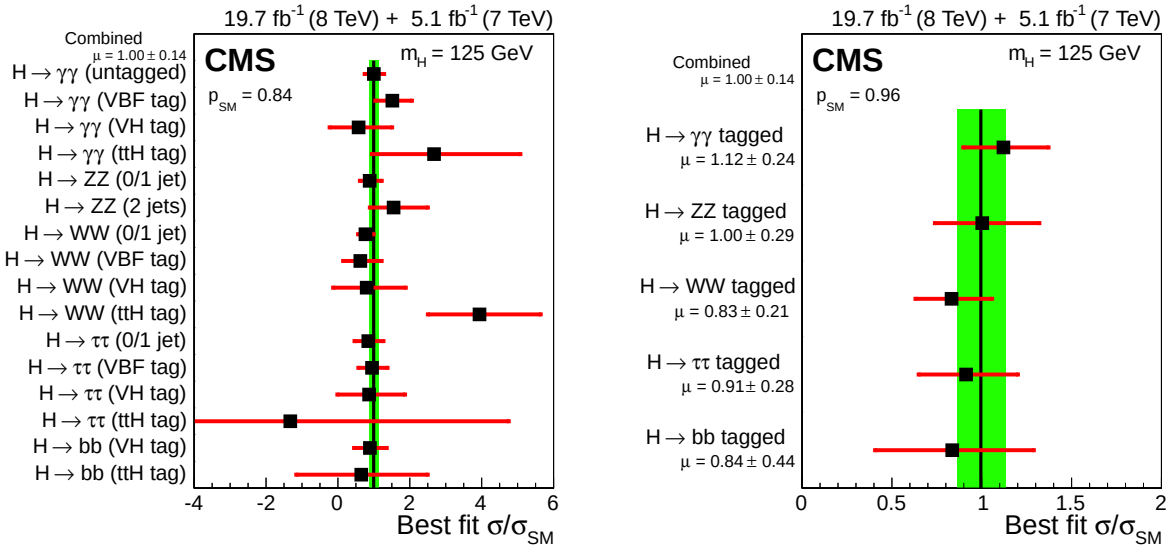

Figure 6-11. Values of the best-fit σ/σ_{SM} for the global combination (vertical bank) and for partial combinations by predominant decay mode and production tag (left) and only by predominant decay mode (right).

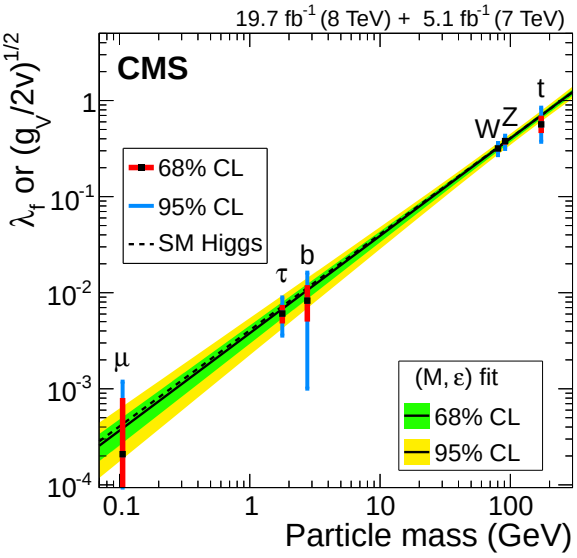

Figure 6-12. Summary of the fits for deviations in the coupling as a function of the particle mass. The dashed line corresponds to the SM expectation. The continuous line shows the result of the coupling-mass fit with the 68% and 95% C.L. regions. The ordinates are different for fermions and massive vector bosons to take into account the expected SM scaling of the coupling with mass, depending on the type of particle.

ATLAS observed an excess of events above the post-fit background that corresponds to a local significance of 1.4 standard deviations. The expectation is 2.6 standard deviations for the SM Higgs boson with $m_H = 125.36\,\text{GeV}$. In comparison to the CMS analysis, ATLAS achieved a very similar expected significance, but a lower observed significance. The signal strength, i.e. the ratio of the observed signal yield to the SM expectation, is found to be $\mu = 0.52 \pm 0.32\,(\text{stat.}) \pm 0.24\,(\text{syst.})$.

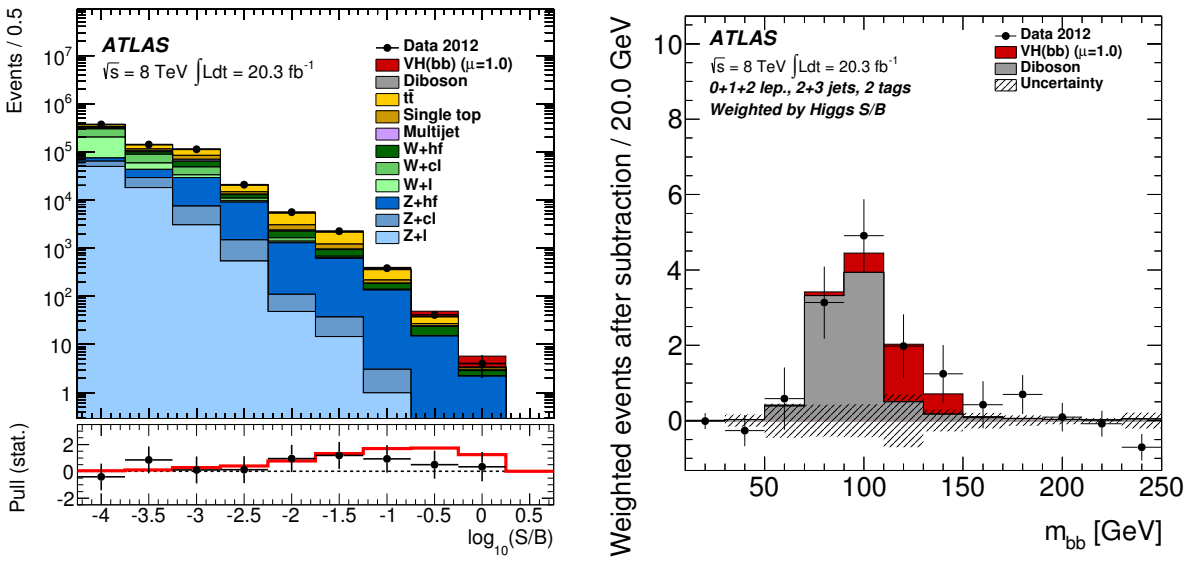

Figure 6-13. Left: Event yields as a function of $\log_{10}(S/B)$ from the ATLAS multivariate analysis on the 8 TeV data. Right: Distribution of $m(\text{jj})$ in data after subtraction of all background except for VV, from the 8 TeV ATLAS cut-based analysis.

CHAPTER 7
CONCLUSIONS AND FUTURE PROSPECTS

7.1 Concluding Remarks

A search for the standard model Higgs boson when produced in association with an electroweak vector boson and decaying to $b\bar{b}$ is reported for the particular $Z(\nu\bar{\nu})H(b\bar{b})$ channel alone, and for the combination of six $VH(b\bar{b})$ channels ($W(\mu\nu)H(b\bar{b})$, $W(e\nu)H(b\bar{b})$, $W(\tau\nu)H(b\bar{b})$, $Z(\mu\mu)H(b\bar{b})$, $Z(ee)H(b\bar{b})$, and $Z(\nu\bar{\nu})H(b\bar{b})$). The search is performed in data samples corresponding to integrated luminosities of up to 5.1 fb^{-1} at $\sqrt{s} = 7$ TeV and up to 18.9 fb^{-1} at $\sqrt{s} = 8$ TeV, recorded by the CMS experiment at the LHC.

Exclusion limits, at the 95% confidence level, on the VH production cross section times the $H \to b\bar{b}$ branching fraction, with respect to the expectations for a standard model Higgs boson, are derived for the Higgs boson mass range of 110–135 GeV. For a Higgs boson mass of 125 GeV, the exclusion limit is expected to be 0.95 in the absence of signal. A mild excess above the background expectation was observed in data, thus the observed exclusion limit is less stringent at 1.89.

The excess of events is quantified to have a local significance of 2.1 standard deviations. This is compatible with the expected significance in the presence of the 125 GeV SM Higgs boson, which is also 2.1 standard deviations. The signal strength corresponding to this excess, relative to that of the standard model Higgs boson, is $\mu = 1.0 \pm 0.5$. The measurements presented here represent the first indication of the $H \to b\bar{b}$ decay at the LHC. These measurements are one of the key components in the global fit of the Higgs couplings and contributed to the CMS Run I "legacy" publication that describes the properties of the discovered Higgs boson.

7.2 Outlook for Run II $VH(b\bar{b})$

For the Run II (2015–2018) $VH(b\bar{b})$ analysis at $\sqrt{s} = 13$ TeV, particularly for the $Z(\nu\bar{\nu})H(b\bar{b})$ channel, there are a few challenges that need to be addressed:

- **Pileup**. To achieve the luminosity goal, LHC will restart in 2015 with bunch spacing of 25 ns and peak instantaneous luminosity exceeding $1 \times 10^{34}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$. Due to the shorter bunch spacing, out-of-time pileup effects will become more severe, particularly in the calorimeter energy measurements. The average number of in-time pileup interactions will be 25 or more. This harsher pileup environment will deteriorate the performance in jet energy resolution, lepton identification, b tagging, $E_{\mathrm{T}}^{\mathrm{miss}}$ reconstruction, etc. A new method that looks promising is the pileup per particle identification (PUPPI) algorithm [146, 147]. The algorithm calculates a weight for each particle based on the charged pileup particle distribution on an event-by-event basis. The weight (between 0 and 1) characterizes how pileup-like a particle is, and is used to rescale its four-momentum before it enters jet clustering. The algorithm has shown significant improvement in the jet p_{T} resolution, jet mass resolution, and $E_{\mathrm{T}}^{\mathrm{miss}}$ resolution at high pileup.

- **Trigger**. At 13 TeV, the parton luminosities increase by about a factor of 2 for the production of heavy resonances, such as W/Z (see Fig. 7-1). The expected increase in the instantaneous luminosity is also close to a factor of 2. Thus, the trigger rates are generally expected to be at least 4 times higher. However, the $E_{\mathrm{T}}^{\mathrm{miss}}$ trigger rates are largely due to the QCD production which has more or less the same production cross section in 8 TeV and in 13 TeV. But their rates are much more sensitive to pileup, thus it is still reasonable to expect the rates to be a few times higher. Increasing the $E_{\mathrm{T}}^{\mathrm{miss}}$ threshold is an effective way to reduce the rates, but it hurts signal acceptance and compromises analysis sensitivity. A better alternative is to improve the online physics object reconstruction and energy calibration by making them more similar to the offline ones. Better online-offline agreement will sharpen the trigger turn-on curves, and then the trigger threshold can be raised while retaining similar signal efficiency. Better noise rejection, jet ID, and pileup jet ID are very powerful in rejecting events with fake or induced $E_{\mathrm{T}}^{\mathrm{miss}}$ (originating from instrumental effects). Usage of track-based $E_{\mathrm{T}}^{\mathrm{miss}}$, i.e. $E_{\mathrm{T}}^{\mathrm{miss}}$ reconstructed by only tracks, may prove very useful as it has little dependence on in-time and out-of-time pileup. Event selection based on jet topology can help removing QCD contamination. Other analysis-level event selection can also be brought to the trigger level to reject more background events. However, a complicated trigger design could make the characterization of its efficiency more difficult.

- **Higgs boson reconstruction**. Boosted massive particle undergoing hadronic decay has a unique signature. For instance, the two b quark daughters from the decay of a very high-p_{T} Higgs boson will be highly collimated. The current anti-k_{T} jet clustering algorithm may fail to resolve them as two individual jets when the boost is particularly high. At the 13 TeV energy regime, this signature may be exploited. A jet substructure technique, which consists of mass drop tagging and jet filtering, was introduced in the BDRS paper [29] to more optimally reconstruct the boosted $H \to b\bar{b}$ decays and distinguish them from the QCD background. The technique aims to capture all the decay products originating from

the Higgs boson, including any gluon radiation, into a merged "fat" jet, analyze the momentum structure of the fat jet constituents, and discard contaminations from underlying event and pileup (see Fig. 7-2 for a simplified illustration of the procedure). Jet substructure has since become an active research area, and many more techniques have been introduced. CMS has explored a number of them in the context of tagging W-jets, i.e. merged jets that originate from boosted hadronically-decaying W bosons [148], and top-jets [149]. Information from jet substructure generally helps in two ways: improving jet p_T and mass resolutions, and increasing background rejection power. Besides jet substructure methods, continual improvement in the b jet energy calibration, e.g. by using multivariate regression technique, remains important.

- b **tagging**. Conventional secondary vertex fitting is usually seeded by reconstructed jets. For boosted $H \rightarrow b\bar{b}$ decays, both b quarks may end up in a single merged jet, and the fit may fail to find one or both of the secondary vertices. This inefficiency can be recovered by using the inclusive vertex finding (IVF) technique [150], which finds secondary vertices using tracks with large impact parameter as seed. IVF is capable of reconstructing secondary vertices even at small opening angles, completely independent of jet reconstruction. In addition, subjet b tagging, e.g. by applying the CSV algorithm on subjets obtained with certain jet substructure technique, has also been implemented in CMS [24]. IVF and subjet CSV are particularly suited for b tagging in boosted topologies.

- $E_T^{\mathbf{miss}}$ **reconstruction**. E_T^{miss} resolution degrades quickly at high pileup. MVA PF E_T^{miss} [23] has been developed to mitigate pileup effects on E_T^{miss} by using a set of multivariate regressions. It uses information that helps identifying energy contributions from different types of particles (charged or neutral, originating from hard scatter or pileup, and clustered or unclustered). MVA PF E_T^{miss} has shown significantly reduced dependency of the resolution on pileup interactions in both data and simulation. Also, PUPPI may further reduce pileup effects.

- **Background modeling**. It goes without saying that better understanding and modeling of the background processes will be important. For $VH(b\bar{b})$ analysis, the p_T spectrum, the jet multiplicity spectrum and the heavy flavor content of the production of W or Z with jets are particularly critical. Better modeling of various Higgs properties, the process of $g \rightarrow b\bar{b}$, the jet fragmentation function in the context of jet substructure, etc will also reduce theoretical/phenomenological uncertainties.

- **Monte Carlo statistics**. Larger amount of data will require a lot more Monte Carlo events to be characterized with good precision. In particular, the raw number of background events that survives in the most sensitive kinematic regime is usually very few. This leads to MC statistical uncertainty which is one of the larger components of the systematic uncertainty. The 8 TeV analysis sensitivity is

currently limited by statistical uncertainty, but the 13 TeV sensitivity may be limited by such systematic uncertainty.

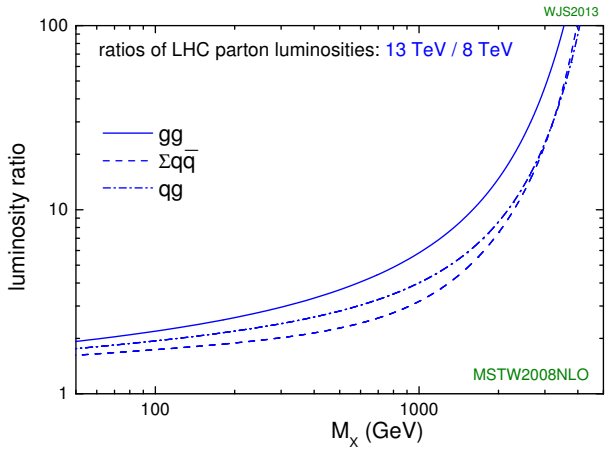

Figure 7-1. Ratios of parton luminosities at 13 TeV to that at 8 TeV at the LHC as a function of the mass of heavy resonance for processes initiated by gg, $q\bar{q}$ (all flavors), or qg [9].

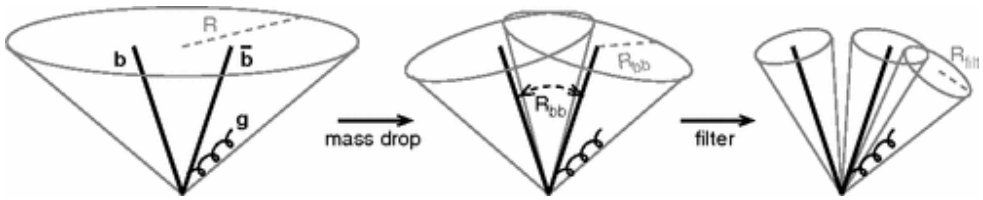

Figure 7-2. Mass drop algorithm starts with jet clustering using a large radius R. The hardest jet is split into two by undoing the last stage of clustering. In this neighbourhood, the three hardest subjets, clustered using a small radius R_{filt}, are selected [29].

APPENDIX A
EVENT DISPLAYS

Recorded events can be visualized using CMS event display software named Fireworks, also known as cmsShow [151]. Figs. A-1–A-6 show the event displays of six $Z(\nu\bar{\nu})H(b\bar{b})$ candidates in the high-boost signal region. In each figure, three views are provided: the close transverse projection in the tracker that shows the secondary vertices and track impact parameters in *b* tagged jets (the axes are in unit of centimeters), the far transverse projection that includes all CMS subdetectors and shows the jets and $E_{\mathrm{T}}^{\mathrm{miss}}$, and the 3D view of the full event. Table A-1 lists the values of the important variables in these events. Note that the values of the event BDT discriminant are in step of 0.05 after repartitioning by the cuts on the background-specific BDT discriminants. In the *VH*-enriched partition, the values are in the range of 0.55–1.00 (inclusive).

Table A-1. The values of the important variables in the displayed events. Kinematic variables are in units of GeV.

Event number (run:lumi:event)	$m(\mathrm{jj})$	$p_{\mathrm{T}}(\mathrm{jj})$	$E_{\mathrm{T}}^{\mathrm{miss}}$	$p_{\mathrm{T}}(\mathrm{j}_1)$	$p_{\mathrm{T}}(\mathrm{j}_2)$	$\mathrm{CSV}_{\mathrm{max}}$	$\mathrm{CSV}_{\mathrm{min}}$	Event BDT discriminant
194108:598:585302653	130.26	199.73	210.67	146.30	89.96	0.988	0.875	0.90
195656:123:113158630	122.55	204.38	215.91	132.22	105.60	0.996	0.928	1.00
198212:263:146829894	129.79	368.99	364.58	270.37	103.17	0.850	0.833	0.95
201278:1819:1951144088	120.13	301.92	311.50	228.15	77.11	0.929	0.843	0.95
205310:520:698472368	120.85	254.40	241.06	160.93	120.50	0.969	0.946	1.00
206246:1070:910620182	137.12	191.57	178.58	185.84	45.58	0.907	0.868	0.80

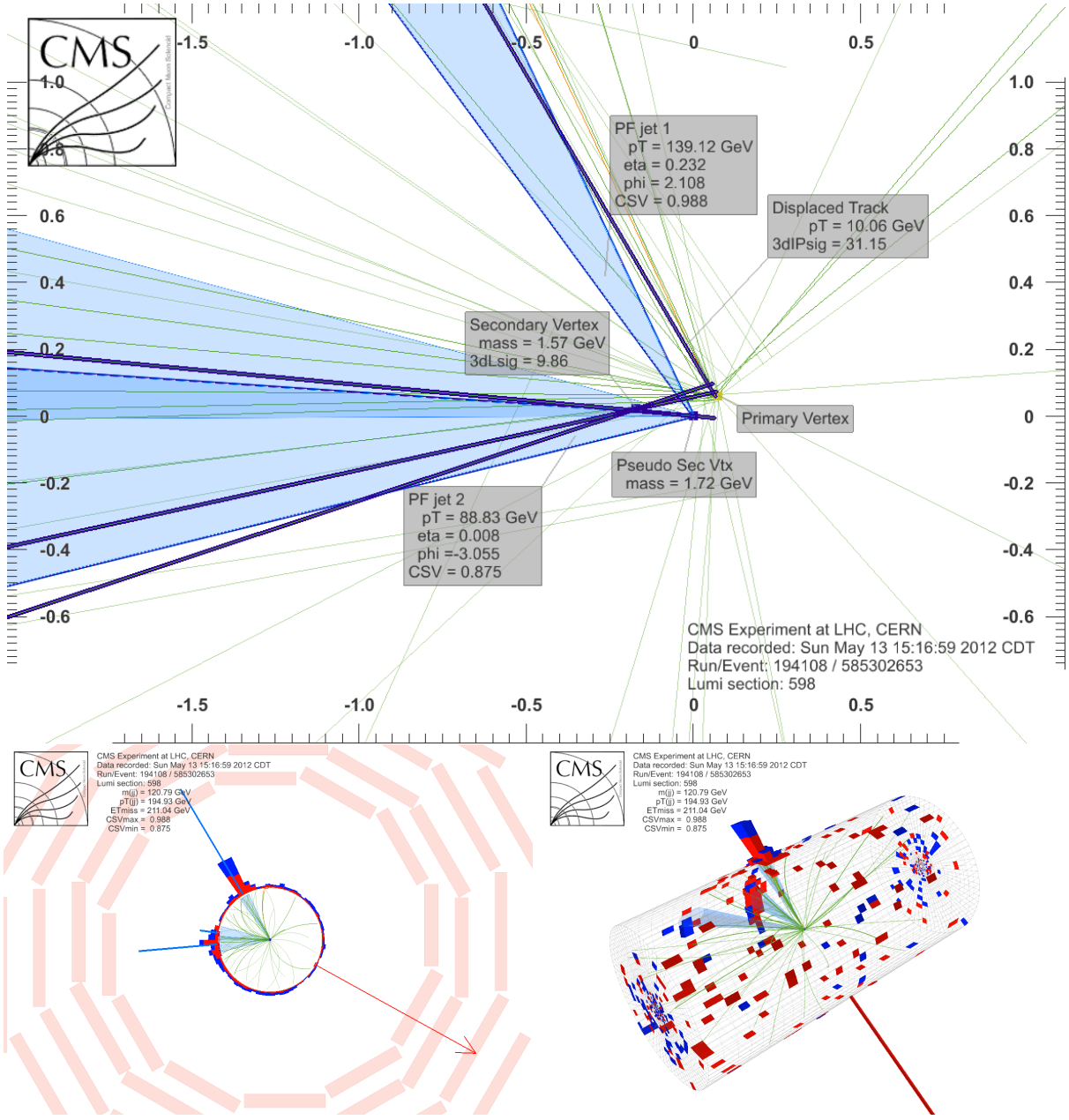

Figure A-1. Event display of $Z(\nu\bar{\nu})H(b\bar{b})$ candidate. Run: 194108 Lumi section: 598 Event: 585302653.

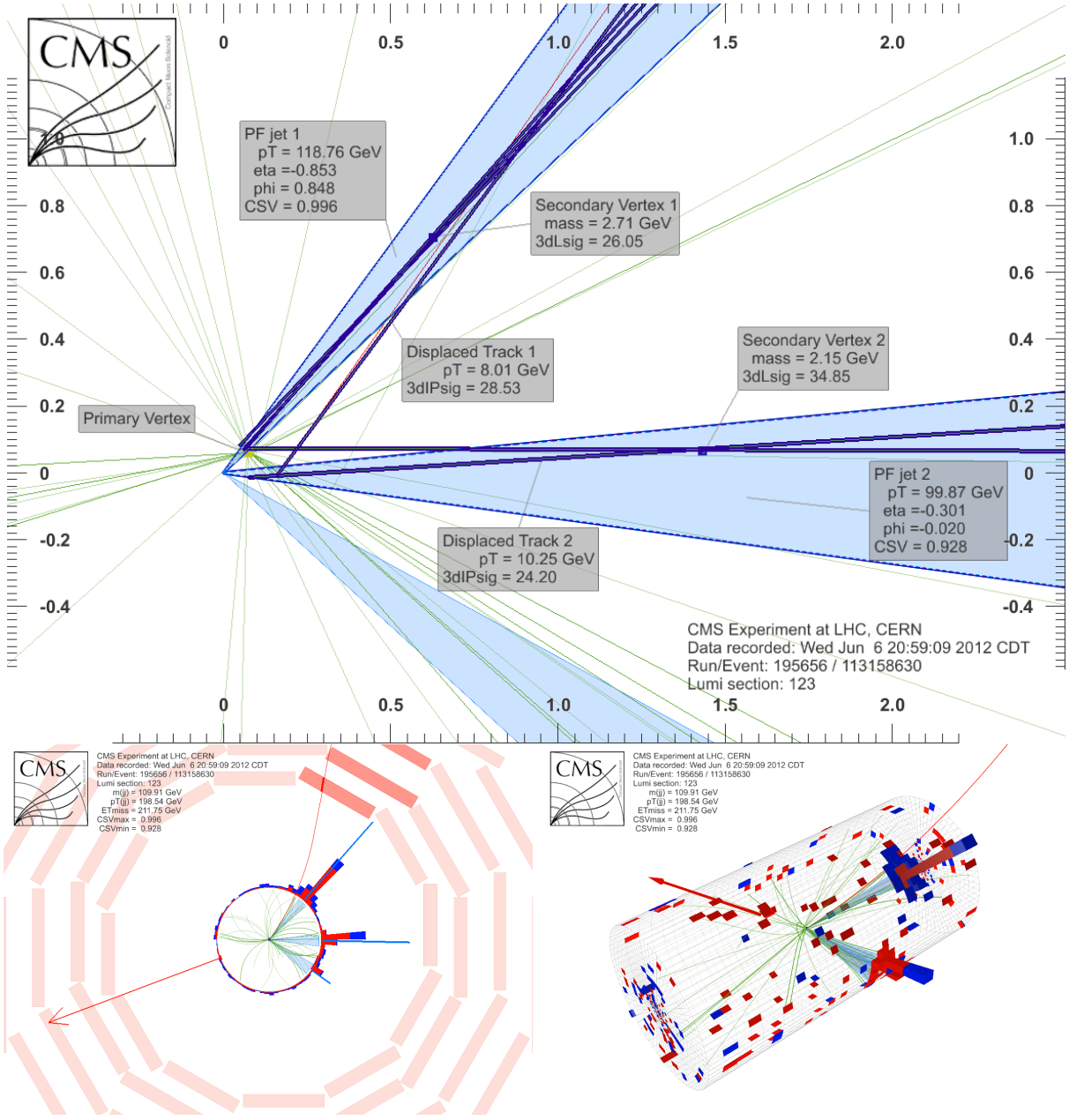

Figure A-2. Event display of $Z(\nu\bar{\nu})H(b\bar{b})$ candidate. Run: 195656 Lumi section: 123 Event: 113158630.

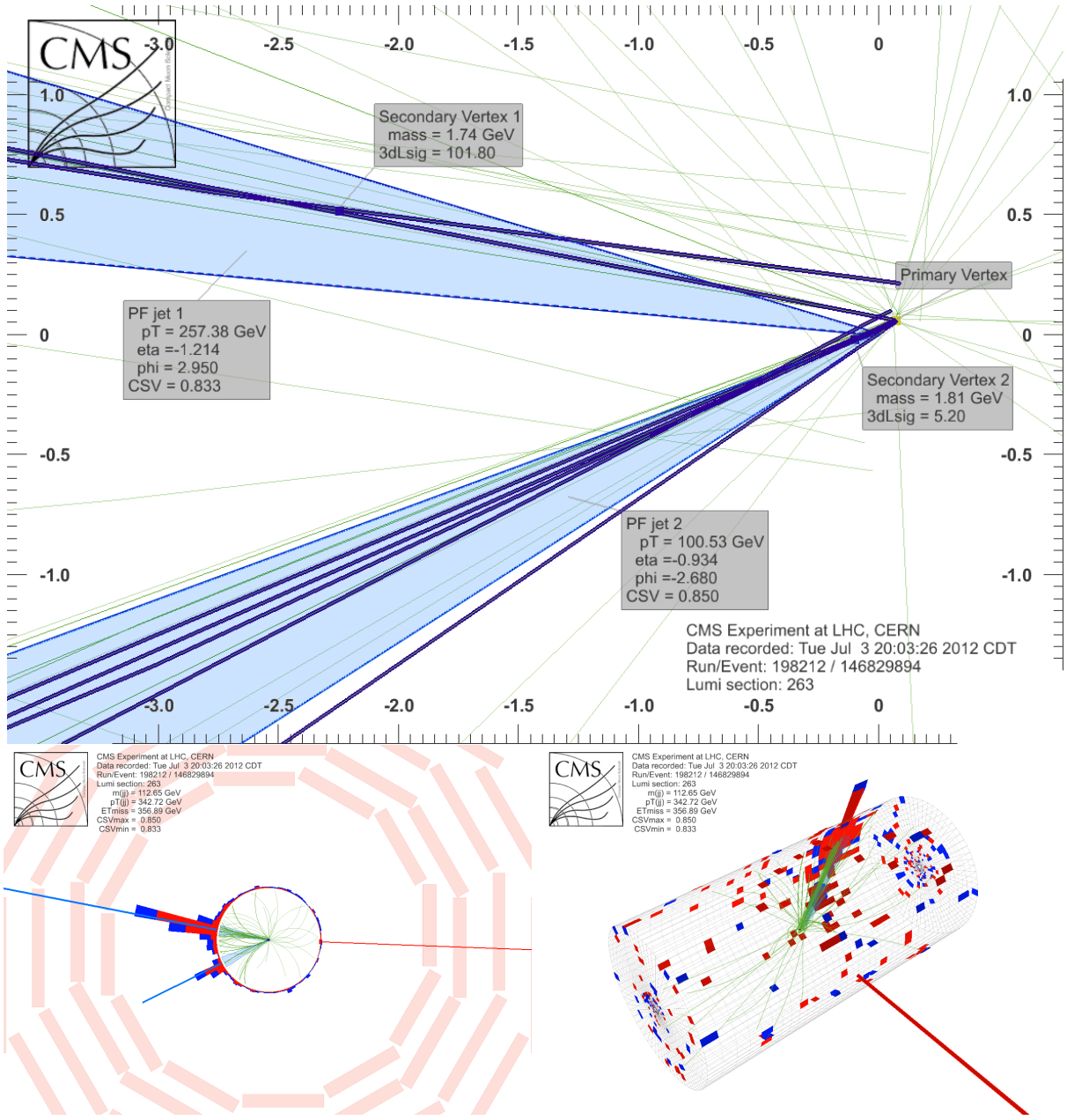

Figure A-3. Event display of $Z(\nu\bar{\nu})H(b\bar{b})$ candidate. Run: 198212 Lumi section: 263 Event: 146829894.

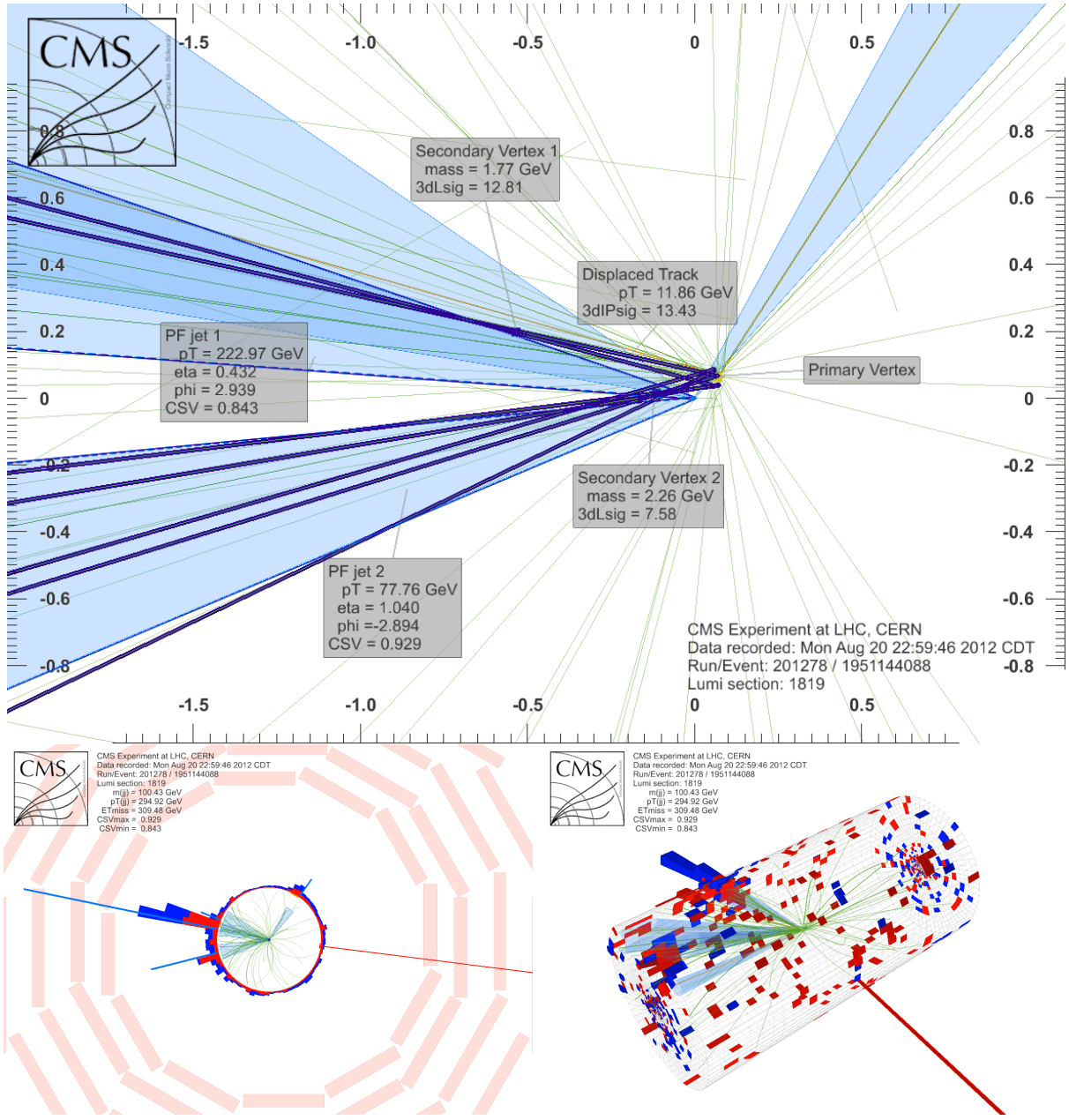

Figure A-4. Event display of $Z(\nu\bar{\nu})H(b\bar{b})$ candidate. Run: 201278 Lumi section: 1819 Event: 1951144088.

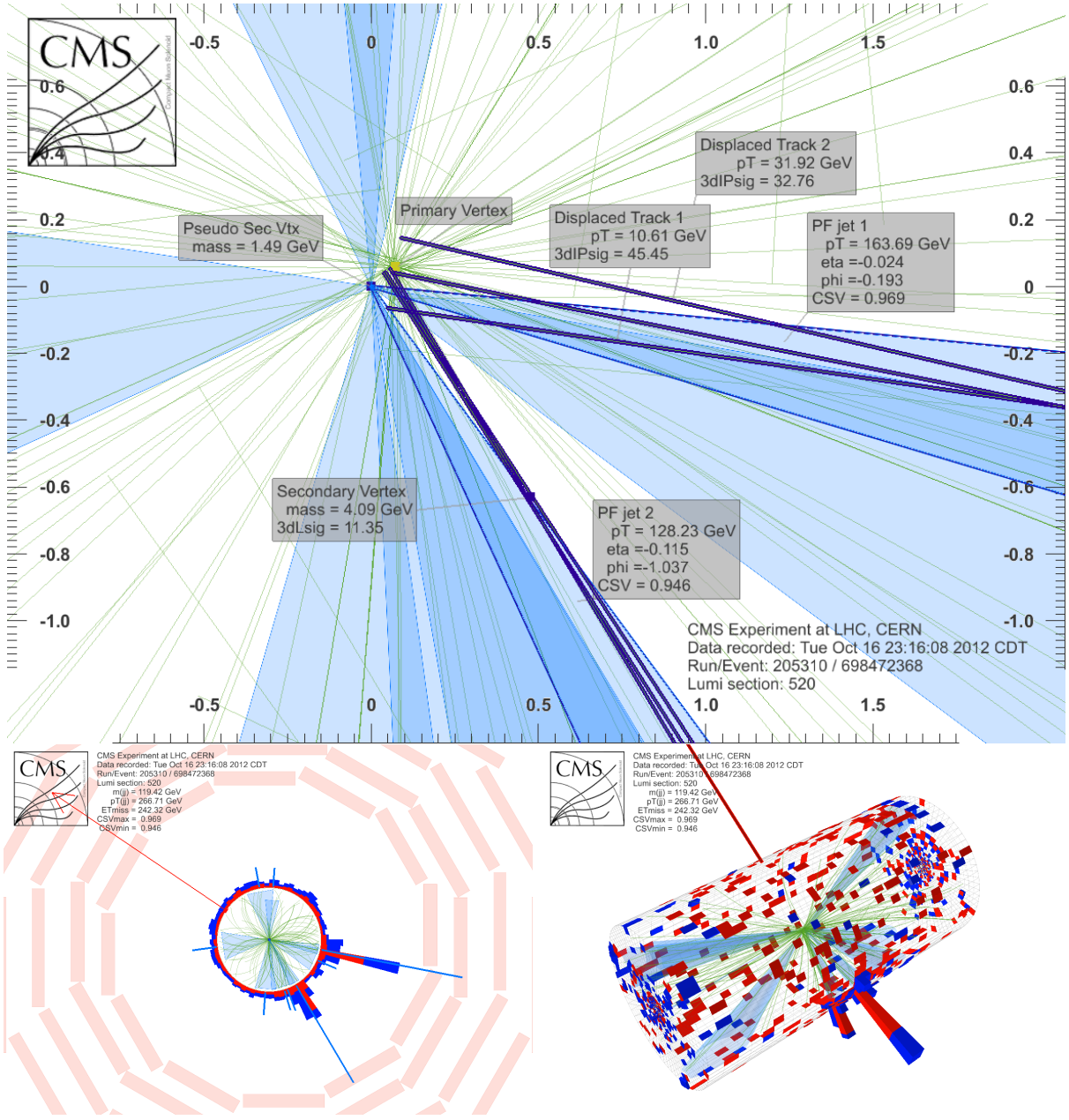

Figure A-5. Event display of $Z(\nu\bar{\nu})H(b\bar{b})$ candidate. Run: 205310 Lumi section: 520 Event: 698472368.

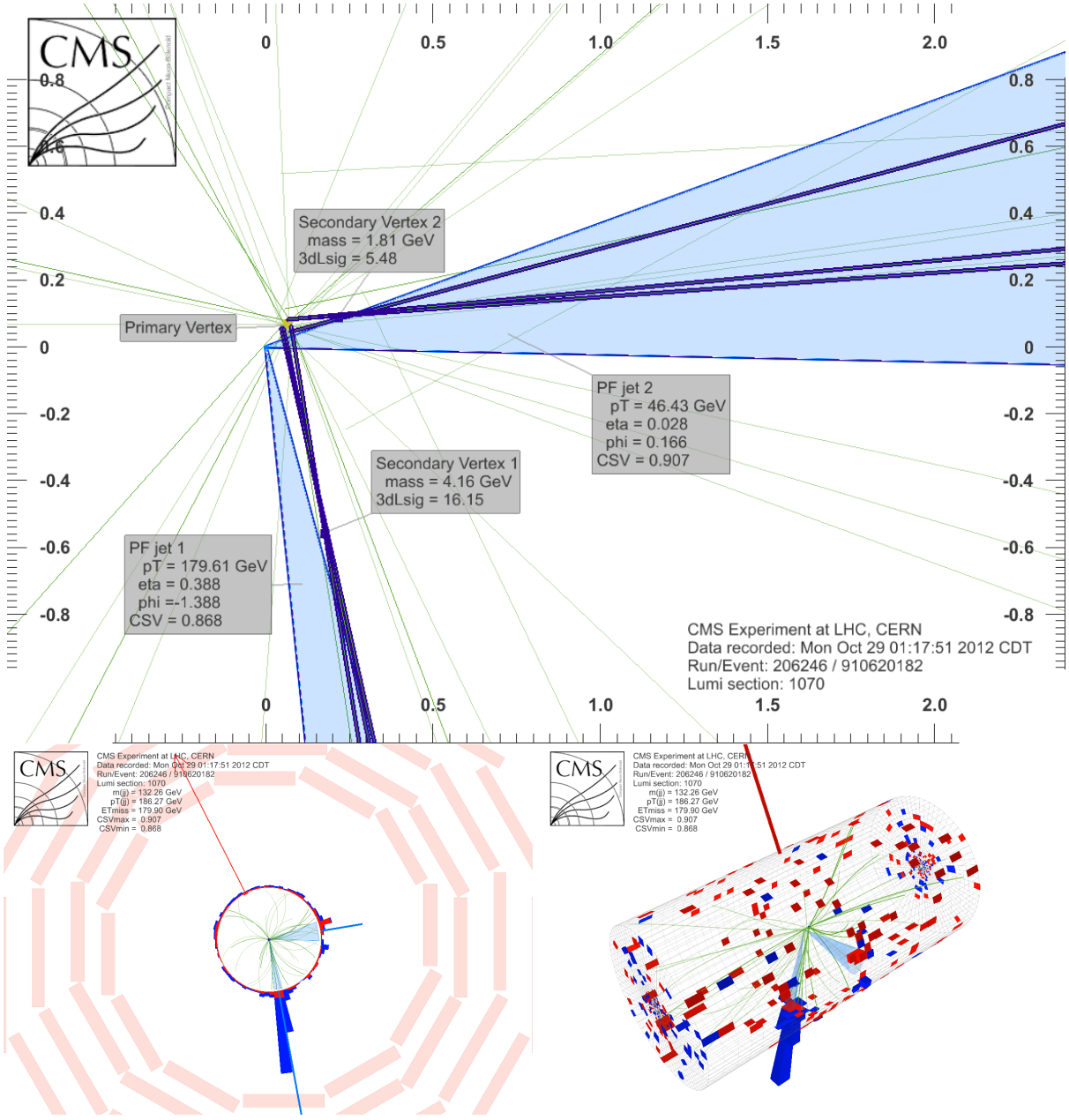

Figure A-6. Event display of $Z(\nu\bar{\nu})H(b\bar{b})$ candidate. Run: 206246 Lumi section: 1070 Event: 910620182.

APPENDIX B
CONTROL REGION DISTRIBUTIONS

Control regions in data are selected to test the accuracy of the modeling of various distributions in the simulated samples. Fig. B-1 checks for agreement between data and simulation in the distributions of various input variables to the b jet energy regression in the high-boost $t\bar{t}$ control region. The scale factors have been applied to the simulated samples. Fig. B-2 checks the distributions of $p_{\mathrm{T}}(\mathrm{jj})$ before and after the regression is applied in different control regions. Figs. B-3–B-7 check the distributions of various input variables to the event BDT discriminant in the high-boost control regions. Figs. B-8–B-12 and B-13–B-17 check the same distributions in the intermediate- and low-boost control regions, respectively. Figs. B-18–B-20 check the outputs of the event BDT discriminants in the high-, intermediate-, and low-boost control regions. Figs. B-21–B-23 check the outputs of the event BDT discriminants that are trained using $VZ(b\bar{b})$ as signal in the high-, intermediate-, and low-boost control regions.

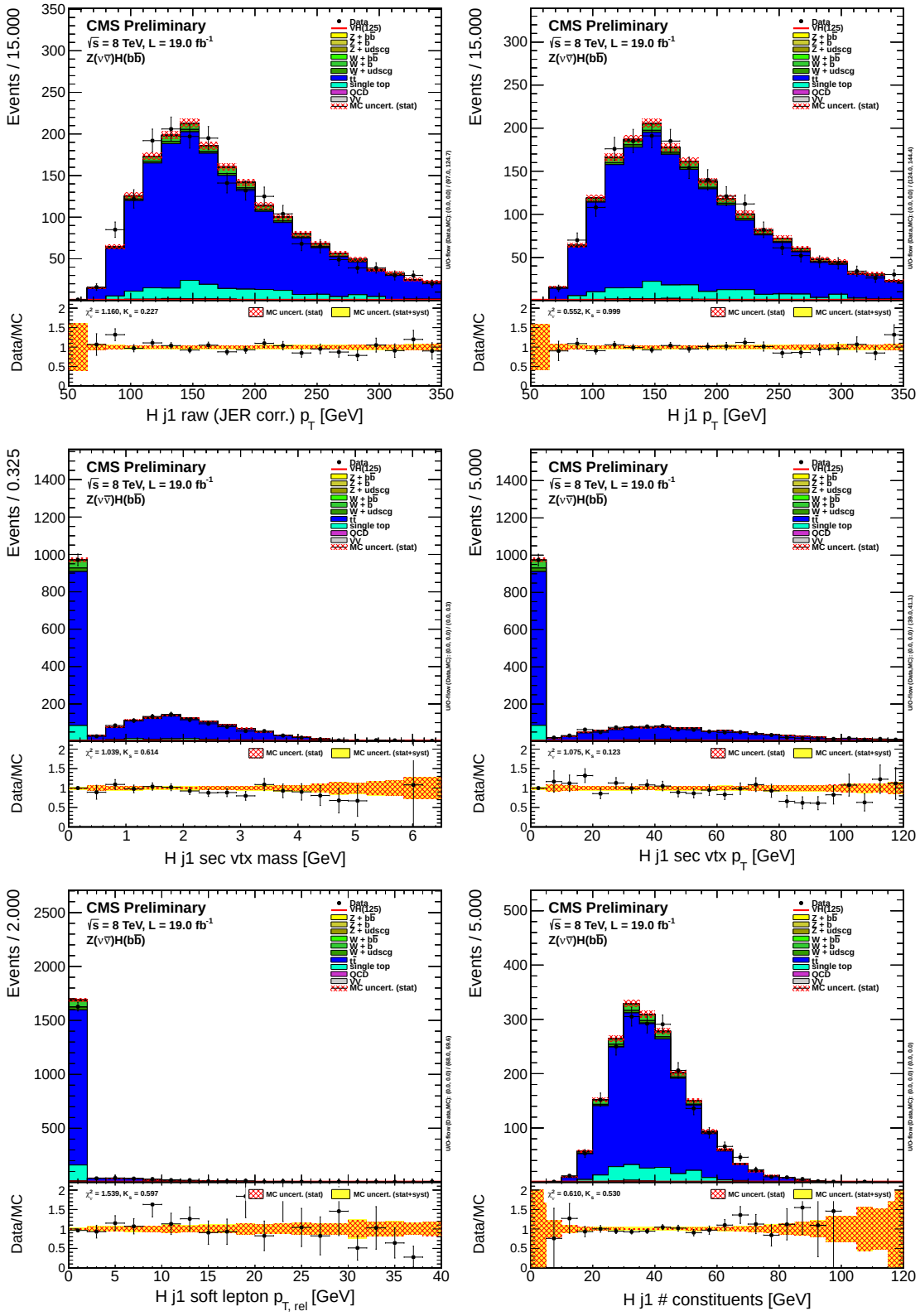

Figure B-1. Distributions of the input variables to the *b* jet energy regression in the high-boost $t\bar{t}$ control region (left to right, top to bottom): raw p_T, p_T, SV mass, SV p_T, SL p_T^{rel}.

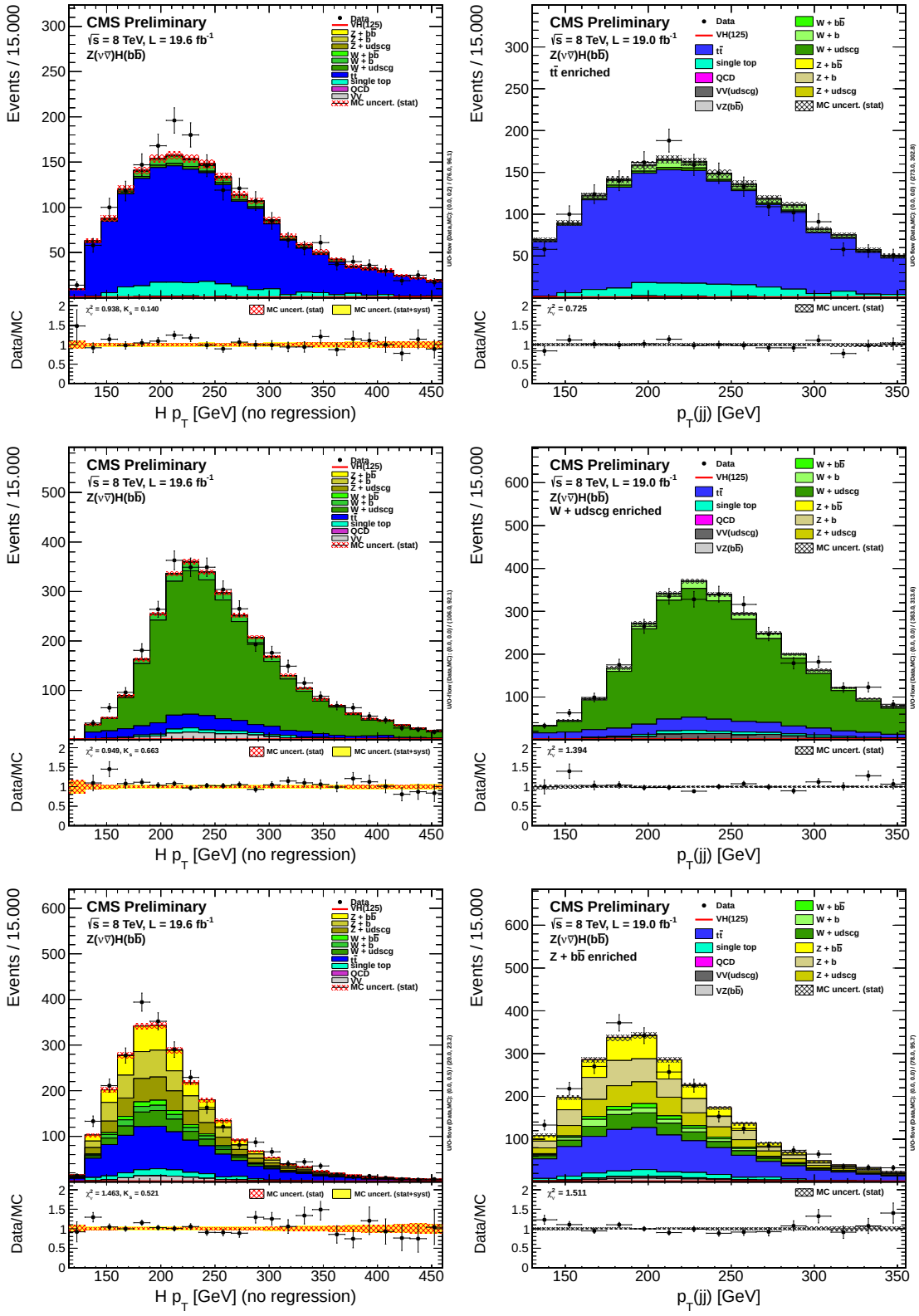

Figure B-2. Distributions of $p_{\mathrm{T}}(\mathrm{jj})$ before (left) and after (right) the regression is applied in the high-boost $t\bar{t}$ (top), $W + \mathrm{LF}$ (middle), in and $Z + \mathrm{HF}$ (bottom) control regions.

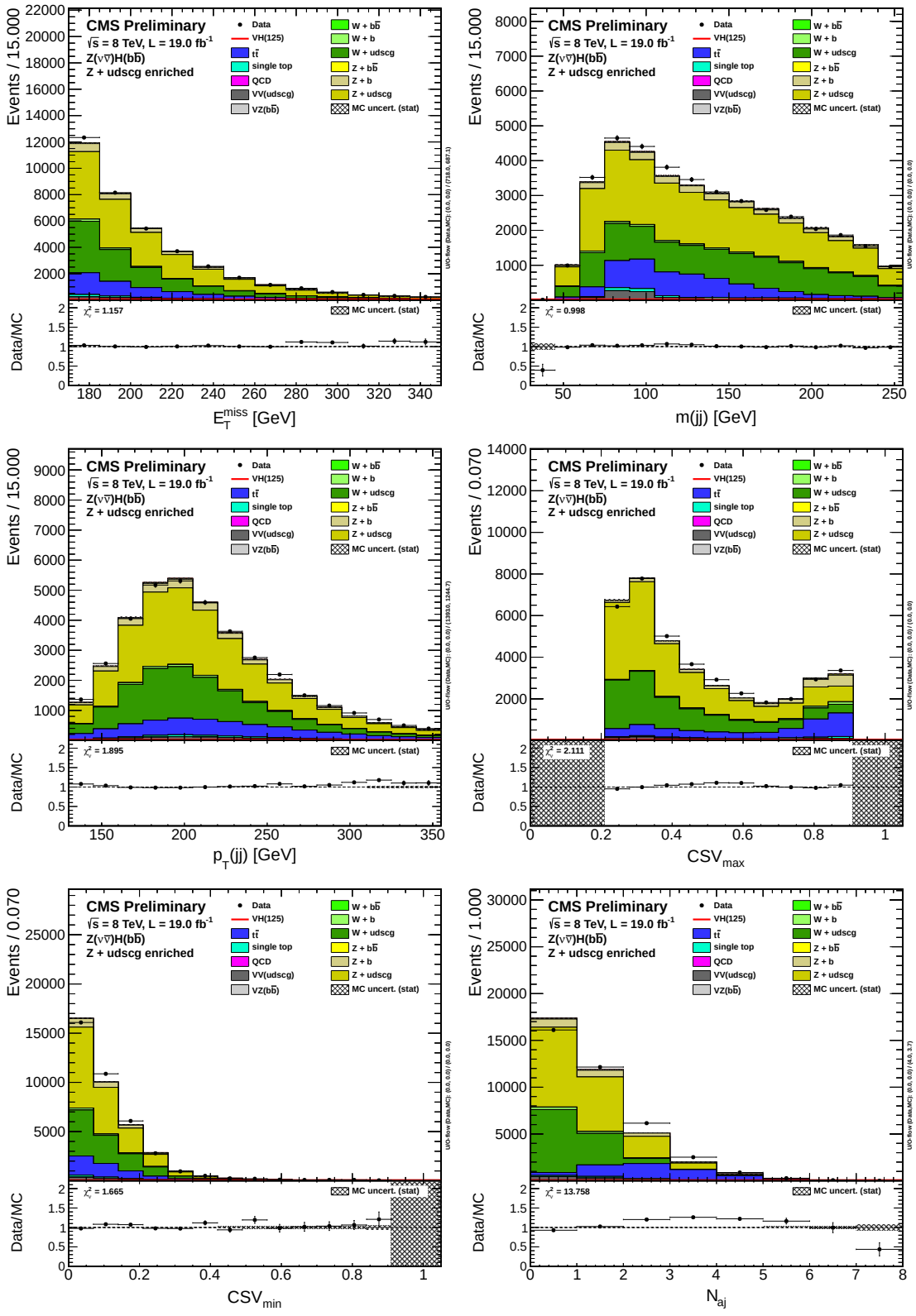

Figure B-3. Distributions of variables in data and simulation in the high-boost $Z + \text{LF}$ control region. From left to right and top to bottom: $E_{\text{T}}^{\text{miss}}$, $m(\text{jj})$, $p_{\text{T}}(\text{jj})$, CSV_{max}, CSV_{min}, and N_{aj}.

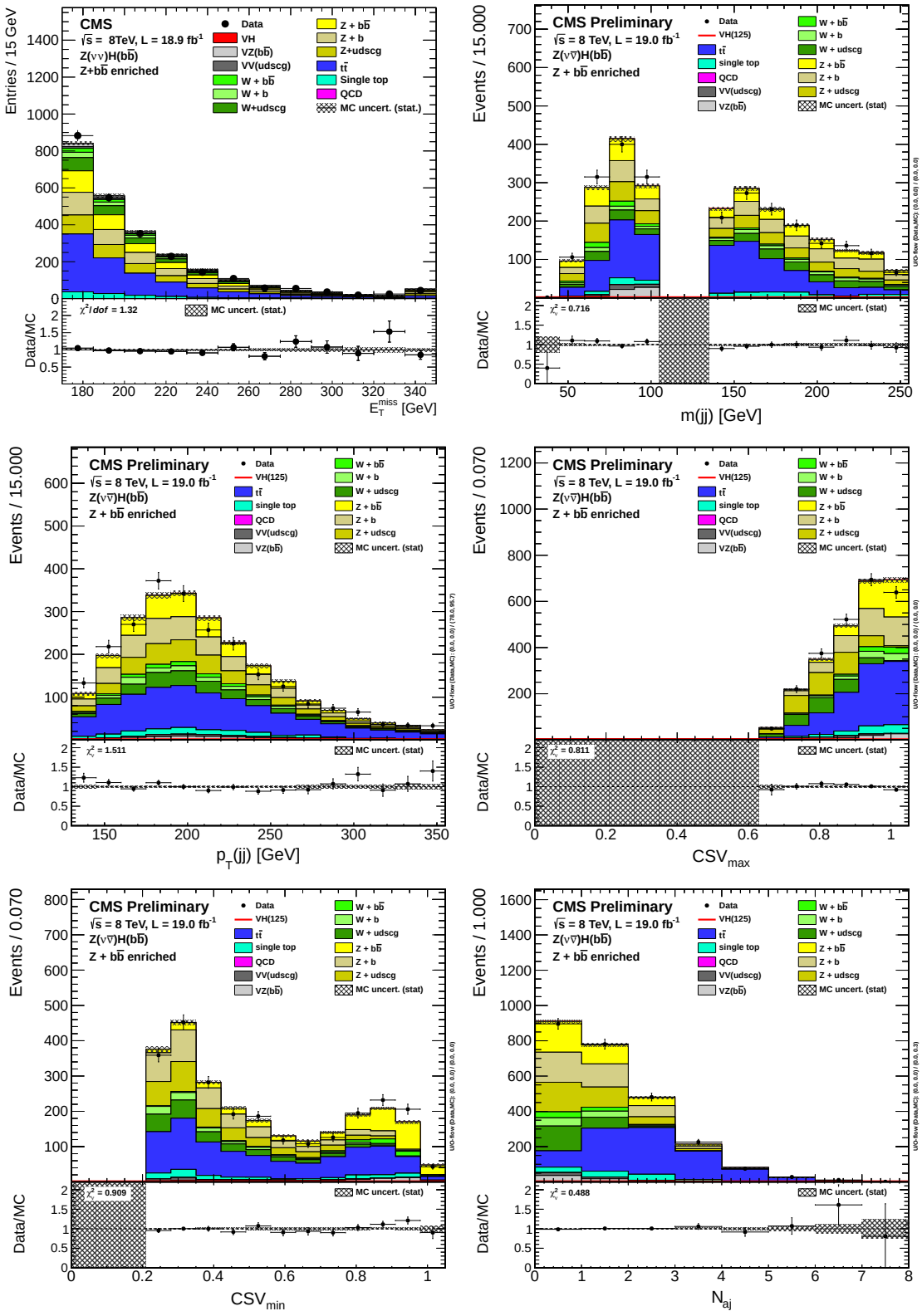

Figure B-4. Distributions of variables in data and simulation in the high-boost $Z + \text{HF}$ control region. From left to right and top to bottom: $E_{\text{T}}^{\text{miss}}$, $m(\text{jj})$, $p_{\text{T}}(\text{jj})$, CSV_{max}, CSV_{min}, and N_{aj}.

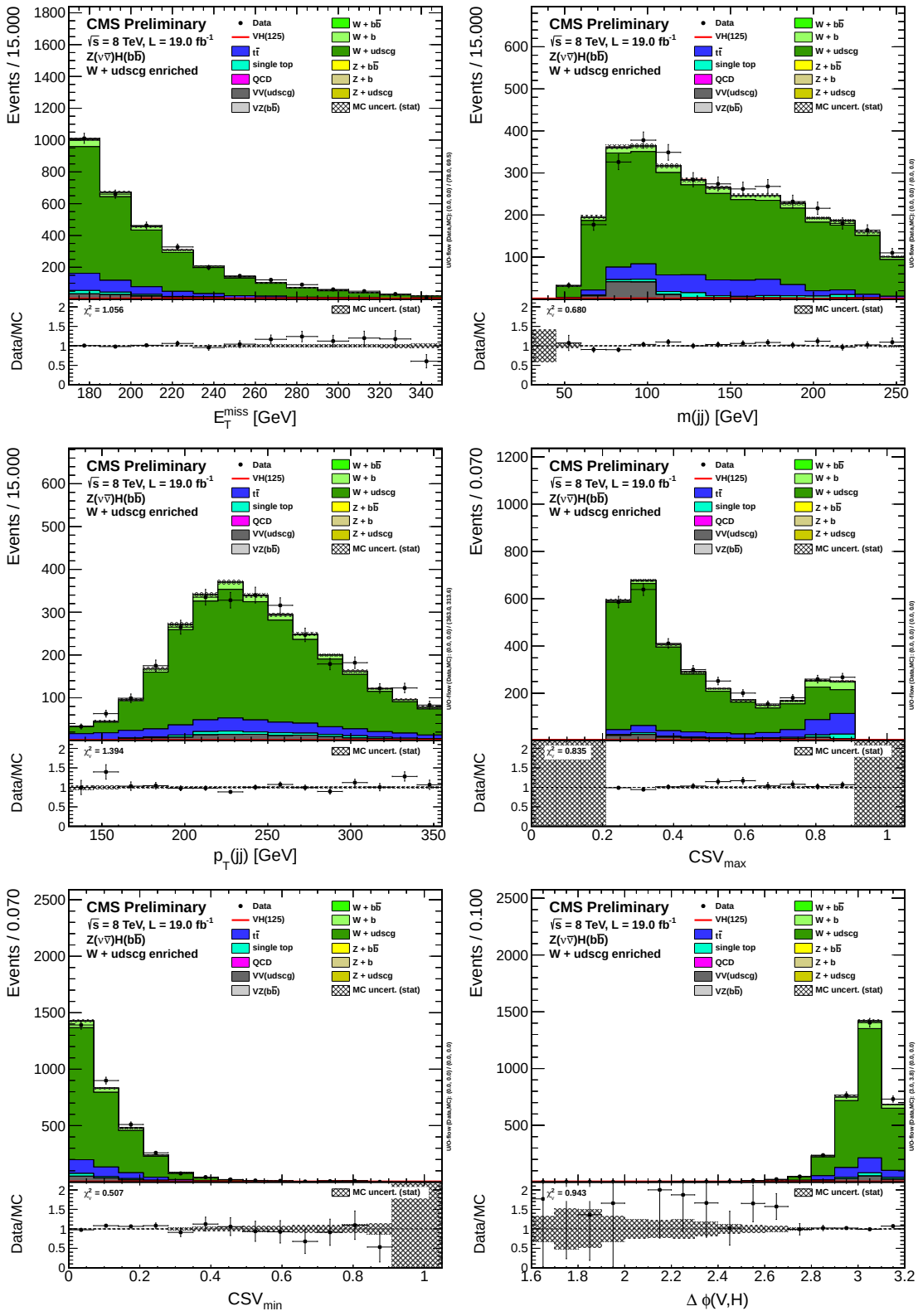

Figure B-5. Distributions of variables in data and simulation in the high-boost $W+$LF control region. From left to right and top to bottom: $E_{\mathrm{T}}^{\mathrm{miss}}$, $m(\mathrm{jj})$, $p_{\mathrm{T}}(\mathrm{jj})$, $\mathrm{CSV}_{\mathrm{max}}$, $\mathrm{CSV}_{\mathrm{min}}$, and $\Delta\phi(V,H)$.

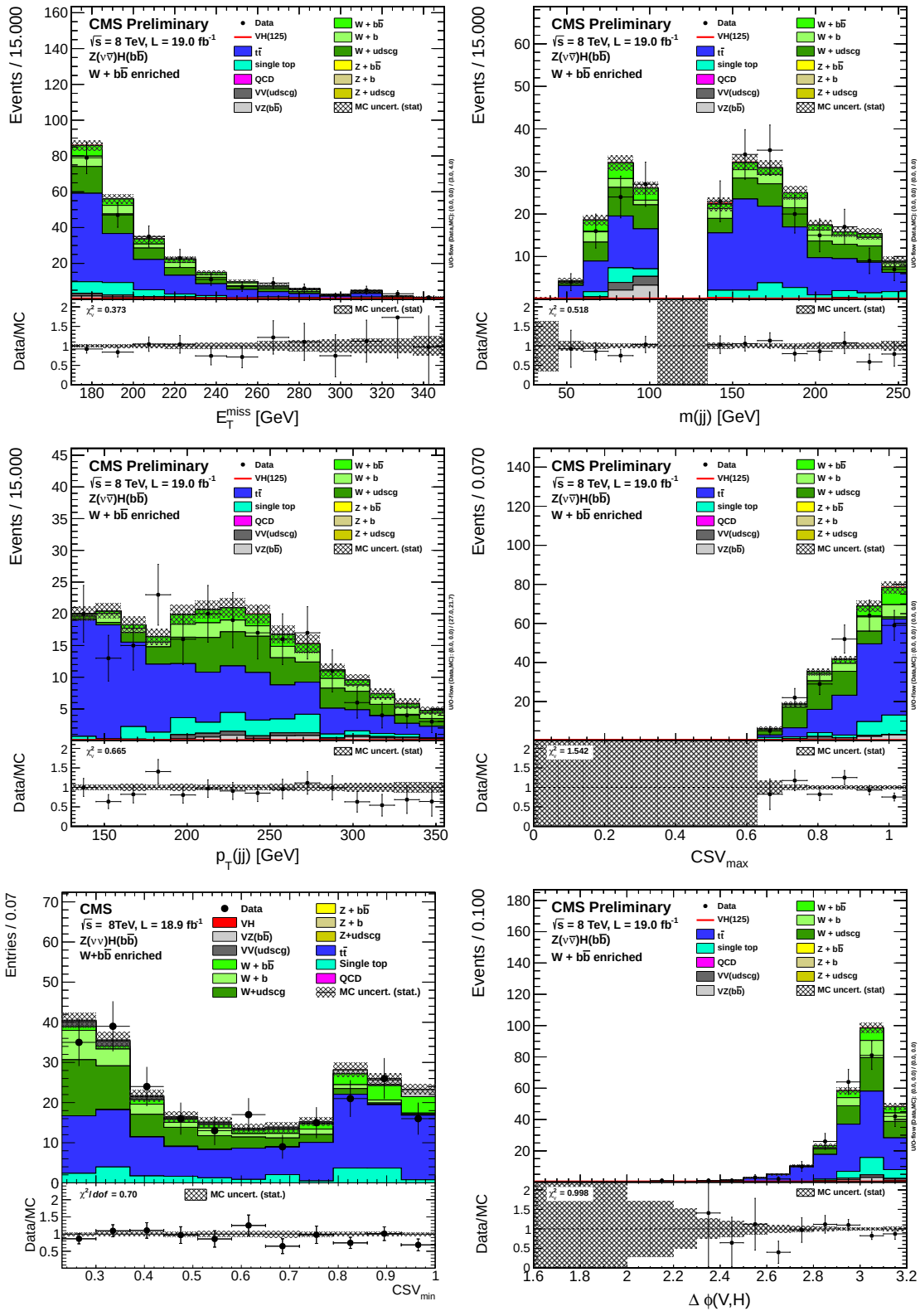

Figure B-6. Distributions of variables in data and simulation in the high-boost $W+$HF control region. From left to right and top to bottom: $E_{\mathrm{T}}^{\mathrm{miss}}$, $m(\mathrm{jj})$, $p_{\mathrm{T}}(\mathrm{jj})$, $\mathrm{CSV}_{\mathrm{max}}$, $\mathrm{CSV}_{\mathrm{min}}$, and $\Delta\phi(V,H)$.

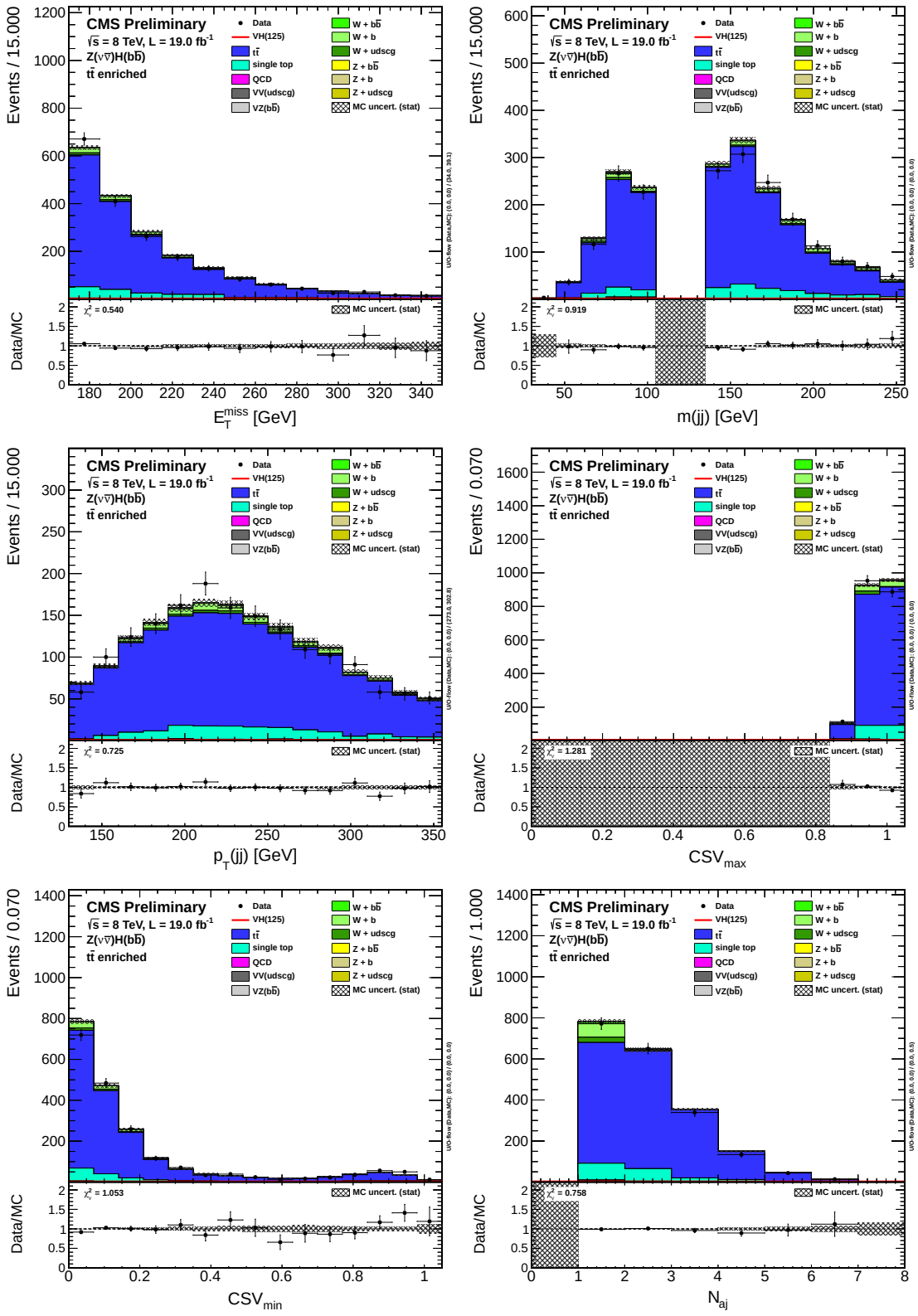

Figure B-7. Distributions of variables in data and simulation in the high-boost $t\bar{t}$ control region. From left to right and top to bottom: $E_{\mathrm{T}}^{\mathrm{miss}}$, $m(\mathrm{jj})$, $p_{\mathrm{T}}(\mathrm{jj})$, $\mathrm{CSV}_{\mathrm{max}}$, $\mathrm{CSV}_{\mathrm{min}}$, and N_{aj}.

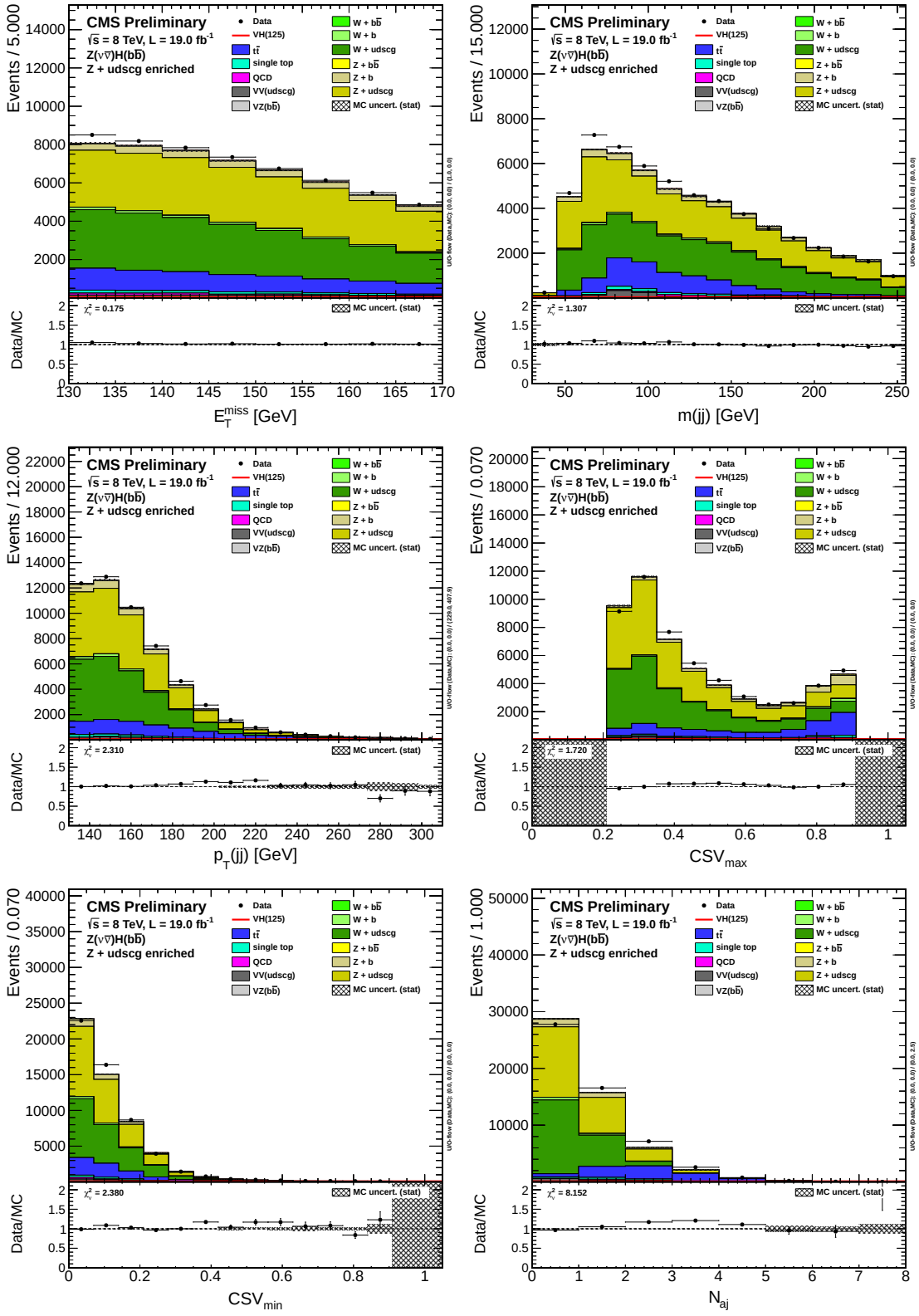

Figure B-8. Distributions of variables in data and simulation in the intermediate-boost $Z + \text{LF}$ control region. From left to right and top to bottom: $E_{\text{T}}^{\text{miss}}$, $m(\text{jj})$, $p_{\text{T}}(\text{jj})$, CSV_{max}, CSV_{min}, and N_{aj}.

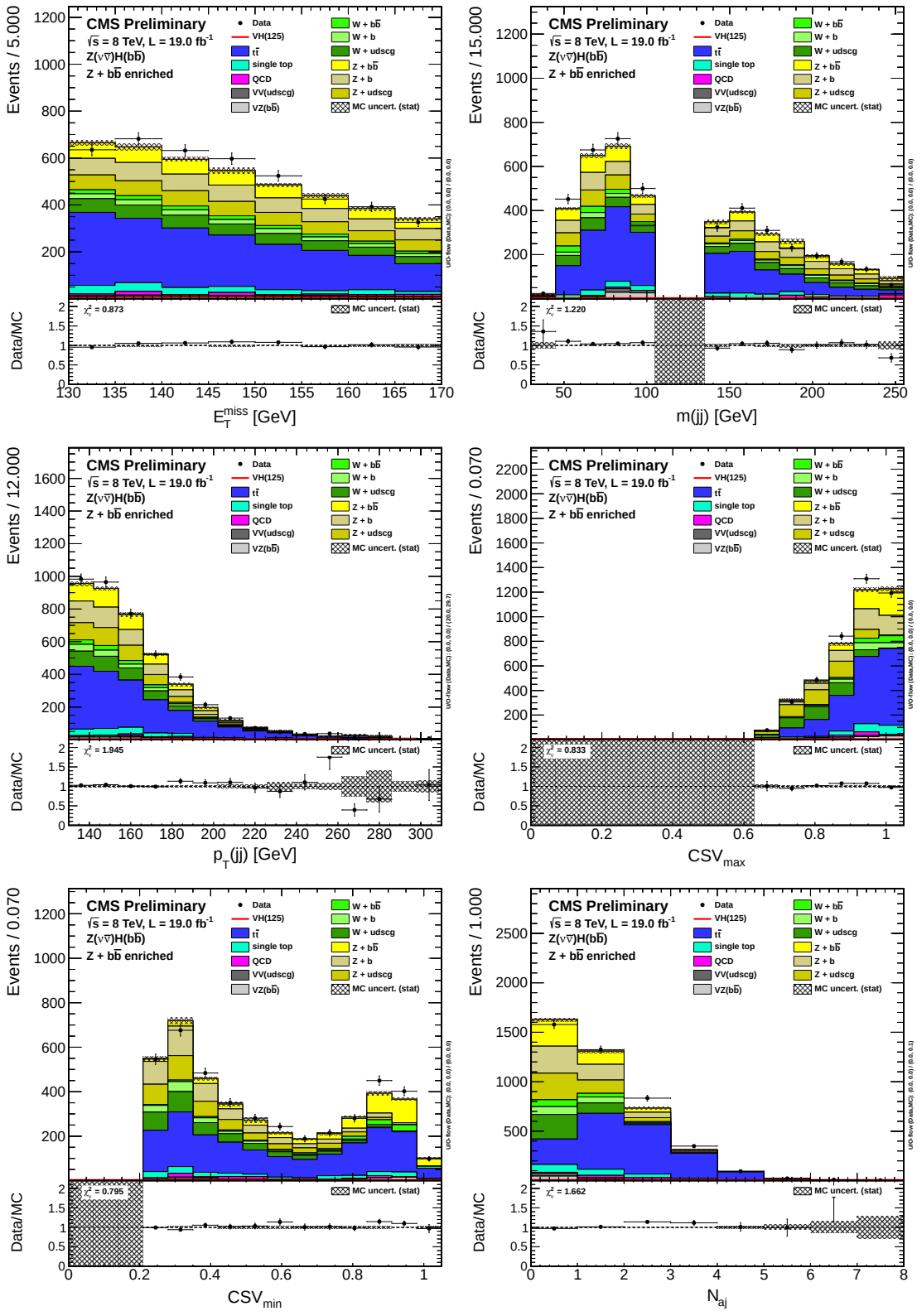

Figure B-9. Distributions of variables in data and simulation in the intermediate-boost $Z + \mathrm{HF}$ control region. From left to right and top to bottom: $E_{\mathrm{T}}^{\mathrm{miss}}$, $m(\mathrm{jj})$, $p_{\mathrm{T}}(\mathrm{jj})$, $\mathrm{CSV}_{\mathrm{max}}$, $\mathrm{CSV}_{\mathrm{min}}$, and N_{aj}.

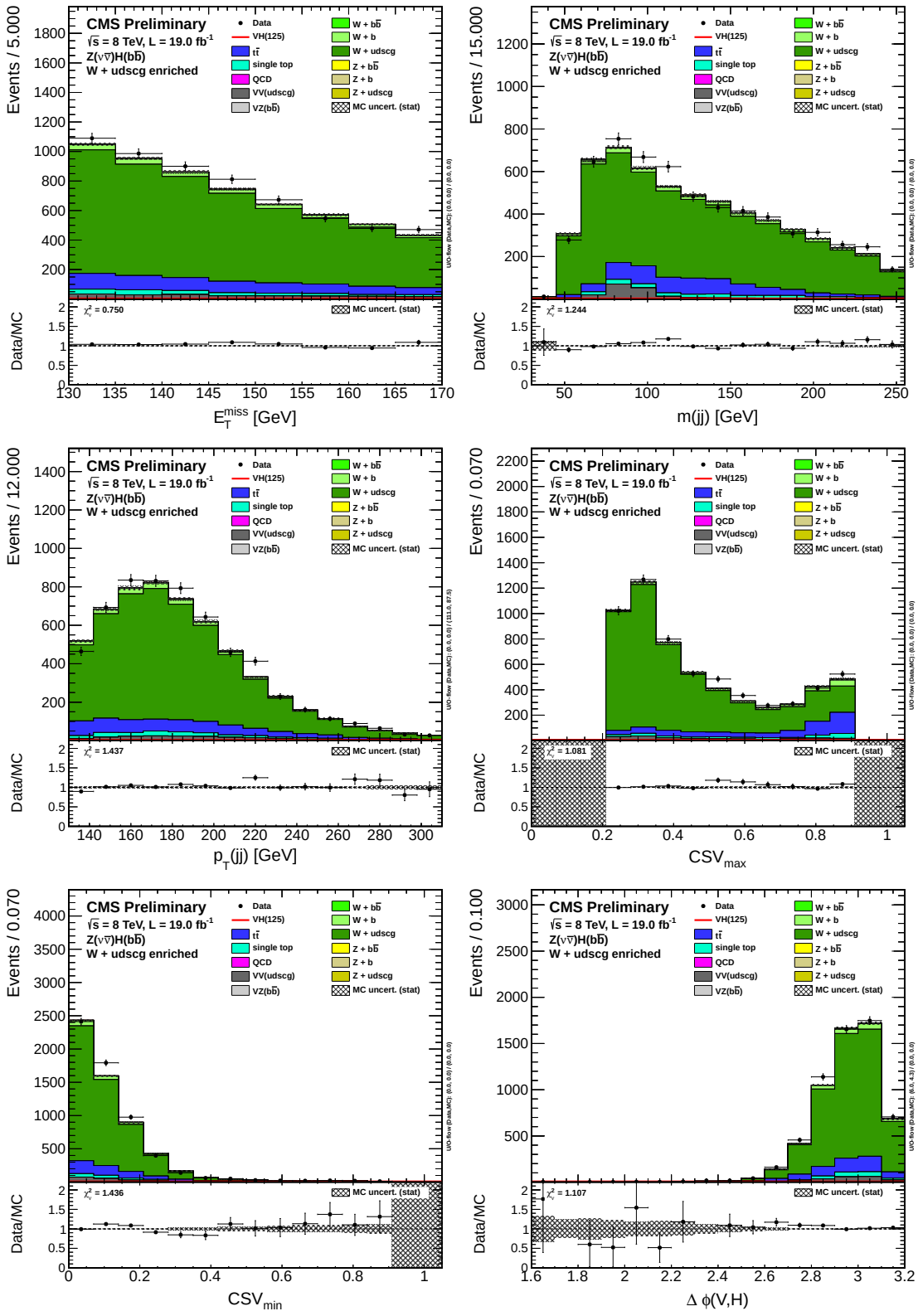

Figure B-10. Distributions of variables in data and simulation in the intermediate-boost $W+\mathrm{LF}$ control region. From left to right and top to bottom: $E_{\mathrm{T}}^{\mathrm{miss}}$, $m(\mathrm{jj})$, $p_{\mathrm{T}}(\mathrm{jj})$, $\mathrm{CSV}_{\mathrm{max}}$, $\mathrm{CSV}_{\mathrm{min}}$, $\Delta\phi(V,H)$.

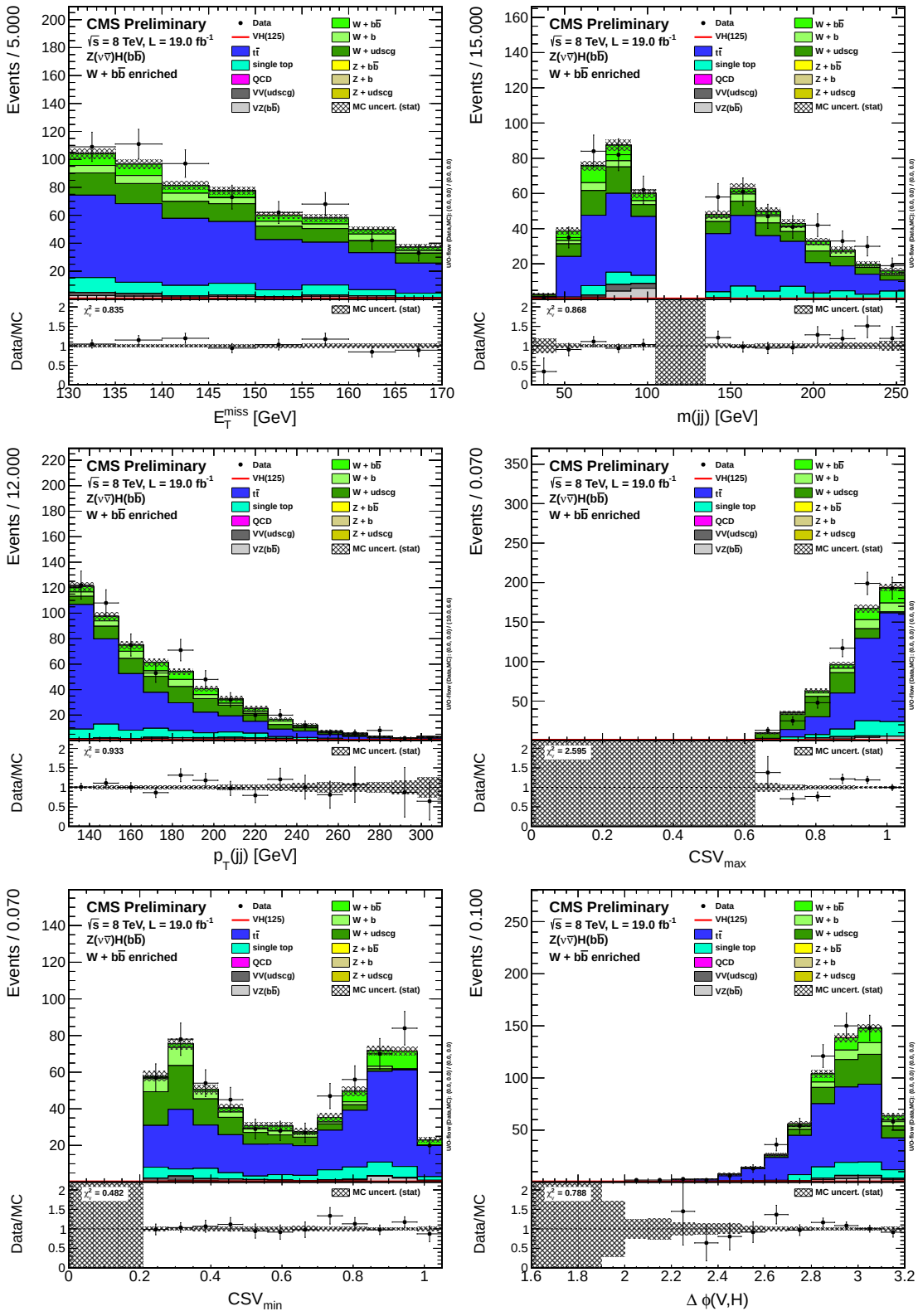

Figure B-11. Distributions of variables in data and simulation in the intermediate-boost $W + \text{HF}$ control region. From left to right and top to bottom: $E_{\text{T}}^{\text{miss}}$, $m(\text{jj})$, $p_{\text{T}}(\text{jj})$, CSV_{max}, CSV_{min}, $\Delta\phi(V, H)$.

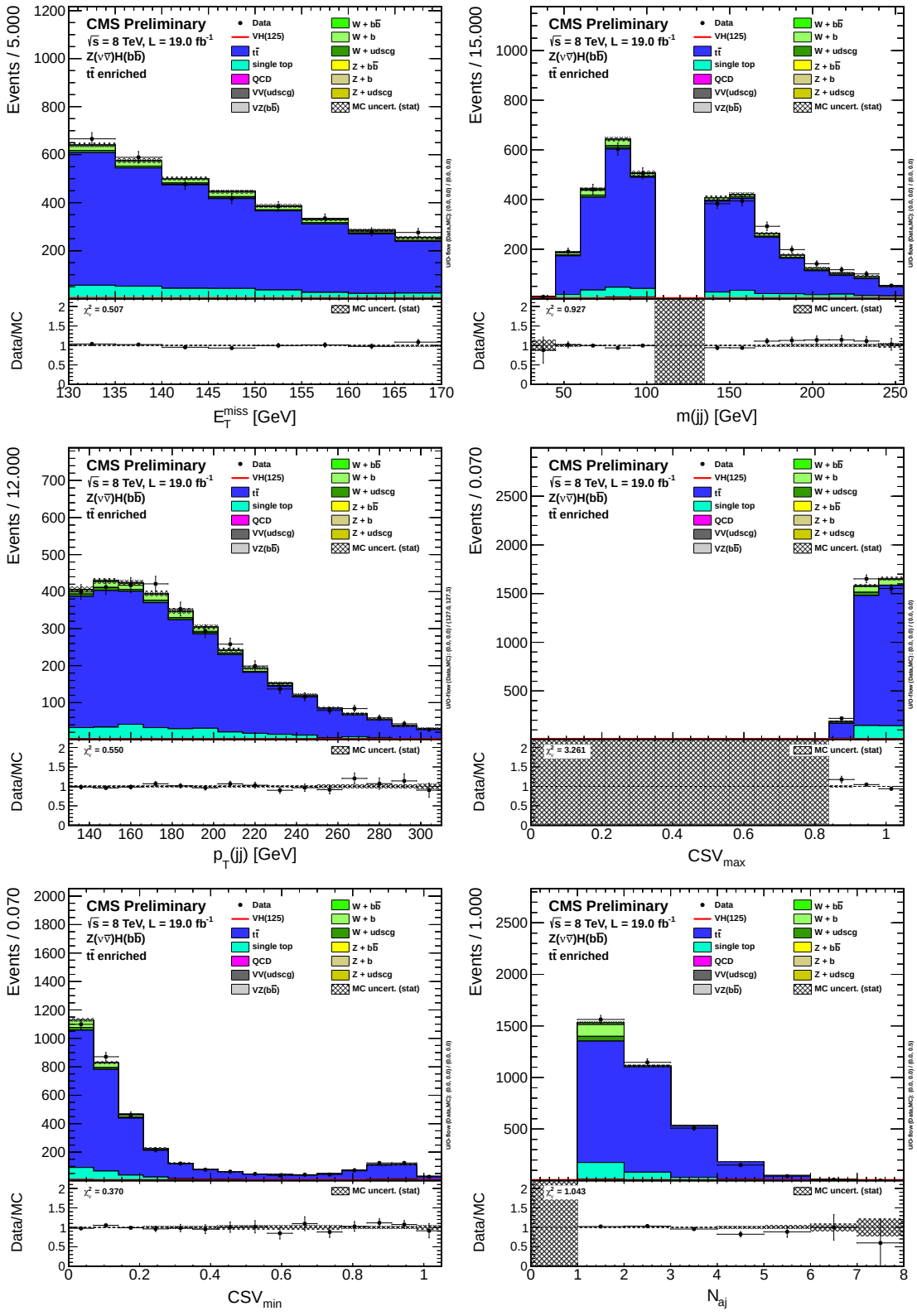

Figure B-12. Distributions of variables in data and simulation in the intermediate-boost $t\bar{t}$ control region. From left to right and top to bottom: $E_{\mathrm{T}}^{\mathrm{miss}}$, $m(\mathrm{jj})$, $p_{\mathrm{T}}(\mathrm{jj})$, $\mathrm{CSV}_{\mathrm{max}}$, $\mathrm{CSV}_{\mathrm{min}}$, and N_{aj}.

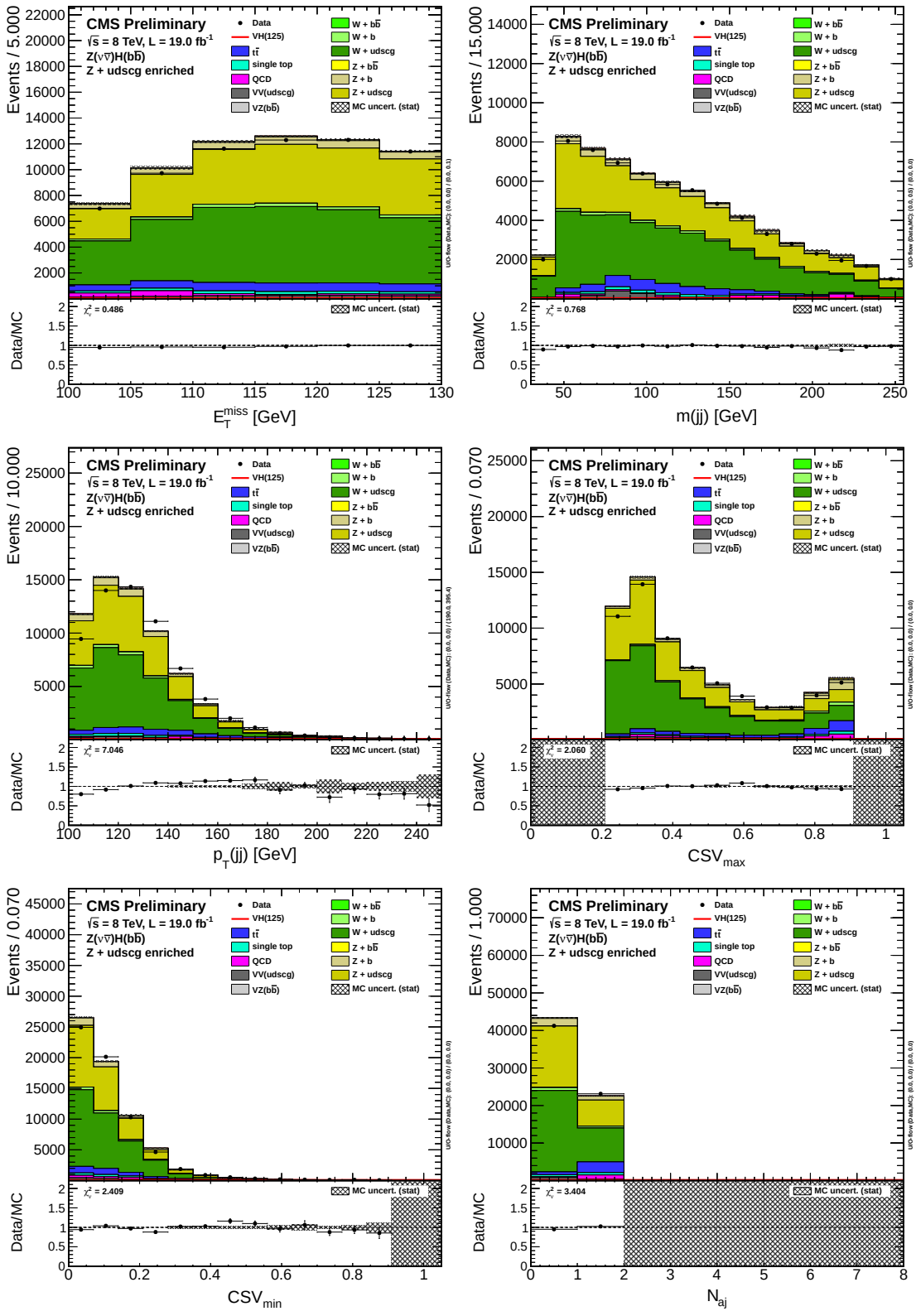

Figure B-13. Distributions of variables in data and simulation in the low-boost $Z + \text{LF}$ control region. From left to right and top to bottom: $E_{\text{T}}^{\text{miss}}$, $m(\text{jj})$, $p_{\text{T}}(\text{jj})$, CSV_{max}, CSV_{min}, and N_{aj}.

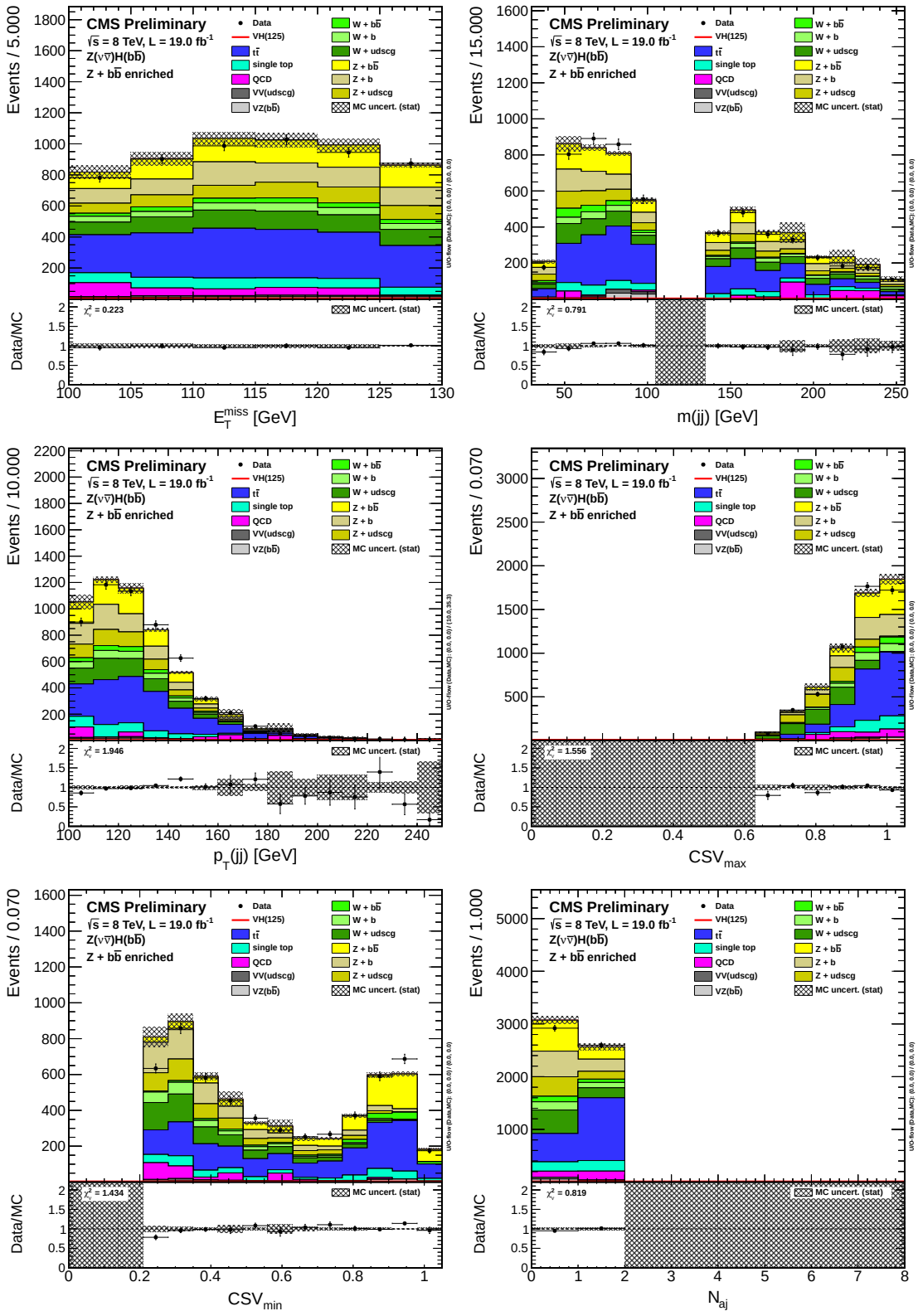

Figure B-14. Distributions of variables in data and simulation in the low-boost $Z + \mathrm{HF}$ control region. From left to right and top to bottom: $E_{\mathrm{T}}^{\mathrm{miss}}$, $m(\mathrm{jj})$, $p_{\mathrm{T}}(\mathrm{jj})$, $\mathrm{CSV}_{\mathrm{max}}$, $\mathrm{CSV}_{\mathrm{min}}$, and N_{aj}.

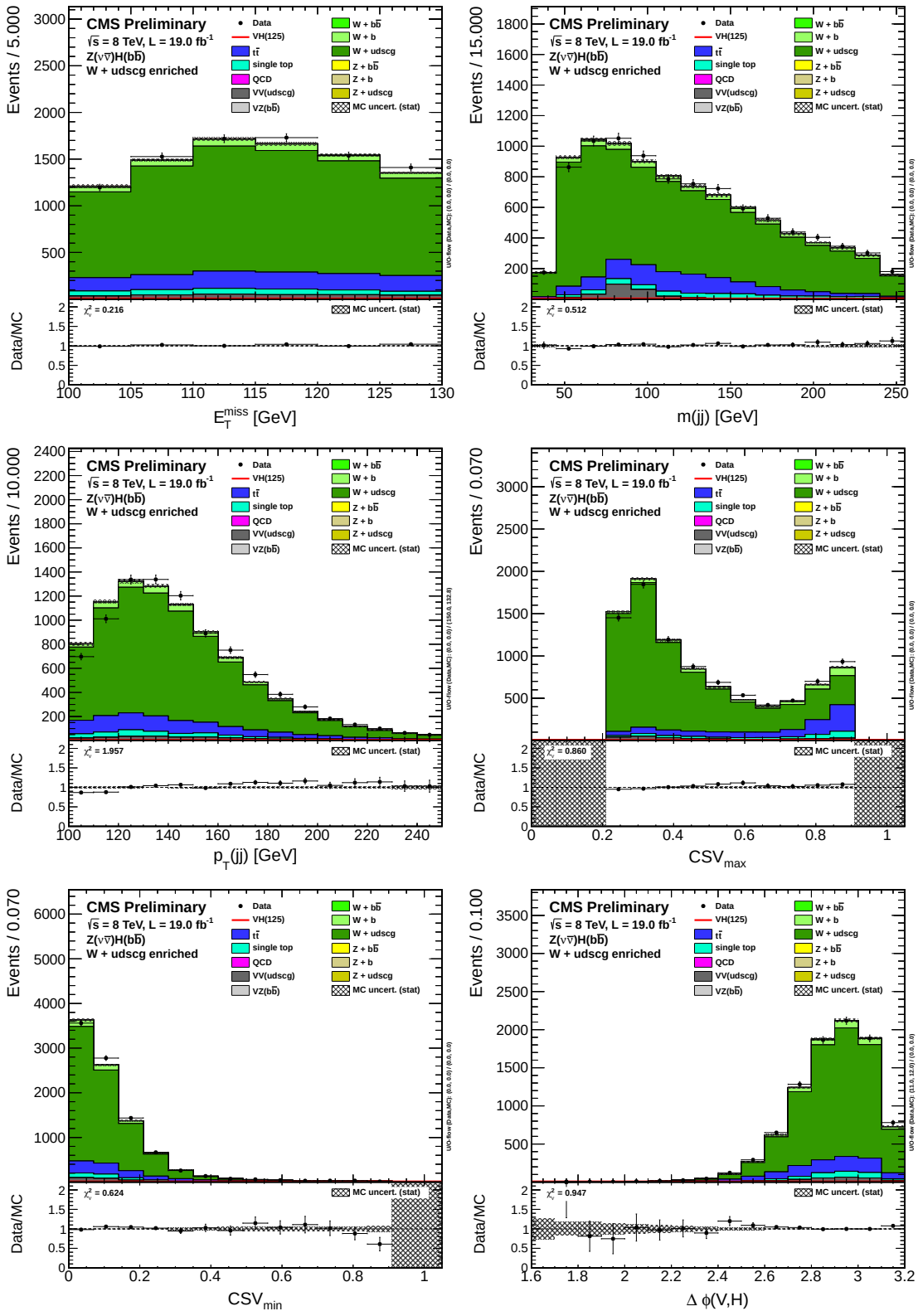

Figure B-15. Distributions of variables in data and simulation in the low-boost W + LF control region. From left to right and top to bottom: $E_{\mathrm{T}}^{\mathrm{miss}}$, $m(\mathrm{jj})$, $p_{\mathrm{T}}(\mathrm{jj})$, $\mathrm{CSV}_{\mathrm{max}}$, $\mathrm{CSV}_{\mathrm{min}}$, $\Delta\phi(V,H)$.

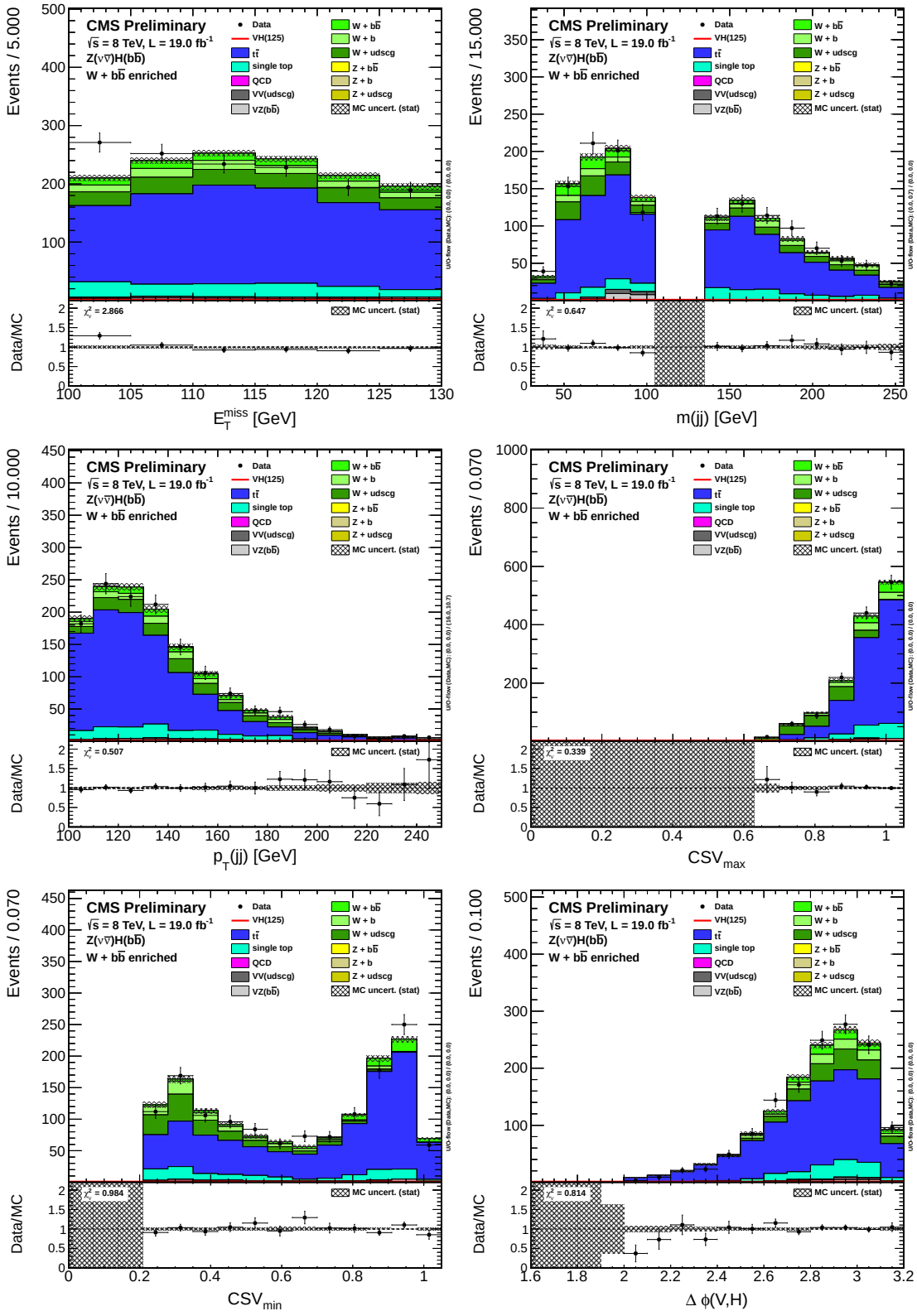

Figure B-16. Distributions of variables in data and simulation in the low-boost $W+\mathrm{HF}$ control region. From left to right and top to bottom: $E_{\mathrm{T}}^{\mathrm{miss}}$, $m(\mathrm{jj})$, $p_{\mathrm{T}}(\mathrm{jj})$, $\mathrm{CSV}_{\mathrm{max}}$, $\mathrm{CSV}_{\mathrm{min}}$, $\Delta\phi(V,H)$.

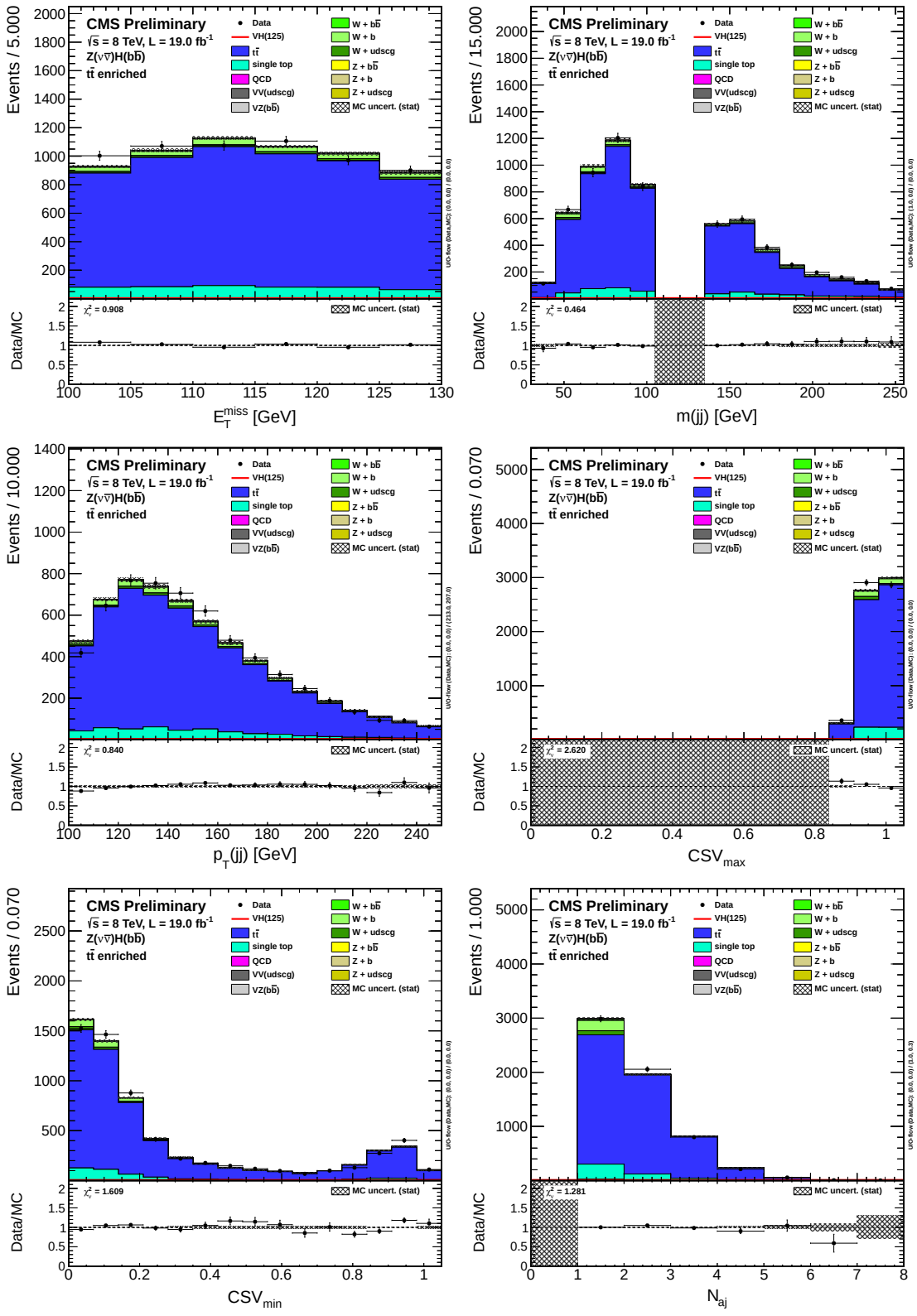

Figure B-17. Distributions of variables in data and simulation in the low-boost $t\bar{t}$ control region. From left to right and top to bottom: $E_{\mathrm{T}}^{\mathrm{miss}}$, $m(\mathrm{jj})$, $p_{\mathrm{T}}(\mathrm{jj})$, $\mathrm{CSV}_{\mathrm{max}}$, $\mathrm{CSV}_{\mathrm{min}}$, and N_{aj}.

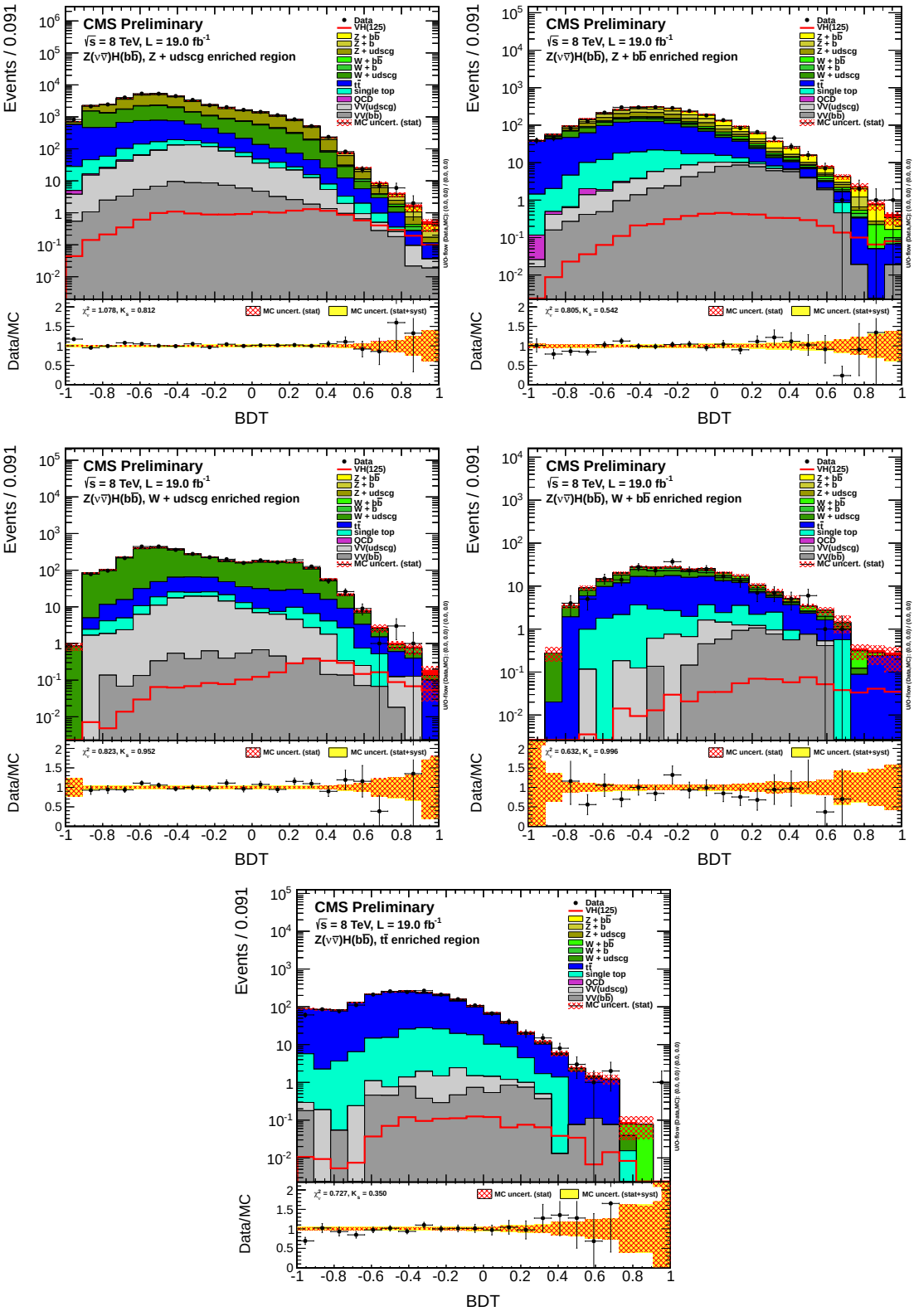

Figure B-18. Distributions of BDT output in data and simulation in the five high-boost control regions. From left to right and top to bottom: $Z+\text{LF}$, $Z+\text{HF}$, $W+\text{LF}$, $W+\text{HF}$, and $t\bar{t}$ control regions.

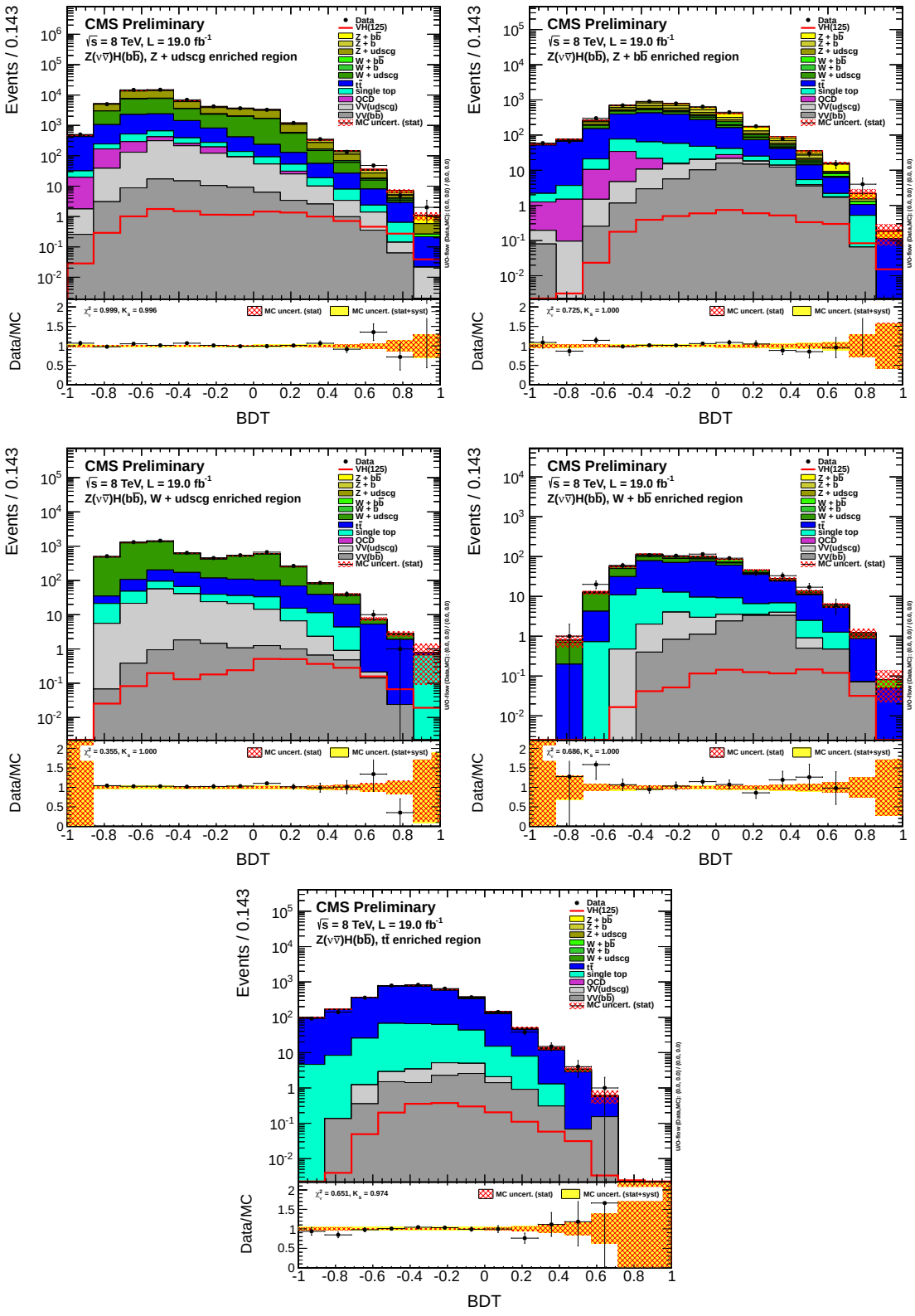

Figure B-19. Distributions of BDT output in data and simulation in the five intermediate-boost control regions. From left to right and top to bottom: $Z+\text{LF}$, $Z+\text{HF}$, $W+\text{LF}$, $W+\text{HF}$, and $t\bar{t}$ control regions.

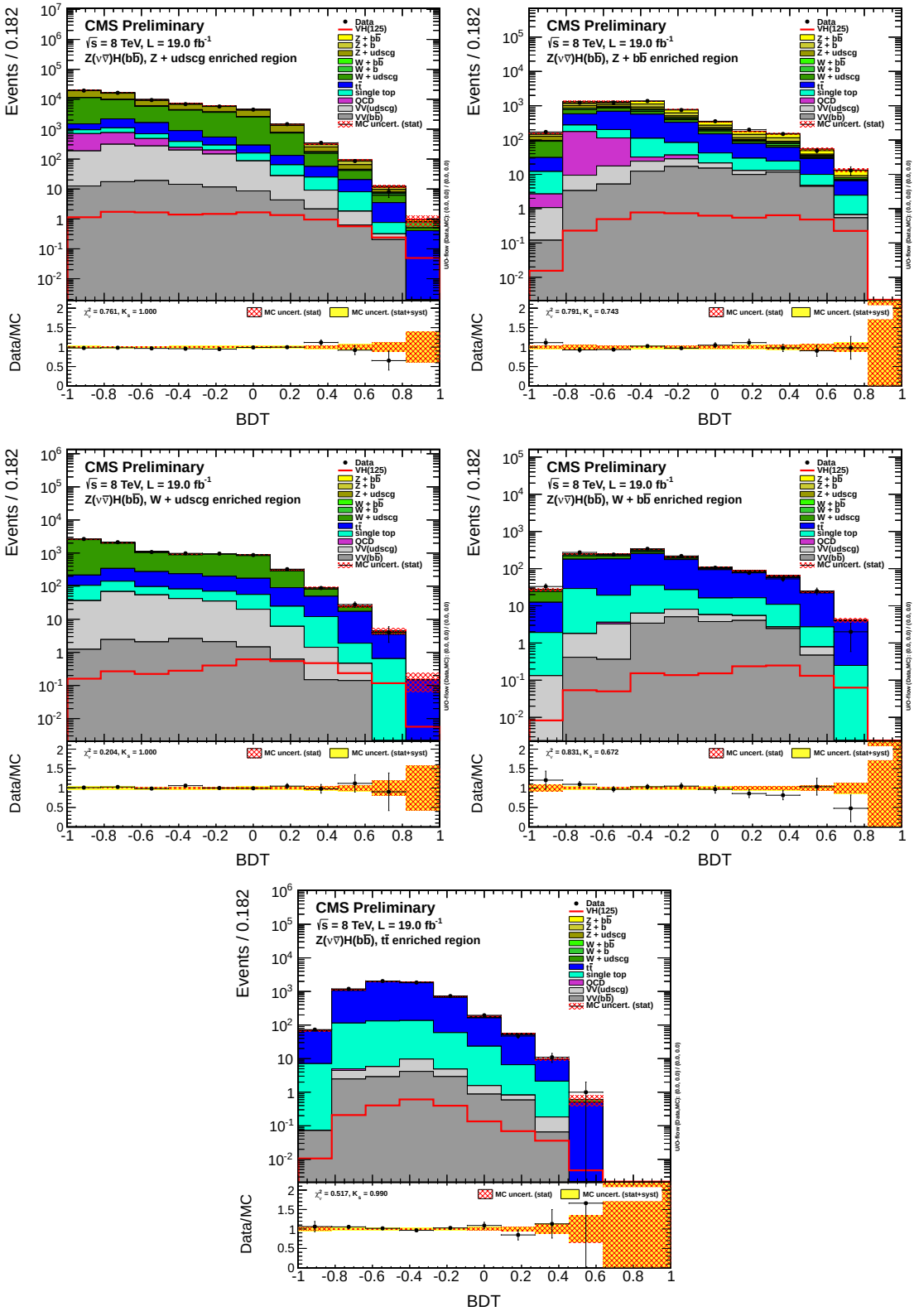

Figure B-20. Distributions of BDT output in data and simulation in the five low-boost control regions. From left to right and top to bottom: $Z+\text{LF}$, $Z+\text{HF}$, $W+\text{LF}$, $W+\text{HF}$, and $t\bar{t}$ control regions.

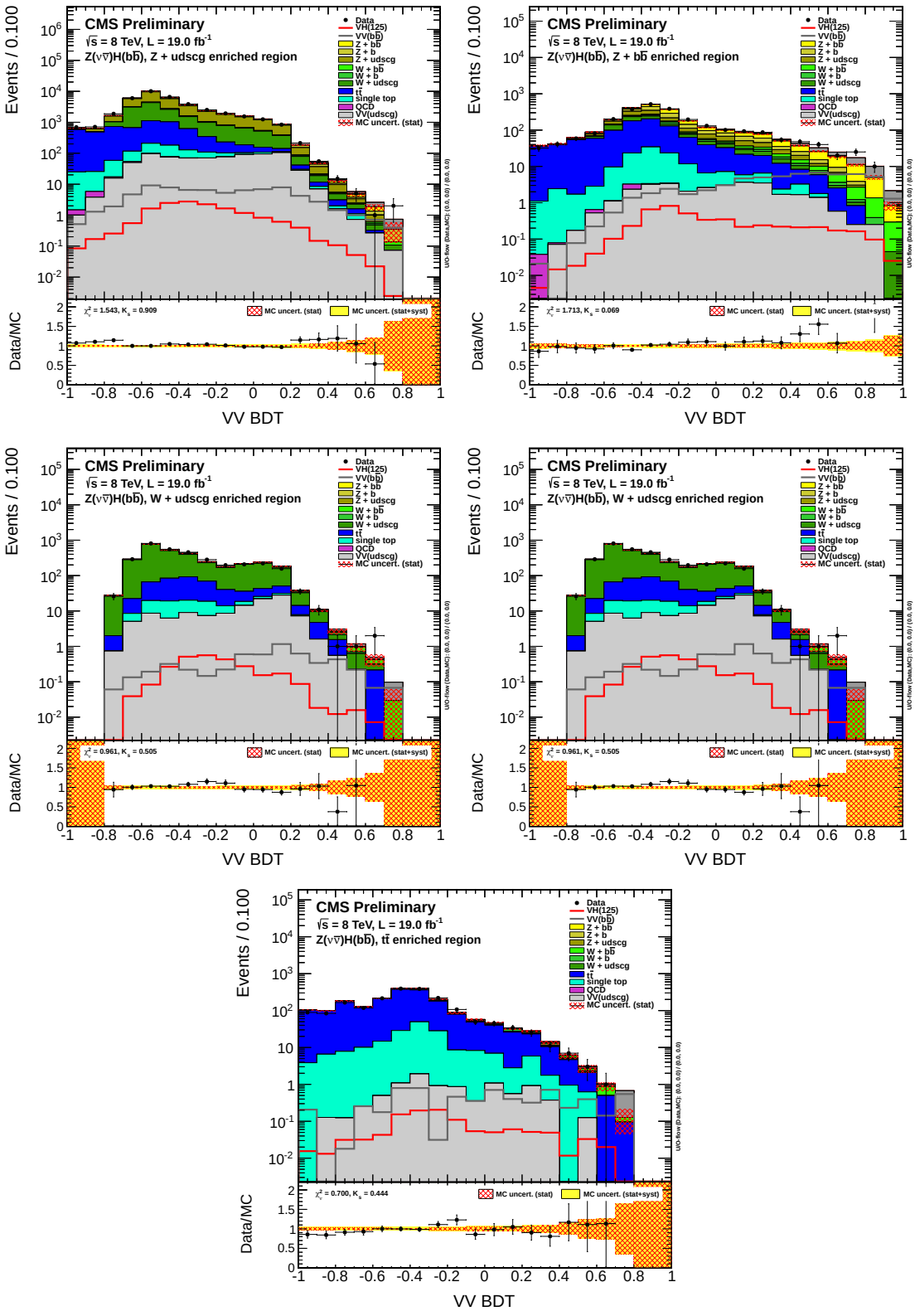

Figure B-21. Distributions of BDT output that is trained using $VZ(b\bar{b})$ as signal in data and simulation in the five high-boost control regions. From left to right and top to bottom: $Z+\text{LF}$, $Z+\text{HF}$, $W+\text{LF}$, $W+\text{HF}$, and $t\bar{t}$ control regions.

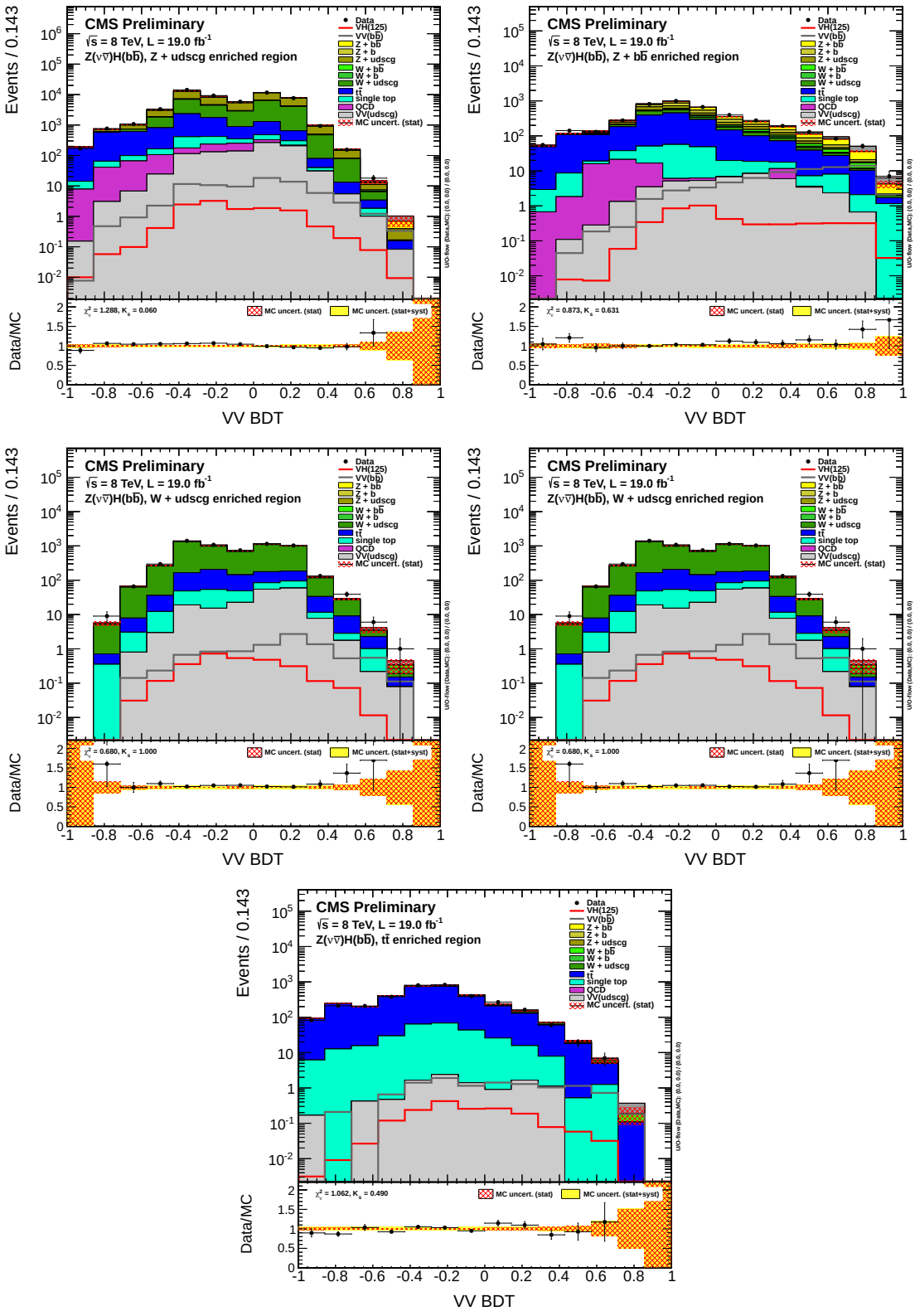

Figure B-22. Distributions of BDT output that is trained using $VZ(b\bar{b})$ as signal in data and simulation in the five intermediate-boost control regions. From left to right and top to bottom: $Z+\text{LF}$, $Z+\text{HF}$, $W+\text{LF}$, $W+\text{HF}$, and $t\bar{t}$ control regions.

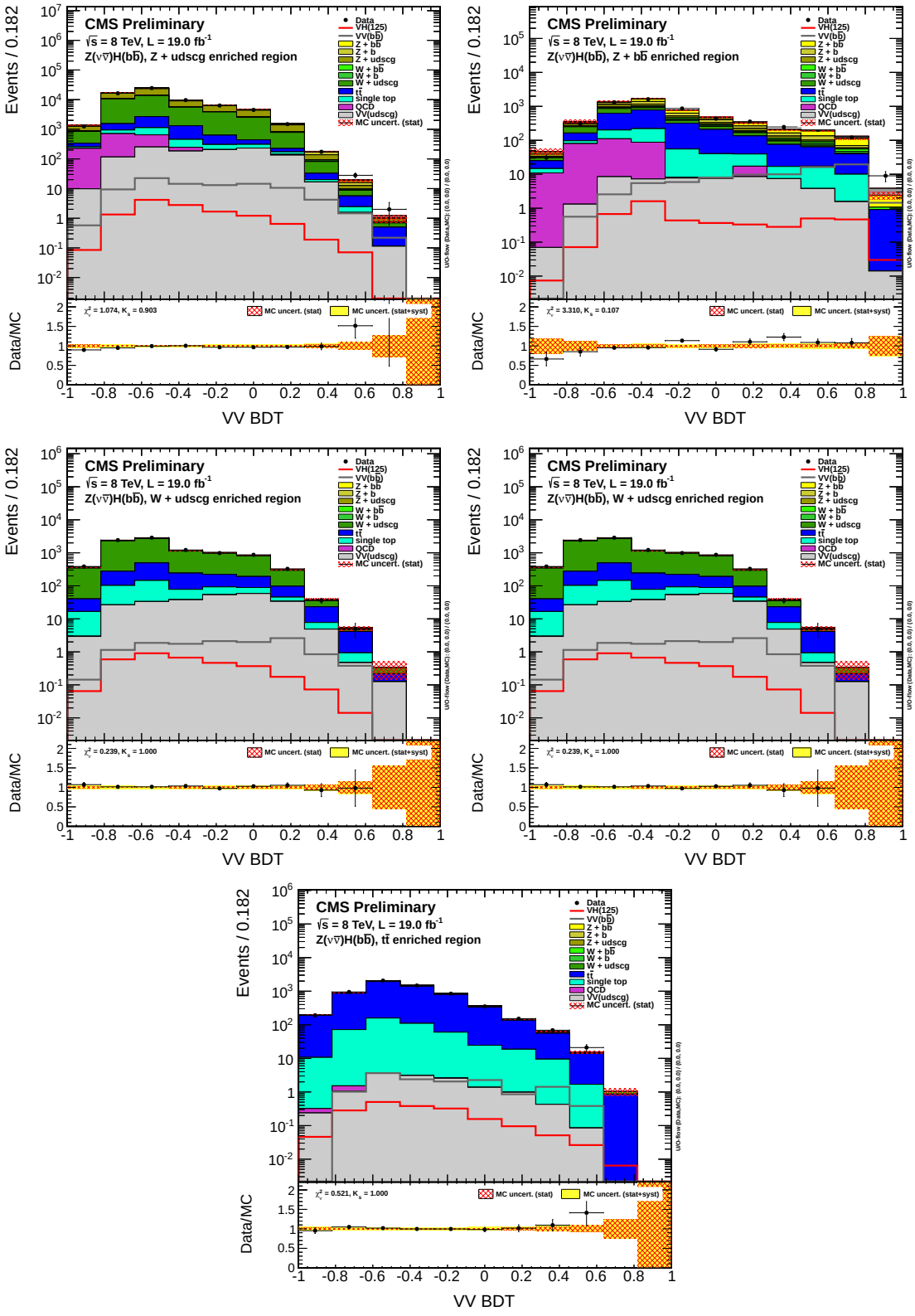

Figure B-23. Distributions of BDT output that is trained using $VZ(b\bar{b})$ as signal in data and simulation in the five low-boost control regions. From left to right and top to bottom: $Z+\text{LF}$, $Z+\text{HF}$, $W+\text{LF}$, $W+\text{HF}$, and $t\bar{t}$ control regions.

APPENDIX C
POST-FIT BDT DISTRIBUTIONS

Figures C-1–C-5 show all the 14 post-fit BDT distributions, for the m_H = 125 GeV training, for all channels and in the 8 TeV analysis. In order to better display the different shapes of the signal and background BDT distributions, Fig. C-6 shows these distributions for the highest-boost region in each channel, normalized to unity.

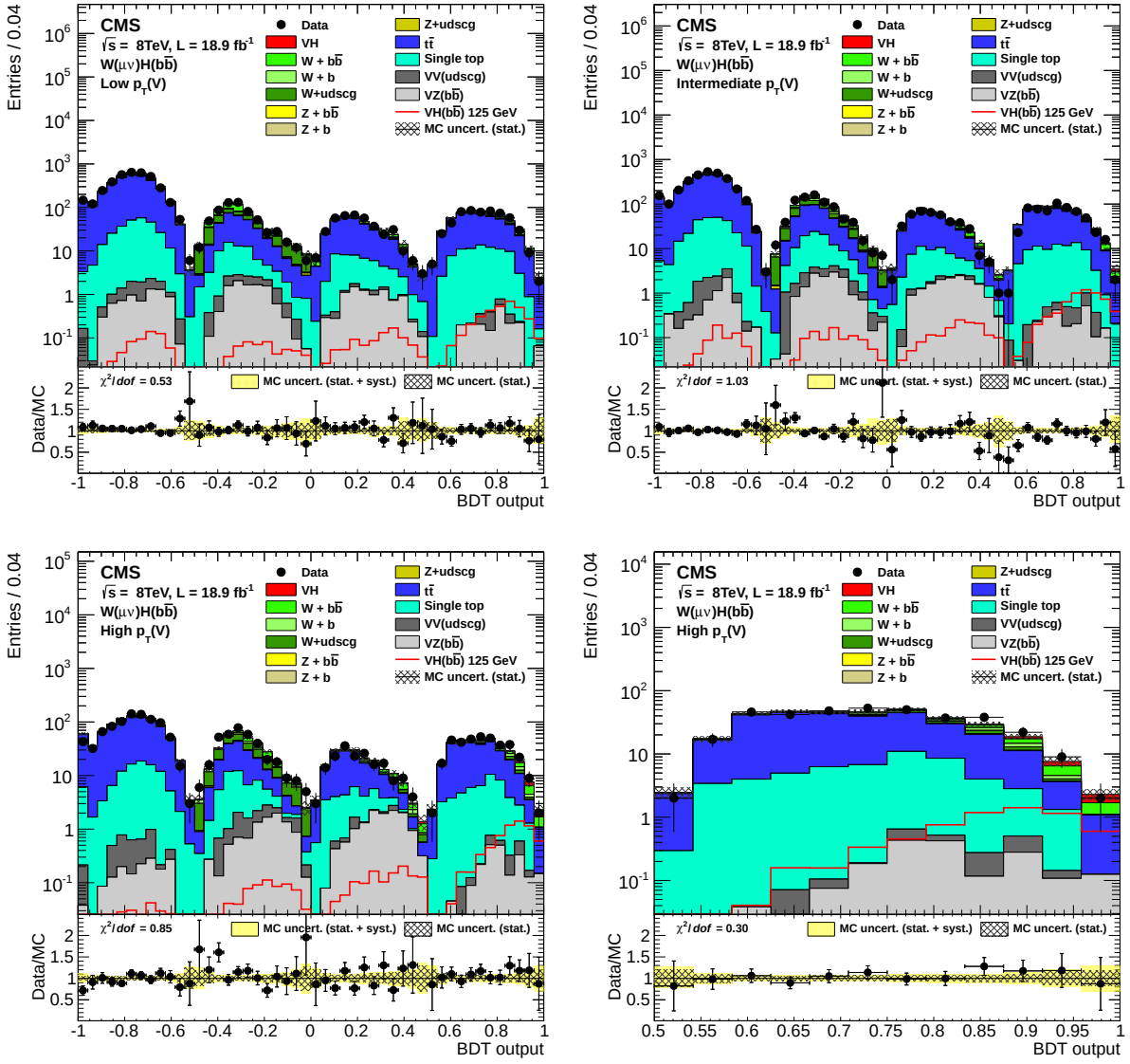

Figure C-1. Post-fit BDT output distributions for $W(\mu\nu)H(b\bar{b})$ in the low-boost region (top left), the intermediate-boost (top right), and the high-boost (bottom left). Bottom right: *VH*-enriched partition of the high-boost region is shown in more detail.

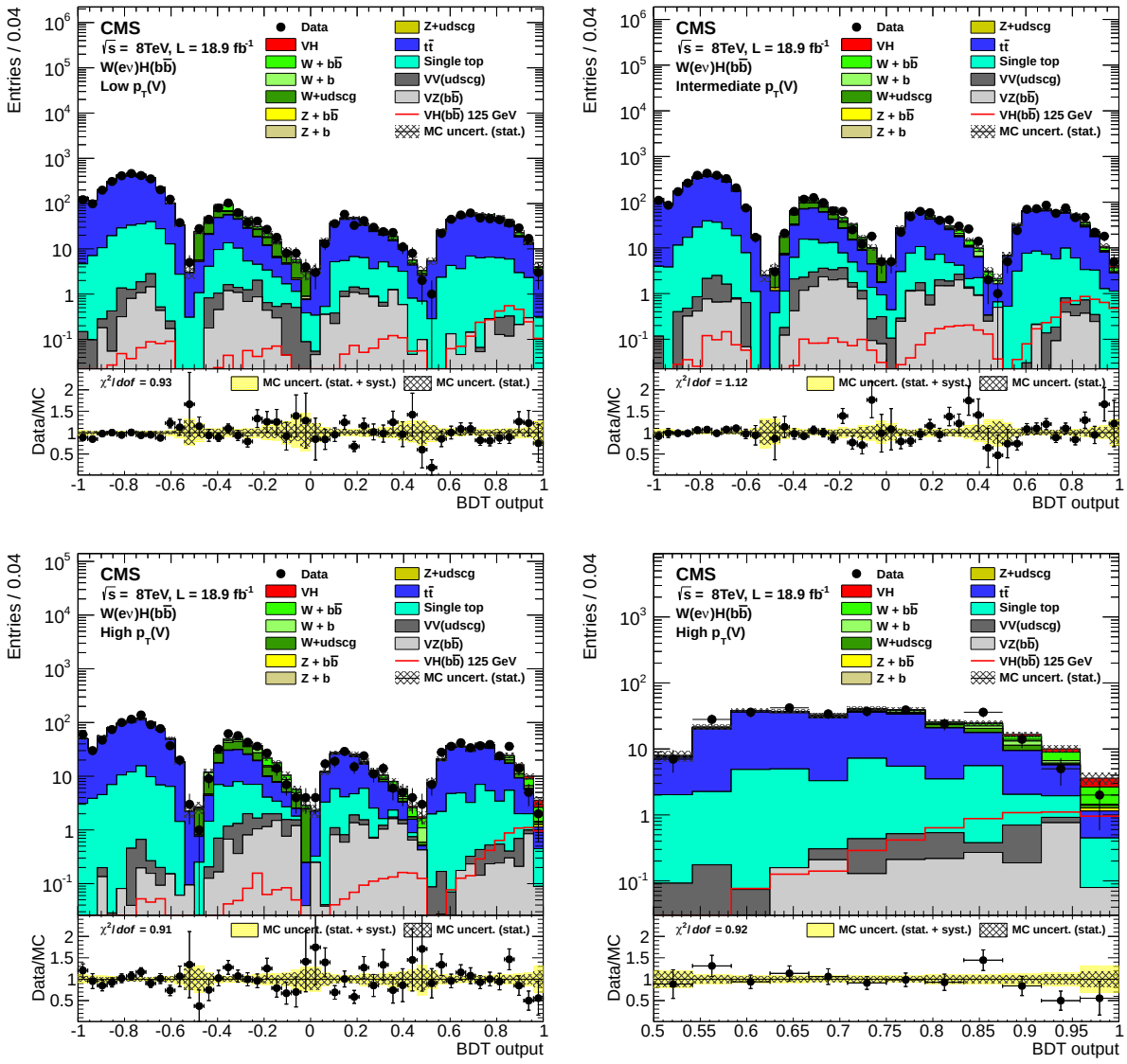

Figure C-2. Post-fit BDT output distributions for $W(e\nu)H(b\bar{b})$ in the low-boost region (top left), the intermediate-boost (top right), and the high-boost (bottom left). Bottom right: *VH*-enriched partition of the high-boost region is shown in more detail.

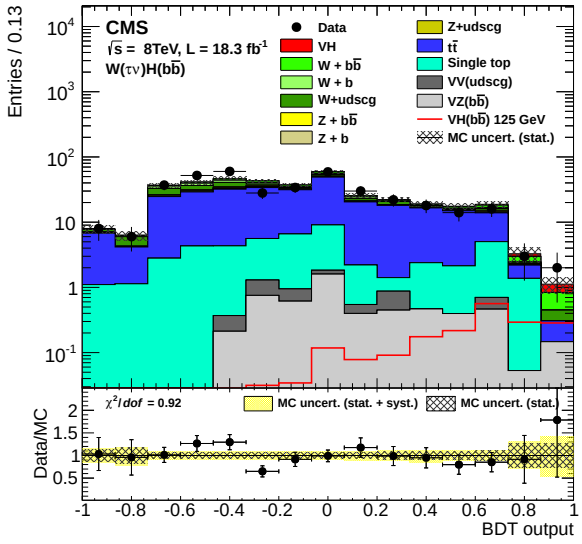

Figure C-3. Post-fit BDT output distributions for $W(\tau\nu)H(b\bar{b})$.

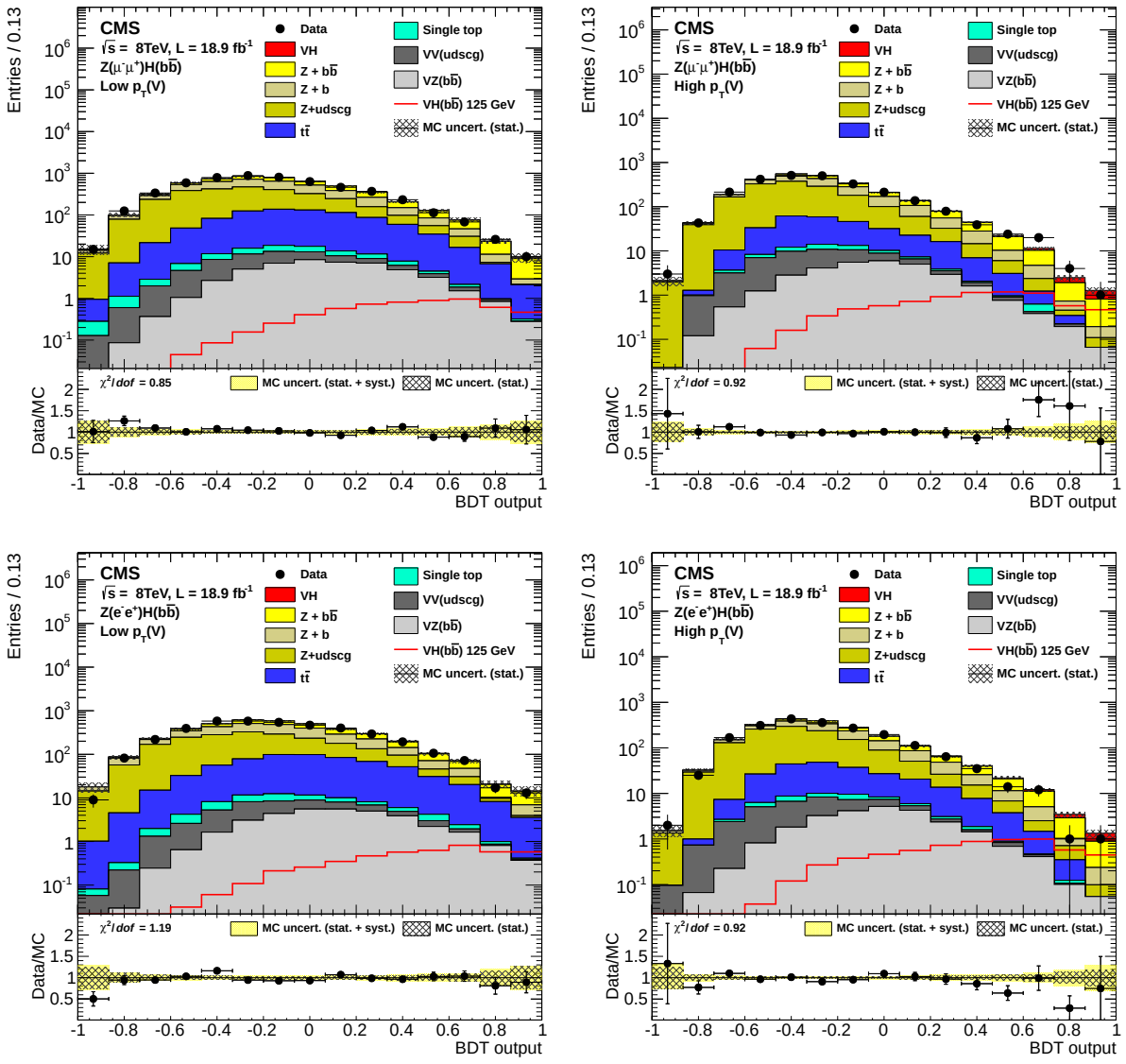

Figure C-4. Post-fit BDT output distributions for $Z(\mu\mu)H(b\bar{b})$ in the low-boost region (top left) and the high-boost (top right), and for $Z(ee)H(b\bar{b})$ in the low-boost region (bottom left) and the high-boost (bottom right)

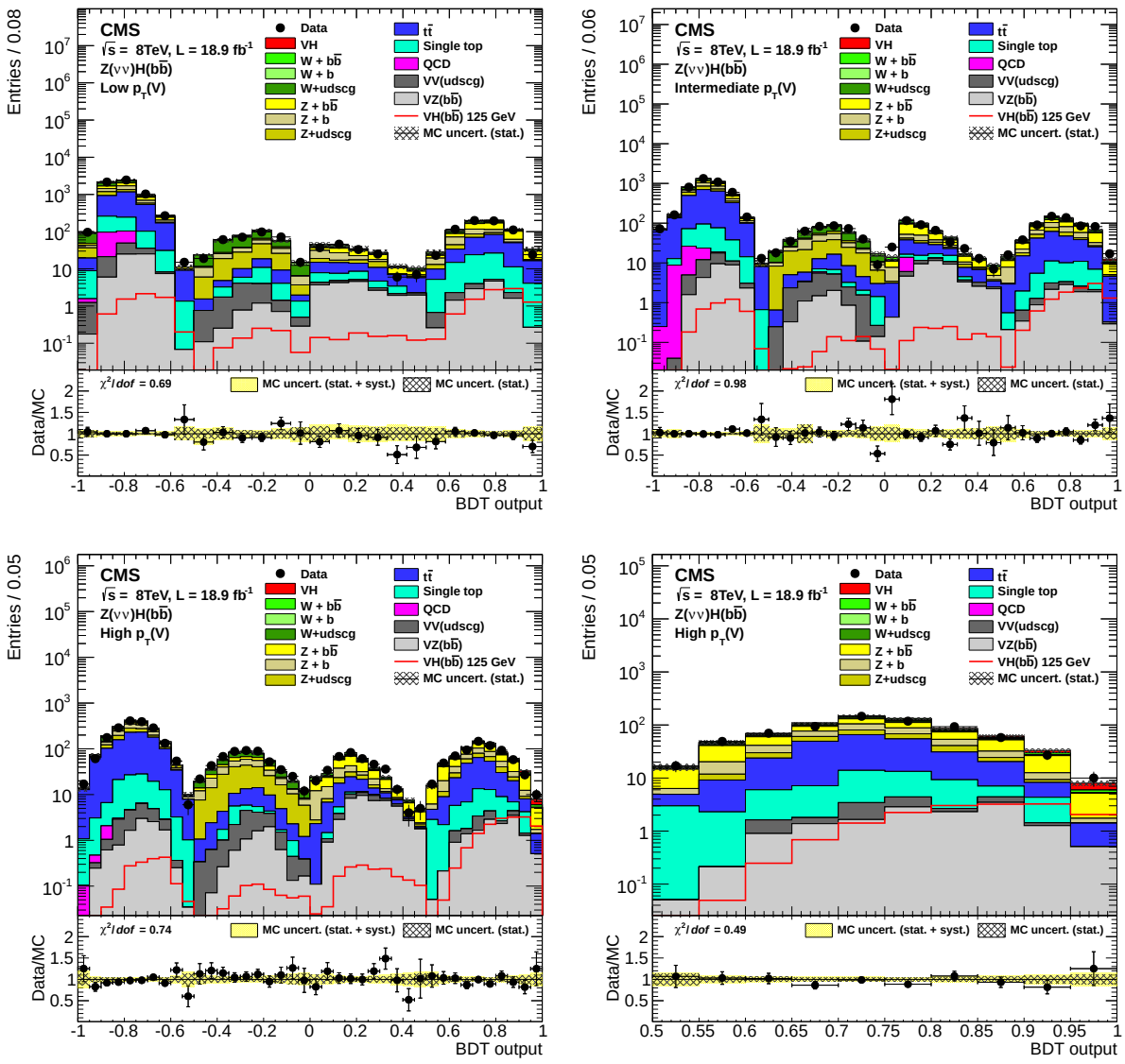

Figure C-5. Post-fit BDT output distributions for $Z(\nu\overline{\nu})H(b\overline{b})$ in the low-boost region (top left), the intermediate-boost (top right), and the high-boost (bottom left). Bottom right: *VH*-enriched partition of the high-boost region is shown in more detail.

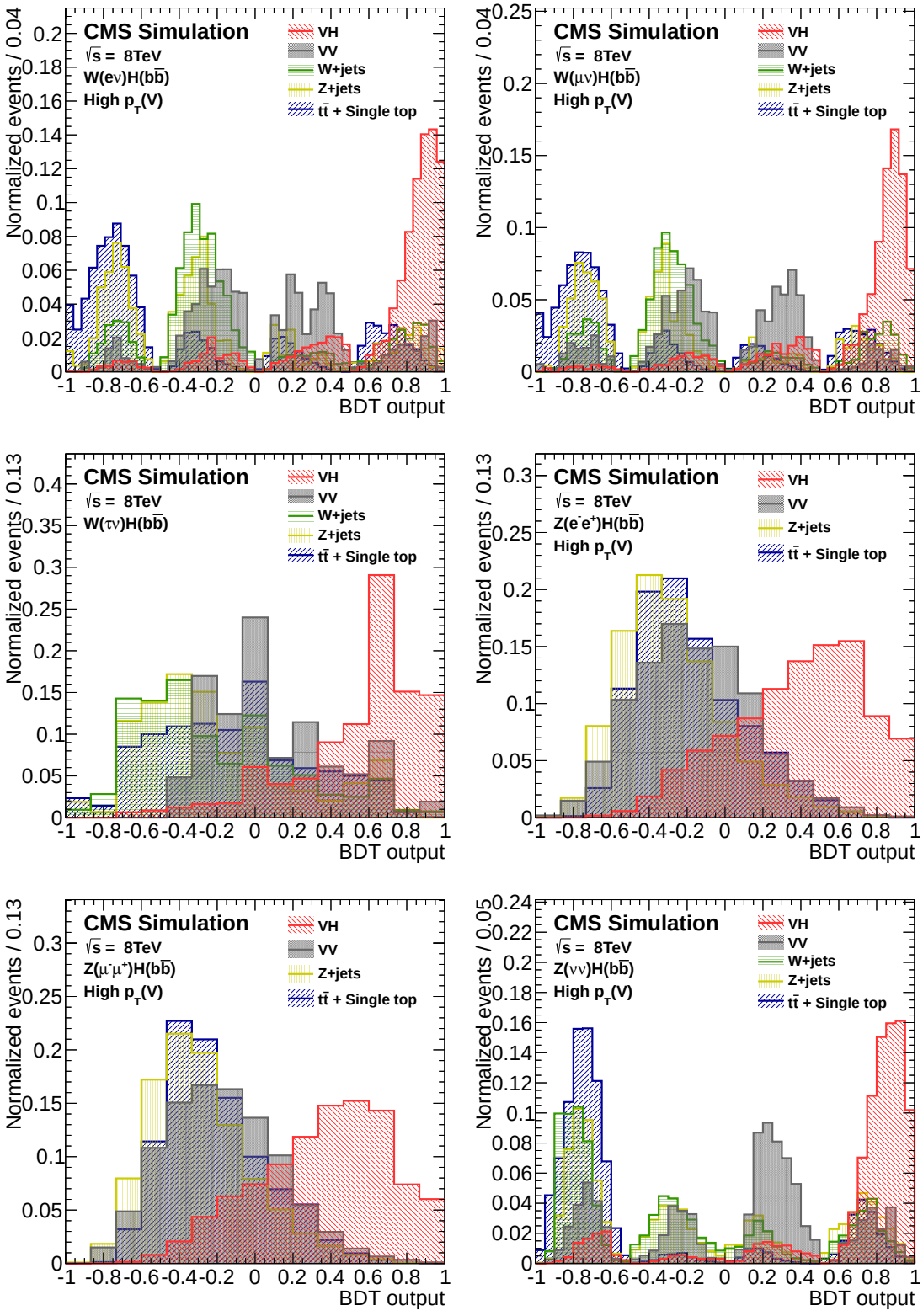

Figure C-6. BDT output distributions, normalized to unity, for the highest-boost region in all the $VH(b\bar{b})$ channels.

APPENDIX D
TRIGGER SCHEMATICS

The schematic diagrams of the $Z(\nu\bar{\nu})H(b\bar{b})$ triggers used in this analysis are shown in Fig. D-1. From left to right, they are:

- `HLT_PFMET150`
- `HLT_DiCentralJetSumpT100_dPhi05_DiCentralPFJet60_25_PFMET100_HBHENoiseCleaned`
- `HLT_DiCentralPFJet30_PFMET80_BTagCSV07`

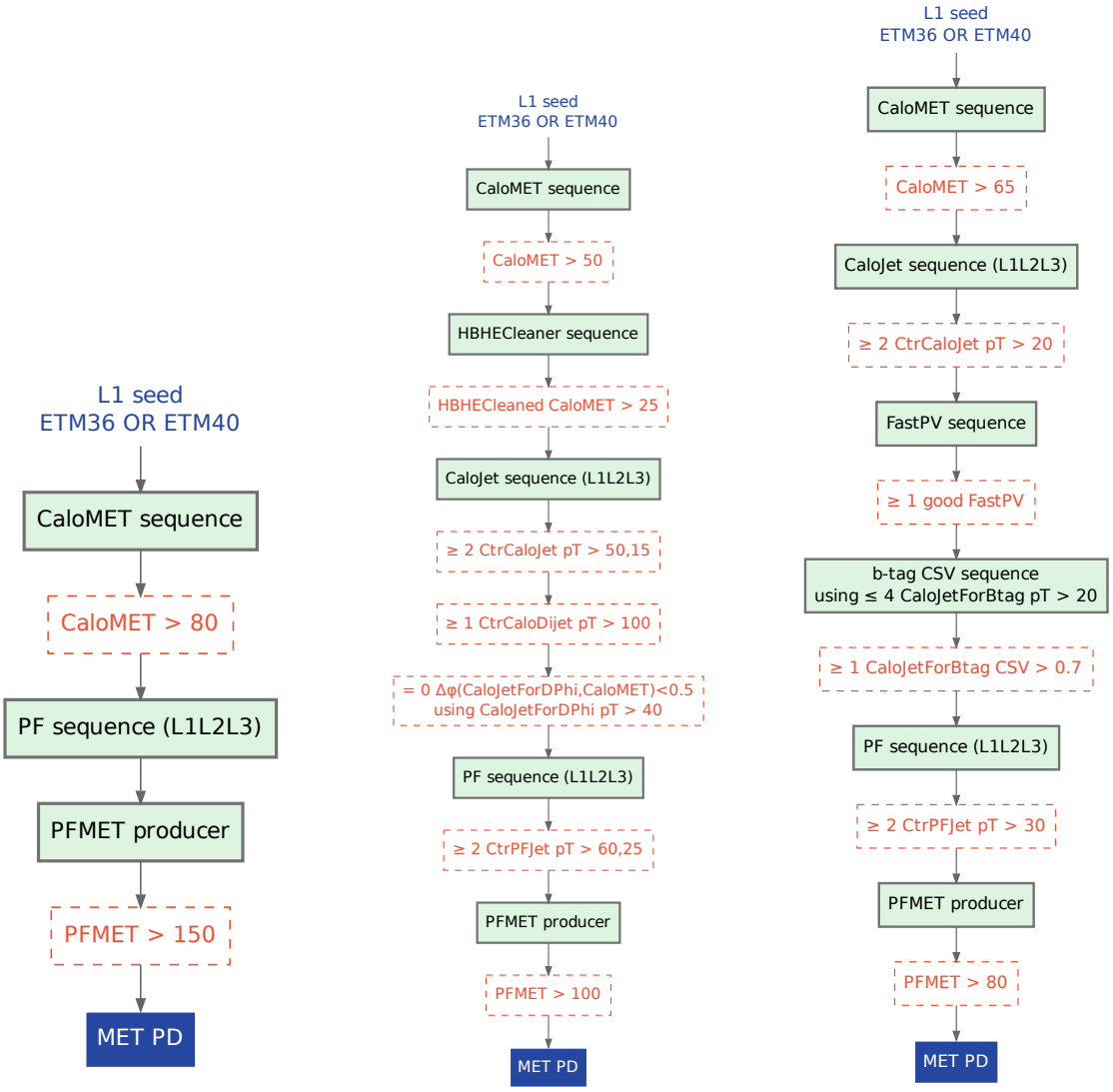

Figure D-1. Schematic diagrams of the three $Z(\nu\bar{\nu})H(b\bar{b})$ triggers.

APPENDIX E
STATISTICAL PROCEDURE

The statistical methodology used in this analysis was developed by the ATLAS and CMS Collaborations in the context of the LHC Higgs Combination Group, as described in Refs. [33, 62, 139, 152]. In CMS, the methodology is implemented in the HIGGSANALYSIS/COMBINEDLIMIT package [153]. Results are obtained using asymptotic formulae from Ref. [154], including routines available in the ROOSTATS package [155].

The chosen test statistic, q, is based on the profile likelihood ratio and is used to determine how signal-like or background-like the data are. Systematic uncertainties are incorporated in the analysis via nuisance parameters that are treated according to the frequentist paradigm.

E.1 Exclusion Limit Calculation

For the calculation of exclusion limits, the test statistic q_μ is defined as:

$$q_\mu = -2 \ln \frac{\mathcal{L}(\text{data} \,|\, \mu \cdot s + b, \hat{\theta}_\mu)}{\mathcal{L}(\text{data} \,|\, \hat{\mu} \cdot s + b, \hat{\theta})}, \text{ with } 0 \leq \hat{\mu} \leq \mu, \tag{E–1}$$

where "data" represents either the actual experimental observation or pseudo data used to construct sampling distributions (toys), s is the expected number and distribution of signal events, b is the number and distribution of background events, μ is the signal strength modifier introduced to accommodate deviations in signal event yields, and θ represents the set of nuisance parameters describing the systematic uncertainties. The value $\hat{\theta}_\mu$ maximizes the likelihood in the numerator for a given μ value, while $\hat{\mu}$ and $\hat{\theta}$ define the point at which the likelihood reaches its global maximum. The range of μ is restricted to the physically meaningful regime, i.e. it is not allowed to be negative.

The likelihood function is constructed as:

$$\mathcal{L}(\text{data} \,|\, \mu \cdot s + b, \theta) = \text{Poisson}(\text{data} \,|\, \mu \cdot s(\theta) + b(\theta)) \cdot p(\hat{\theta} \,|\, \theta), \tag{E–2}$$

where $p(\hat{\theta} \mid \theta)$ is a fictional auxiliary "measurement" probability density function (pdf) that allows systematic error pdf $\rho(\theta \mid \hat{\theta})$ to be written as:

$$\rho(\theta \mid \hat{\theta}) \sim p(\hat{\theta} \mid \theta) \cdot \pi_\theta(\theta), \tag{E–3}$$

where $\pi_\theta(\theta)$ functions are hyper-priors for those "measurements".

The CL_s value is then defined as the ratio of two probabilities:

$$CL_s(\mu) = \frac{P(q_\mu \geq q_\mu^{\text{data}} \mid \mu \cdot s + b)}{P(q_\mu \geq q_\mu^{\text{data}} \mid b)}, \tag{E–4}$$

where q_μ^{data} is the value of the test statistic observed in data. The numerator probability is evaluated under the signal+background hypothesis, whereas the denominator is under the background-only hypothesis ($\mu = 0$). $CL_s(\mu) \leq \alpha$ is used as the criterion for excluding the presence of a signal at the $1 - \alpha$ confidence level. For instance, the Higgs boson signal with a cross section $\sigma = \mu \cdot \sigma_{\text{SM}}$ is excluded at 95% confidence level if $CL_s(\mu) \leq 0.05$. Here, σ_{SM} stands for the SM Higgs boson cross section.

E.2 *p*-value and Significance Calculation

To quantify the presence of an excess of events over the expected background, the test statistic q_0 is defined as:

$$q_0 = -2 \ln \frac{\mathcal{L}(\text{data} \mid b, \hat{\theta}_0)}{\mathcal{L}(\text{data} \mid \hat{\mu} \cdot s + b, \hat{\theta})}, \text{ with } \hat{\mu} > 0, \tag{E–5}$$

where the likelihood in the numerator corresponds to the background-only hypothesis. The value $\hat{\theta}_0$ maximizes the likelihood in the numerator under the background-only hypothesis ($\mu = 0$), while $\hat{\mu}$ and $\hat{\theta}$ define the point at which the likelihood reaches its global maximum.

The quantity p_0, henceforth referred to as the local *p*-value, is defined as the probability, under the background-only hypothesis, to obtain a value of q_0 at least as

large as that observed in data, q_0^{data}:

$$p_0 = \mathrm{P}\left(q_0 \geq q_0^{\text{data}} \,\middle|\, b\right). \tag{E–6}$$

The local significance Z of a signal-like excess is computed using the one-sided Gaussian tail convention:

$$p_0 = \int_Z^{+\infty} \frac{1}{\sqrt{2\pi}} \exp(-x^2/2)\, \mathrm{d}x. \tag{E–7}$$

For instance, the 5σ significance (Z = 5) corresponds to $p_0 = 2.8 \times 10^{-7}$. Note that very small p-values should be interpreted with caution, since systematic biases and uncertainties in the underlying model are only known to a given precision.

E.3 Signal-Model Parameter Extraction

Signal-model parameters a, such as the signal strength modifier μ, are evaluated from a scan of the profile likelihood ratio $q(a)$:

$$q(a) = -2 \ln \frac{\mathcal{L}(\text{data} \,|\, s(a) + b,\, \hat{\theta}_a)}{\mathcal{L}(\text{data} \,|\, s(\hat{a}) + b,\, \hat{\theta})}. \tag{E–8}$$

The values of the parameters $\hat{a}$ and $\hat{\theta}$ that maximize the likelihood $\mathcal{L}(\text{data} \,|\, s(\hat{a}) + b,\, \hat{\theta})$ are called the best-fit set. The one-dimensional (1D) 68% and 95% C.L. confidence intervals for a given signal-model parameter, a_i, are evaluated from $q(a_i) = 1$ and $q(a_i) = 3.84$, respectively, with all other unconstrained model parameters treated in the same way as the nuisance parameters. The two-dimensional (2D) 68% and 95% C.L. confidence regions for pairs of parameters are derived from $q(a_i, a_j) = 2.30$ and $q(a_i, a_j) = 5.99$, respectively. This implies that boundaries of 2D confidence regions projected onto either parameter axis are not identical to the 1D confidence interval for that parameter.

REFERENCES

[1] Wikipedia, "Standard Model — Wikipedia, The Free Encyclopedia", 2014. [Online; accessed 12-December-2014].

[2] L. Álvarez-Gaumé and J. Ellis, "Eyes on a prize particle", *Nat. Phys.* **7** (2011) 2–3, doi:10.1038/nphys1874.

[3] LHC Higgs Cross Section Working Group et al., "Handbook of LHC Higgs Cross Sections: 1. Inclusive Observables", *CERN-2011-002* (2011) arXiv:1101.0593.

[4] LHC Higgs Cross Section Working Group et al., "Handbook of LHC Higgs Cross Sections: 2. Differential Distributions", *CERN-2012-002* (2012) arXiv:1201.3084.

[5] A. Heinson, "Feynman Diagrams for Top Physics Talks and Notes", (2011).

[6] CMS Collaboration, "Summaries of CMS cross section measurements (CMS TWiki)", (2015).

[7] F. Marcastel, "CERN's Accelerator Complex", (2013). General Photo.

[8] D. Dominguez, "3D cut of the LHC dipole", (2014). General Photo.

[9] W. J. Stirling, 2012. private communication.

[10] "LHC Luminosity Plots for the 2012 Proton Run", (2013).

[11] "CMS Luminosity — Public Results", (2013).

[12] T. Sakuma and T. McCauley, "Detector and Event Visualization with SketchUp at the CMS Experiment", *J. Phys. Conf. Ser.* **513** (2014) 022032, doi:10.1088/1742-6596/513/2/022032, arXiv:1311.4942.

[13] CMS Collaboration, "Description and performance of track and primary-vertex reconstruction with the CMS tracker", *JINST* **9** (2014) P10009, doi:10.1088/1748-0221/9/10/P10009, arXiv:1405.6569.

[14] CMS Collaboration, "CMS technical design report, volume I: Detector performance and software",.

[15] P. Adzic et al., "Energy resolution of the barrel of the CMS electromagnetic calorimeter", *JINST* **2** (2007) P04004, doi:10.1088/1748-0221/2/04/P04004.

[16] CMS Collaboration, "The CMS experiment at the CERN LHC", *JINST* **3** (2008) S08004, doi:10.1088/1748-0221/3/08/S08004.

[17] CMS Collaboration, "Aligning the CMS Muon Chambers with the Muon Alignment System during an Extended Cosmic Ray Run", *JINST* **5** (2010) T03019, doi:10.1088/1748-0221/5/03/T03019, arXiv:0911.4770.

[18] CMS Collaboration, "CMS: The TriDAS project. Technical design report, Vol. 2: Data acquisition and high-level trigger",.

[19] Particle Data Group, "Review of Particle Physics (RPP)", *Phys. Rev.* **D86** (2012) 010001, doi:10.1103/PhysRevD.86.010001.

[20] T. Gleisberg et al., "Event generation with SHERPA 1.1", *JHEP* **0902** (2009) 007, doi:10.1088/1126-6708/2009/02/007, arXiv:0811.4622.

[21] CMS Collaboration, "Jet Energy Scale performance in 2011", CMS Detector Performance Summary CMS-DP-2012-006, 2012.

[22] CMS Collaboration, "Status of the 8 TeV Jet Energy Corrections and Uncertainties based on 11 fb^{-1} of data in CMS", CMS Detector Performance Summary CMS-DP-2013-011, 2013.

[23] CMS Collaboration, "Performance of Missing Transverse Momentum Reconstruction Algorithms in Proton-Proton Collisions at $\sqrt{s} = 8$ TeV with the CMS Detector", CMS Physics Analysis Summary CMS-PAS-JME-12-002, 2012.

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

[25] CMS Collaboration, "Identification of b-quark jets with the CMS experiment", *JINST* **8** (2013) P04013, doi:10.1088/1748-0221/8/04/P04013, arXiv:1211.4462.

[26] A. Hocker et al., "TMVA - Toolkit for Multivariate Data Analysis", *PoS* **ACAT** (2007) 040, arXiv:physics/0703039.

[27] W. Verkerke and D. P. Kirkby, "The RooFit toolkit for data modeling", *eConf* **C0303241** (2003) MOLT007, arXiv:physics/0306116.

[28] CMS Collaboration, "Search for the standard model Higgs boson produced in association with a W or a Z boson and decaying to bottom quarks", *Phys. Rev.* **D89** (2014), no. 1, 012003, doi:10.1103/PhysRevD.89.012003, arXiv:1310.3687.

[29] J. M. Butterworth, A. R. Davison, M. Rubin, and G. P. Salam, "Jet substructure as a new Higgs search channel at the LHC", *Phys. Rev. Lett.* **100** (2008) 242001, doi:10.1103/PhysRevLett.100.242001, arXiv:0802.2470.

[30] CMS Collaboration, "Evidence for the direct decay of the 125 GeV Higgs boson to fermions", *Nature Phys.* **10** (2014) 557–560, doi:10.1038/nphys3005, arXiv:1401.6527.

[31] CMS Collaboration, "Measurement of WZ and ZZ production in pp collisions at $\sqrt{s} = 8$ TeV in final states with b-tagged jets", *Eur. Phys. J.* **C74** (2014), no. 8, 2973, doi:10.1140/epjc/s10052-014-2973-5, arXiv:1403.3047.

[32] CMS Collaboration, "Search for invisible decays of Higgs bosons in the vector boson fusion and associated ZH production modes", *Eur. Phys. J.* **C74** (2014), no. 8, 2980, doi:10.1140/epjc/s10052-014-2980-6, arXiv:1404.1344.

[33] CMS Collaboration, "Precise determination of the mass of the Higgs boson and tests of compatibility of its couplings with the standard model predictions using proton collisions at 7 and 8 TeV", arXiv:1412.8662. Submitted to Eur. Phys. J. C.

[34] P. W. Higgs, "Broken symmetries, massless particles and gauge fields", *Phys. Lett.* **12** (1964) 132, doi:10.1016/0031-9163(64)91136-9.

[35] P. W. Higgs, "Broken symmetries and the masses of gauge bosons", *Phys. Rev. Lett.* **13** (1964) 508, doi:10.1103/PhysRevLett.13.508.

[36] P. W. Higgs, "Spontaneous symmetry breakdown without massless bosons", *Phys. Rev.* **145** (1966) 1156, doi:10.1103/PhysRev.145.1156.

[37] F. Englert and R. Brout, "Broken symmetry and the mass of gauge vector mesons", *Phys. Rev. Lett.* **13** (1964) 321, doi:10.1103/PhysRevLett.13.321.

[38] G. S. Guralnik, C. R. Hagen, and T. W. B. Kibble, "Global conservation laws and massless particles", *Phys. Rev. Lett.* **13** (1964) 585, doi:10.1103/PhysRevLett.13.585.

[39] T. W. B. Kibble, "Symmetry breaking in non-Abelian gauge theories", *Phys. Rev.* **155** (1967) 1554, doi:10.1103/PhysRev.155.1554.

[40] S. Glashow, "Partial Symmetries of Weak Interactions", *Nucl. Phys.* **22** (1961) 579–588, doi:10.1016/0029-5582(61)90469-2.

[41] S. Weinberg, "A Model of Leptons", *Phys. Rev. Lett.* **19** (1967) 1264–1266, doi:10.1103/PhysRevLett.19.1264.

[42] A. Salam, "Weak and Electromagnetic Interactions", *Conf. Proc.* **C680519** (1968) 367–377.

[43] G. 't Hooft, "Renormalizable Lagrangians for Massive Yang-Mills Fields", *Nucl. Phys.* **B35** (1971) 167–188, doi:10.1016/0550-3213(71)90139-8.

[44] G. 't Hooft and M. Veltman, "Regularization and Renormalization of Gauge Fields", *Nucl. Phys.* **B44** (1972) 189–213, doi:10.1016/0550-3213(72)90279-9.

[45] S. Dawson, "Introduction to electroweak symmetry breaking", arXiv:hep-ph/9901280.

[46] Y. Nambu and G. Jona-Lasinio, "Dynamical Model of Elementary Particles Based on an Analogy with Superconductivity. 1.", *Phys. Rev.* **122** (1961) 345–358, doi:10.1103/PhysRev.122.345.

[47] Y. Nambu and G. Jona-Lasinio, "Dynamical Model of Elementary Particles Based on an Analogy with Superconductivity. 2.", *Phys. Rev.* **124** (1961) 246–254, doi:10.1103/PhysRev.124.246.

[48] B. Cleveland et al., "Measurement of the solar electron neutrino flux with the Homestake chlorine detector", *Astrophys. J.* **496** (1998) 505–526, doi:10.1086/305343.

[49] Super-Kamiokande Collaboration, "Evidence for oscillation of atmospheric neutrinos", *Phys. Rev. Lett.* **81** (1998) 1562–1567, doi:10.1103/PhysRevLett.81.1562, arXiv:hep-ex/9807003.

[50] SNO Collaboration, "Measurement of the rate of $\nu_e + d \to p + p + e^-$ interactions produced by ^8B solar neutrinos at the Sudbury Neutrino Observatory", *Phys. Rev. Lett.* **87** (2001) 071301, doi:10.1103/PhysRevLett.87.071301, arXiv:nucl-ex/0106015.

[51] KamLAND Collaboration, "First results from KamLAND: Evidence for reactor anti-neutrino disappearance", *Phys. Rev. Lett.* **90** (2003) 021802, doi:10.1103/PhysRevLett.90.021802, arXiv:hep-ex/0212021.

[52] LEP Working Group for Higgs boson searches, ALEPH Collaboration, DELPHI Collaboration, L3 Collaboration, OPAL Collaboration, "Search for the standard model Higgs boson at LEP", *Phys. Lett.* **B565** (2003) 61–75, doi:10.1016/S0370-2693(03)00614-2, arXiv:hep-ex/0306033.

[53] CDF Collaboration, D0 Collaboration, "Higgs Boson Studies at the Tevatron", *Phys. Rev.* **D88** (2013), no. 5, 052014, doi:10.1103/PhysRevD.88.052014, arXiv:1303.6346.

[54] CDF Collaboration, D0 Collaboration, "Evidence for a particle produced in association with weak bosons and decaying to a bottom-antibottom quark pair in Higgs boson searches at the Tevatron", *Phys. Rev. Lett.* **109** (2012) 071804, doi:10.1103/PhysRevLett.109.071804, arXiv:1207.6436.

[55] CMS Collaboration, "Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC", *Phys. Lett.* **B716** (2012) 30–61, doi:10.1016/j.physletb.2012.08.021, arXiv:1207.7235.

[56] ATLAS Collaboration, "Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC", *Phys. Lett.* **B716** (2012) 1–29, doi:10.1016/j.physletb.2012.08.020, arXiv:1207.7214.

[57] ATLAS Collaboration, "Measurements of the Higgs boson production and decay rates and coupling strengths using pp collision data at sqrts = 7 and 8 TeV in the ATLAS experiment", ATLAS Conference Note ATLAS-CONF-2015-007, 2015.

[58] ATLAS, "Measurement of the Higgs boson mass from the $H \to \gamma\gamma$ and $H \to ZZ^* \to 4\ell$ channels with the ATLAS detector using 25 fb^{-1} of *pp* collision data", *Phys.Rev.* **D90** (2014), no. 5, 052004, doi:10.1103/PhysRevD.90.052004, arXiv:1406.3827.

[59] ATLAS, "Evidence for the Higgs-boson Yukawa coupling to tau leptons with the ATLAS detector", arXiv:1501.04943. Submitted to JHEP.

[60] ATLAS Collaboration, "Search for the $b\bar{b}$ decay of the Standard Model Higgs boson in associated $(W/Z)H$ production with the ATLAS detector", *JHEP* **1501** (2015) 069, doi:10.1007/JHEP01(2015)069, arXiv:1409.6212.

[61] D. Binosi and L. Theussl, "JaxoDraw: A Graphical user interface for drawing Feynman diagrams", *Comput. Phys. Commun.* **161** (2004) 76–86, doi:10.1016/j.cpc.2004.05.001, arXiv:hep-ph/0309015.

[62] CMS Collaboration, "Observation of a new boson with mass near 125 GeV in pp collisions at $\sqrt{s}$ = 7 and 8 TeV", *JHEP* **1306** (2013) 081, doi:10.1007/JHEP06(2013)081, arXiv:1303.4571.

[63] M. Bachtis, "Heavy Neutral Particle Decays to Tau Pairs in Proton Collisions at $\sqrt{s}$ = 7 TeV with CMS at the CERN Large Hadron Collider". PhD thesis, University of Wisconsin-Madison, 2012.

[64] S. A. Koay, "A Search for Dark Matter Production with Jets and Missing Momentum Signature in Proton-Proton Collisions at 7 TeV". PhD thesis, University of California, Santa Barbara, 2011.

[65] P. Bortignon, "Search for the standard model Higgs boson produced in association with a Z boson with the CMS detector at the LHC". PhD thesis, ETH Zurich, 2014.

[66] G. Petrucciani, "The search for the Higgs boson at CMS". PhD thesis, Scuola Normale Superiore di Pisa, 2013.

[67] C. Lefevre, "LHC: the guide", (2009).

[68] TOTEM Collaboration, "Luminosity-Independent Measurement of the Proton-Proton Total Cross Section at sqrt(s)=8 TeV", *Phys. Rev. Lett.* **111** (2013), no. 1, 012001, doi:10.1103/PhysRevLett.111.012001.

[69] M. Lamont, "Status of the LHC", *J. Phys. Conf. Ser.* **455** (2013) 012001, doi:10.1088/1742-6596/455/1/012001.

[70] L. Evans and P. Bryant, "LHC Machine", *JINST* **3** (2008) S08001, doi:10.1088/1748-0221/3/08/S08001.

[71] CMS Collaboration, "CMS technical design report, volume II: Physics performance", *J. Phys.* **G34** (2007) 995–1579, doi:10.1088/0954-3899/34/6/S01.

[72] USCMS, ECAL/HCAL, "The CMS barrel calorimeter response to particle beams from 2-GeV/c to 350-GeV/c", *Eur. Phys. J.* **C60** (2009) 359–373, doi:10.1140/epjc/s10052-009-0959-5,10.1140/epjc/s10052-009-1024-0.

[73] CMS Collaboration, "CMS. The TriDAS project. Technical design report, vol. 1: The trigger systems",.

[74] CMS Trigger and Data Acquisition Group, "The CMS high level trigger", *Eur. Phys. J.* **C46** (2006) 605–667, doi:10.1140/epjc/s2006-02495-8, arXiv:hep-ex/0512077.

[75] I. Bird et al., "LHC computing Grid. Technical design report",.

[76] CMS Collaboration, "CMS: The computing project. Technical design report",.

[77] J. C. Collins, D. E. Soper, and G. F. Sterman, "Factorization of Hard Processes in QCD", *Adv. Ser. Direct. High Energy Phys.* **5** (1988) 1–91, arXiv:hep-ph/0409313.

[78] T. Sjostrand, "Monte Carlo Tools", arXiv:0911.5286.

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

[80] CMS Collaboration, "Particle-Flow Event Reconstruction in CMS and Performance for Jets, Taus, and $E_{\mathrm{T}}^{\mathrm{miss}}$", CMS Physics Analysis Summary CMS-PAS-PFT-09-001, 2009.

[81] CMS Collaboration, "Commissioning of the Particle-flow Event Reconstruction with the first LHC collisions recorded in the CMS detector", CMS Physics Analysis Summary CMS-PAS-PFT-10-001, 2010.

[82] R. Fruhwirth, "Application of Kalman filtering to track and vertex fitting", *Nucl. Instrum. Meth.* **A262** (1987) 444–450, doi:10.1016/0168-9002(87)90887-4.

[83] CMS Collaboration, "SWGuideIterativeTracking (CMS TWiki)", (2012).

[84] K. Rose, "Deterministic annealing for clustering, compression, classification, regression, and related optimization problems", *Proceedings of the IEEE* **86** (1998), no. 11, 2210–2239, doi:10.1109/5.726788.

[85] R. Fruhwirth, W. Waltenberger, and P. Vanlaer, "Adaptive vertex fitting", *J. Phys.* **G34** (2007) N343, doi:10.1088/0954-3899/34/12/N01.

[86] CMS Collaboration, "Performance of CMS muon reconstruction in *pp* collision events at $\sqrt{s} = 7$ TeV", *JINST* **7** (2012) P10002, doi:10.1088/1748-0221/7/10/P10002, arXiv:1206.4071.

[87] CMS Collaboration, "Electron Reconstruction and Identification at $\sqrt{s} = 7$ TeV", CMS Physics Analysis Summary CMS-PAS-EGM-10-004, 2010.

[88] CMS Collaboration, "Electron performance with 19.6 fb^{-1} of data collected at $\sqrt{s} = 8$ TeV with the CMS detector.", CMS Detector Performance Summary CMS-DP-2013-003, 2013.

[89] CMS Collaboration, "Performance of tau-lepton reconstruction and identification in CMS", *JINST* **7** (2012) P01001, doi:10.1088/1748-0221/7/01/P01001, arXiv:1109.6034.

[90] M. Cacciari, G. P. Salam, and G. Soyez, "The anti-k_t jet clustering algorithm", *JHEP* **04** (2008) 063, doi:10.1088/1126-6708/2008/04/063, arXiv:0802.1189.

[91] M. Cacciari, G. P. Salam, and G. Soyez, "FastJet User Manual", *Eur. Phys. J. C* **72** (2012) 1896, doi:10.1140/epjc/s10052-012-1896-2, arXiv:1111.6097.

[92] CMS Collaboration, "Determination of jet energy calibration and transverse momentum resolution in CMS", *JINST* **6** (2011) P11002, doi:10.1088/1748-0221/6/11/P11002, arXiv:1107.4277.

[93] M. Cacciari and G. P. Salam, "Pileup subtraction using jet areas", *Phys. Lett.* **B659** (2008) 119–126, doi:10.1016/j.physletb.2007.09.077, arXiv:0707.1378.

[94] CMS Collaboration, "Jet Energy Resolution (CMS TWiki)", (2014).

[95] CMS Collaboration, "Jet Identification (CMS TWiki)", (2014).

[96] CMS Collaboration, "Pileup Jet Identification", CMS Physics Analysis Summary CMS-PAS-JME-13-005, 2013.

[97] CMS Collaboration, "Missing transverse energy performance of the CMS detector", *JINST* **6** (2011) P09001, doi:10.1088/1748-0221/6/09/P09001, arXiv:1106.5048.

[98] CMS Collaboration, "MET Optional Filters (CMS TWiki)", (2013).

[99] Heavy Flavor Averaging Group (HFAG), "Averages of *b*-hadron, *c*-hadron, and τ-lepton properties as of summer 2014", arXiv:1412.7515.

[100] ALEPH Collaboration, DELPHI Collaboration, L3 Collaboration, OPAL Collaboration, SLD Collaboration, LEP Electroweak Working Group, SLD Electroweak Group, SLD Heavy Flavour Group, "Precision electroweak measurements on the *Z* resonance", *Phys. Rept.* **427** (2006) 257–454, doi:10.1016/j.physrep.2005.12.006, arXiv:hep-ex/0509008.

[101] W. Waltenberger, "Adaptive vertex reconstruction", CMS Note CERN-CMS-NOTE-2008-033, 2008.

[102] CMS Collaboration, "Usage of b Tag Objects for 8 TeV Data with the 53X PromptReco (CMS TWiki)", (2014).

[103] CMS VHbb Team, "Search for the Standard Model Higgs Boson Produced in Association with W and Z and Decaying to Bottom Quarks (LHCP 2013)", CMS Note 2013/069, 2013.

[104] S. Frixione, P. Nason, and C. Oleari, "Matching NLO QCD computations with parton shower simulations: the POWHEG method", *JHEP* **11** (2007) 070, doi:10.1088/1126-6708/2007/11/070, arXiv:0709.2092.

[105] LHC Physics, "SM Higgs production cross sections at $\sqrt{s}$ = 8 TeV (2012 update, used until summer 2013) (LHCPhysics TWiki)", (2014).

[106] J. Alwall et al., "MadGraph 5: going beyond", *JHEP* **06** (2011) 128, doi:10.1007/JHEP06(2011)128, arXiv:1106.0522.

[107] T. Sjöstrand, S. Mrenna, and P. Z. Skands, "PYTHIA 6.4 physics and manual", *JHEP* **0605** (2006) 026, doi:10.1088/1126-6708/2006/05/026, arXiv:hep-ph/0603175.

[108] J. M. Campbell and R. K. Ellis, "MCFM for the Tevatron and the LHC", *Nucl. Phys. Proc. Suppl.* **205-206** (2010) 10, doi:10.1016/j.nuclphysbps.2010.08.011, arXiv:1007.3492.

[109] R. Gavin, Y. Li, F. Petriello, and S. Quackenbush, "FEWZ 2.0: A code for hadronic Z production at next-to-next-to-leading order", *Comput. Phys. Commun.* **182** (2011) 2388, doi:10.1016/j.cpc.2011.06.008, arXiv:1011.3540.

[110] Y. Li and F. Petriello, "Combining QCD and electroweak corrections to dilepton production in FEWZ", *Phys. Rev. D* **86** (2012) 094034, doi:10.1103/PhysRevD.86.094034, arXiv:1208.5967.

[111] R. Gavin, Y. Li, F. Petriello, and S. Quackenbush, "W Physics at the LHC with FEWZ 2.1", *Comput. Phys. Commun.* **184** (2013) 208, doi:10.1016/j.cpc.2012.09.005, arXiv:1201.5896.

[112] A. D. Martin, W. J. Stirling, R. S. Thorne, and G. Watt, "Parton distributions for the LHC", *Eur. Phys. J. C* **63** (2009) 189, doi:10.1140/epjc/s10052-009-1072-5, arXiv:0901.0002.

[113] J. Pumplin et al., "New generation of parton distributions with uncertainties from global QCD analysis", *JHEP* **07** (2002) 012, doi:10.1088/1126-6708/2002/07/012, arXiv:hep-ph/0201195.

[114] M. Bähr et al., "Herwig++ physics and manual", *Eur. Phys. J. C* **58** (2008) 639, doi:10.1140/epjc/s10052-008-0798-9, arXiv:0803.0883.

[115] CMS Collaboration, "Measurement of the underlying event activity at the LHC with $\sqrt{s} = 7$ TeV and comparison with $\sqrt{s} = 0.9$ TeV", *JHEP* **09** (2011) 109, doi:10.1007/JHEP09(2011)109, arXiv:1107.0330.

[116] S. Jadach, J. H. Kühn, and Z. Wąs, "TAUOLA—a library of Monte Carlo programs to simulate decays of polarized tau leptons", *Comput. Phys. Commun.* **64** (1991) 275, doi:10.1016/0010-4655(91)90038-M.

[117] M. Ciccolini, A. Denner, and S. Dittmaier, "Strong and electroweak corrections to the production of Higgs+2jets via weak interactions at the LHC", *Phys. Rev. Lett.* **99** (2007) 161803, doi:10.1103/PhysRevLett.99.161803, arXiv:0707.0381.

[118] M. Ciccolini, A. Denner, and S. Dittmaier, "Electroweak and QCD corrections to Higgs production via vector-boson fusion at the LHC", *Phys. Rev. D* **77** (2008) 013002, doi:10.1103/PhysRevD.77.013002, arXiv:0710.4749.

[119] A. Denner, S. Dittmaier, S. Kallweit, and A. Muck, "Electroweak corrections to Higgs-strahlung off W/Z bosons at the Tevatron and the LHC with HAWK", *JHEP* **03** (2012) 075, doi:10.1007/JHEP03(2012)075, arXiv:1112.5142.

[120] G. Ferrera, M. Grazzini, and F. Tramontano, "Associated WH production at hadron colliders: a fully exclusive QCD calculation at NNLO", *Phys. Rev. Lett.* **107** (2011) 152003, doi:10.1103/PhysRevLett.107.152003, arXiv:1107.1164.

[121] CMS Collaboration, "Pileup Studies (CMS TWiki)", (2013).

[122] CMS Collaboration, "Pileup Reweighting (CMS TWiki)", (2011).

[123] CDF Collaboration, D0 Collaboration, "Improved b-jet Energy Correction for $H \to b\bar{b}$ Searches at CDF", arXiv:1107.3026.

[124] J. Gallicchio and M. D. Schwartz, "Seeing in Color: Jet Superstructure", *Phys. Rev. Lett.* **105** (2010) 022001, doi:10.1103/PhysRevLett.105.022001, arXiv:1001.5027.

[125] CDF Collaboration, "Search for the standard model Higgs boson decaying to a bb pair in events with two oppositely-charged leptons using the full CDF data set", *Phys. Rev. Lett.* **109** (2012) 111803, doi:10.1103/PhysRevLett.109.111803, arXiv:1207.1704.

[126] ATLAS Collaboration, "Measurement of the cross-section for W boson production in association with b-jets in pp collisions at $\sqrt{s} = 7$ TeV with the ATLAS detector", *JHEP* **06** (2013) 084, doi:10.1007/JHEP06(2013)084, arXiv:1302.2929.

[127] 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", *JHEP* **1312** (2013) 039, doi:10.1007/JHEP12(2013)039, arXiv:1310.1349.

[128] CMS Collaboration, "Measurement of the Z/γ^{**}+b-jet cross section in pp collisions at 7 TeV", *JHEP* **06** (2012) 126, doi:10.1007/JHEP06(2012)126, arXiv:1204.1643.

[129] CMS Collaboration, "CMS Luminosity Based on Pixel Cluster Counting — Summer 2013 Update", CMS Physics Analysis Summary CMS-PAS-LUM-13-001, 2013.

[130] S. Alekhin et al., "The PDF4LHC Working Group Interim Report", (2011). arXiv:1101.0536.

[131] M. Botje et al., "The PDF4LHC Working Group Interim Recommendations", (2011). arXiv:1101.0538.

[132] H.-L. Lai et al., "New parton distributions for collider physics", *Phys. Rev. D* **82** (2010) 074024, doi:10.1103/PhysRevD.82.074024, arXiv:1007.2241.

[133] R. D. Ball et al., "Impact of Heavy Quark Masses on Parton Distributions and LHC Phenomenology", *Nucl. Phys. B* **849** (2011) 296, doi:10.1016/j.nuclphysb.2011.03.021, arXiv:1101.1300.

[134] S. Frixione and B. R. Webber, "Matching NLO QCD computations and parton shower simulations", *JHEP* **06** (2002) 029, doi:10.1088/1126-6708/2002/06/029, arXiv:hep-ph/0204244.

[135] CMS Collaboration, "Measurement of the single-top-quark *t*-channel cross section in *pp* collisions at $\sqrt{s} = 7$ TeV", *JHEP* **12** (2012) 035, doi:10.1007/JHEP12(2012)035, arXiv:1209.4533.

[136] CMS Collaboration, "Measurement of W+W- and ZZ production cross sections in pp collisions at sqrt(s) = 8 TeV", *Phys. Lett.* **B721** (2013) 190–211, doi:10.1016/j.physletb.2013.03.027, arXiv:1301.4698.

[137] A. L. Read, "Presentation of search results: The CL_s technique", *J. Phys. G* **28** (2002) 2693, doi:10.1088/0954-3899/28/10/313.

[138] T. Junk, "Confidence level computation for combining searches with small statistics", *Nucl. Instrum. Meth.* **A434** (1999) 435–443, doi:10.1016/S0168-9002(99)00498-2, arXiv:hep-ex/9902006.

[139] ATLAS Collaboration, CMS Collaboration, and LHC Higgs Combination Group, "Procedure for the LHC Higgs boson search combination in Summer 2011", CMS Note CMS-NOTE-2011-005. ATL-PHYS-PUB-2011-011, 2011.

[140] LHC Higgs Cross Section Working Group et al., "Handbook of LHC Higgs Cross Sections: 3. Higgs Properties", *CERN-2013-004* (2013) arXiv:1307.1347.

[141] L. Altenkamp et al., "Gluon-induced Higgs-strahlung at next-to-leading order QCD", *J. High Energy Phys.* **02** (2013) 078, doi:10.1007/JHEP02(2013)078, arXiv:1211.5015.

[142] C. Englert, M. McCullough, and M. Spannowsky, "Gluon-initiated associated production boosts Higgs physics", *Phys.Rev.* **D89** (2014), no. 1, 013013, doi:10.1103/PhysRevD.89.013013, arXiv:1310.4828.

[143] G. Luisoni, P. Nason, C. Oleari, and F. Tramontano, "$HW^{\pm}/HZ$ + 0 and 1 jet at NLO with the POWHEG BOX interfaced to GoSam and their merging within MiNLO", *J. High Energy Phys.* **10** (2013) 083, doi:10.1007/JHEP10(2013)083, arXiv:1306.2542.

[144] G. Ferrera, M. Grazzini, and F. Tramontano, "Associated ZH production at hadron colliders: The fully differential NNLO QCD calculation", *Phys. Lett. B* **740** (2015) 51, doi:10.1016/j.physletb.2014.11.040, arXiv:1407.4747.

[145] LHC Physics, "SM Higgs production cross sections at $\sqrt{s}$ = 8 TeV (update in CERN Report3) (LHCPhysics TWiki)", (2014).

[146] D. Bertolini, P. Harris, M. Low, and N. Tran, "Pileup Per Particle Identification", *JHEP* **1410** (2014) 59, doi:10.1007/JHEP10(2014)059, arXiv:1407.6013.

[147] CMS Collaboration, "Pileup Removal Algorithms", CMS Physics Analysis Summary CMS-PAS-JME-14-001, 2014.

[148] CMS Collaboration, "Identifying Hadronically Decaying W Bosons Merged into a Single Jet", CMS Physics Analysis Summary CMS-PAS-JME-13-006, 2013.

[149] CMS Collaboration, "Boosted Top Jet Tagging at CMS", Technical Report CMS-PAS-JME-13-007, 2014.

[150] CMS Collaboration, "Measurement of $B\bar{B}$ Angular Correlations based on Secondary Vertex Reconstruction at $\sqrt{s} = 7$ TeV", *JHEP* **1103** (2011) 136, doi:10.1007/JHEP03(2011)136, arXiv:1102.3194.

[151] CMS Collaboration, "Physics Analysis Oriented Event Display (Fireworks / cmsShow) (CMS TWiki)", (2015).

[152] CMS Collaboration, "Combined results of searches for the standard model Higgs boson in *pp* collisions at $\sqrt{s} = 7$ TeV", *Phys.Lett.* **B710** (2012) 26–48, doi:10.1016/j.physletb.2012.02.064, arXiv:1202.1488.

[153] CMS Collaboration, "Documentation of the RooStats-based statistics tools for Higgs PAG (CMS TWiki)", (2014).

[154] G. Cowan, K. Cranmer, E. Gross, and O. Vitells, "Asymptotic formulae for likelihood-based tests of new physics", *Eur.Phys.J.* **C71** (2011) 1554, doi:10.1140/epjc/s10052-011-1554-0,10.1140/epjc/s10052-013-2501-z, arXiv:1007.1727.

[155] L. Moneta et al., "The RooStats Project", *PoS* **ACAT2010** (2010) 057, arXiv:1009.1003.

BIOGRAPHICAL SKETCH

Jia Fu grew up in a small town called Triang in Malaysia. After high school, he went to Taylor's College at Subang Jaya, Malaysia and enrolled in a program where one can take the first two years of U.S. university courses in Malaysia, and transfer the course credits to an U.S. university. He joined the University of Nebraska at Lincoln as an international transfer student. There, he worked with Dr. Roger Kirby and Dr. Kenneth Bloom for undergraduate physics research and obtained his Bachelor of Science degree in physics. After that, he joined the summer student program at CERN laboratory near Geneva, Switzerland and was placed in Dr. Joe Incandela's group from the Universiy of California at Santa Barbara. He continued to work with the group after the program had concluded until the next graduate school intake. He applied to and was accepted by the University of Florida for graduate study in experimental high energy physics. After taking the graduate courses, he moved to Fermi National Accelerator Laboratory near Chicago to work closely with his advisor, Dr. Jacobo Konigsberg, in the Compact Muon Solenoid experiment. He received his Ph.D. from the University of Florida in the spring of 2015.