# Search for the Standard Model Higgs Boson Produced in Association with a Vector Boson and Decaying to Bottom Quarks at the CMS Detector

Michael Ryan Mooney

A Dissertation

Presented to the Faculty

of Princeton University

in Candidacy for the Degree

of Doctor of Philosophy

Recommended for Acceptance

by the Department of

Physics

Advisor: James David Olsen

September, 2014

# Abstract

A search for the Standard Model Higgs boson produced in association with a W or Z boson and decaying to bottom quarks is presented. A sample of approximately 24 $\mathrm{fb}^{-1}$ of data recorded by the CMS experiment at the Large Hadron Collider, operating at center-of-mass energies of 7 TeV and 8 TeV in 2011 and 2012, respectively, is used to search for events consistent with the signature of two b jets recoiling with high momentum from a $\mathrm{W}(\ell\nu)$, $\mathrm{W}(\tau\nu)$, $\mathrm{Z}(\ell\ell)$, or $\mathrm{Z}(\nu\nu)$ decay, where $\ell$ refers to either an electron or a muon. One-prong hadronic decays of the tau particle are selected for in the case of $\mathrm{W}(\tau\nu)$. Observed signal significance and 95% confidence level upper limits on the production cross section times $\mathrm{H} \rightarrow \mathrm{b\bar{b}}$ branching fraction, relative to the Standard Model prediction, are presented for the 110-135 GeV Higgs mass range.

# Acknowledgements

There are many people that have professionally or personally influenced the work presented in this dissertation. First, I would like to extend my deepest gratitude to my advisor, Jim Olsen, for guiding me through six years of difficult academic work. I admire Jim not only for his immense physics knowledge and intuition, but also for his attitude in working with others – he always treated me as an equal when we worked together and inspired confidence in me when I needed it.

I would like to thank the other Princeton professors involved with the CMS experiment, namely Chris Tully, Dan Marlow, and Pierre Piroue, for being helpful resources in both physics and more practical matters. Many thanks to Mariangela Lisanti for reading my thesis, which was longer than I promised her it would be (sorry about that). Thanks to Peter Meyers for discussions on neutrino physics that were very helpful in planning my future. I would also like to thank Kim Dawidowski for her considerable help in keeping my life organized over the past six years.

It has been a pleasure to work with many great physicists at CERN. Of these physicists, I would first like to thank David Lopes-Pegna for his mentorship. I have learned a great deal from David in our discussions, which often went quite late into the night (at least in my time zone). His energy and determination was inspirational and undoubtedly a crucial ingredient of my success. I would also like to thank Jacobo Konigsberg, Andrea Rizzi, Michele de Gruttola, Niklas Mohr, Seth Zenz, Pierluigi Bortignon, Matt Fisher, and Jia Fu Low for their collaboration on the VH(b$\bar{\text{b}}$) analysis. Together we have achieved an impressive feat that I am really proud of. Thanks also to Giovanni Abbiendi, Silvia Goy Lopez, Ivan Mikulec, Kevin Stenson, and Slava Valuev for helpful

discussions about tracking and muon physics.

To my friends in Princeton, Geneva, and elsewhere, I owe much gratitude for helping to keep my life relatively sane while in graduate school. There are too many to name here, but in particular I would like to thank (in alphabetical order) Ali Altug, Farzaneh Badii, Rachel Bartek, Isa de Castro, Anushya Chandran, Seth Chizeck, Ryan Cockerham, Julie Cross, Mihai Cucuringu, Jigisha Dharba, Ben Fong, Davide Gerbaudo, Robert (Jason) Harris, Andrew Hartnett, Kevin Hughes, Arie Israel, Matt Johnson, Darya Krasilnikov, Ted Laird, Chris Lester, David McGady, Alex Mott, Elina Nalibotski, Guilherme Pimentel, Xiaohang Quan, Jason Rogers, Halil Saka, Rishika Samant, Blake Sherwin, Eduardo da Silva Neto, Jon Stahlman, Rami Vanguri, and Andrzej Zuranski. I am incredibly lucky for this list, incomplete as it is, to be so long.

Finally, I want to thank my parents, Paula and Mitchell. It is not surprising that I have made it so far in life because these two people are simply amazing. Having few resources to work with, my parents made many sacrifices in order to ensure that I had a comfortable life and a great education. Furthermore, they taught me to be thoughtful and understanding of others, and to always remain humble, qualities that I think are extremely important. Above all else, they raised me to be a passionate individual. It is this passion, for learning and trying to make a difference in the world, that has ultimately made me the person I am today.

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

The Standard Model is a theoretical framework that describes all known fundamental particles and their interactions through the electromagnetic, weak, and strong forces. This framework has been remarkably successful over the last century in making predictions, both qualitative and quantitative in nature, that have been verified experimentally. Such predictions have led to a number of important discoveries in physics over the years, including the discovery of the W and Z bosons at the Super Proton Synchrotron (SPS) in 1983 [16–19] and the discovery of the top quark at the Tevatron in 1995 [20, 21]. Except for the surprise observation of neutrino oscillations [22], the Standard Model has never failed a single experimental test.

The most recently discovered fundamental particle is the Higgs boson, observed at the Large Hadron Collider (LHC) in 2012 [23, 24]. Its discovery followed hints for the particle's existence at the LHC in late 2011 [25, 26] and shortly after similar evidence was found at the Tevatron [27, 28]. This observation was made using just a small fraction of over 500,000 proton-proton collisions in which a Standard Model Higgs boson was produced – many signal events are not analyzed due to being very similar to background processes. The existence of the Higgs boson is necessary to account for the unification of the electromagnetic and weak forces at the very high energy densities of the early universe while preserving local gauge invariance for the interactions (see Section 1.1). The observation of the Higgs boson at a mass of roughly 125 GeV, along with several follow-up measurements to pin down the properties of the particle [29–39], has essentially "completed" the Standard Model, as we now have what we believe to be a theoretically-consistent description of the electromagnetic, weak, and strong interactions that has been verified experimentally.

Despite this achievement, there are still remaining questions within Higgs sector physics. Of principal interest is to ascertain whether or not the Higgs boson couples to bottom quarks, as this has not yet been firmly established. Even if such decays of the Higgs boson occur, any discrepancies in the branching fractions of the Higgs boson decaying to these particles is a possible sign of the existence of new physical processes beyond those explained by the Standard Model, if not attributed to discrepancies in theoretical predictions. At the very least, as the total mass width of the Higgs boson resonance is dominated by the contribution from decays to bottom quarks, constraining this decay rate would further constrain all other branching fractions – an important contribution in its own right.

Described in this dissertation is a search for the Standard Model Higgs boson decaying into bottom quarks (more specifically a $\mathrm{b\bar{b}}$ pair) and produced in association with a leptonically-decaying vector boson (either W or Z) at the CMS detector, one of the two general-purpose detectors associated with the LHC. This search, also referred to as the $\mathrm{VH(b\bar{b})}$ search, was included in the CMS Higgs boson discovery paper [24] and has also led to two dedicated publications [40, 41] and more recently a publication that shows evidence for the Standard Model Higgs boson decaying to fermions [42], combining the most recent CMS $\mathrm{VH(b\bar{b})}$ analysis results with those of the CMS $\mathrm{H} \rightarrow \tau^+\tau^-$ analysis [43]. The results presented here reflect the use of the complete LHC proton-proton collision dataset collected in 2011 at $\sqrt{s} = 7$ TeV and 2012 at $\sqrt{s} = 8$ TeV. However, the analysis methodology discussed in this work describes only the most recent round of the analysis using strictly the 2012 dataset.

We begin in this chapter by discussing the theoretical motivation for the $\mathrm{VH(b\bar{b})}$ search as well as a general overview of the Standard Model of particle physics. In Section 1.1 the Standard Model is described in detail, including its limitations in describing certain phenomena observed in the universe. This is followed by a discussion of the Higgs mechanism, which explains how the electromagnetic and weak forces split apart as well as how the W and Z bosons acquired mass as the expanding early universe cooled, in Section 1.2. Finally, in Section 1.3, we give an overview of the CMS VH($\mathrm{b\bar{b}}$) analysis methodology, discussing the organization of this dissertation in the process.

## 1.1 Standard Model

The Standard Model of particle physics, as discussed above, was constructed to explain the electromagnetic, weak, and strong interactions between all known fundamental particles. This theoretical model makes use of the mathematical framework of relativistic quantum field theory (QFT), which represents particles as excitations of relativistic quantum fields. While the construction of the Standard Model was a process that began formally in the early 1960's, the Standard Model represents over a century of attained knowledge about the constituents of the universe and their interactions, beginning with J. J. Thomson's discovery of the electron in 1898 [44]. Of the four fundamental forces known today, only the gravitational force is not incorporated into the Standard Model. We briefly discuss this and other shortcomings of the Standard Model at the end of this section.

The fundamental particles of the Standard Model (SM) that exist in the present-day universe are summarized in Figure 1.1. This includes both stable matter (such as electrons) and particles that only exist briefly after high-energy particle collisions (such as Higgs bosons). The fundamental constituents of the universe can be broken up into two broad categories: "fermions" that have half-integer spin (in particular, spin-$\frac{1}{2}$) and "bosons" that have integer spin. Fermions include both "leptons" and "quarks" – collectively these massive particles account for all visible matter in the universe. Leptons include the electrically-charged electron (e) and its heavier counterparts, the muon ($\mu$) and the tau particle ($\tau$), as well as their corresponding electrically-neutral analogues, the neutrinos ($\nu_{\mathrm{e}}$, $\nu_\mu$, and $\nu_\tau$). The original formulation of the Standard Model identified neutrinos as massless particles, but attempts have been made to extend the SM in order to allow for massive neutrinos [45], as is necessary given the observation of neutrino oscillations. The current upper bounds on the masses are listed in Figure 1.1. Quarks, of which hadrons such as protons and neutrons are composed, come in six different "flavors" or varieties: up (u), down (d), strange (s), charm (c), bottom (b), and top (t). Bosons are force-mediating particles and include the photon ($\gamma$), the W boson (W), the Z boson (Z), and the gluon (g) – of these particles, only the W boson is electrically charged, and all have spin-1. The electromagnetic force is carried by the massless photon, the weak force by the massive W and Z bosons, and the strong force by the massless gluon. It has been suggested that the gravitational force is mediated by a hypothetical spin-2 particle called the graviton, but as quantum gravity has not yet been satisfactorally unified with the other forces, the graviton is not included in the Standard Model. The last particle of the SM is the Higgs boson (H), a massive spin-0 particle that is the quantum of a field involved in electroweak symmetry breaking (see below).

While the fundamental particle spectrum described above and presented in Figure 1.1 holds for the present-day universe, in the very early universe (within the first $10^{-12}$ s after the Big Bang) when temperatures exceeded $10^{15}$ K, the picture was slightly different. In this epoch, all particles were massless and the electromagnetic and weak forces we observe today were unified into one "electroweak" force. Aside from mass, the properties of the fundamental particles were very similar to those we observe today (with possible differences due only to new physical processes beyond the SM). However, the symmetries observed by nature were different. We explain this below in the context of quantized gauge theories.

A quantized gauge theory is a type of quantum field theory in which the Lagrangian is invariant under a continuous group of local transformations [46], with the additional consequence that the interaction described by the theory is mediated by a massless vector boson. The local transformations, called gauge transformations, reflect the redundancy in the description of a physical system in terms of the Lagrangian, from which all particle dynamics can be determined. An example is quantum electrodynamics (QED) [47–51], a quantized abelian gauge theory describing electromagnetism in which the massless

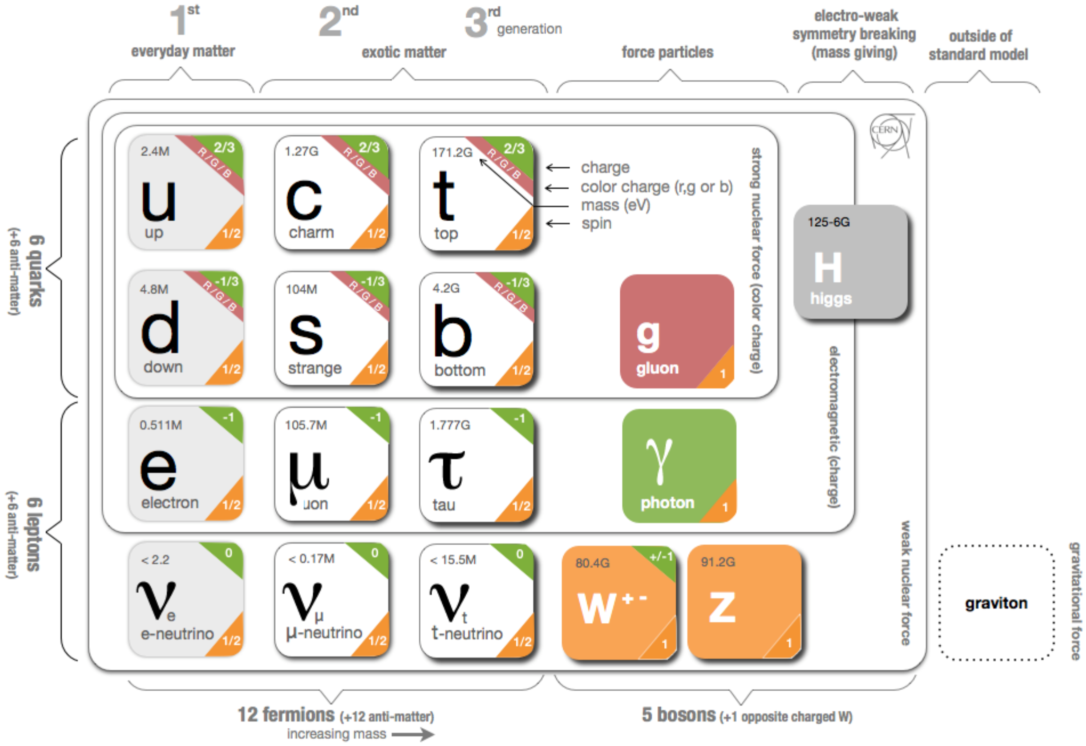

Figure 1.1: List of all fundamental particles that can be found in the present-day universe [1]. The particles are sectioned off by which forces they are affected by: neutrinos feel only the weak force, electrically-charged leptons feel both the electromagnetic and weak forces, and quarks feel all three (electromagnetic, weak, and strong forces). The massive Higgs boson (H) is a product of electroweak symmetry breaking and is explained in more detail in Sections 1.2 and 1.3.

photon carries the electromagnetic force. Representing the fundamental interactions of the universe in terms of quantized gauge theories has many advantages, as requiring a theory to be invariant under local gauge transformations greatly constrains the form of the theory, elucidates subtle symmetries of the system, and predicts the existence of additional particles necessary to ensure gauge invariance of the system.

The most impressive feature of local gauge invariance is its remarkable success in helping build a consistent formulation of the Standard Model with much predictive power. In addition to being incorporated into QED as described above with a U(1) gauge group describing the underlying symmetry, it was successfully integrated into a theory of the strong interaction, quantum chromodynamics (QCD) [52–54]. This formulation makes use of Yang-Mills theory [55] to describe the strong interaction in terms of a non-abelian gauge theory. QCD was developed after deep inelastic scattering (DIS) experiments [56] involving scattering electrons off of protons revealed a substructure of the proton consistent with the earlier proposed quark model [57, 58], which was developed to explain the experimentally observed spectrum of hadrons. The charge of QCD is referred to as "color" (either red, green, or blue) and reflects an internal symmetry with a SU(3) gauge group. The strong force is mediated by the massless gluon, which is itself color-charged, and requires the confinement of quarks into hadrons. Together, the independent symmetries offered by QED and QCD describe a complete picture of the gauge symmetries held in the *present-day* universe, with

$$\mathrm{SU}(3)_{\mathrm{C}} \times \mathrm{U}(1)_{\mathrm{EM}} \tag{1.1}$$

representing the total symmetry group. Here, "C" refers to the "color interaction" or strong interaction described by QCD and "EM" refers to the electromagnetic interaction described by QED. Notably, the weak interaction is not included in this gauge group. Even before QCD fully matured as a gauge theory describing the strong interaction, progress had already begun in preserving local gauge invariance in the weak sector with implications for symmetries in the *early* universe, as is discussed below.

Efforts to reformulate the weak interaction as a gauge theory, as for QED and QCD, is problematic as the W and Z bosons that mediate the force are massive. This was expected even before the mass of the W and Z bosons were measured at collider experiments due to the short-range nature of the weak interaction as observed from radioactive nuclear decay. In order to introduce gauge invariance in the weak sector and allow for massive force carriers one had to invoke the Higgs mechanism [59–61], a form of spontaneous symmetry breaking that produces massive particles (instead of massless Goldstone bosons) due to the coupling of a complex-valued scalar field doublet called the "Higgs field" to the gauge fields of a symmetry group. The Higgs mechanism is described in more detail in Section 1.2. This mechanism allowed for the construction of a gauge theory describing the electroweak force, a unification of the electromagnetic and weak forces that existed in the early universe, with a corresponding gauge group of SU(2) × U(1) [62–64]. Thus we have

$$\mathrm{SU(3)_C} \times \mathrm{SU(2)_L} \times \mathrm{U(1)_Y} \tag{1.2}$$

as the total symmetry group of the "unbroken" Standard Model Lagrangian, which we often refer to as *the* Standard Model. Here, "L" and "Y" refer to weak isospin and weak hypercharge, which are the charges of the SU(2) and U(1) components of the electroweak gauge group, respectively.

Weak isospin and weak hypercharge are no longer explicitly conserved by the vacuum after electroweak symmetry breaking, as interactions with the Higgs condensate (formed upon the Higgs field obtaining a non-zero vacuum expectation value) lead to the fermions not having well-defined weak isospin or weak hypercharge. However, the quantities are still conserved in particle interactions and are used to describe the weak interaction after electroweak symmetry breaking. Note that this interaction of the Higgs condensate with the fermions is also responsible for giving fermions mass while the W and Z bosons acquire their mass directly from the Higgs mechanism. We often say that the $\mathrm{SU(3)_C} \times \mathrm{SU(2)_L} \times \mathrm{U(1)_Y}$ symmetry group of the Standard Model was "broken" down to a lower symmetry group of $\mathrm{SU(3)_C} \times \mathrm{U(1)_{EM}}$ by electroweak symmetry breaking as the universe cooled. However, a better term may be "hidden" as the symmetries are still obeyed by the Lagrangian, yet no longer explicitly observed by the ground state of the vacuum in which the Higgs field has a non-zero vacuum expectation value.

We end this section with a discussion on some of the shortcomings of the Standard Model. While the SM is very good at describing three of the four principal interactions that we observe today, it is unable to account for the gravitational force in the sense that it cannot describe interactions at the Planck scale (roughly $1.22 \times 10^{19}$ GeV), at which point the quantum effects of gravity are too strong to ignore. String theory [65] attempts to explain these effects though it is unable to be tested by current experiments due to the energy scale being inaccessible by several orders of magnitude. The SM also does not address the hierarchy problem, the large discrepancy in strength of the different forces with the gravitational force being much weaker than the others, leading to a discrepancy in the cosmological constant of the universe with respect to expectations [66, 67]. Also left unexplained is the naturalness problem: there should be extremely large quantum corrections to the Higgs boson mass unless there are cancellations from e.g. supersymmetry (SUSY) [68] or the presence of extra dimensions [69, 70]. Additionally, the Standard Model does not account for the merging of the strong force with the electroweak force, which may have occurred in the early universe (until roughly $10^{-36}$ s after the Big Bang). Grand unified theories (GUT's) [71, 72] attempt to explain this feature of the early universe. Dark matter is still left unexplained at present [73], as is dark energy [74, 75], which account for roughly 27% and 68% of the energy density of the universe, respectively [76]. Finally, the matter-antimatter asymmetry in the universe [77], neutrinos being massive [22, 45], and the existence of exactly three generations of fermions [78] are all open questions that require physics beyond the Standard Model to explain.

## 1.2 Higgs Mechanism

We give a simplified introduction to the Higgs mechanism that is responsible for electroweak symmetry breaking in the Standard Model – for a more comprehensive discussion, please refer to [79, 80]. As discussed in Section 1.1, the Higgs mechanism is introduced in the SM in order to reformulate the weak interaction as a gauge theory, in the process unifying the electromagnetic and weak forces into one electroweak force.

These forces separate as the universe cools and the Higgs field acquires a vacuum expectation value, with the W and Z bosons gaining mass as electroweak gauge symmetry is broken. The properties of the vacuum state dictate the physics, as is seen below.

We begin by presuming the existence of a new two-component complex scalar field, the Higgs field. The Higgs field $\hat{\phi}$ is a SU(2) doublet

$$\hat{\phi} = \begin{pmatrix} \hat{\phi}^+ \\ \hat{\phi}^0 \end{pmatrix} \tag{1.3}$$

with four degrees of freedom in total (the real and imaginary parts of each component of the doublet). The effective Higgs potential, illustrated in Figure 1.2, is given by

$$V(\hat{\phi}) = \frac{\lambda}{4}(\hat{\phi}^\dagger\hat{\phi})^2 - \mu^2\hat{\phi}^\dagger\hat{\phi} \tag{1.4}$$

with $\lambda, \mu^2 > 0$. We assume the following vacuum expectation value for the Higgs field (in unitary gauge):

$$\langle 0| \hat{\phi} |0\rangle = \begin{pmatrix} 0 \\ v/\sqrt{2} \end{pmatrix}. \tag{1.5}$$

Here, $v/\sqrt{2}$ is the tree-level Higgs vacuum value. We then parameterize fluctuations about this value as

$$\hat{\phi} = \begin{pmatrix} 0 \\ \frac{1}{\sqrt{2}}(v + \hat{H}) \end{pmatrix} \tag{1.6}$$

where $\hat{H}$ is the *physical* Higgs field with $\langle 0| \hat{H} |0\rangle = 0$. In QFT, particles are expressed as excitations from a ground state, which is the vacuum. Accordingly, the expansion of the Higgs field as done in Equation 1.6 yields a simple interpretation in terms of emergent particles. This is easily seen in Figure 1.2, where the blue ball represents the value of $V(\hat{\phi})$ (with an additional factor of $i$) as the universe cools. In the broken phase, $V(\hat{\phi})$ is at a local minimum. In this picture, using the parametrization of Equation 1.6, $\hat{H}$ represents a massive particle oscillating "radially" while the "angular mode" represents what would normally be a massless Goldstone boson. This mode and the other two degrees of freedom (from the first component of the Higgs field) will instead be used to give the W and Z bosons mass due to the Higgs field $\hat{\phi}$ being coupled to gauge fields.

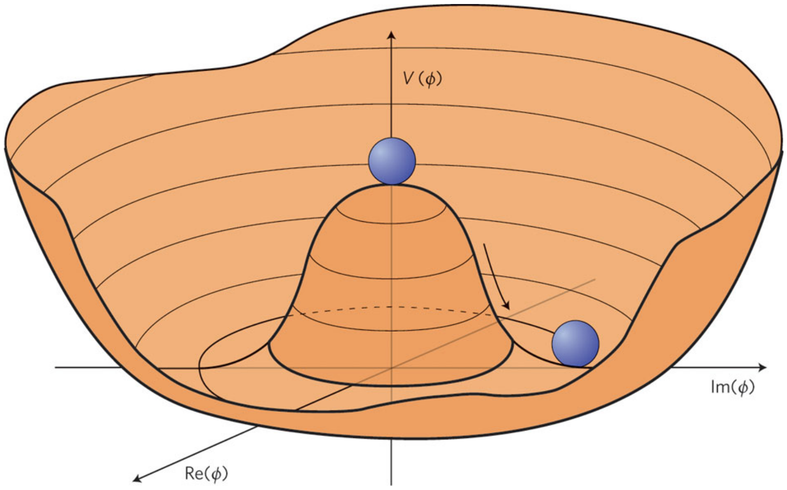

Figure 1.2: Effective potential of the Higgs field [2]. For illustration purposes, only one component of the SU(2) doublet is shown (real and imaginary parts). The blue ball represents the evolution of the ground state as the universe cools and the Higgs field acquires a vacuum expectation value.

Now we construct a gauge theory with massless gauge fields that respects a SU(2) × U(1) local gauge symmetry. Three massless gauge fields $\hat{\boldsymbol{W}}^{\mu}$ are associated with the SU(2) gauge group and one massless gauge field $\hat{B}^{\mu}$ is associated with the U(1) gauge group. The Lagrangian for the sector consisting of the gauge fields and the Higgs fields is

$$\hat{\mathcal{L}}_{\mathrm{G}\Phi} = (\hat{D}_{\mu}\hat{\phi})^{\dagger}(\hat{D}^{\mu}\hat{\phi}) - \frac{\lambda}{4}(\hat{\phi}^{\dagger}\hat{\phi})^{2} + \mu^{2}\hat{\phi}^{\dagger}\hat{\phi} - \frac{1}{4}\hat{\boldsymbol{F}}_{\mu\nu}\cdot\hat{\boldsymbol{F}}^{\mu\nu} - \frac{1}{4}\hat{G}_{\mu\nu}\hat{G}^{\mu\nu} \tag{1.7}$$

where $\hat{\boldsymbol{F}}_{\mu\nu}$ is the SU(2) field strength tensor for the gauge fields $\hat{\boldsymbol{W}}^{\mu}$, $\hat{G}_{\mu\nu}$ is the U(1) field strength tensor for the gauge field $\hat{B}^{\mu}$, and the covariant derivative $\hat{D}^{\mu}\hat{\phi}$ is given by

$$\hat{D}^{\mu}\hat{\phi} = (\partial^{\mu} + ig\boldsymbol{\tau}\cdot\hat{\boldsymbol{W}}^{\mu}/2 + ig'\hat{B}^{\mu}/2)\hat{\phi} \tag{1.8}$$

where $g$ and $g'$ are the coupling strengths of the interactions represented by the SU(2) and U(1) local gauge symmetries, respectively, and $\boldsymbol{\tau}$ are the Pauli matrices. After symmetry breaking, i.e. once the Higgs field has shifted to the ground state represented by Equation 1.6, the quadratic parts of Equation 1.7 can be written in unitary gauge as

$$\begin{aligned}\hat{\mathcal{L}}^{\text{broken}}_{\text{G}\Phi} = {} & \frac{1}{2}\partial_\mu \hat{H} \partial^\mu \hat{H} - \mu^2 \hat{H}^2 \\ & - \frac{1}{4}(\partial_\mu \hat{W}_{1\nu} - \partial_\nu \hat{W}_{1\mu})(\partial^\mu \hat{W}_1^\nu - \partial^\nu \hat{W}_1^\mu) + \frac{1}{8} g^2 v^2 \hat{W}_{1\mu} \hat{W}_1^\mu \\ & - \frac{1}{4}(\partial_\mu \hat{W}_{2\nu} - \partial_\nu \hat{W}_{2\mu})(\partial^\mu \hat{W}_2^\nu - \partial^\nu \hat{W}_2^\mu) + \frac{1}{8} g^2 v^2 \hat{W}_{2\mu} \hat{W}_2^\mu \\ & - \frac{1}{4}(\partial_\mu \hat{Z}_\nu - \partial_\nu \hat{Z}_\mu)(\partial^\mu \hat{Z}^\nu - \partial^\nu \hat{Z}^\mu) + \frac{v^2}{8}(g^2 + g'^2)\hat{Z}_\mu \hat{Z}^\mu \\ & - \frac{1}{4}\hat{F}_{\mu\nu}\hat{F}^{\mu\nu}\end{aligned} \tag{1.9}$$

where

$$\hat{Z}^\mu = \cos\theta_{\text{W}} \hat{W}_3^\mu - \sin\theta_{\text{W}} \hat{B}^\mu \tag{1.10}$$

$$\hat{A}^\mu = \sin\theta_{\text{W}} \hat{W}_3^\mu + \cos\theta_{\text{W}} \hat{B}^\mu \tag{1.11}$$

and

$$\hat{F}^{\mu\nu} = \partial^\mu \hat{A}^\nu - \partial^\nu \hat{A}^\mu \tag{1.12}$$

with

$$\cos\theta_{\text{W}} = g/(g^2 + g'^2)^{1/2} \qquad \sin\theta_{\text{W}} = g'/(g^2 + g'^2)^{1/2}. \tag{1.13}$$

It is relatively straightforward to interpret the broken phase Lagrangian in the case of electroweak symmetry breaking, as we discuss below.

We now associate the SU(2) and U(1) gauge groups discussed above to the symmetry group of the electroweak interaction in the unbroken phase, $\text{SU}(2)_{\text{L}} \times \text{U}(1)_{\text{Y}}$. With this in mind, we can interpret Equation 1.9 line-by-line. The first line tells us that the mass of the Higgs boson, the quantum of the physical Higgs field $\hat{H}$, is

$$m_{\text{H}} = \sqrt{2}\mu = \sqrt{2\lambda} v. \tag{1.14}$$

The second and third lines show that the charged W bosons, represented by $W_1^\mu$ and $W_2^\mu$, have a mass

$$M_{\text{W}} = gv/2 \tag{1.15}$$

where $g$ is now interpreted as the coupling strength of the charged weak current. The fourth line gives the mass of the Z boson, represented by $Z^\mu$, as

$$M_{\text{Z}} = M_{\text{W}}/\cos\theta_{\text{W}} \tag{1.16}$$

and the fifth line shows that the $A^\mu$ field describes a massless particle, which we identify with the photon. Thus, by coupling to gauge fields, the four degrees of the Higgs field $\hat{\phi}$ turn into masses for the three weak vector bosons (two W bosons and one Z boson) and one new particle, the Higgs boson, after spontaneous symmetry breaking.

The observation of the Higgs boson discussed at the beginning of the chapter validates the Higgs mechanism as the source of electroweak symmetry breaking and as the mechanism responsible for the heavy vector bosons associated with the weak force acquiring mass. This last ingredient of the Standard Model has been sought for over four decades in collider experiments, beginning with the Large Electron-Positron (LEP) collider [81], continuing at the Tevatron [27, 28], and at last ending with the discovery at the LHC primarily using the bosonic decay modes of the Higgs boson [23, 24]. In the next section we discuss an overview of the search at the CMS detector for $\mathrm{H} \to \mathrm{b\bar{b}}$ decays, one of the least experimentally constrained decays of the Higgs boson.

## 1.3 VH($\mathrm{b\bar{b}}$) Analysis Strategy

The backbone of the Standard Model is the manner in which the weak gauge bosons become massive while the photon remains massless. It is widely believed that the most simple and elegant way for this to occur is through the Higgs mechanism described in Section 1.2. However, as we have seen, this mechanism predicts the production of an additional massive spin-0 particle, the Higgs boson, which must be experimentally observed. The discovery of the SM Higgs boson discussed at the beginning of this chapter was a monumental achievement for the Standard Model, but we must be sure that the Higgs boson decays to all possible final states predicted in the Standard Model with the appropriate branching fractions before we can claim that "the SM Higgs boson" has been found with certainty. We outline our procedure for searching for $\mathrm{H} \to \mathrm{b\bar{b}}$ decays, as of yet unobserved at any collider experiment, at the CMS detector below.

The principal production mechanisms for Higgs bosons at the LHC are shown in Figure 1.3. In addition, we show the cross sections associated with each production mechanism in Figure 1.4. The total production cross section for 125 GeV Higgs bosons at the LHC with $\sqrt{s} = 8$ TeV is 22 pb. While the gluon-gluon fusion production mechanism yields the highest cross section of any Higgs boson production process (roughly 19 pb for 125 GeV Higgs bosons), it has the unfortunate feature that there are no additional "tags" or easily identified particles in the signal event. This translates to very high levels of background in the event, especially from QCD processes with two or more jets (collectively termed "QCD multi-jet" processes). With the signal-to-background ratio too low from the overwhelming QCD multi-jet background, we instead choose in the VH production mechanism [82], also called "Higgs-strahlung" in analogy to Bremsstrahlung photon radiation from a decelerating charged particle, where "V" refers to either a W or Z boson. This production mechanism has a higher signal-to-background ratio than the other signal production processes, such as vector boson fusion (VBF) and $\mathrm{t\bar{t}H}$ production, and benefits from other advantages as well. First, by requiring leptonic decays of the W or Z boson produced in association with the Higgs boson, we can use simple and highly efficient triggers to retain signal events. Additionally, the W/Z boson and the Higgs boson are expected to be well-separated ("back-to-back") and found in the central part of the detector, so the acceptance rate should be relatively high. Finally, QCD multi-jet background levels should be low due to the additional leptonic activity in the event, or the presence of high missing transverse energy in the detector (see Section 3.4.5).

With these considerations in mind, in searching for $\mathrm{H \to b\bar{b}}$ decays, the best sensitivity can be obtained at the LHC in the $\mathrm{VH(b\bar{b})}$ mode. We note that despite the lower cross section of the VH production process (roughly 1 pb for 125 GeV Higgs bosons), the analysis still benefits from the high branching fraction of the $\mathrm{H \to b\bar{b}}$ decay as seen in Figure 1.4 (at $\sqrt{s} = 8$ TeV, this is roughly 57% for a 125 GeV Higgs boson). We look for $\mathrm{VH(b\bar{b})}$ events using a variety of different search channels, requiring only leptonic decays of the W or Z boson: Z(ee)H, $\mathrm{Z(\mu\mu)H}$, $\mathrm{Z(\nu\nu)H}$, $\mathrm{W(e\nu)H}$, $\mathrm{W(\mu\nu)H}$, and $\mathrm{W(\tau\nu)H}$ (with H decaying to a $\mathrm{b\bar{b}}$ pair in all cases). Occasionally we refer to $\mathrm{Z(\ell\ell)H}$ or $\mathrm{W(\ell\nu)H}$ to group together the electron and muon modes (the tau lepton is not included here). In the case of $\mathrm{W(\tau\nu)H}$, we require one-prong hadronic decays of the tau lepton (see Section 3.4.3).

The clean signal topology suggests a simple strategy for selecting signal events. We

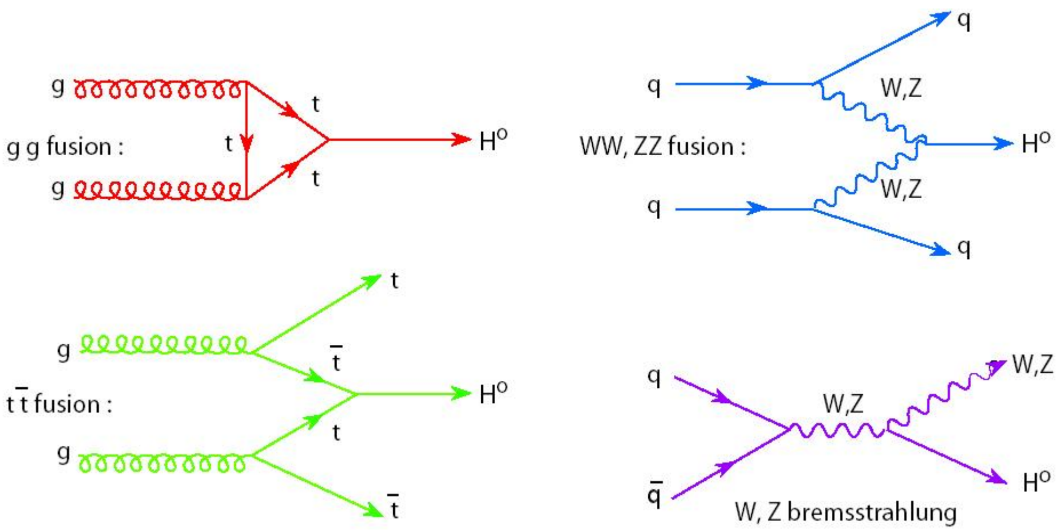

Figure 1.3: The various Higgs boson production mechanisms at the LHC [3]. Shown processes include gluon-gluon fusion (top left), vector boson fusion (top right), $t\bar{t}$H production (bottom left), and VH production (bottom right).

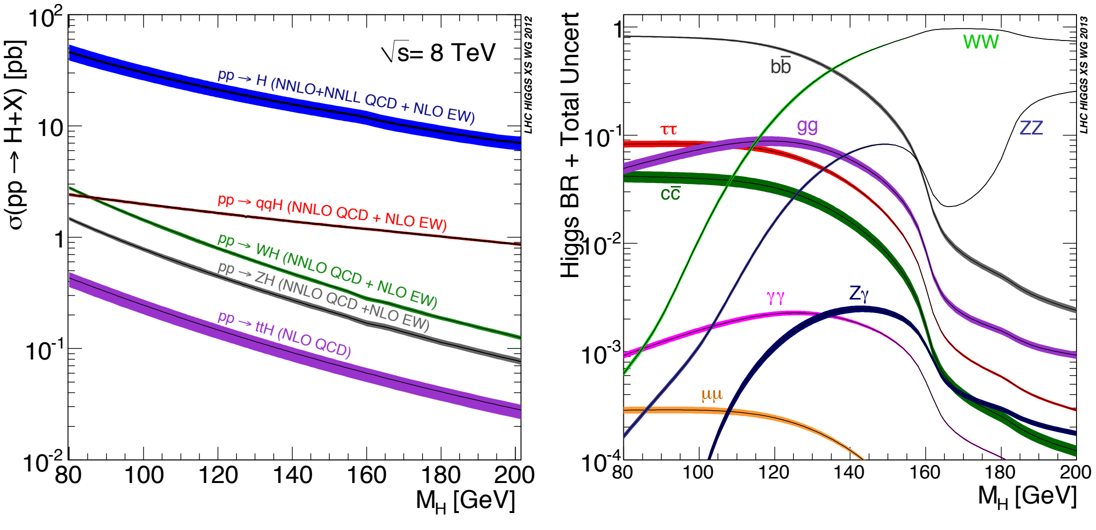

Figure 1.4: Production cross sections (left) and Higgs decay branching fractions (right) for SM Higgs bosons produced in proton-proton collisions at $\sqrt{s} = 8$ TeV [4]. We zoom in on the mass range most applicable for the VH($b\bar{b}$) analysis (what is referred to as a "low mass Higgs search").

use single or double lepton triggers to retain signal-like events for further analysis (with triggers making use of the missing transverse energy of the event, with possible additional jet activity, to tag Z($\nu\nu$)H events – see Section 4.3). We then look for a di-jet candidate associated with the H → $\mathrm{b\bar{b}}$ decay with both jets consistent with b quark production (referred to as "b-tagging"), as well as a vector boson candidate built from the leptonic activity and/or missing transverse energy in the event. These are required to be back-to-back in the detector, consistent with the signal topology described above. The final requirement we add to this baseline selection is that the Higgs boson candidate is "boosted" – i.e. it has considerable transverse momentum ($p_\mathrm{T}$) [82]. This requirement, which translates to significant $p_\mathrm{T}$ cuts on all physics objects in the event, greatly reduces background levels from a variety of different processes (such as the W+jets and Z+jets background) and improves the di-jet mass resolution. This increases the sensitivity of the analysis significantly.

This dissertation discusses the latest version of the VH($\mathrm{b\bar{b}}$) analysis carried out at CMS using the complete LHC dataset [41]. In this chapter, we have presented the theoretical motivation and analysis strategy for the search. In Chapter 2, we discuss the LHC accelerator and the CMS detector, giving a brief overview of each. Chapter 3 details the reconstruction of physics objects at CMS, which represent particle hypotheses using raw data collected by the detector. In Chapter 4, we discuss the datasets and simulation samples used in the analysis along with the triggers used to keep events in data for further analysis. Chapter 5 summarizes the selection cuts used to define the signal region in the VH($\mathrm{b\bar{b}}$) analysis. Chapter 6 highlights our use of a dedicated b-jet energy regression that is used to correct the energy of the Higgs boson candidate jets and consequently improve the di-jet mass resolution. In Chapter 7, we discuss the various background control samples used in the analysis, which are used to isolate and better understand the various background processes that contribute in the signal region. Chapter 8 details our methodology for separating out and extracting hypothesized signal events from background within the signal region, also going into detail on the various cross-check analyses used to validate the VH($\mathrm{b\bar{b}}$) search methodology. In Chapter 9, we present the results of the signal extraction, including the cross-check results, and finally, in Chapter 10, we summarize the conclusions of the CMS VH($b\bar{b}$) search.

# Chapter 2

# Experiment Overview

In order to recreate the conditions necessary to produce on-shell Higgs bosons, which were created in abundance in the early universe, we must make use of some of the most advanced technology available. This is necessary to accelerate particles to the energy required to produce Higgs bosons within the mass range of interest, as well as to have collisions occur often enough (i.e. at high enough instantaneous luminosity) in order to detect a statistically significant sample of Higgs bosons decaying to other known particles. The Large Hadron Collider (LHC) [83], a proton-proton collider based near Geneva, Switzerland and operating since 2009, was built with these considerations in mind. This machine is capable of accelerating individual protons to energies of up to 4 TeV (and eventually 7 TeV, the design energy) with instantaneous luminosity of up to $7 \times 10^{33}$ $\mathrm{cm^{-2}\,s^{-1}}$ (with a design instantaneous luminosity of $10^{34}$ $\mathrm{cm^{-2}\,s^{-1}}$). Such design parameters were used in order to enable the detection of the SM Higgs boson in the full mass range of interest within the lifetime of the experiment.

The proton-proton collisions must occur inside of an active detector in order for the decay products of the SM Higgs boson, or other hypothetical short-lived particles associated with previously unidentified physics processes, to be observed. The decay products, obtaining momentum transverse to the beam-line from the hard interaction of partons within the two colliding protons, are detected by interacting with the active detector elements that surround the interaction point. These interactions create digitized signals from the detector electronics that propagate to nearby computing mainframes that determine whether or not to keep the proton-proton collision event. If the event is kept, the raw information associated with the collision event is shipped to a computing cluster to be fully reconstructed into a collection of hypothesized particles (referred to as "physics objects" – these are described in Chapter 3), either immediately or at a later point in time. The fully reconstructed events are then analyzed by physicists, who attempt to search for new particles or characterize previously discovered particles using this collected data.

There are four principal particle detectors at the LHC: ALICE, LHCb, ATLAS, and CMS. The first two are experiments dedicated to exploring a narrow set of physics processes, in particular, ALICE (A Large Ion Collider Experiment) [84] principally seeks to better understand properties of strongly interacting matter at extreme energy densities using collision of lead ions, while LHCb (Large Hadron Collider beauty) [85] focuses on using proton-proton collisions to study flavor physics, one important aspect of which is to better understand the matter-antimatter asymmetry in the early universe via CP-violation in the bottom quark sector. The last two experiments, ATLAS (A Toroidal LHC ApparatuS) [86] and CMS (Compact Muon Solenoid) [8], are general-purpose detectors that were built primarily to search for the SM Higgs boson, but also to be flexible enough to explore a wider range of open questions in modern physics, many of which are discussed briefly in Chapter 1. The primary physics agenda of these two general-purpose experiments makes use of proton-proton collisions, though both collaborations analyze collisions of lead ions occurring within the respective detectors in order to aid ALICE in the quest to better understand the physics of highly dense and strongly interacting matter.

In this chapter, we discuss in detail the experimental setup used to collect the data analyzed within the CMS VH($\mathrm{b\bar{b}}$) search. We begin by discussing the Large Hadron Collider in Section 2.1. The various sub-detectors of the CMS experiment are explained in Section 2.2, starting from the innermost component and moving outward, ending with a discussion of the setup used to trigger on proton-proton collision events of interest.

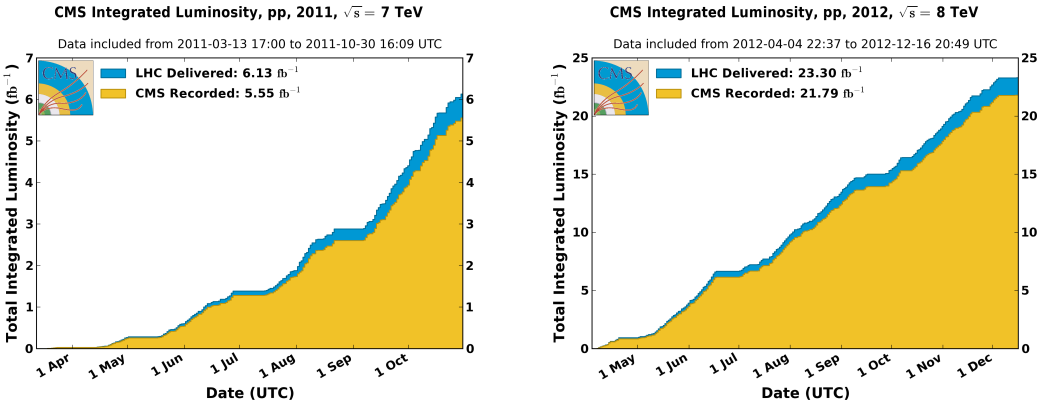

Figure 2.1: Total integrated luminosity delivered by the LHC and recorded at CMS in both 2011 (left) and 2012 (right) as a function of time [5].

## 2.1 Large Hadron Collider

The Large Hadron Collider is a two-ring superconducting synchrotron, built by the European Organization for Nuclear Research (CERN) near Geneva, Switzerland, capable of colliding either protons or lead ions. The machine, 27 km in circumference, lies in a tunnel about 100 m below the surface on both sides of the French-Swiss border. This tunnel, dug between 1983 and 1988, was originally used for the LEP collider that operated between 1989 and 2000. The LHC was first conceptualized in 1984 with construction beginning roughly a decade later. The construction of the collider and its associated detectors finished in 2008, with successful collisions of protons recorded for the first time in 2009. Due to a magnet quench during test runs in 2008, the machine has operated so far at energies lower than the design energy ($\sqrt{s} = 14$ TeV, or 7 TeV per proton) for proton-proton collisions. In the major data-taking runs of 2011 and 2012, roughly 5 fb$^{-1}$ of data was taken at $\sqrt{s} = 7$ TeV and roughly 20 fb$^{-1}$ at $\sqrt{s} = 8$ TeV, respectively. The integrated luminosity recorded by CMS is shown in Figure 2.1 as a function of time for both 2011 and 2012. The LHC will operate much closer to design energy beginning in the data-taking runs of 2015.

In comparison to LEP [87–90], an electron-positron collider, and the Tevatron [91–93], a proton-antiproton collider located at Fermilab in Batavia, Illinois, the LHC was chosen to collide beams of protons to search for the Higgs boson, the chief physics goal of the machine. As the power emitted (and thus lost) from a particle moving in a magnetic field due to synchrotron radiation is proportional to $1/m^4$, it is much easier to accelerate protons and antiprotons to high energies than it is for electrons or positrons. Proton-proton collisions are picked over proton-antiproton collisions largely because the high instantaneous luminosity desired at the LHC is very difficult to achieve with the latter setup, which is due to antiprotons being difficult to accumulate and store. Additionally, note that gluon-gluon fusion is the dominant partonic process at the center-of-mass energy of the LHC, independent of the use of either protons or antiprotons. As gluon-gluon fusion is a principal production mechanism for Higgs bosons, and because antiprotons are more difficult to produce at high instantaneous luminosities as discussed, the use of strictly proton beams at the LHC is a natural choice. The drawback of colliding hadrons is the high frequency of events from QCD multi-jet processes, and contamination from hadronic processes in the underlying event of the collision. These issues pose difficult challenges regarding event triggering and reconstruction for the LHC detectors.

Protons are grouped into bunches of roughly $1.5 \times 10^{11}$ within each of the two beams, which travel in opposite directions around the collider. The proton bunches undergo a series of acceleration steps before being injected into the main ring of the LHC, as shown in Figure 2.2. The protons are first accelerated in the linear accelerator (LINAC) and injected into the Proton Synchrotron Booster (PSB), also referred to simply as the Booster, to obtain kinetic energies of roughly 1.4 GeV. The protons are then injected into the Proton Synchrotron (PS) to be arranged into bunches spaced by either 25 ns or 50 ns in flight time, reaching energies of roughly 25 GeV. Next the protons are injected into the Super Proton Synchrotron (SPS) where energies of roughly 450 GeV are achieved for each proton. Finally, the proton bunches are injected into the LHC beam pipe. Once the LHC beam contains the desired number of bunches, the proton bunches are brought to maximal energy via accelerating radio-frequency cavities. The two beams are forced to collide within only four sections around the ring so that proton-proton collisions may occur in the center of one of the four detectors discussed above: ALICE, LHCb, ATLAS, and CMS. The organization of these experiments around the LHC ring is illustrated in Figure 2.3.

The LHC steers the proton beams using a series of 1,232 dipole magnets, which can produce magnetic fields upward of 8 T. For the lower energy runs in 2012 ($\sqrt{s} = 8$ TeV), the magnets were operating at roughly 5.5 T. An additional 392 quadrupole magnets are used to keep the proton beams as focused as possible in order to reduce proton-proton interactions within the same beam and to keep the beam from interacting with the beam pipe. Just prior to collision, yet another type of magnet is used to "squeeze" the protons closer together in order to maximize the instantaneous luminosity of the machine. The large magnetic fields necessary to steer and squeeze the beams in the LHC beam pipe require the use of superconducting electromagnets made of copper-clad niobium-titanium cables, operating at currents of up to 12 kA. These electromagnets must be cooled down to temperatures of roughly 1.9 K using liquid helium in order to achieve superconductivity.

While proton-proton collisions are the main focus of the physics program at the LHC, roughly one month a year is dedicated to the analysis of lead ion collisions in order to study properties of quark-gluon plasma (QGP) [94–96], an extremely hot and dense phase of quantum chromodynamics that is hypothesized to have existed in the first few milliseconds after the Big Bang. In this effort, lead ion collisions are studied at ALICE as well as both ATLAS and CMS. At design conditions, the lead ion beams attain an energy of 2.76 TeV per nucleon and an instantaneous luminosity of up to $10^{27}$ cm$^{-2}$ s$^{-1}$.

## 2.2 Compact Muon Solenoid

The Compact Muon Solenoid is one of two general-purpose detectors at the Large Hadron Collider. Along with the ATLAS detector, CMS was built primarily to search for the Standard Model Higgs boson in the full mass range allowed by theoretical predictions (up to roughly 1 TeV) by using proton-proton collisions. As a general-purpose experiment, the CMS detector is also well-suited to search for other signatures of new physics

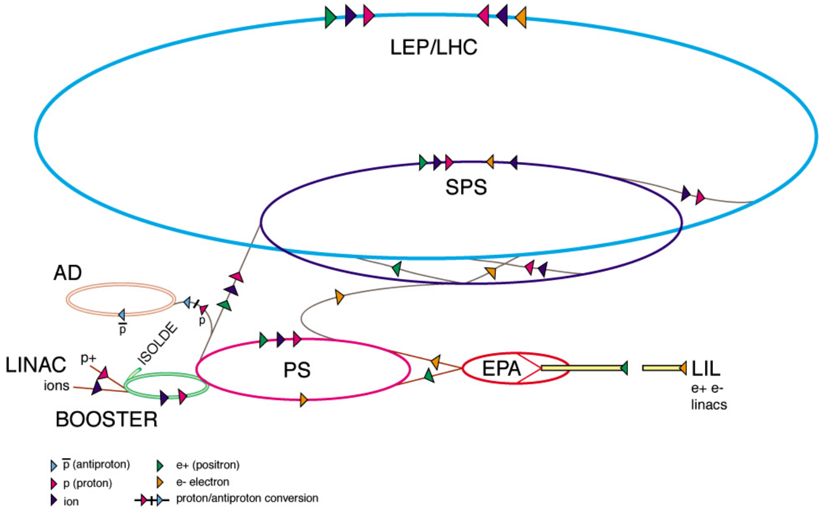

Figure 2.2: The LHC accelerator complex [6]. Individual protons are accelerated up to energies of 7 TeV (LHC design energy) in the largest tunnel.

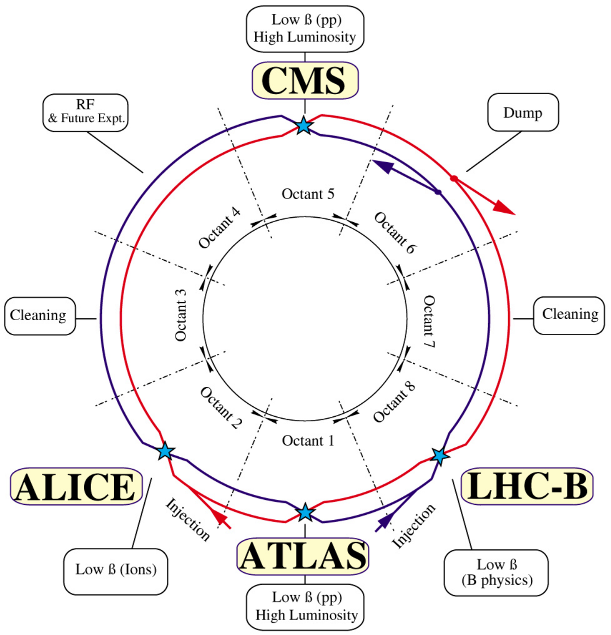

Figure 2.3: The organization of the various detectors with respect to the LHC interaction points [7]. Only four octants are used for experiments at present, with the others used for beam-cleaning or dumping of the beam.

discussed in Chapter 1, including supersymmetry, dark matter, large extra dimensions, black holes, and leptoquarks, as well as to further constrain the properties of the previously discovered Standard Model particles. The powerful tracking sub-detectors and muon system of CMS make this an especially successful enterprise, with the former enabling robust particle identification via the CMS Particle Flow algorithm (see Section 3.3) and the latter (after which the CMS detector is named) providing precise characterization of some of the cleanest final states associated with both new physics signals and SM processes – those containing muons of moderate to high $p_{\mathrm{T}}$.

We describe the CMS detector, illustrated in Figure 2.4, one sub-detector at a time, starting with the sub-detector located closest to the beam line and moving outward. Thus, we begin with the silicon tracker in Section 2.2.1. Next the electromagnetic and hadronic calorimeters are described in Sections 2.2.2 and 2.2.3, respectively. The solenoid magnet that generates the magnetic field in CMS, necessary for the tracker and muon system particle $p_{\mathrm{T}}$ measurements, is discussed in Section 2.2.4. The outermost sub-detector, the muon system, is described in Section 2.2.5. Finally, in Section 2.2.6 we discuss the trigger system (both hardware and software) that is used at CMS to keep interesting proton-proton collision events for further analysis.

As the CMS detector is an exceptionally complex piece of hardware, the discussion in this section is limited, and one should refer to [8, 9] for more detail. The coordinate system used at CMS, and in the following discussion, is as follows: the $x$-axis points radially inward toward the LHC center, the $y$-axis points upward away from the center of the Earth, and the $z$-axis points along the clockwise beam direction, looking from above. The origin is set at the very center of the detector, on the beam line. Translating to cylindrical coordinates, the radial distance $r$ is measured from the center of the beam line, the azimuthal angle $\phi$ is measured in the $x-y$ plane from the $x$-axis, and the polar angle $\theta$ is measured from the $z$-axis. Instead of $\theta$, physicists at collider experiments typically use the “pseudorapidity” $\eta$:

$$\eta = -\ln\left(\tan\frac{\theta}{2}\right). \tag{2.1}$$

The pseudorapidity is used instead of the polar angle as pseudorapidity differences

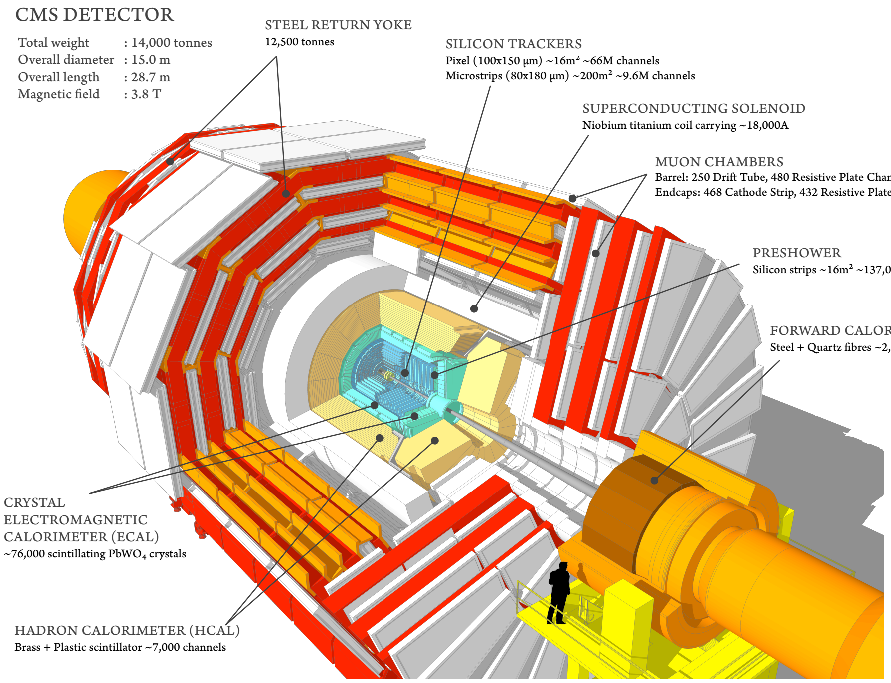
CMS DETECTOR
Total weight : 14,000 tonnes
Overall diameter : 15.0 m
Overall length : 28.7 m
Magnetic field : 3.8 T
STEEL RETURN YOKE
12,500 tonnes
SILICON TRACKERS
Pixel (100x150 μm) ~16m² ~66M channels
Microstrips (80x180 μm) ~200m² ~9.6M channels
SUPERCONDUCTING SOLENOID
Niobium titanium coil carrying ~18,000A
MUON CHAMBERS
Barrel: 250 Drift Tube, 480 Resistive Plate Chambers
Endcaps: 468 Cathode Strip, 432 Resistive Plate Chambers
PRESHOWER
Silicon strips ~16m² ~137,000 channels
FORWARD CALORIMETER
Steel + Quartz fibres ~2,000 Channels
CRYSTAL ELECTROMAGNETIC CALORIMETER (ECAL)
~76,000 scintillating PbWO₄ crystals
HADRON CALORIMETER (HCAL)
Brass + Plastic scintillator ~7,000 channels

Figure 2.4: Layout of the CMS detector. For a view of a central slice of the CMS detector in the $r - \phi$ plane, please refer to Figure 3.3.

(e.g. $\Delta\eta$ between the jets associated to the two bottom quarks from a $\mathrm{H} \to \mathrm{b\bar{b}}$ decay) are Lorentz invariant in the relativistic limit. The central CMS barrel occupies lower values of pseudorapidity, while the two endcaps are located at higher values of pseudorapidity.

### 2.2.1 Silicon Tracker

The innermost sub-detector of the CMS experiment is the silicon tracker [8, 9, 97], a system that uses semiconductor detectors (silicon pixels and strips) to measure the effect of incident charged particles with the goal of reconstructing "tracks" that represent the trajectories of the particles. Individual interactions of charged particles with the silicon pixels or strips are referred to as "hits" and can be built into tracks using comprehensive reconstruction techniques, discussed in Chapter 3. The silicon tracker, which covers pseudorapidities of $|\eta| < 2.5$, is capable of measuring the $p_{\mathrm{T}}$ of a particle by measuring the radius of curvature of the particle's track, which is helical due to the presence of a

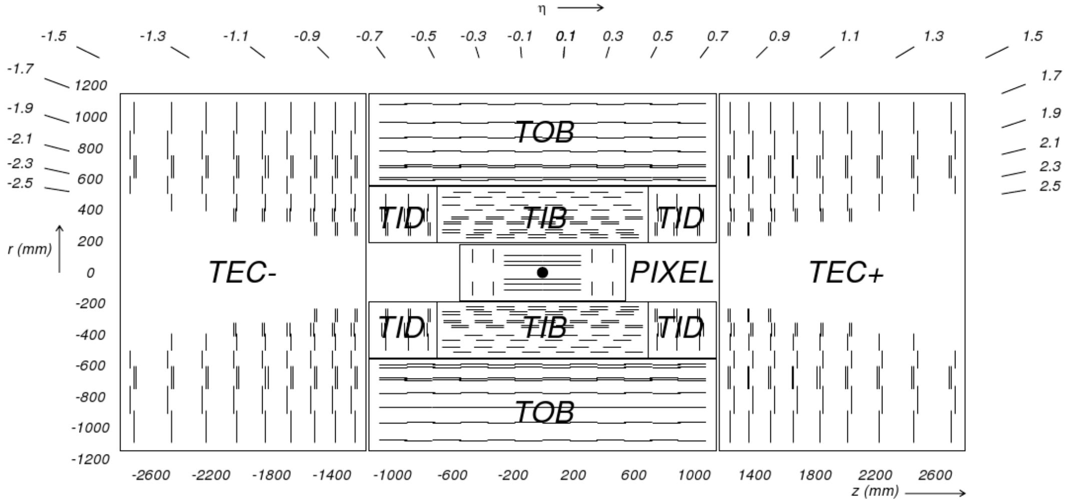

Figure 2.5: Schematic of the CMS tracker in the $r-z$ plane [8]. The lines in the strip detector indicate silicon strips, whereas lines in the pixel detector represent ladders and petals on which the silicon pixels are mounted in the barrel and endcaps, respectively.

large magnetic field provided by the solenoid magnet (discussed in Section 2.2.4).

Semiconductor detectors are both fast to read out (fast electrons from ionization currents) and very dense compared to gaseous detectors, the latter feature allowing charged particles to deposit energy in a relatively small volume. The great timing and spatial resolution of the semiconductor detectors allow the silicon tracker to reconstruct tracks with very good $p_{\mathrm{T}}$ resolution in busy detector environments resulting from proton-proton collisions at high instantaneous luminosity. Despite the excellent tracking performance, there are some drawbacks in using a silicon tracking system in collider experiments: the semiconductor detectors are very expensive and degrade over time due to radiation damage, requiring maintenance and/or replacement.

The silicon tracker is further divided up into two very different components based on both design and function: the pixel detector and the strip detector. The pixel detector is located closer to the beam line and is comprised of tightly packed silicon pixels with two-dimensional segmentation, while the larger strip detector is located outside of the pixel detector and is made out of silicon strips with one-dimensional segmentation. Both the pixel detector and strip detector are illustrated in Figure 2.5.

The pixel detector consists of approximately 66 million silicon pixels, each with an area of approximately $100 \times 150\ \mu\text{m}^2$. These are arranged in three layers in the barrel (at $r$ values of 4.4 cm, 7.3 cm, and 10.2 cm) and two layers in both forward regions of the detector (at $z$ values of $\pm 34.5$ cm and $\pm 46.5$ cm). This geometry yields coverage within the pseudorapidity range of $|\eta| < 2.5$. The pixel detector is used in seeding the iterative tracking done at CMS, helping to provide excellent tracking performance with minimal computation time. Furthermore, the pixel detector is essential in reconstructing primary vertices, reconstructing secondary vertices (arising from the decay of relatively long-lived particles), and measuring the displacement of tracks from those vertices. As a result, the pixel detector is key in b-tagging at CMS, which is discussed in Section 3.4.4.

The strip detector is composed of roughly 9.3 million silicon microstrips, forming an active area of roughly 198 $\text{m}^2$ and covering the same pseudorapidity range as the pixel detector ($|\eta| < 2.5$). The detector is divided into four distinct units: the Tracker Inner Barrel (TIB), the Tracker Inner Disks (TID), the Tracker Outer Barrel (TOB), and the Tracker EndCaps (TEC). The TIB and TID are located closest to the pixel detector, between in 20 cm and 55 cm in $r$. The TIB has four barrel layers, while the TID has three disks on each end of the TIB. The TOB extends to $r = 116$ cm and provides an additional six layers of silicon strips. The TEC has nine disks on each end of the TIB/TID/TOB, located between 124 cm and 282 cm in $z$. The pitch of the silicon strips varies between 80 $\mu$m and 190 $\mu$m, depending on the location in the tracker. The TIB and TOB make use of silicon strips parallel to $z$ while the TID and TEC use radial strips. Some of the layers make use of double-sided sensors in stereo, enabling measurement of two coordinates simultaneously. Together, the units of the strip detector provide the greatest utility in measuring track $p_\text{T}$, making use of hits in many different layers to tightly constrain the $p_\text{T}$ of each track. The track $p_\text{T}$ resolution from this measurement is roughly 0.5–2% for most of the relevant kinematic range, with less good track $p_\text{T}$ resolution (up to 5%) for low $p_\text{T}$ tracks (less than 1 GeV) at high pseudorapidities.

### 2.2.2 Electromagnetic Calorimeter

The next sub-detector is the electromagnetic calorimeter (ECAL) [8, 9, 98], located just outside of the silicon tracker. The CMS ECAL, which is primarily composed of lead tungstate ($PbWO_4$) crystals, is used to perform an energy measurement of the electromagnetic showers associated with impinging particles (and thus an energy measurement of the particles themselves). While most particles with significant transverse momentum ($p_T$ of 1 GeV or higher) make it through the silicon tracker, the short radiation length of the lead tungstate in the ECAL leads to photons and electrons being stopped within the volume of the sub-detector, dissipating their remaining energy in the form of electromagnetic showers. The large number of radiation lengths in the ECAL (roughly 25, averaging over pseudorapidity) allows for discrimination between electrons and jets of hadrons from the production of quarks or gluons in the detector. Hadrons such as charged pions and neutral pions also radiate energy in the ECAL in the form of electromagnetic showers but typically pass through to the hadronic calorimeter, described in the next section. This is because the number of interaction lengths in the ECAL is quite small – roughly 1.5 at maximum. The layout of the ECAL sub-detector is shown in Figure 2.6.

The CMS ECAL is divided into three units: the ECAL Barrel (EB), the ECAL Endcap (EE), and the ECAL preShower (ES). The EB and EE, covering pseudorapidities of $|\eta| < 1.479$ and $1.479 < |\eta| < 3.0$, respectively, are constructed out of lead tungstate scintillating crystals. In addition to stopping photons and electrons by virtue of having a short radiation length (0.89 cm), these radiation-hard crystals well confine an electromagnetic shower in the transverse direction due to having a small Moliere radius (2.19 cm), providing good spatial resolution when reconstructing an impinging particle. The EB is composed of more than 60 thousand individual crystals, each roughly $25 \times 25$ mm$^2$ in terms of cross-sectional area and 230 mm long, while the EE in total uses over 14 thousand crystals with dimensions of roughly $30 \times 30$ mm$^2$ in cross-sectional area and 220 mm in length. Groups of 25 crystals form "trigger towers" that contribute to the L1 trigger decision, as discussed in Section 2.2.6. The ES, a sampling calorimeter

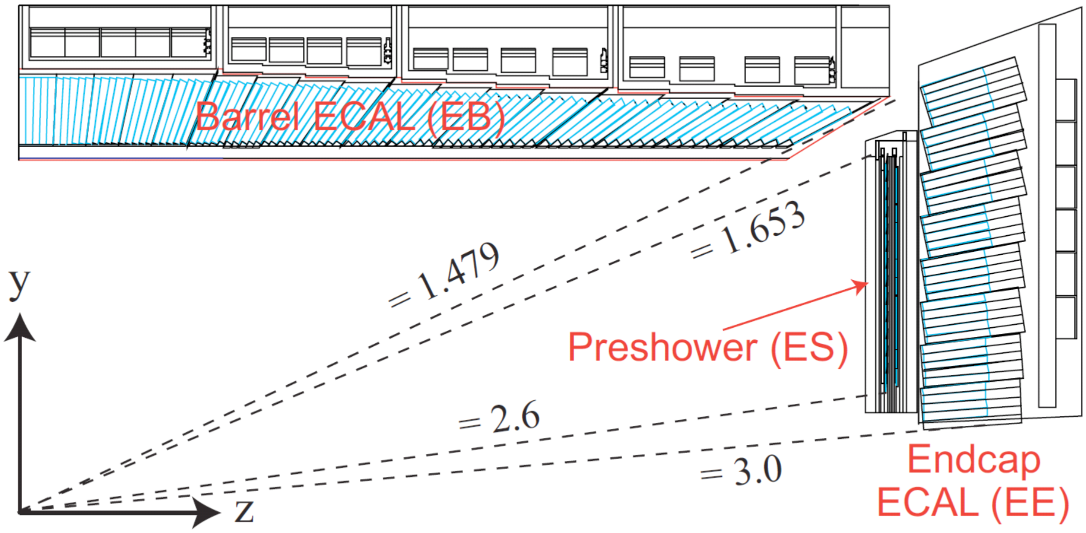

Figure 2.6: Layout of the CMS ECAL sub-detector as viewed in the $r - z$ plane [9]. Note the small gap between the EB and ES detectors where photon/electron identification is less good.

comprised of alternating layers of lead radiator and silicon strip sensors (two of each), provides additional coverage for pseudorapidities of $1.653 < |\eta| < 2.6$. This unit helps to discriminate between single photons and di-photons (e.g. from $\pi^0 \to \gamma\gamma$ or $\mathrm{H} \to \gamma\gamma$ decays) in the forward regions of the detector. The ECAL energy resolution for electrons and photons is roughly 0.5–3%, depending on both pseudorapidity and energy, for electron/photon energies greater than 10 GeV.

### 2.2.3 Hadronic Calorimeter

Located just outside of the ECAL is the hadronic calorimeter (HCAL) [8, 9, 99], a sampling calorimeter that primarily serves in making an energy measurement for hadrons that are either formed from jet fragmentation or are decay products of other particles. The HCAL provides a large number of interaction lengths to the total material budget of the detector, with more than 10 interaction lengths at higher values of $|\eta|$. This results in the majority of hadronic activity in the detector stopping within the HCAL, with hadron “punch-through” to the muon system relatively uncommon. A schematic of the

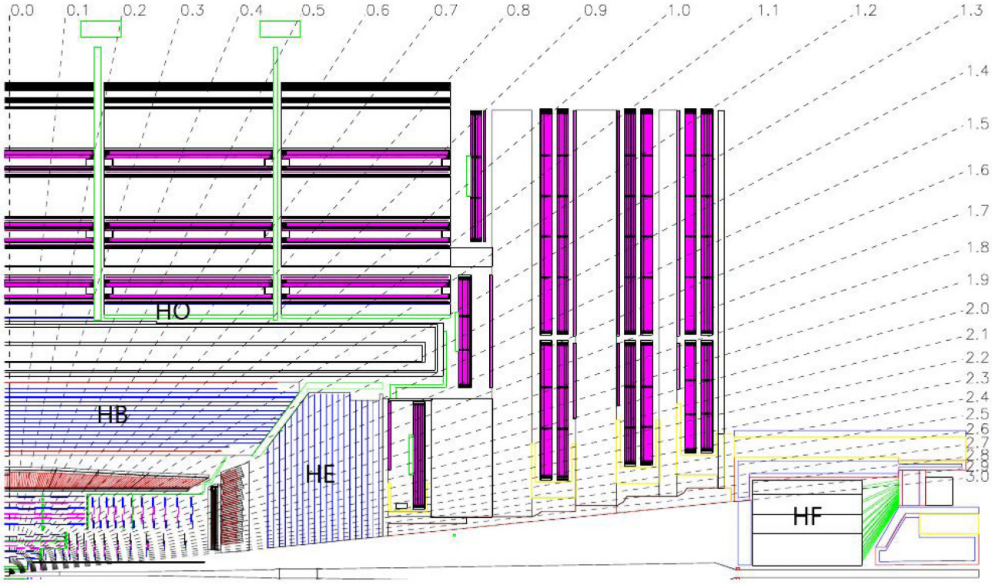

Figure 2.7: A view of one quarter of the CMS hadronic calorimeter in the $r - z$ plane, showing the positions of the various sub-detector units [8].

CMS HCAL sub-detector is displayed in Figure 2.7.

The CMS HCAL is primarily composed of alternating layers of plastic scintillator and brass absorber. The sub-detector is divided into four smaller units: the HCAL Barrel (HB), the HCAL Outer barrel (HO), the HCAL Endcap (HE), and the HCAL Forward calorimeter (HF). The HB and HO are located in the central part of CMS at $|\eta| < 1.3$, with the HE covering $1.3 < |\eta| < 3.0$ and the HF more forward at $3.0 < |\eta| < 5.0$. The HCAL is segmented into "towers" of $\Delta\eta \times \Delta\phi = 0.087 \times 5^\circ$ for the HB/HO and roughly $\Delta\eta \times \Delta\phi = 0.175 \times 10^\circ$ (on average) for the HE/HF, with the longitudinal depth segmentation ranging between one and three segments. The HB begins at roughly $r = 1.77$ m and extends to the beginning of the solenoid magnet at $r = 2.95$ m. The HO is actually located outside of the solenoid magnet and serves to catch potential punch-through hadrons that by chance make it through the HB. The HB, HO, and HE are composed of plastic scintillator and brass as described above (with only scintillator for the HO), while the HF instead makes use of quartz fiber and steel absorber. This is because the radiation dose is especially high in the forward regions of the detector, necessitating the use of radiation-hard quartz for the active medium. Unlike the other HCAL units, the HF is located far away from the interaction point as seen in Figure 2.7, beginning at $z$ values of roughly $\pm 11.2$ m. The energy resolution for hadronic jets, dominated by the HCAL contribution to the resolution, is at most 25% for jets produced at energies greater than 30 GeV. With the use of the CMS Particle Flow algorithm (see Section 3.3), which makes use of the tracking system to help measure the total jet energy, this number is closer to 15%.

### 2.2.4 Solenoid Magnet

Surrounding the HCAL (except for the HO) is the solenoid magnet of the CMS detector [8, 9, 100]. The CMS magnet is the world's largest superconducting magnet, being 12.5 m long and 3 m in radius. The solenoid magnet provides a nearly uniform magnetic field of 3.8 T inside of the magnet that must be present in order to measure the $p_{\mathrm{T}}$ of a particle track, as described in Section 2.2.1. The magnetic field is returned within the outer CMS muon system via the iron yoke that is interleaved between the active muon system components (see Section 2.2.5). The solenoid itself is made out of niobium-titanium that is mechanically reinforced with an aluminum alloy, and is placed in a cryostat cooled down to temperatures of 4.5 K in order to maintain superconductivity necessary to produce the large magnetic field. The produced magnetic field stores roughly 2.35 GJ of energy.

### 2.2.5 Muon System

The outermost sub-detector of the CMS detector is the muon system [8, 9, 101]. While most particles radiate the totality of their energy within one of the sub-detectors closer to the beam line via electromagnetic or nuclear interactions, muons, as minimum-ionizing particles, typically do not heavily interact with the materials of the other sub-detectors. Furthermore, as $c\tau = 658.6$ m for muons, the vast majority of the particles do not decay to electrons within the CMS detector volume. Thus, muons are expected to escape from the detector and can be detected in the relatively low radiation environment present

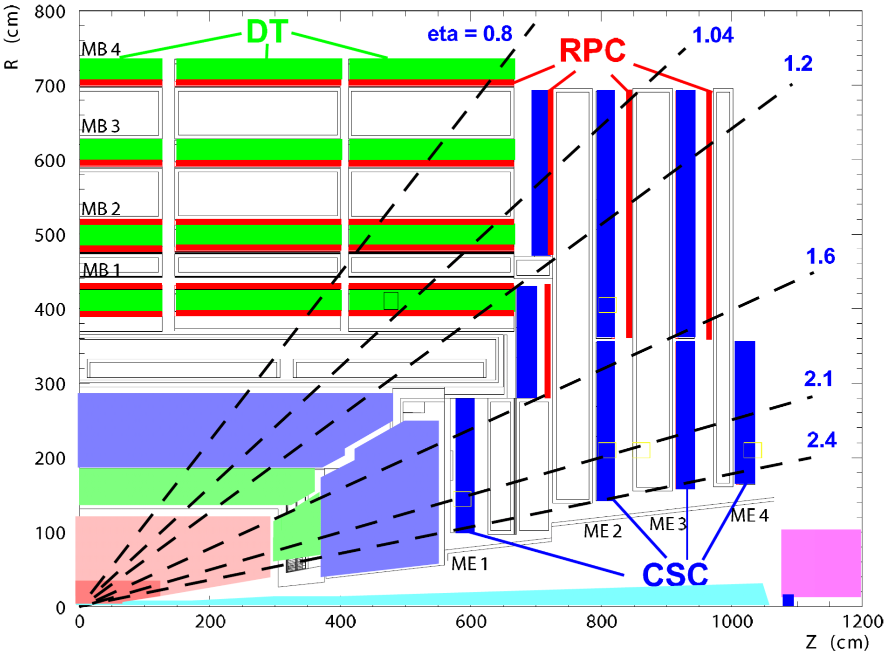

Figure 2.8: A cross sectional view of the CMS muon system, showing the organization of one quarter of the sub-detector in the $r - z$ plane [8].

in the outer muon system. Only very rarely does hadron punch-through lead to fake muons. The organization of the CMS muon system is shown in Figure 2.8.

The CMS muon system is constructed out of three different types of detectors: the Drift Tube (DT) detector, the Cathode Strip Chamber (CSC) detector, and the Resistive Plate Chamber (RPC) detector. The DT detector is located in the CMS barrel ($|\eta| < 1.2$), the CSC detector in the endcaps ($0.9 < |\eta| < 2.4$), and the RPC detector providing additional coverage in both the barrel and the endcaps ($|\eta| < 1.6$). The DT detectors are gas detectors split into drift cells with a cross sectional area of $13 \times 42$ mm$^2$. The drift cells are filled with a gas mixture of 85% Ar and 15% $CO_2$ at one atmosphere of pressure, with an electric field in the cell pushing ionization electrons, created by an impinging muon, toward anode wires for signal detection. The CSC detector is comprised of multi-wire proportional counters that function similarly to the DT detector drift cells, but with greater segmentation to handle the higher flux of particles in the forward regions of the detector. Each individual chamber of the CSC detector is constructed out of six planes of cathode strips and six planes of anode wires, with the strips ranging from 2.2 mrad to 4.7 mrad in terms of width. Finally, the RPC detector provides additional trigger information to the muon system in both the barrel and endcaps, with a very fast response time compared to the other detector types, though providing coarser spatial resolution for detected hits. The gas detectors of the RPC are double-gap chambers with common strip pick-ups.

Hits from all of the units of the muon system are combined into “standalone muon tracks” that are combined with information from the inner silicon tracker to fully reconstruct a muon at CMS, as described further in Section 3.1.2. The standalone muon tracks are combined with tracks from the inner tracking system to form “global muons” that are used in most CMS analyses that expect muons in the final state. For global muons with $p_{\mathrm{T}} < 100$ GeV, the $p_{\mathrm{T}}$ resolution is 1–2% in the barrel and less than 6% in the endcaps. The global muon $p_{\mathrm{T}}$ resolution is better than 10% in the barrel for muons up to 1 TeV in $p_{\mathrm{T}}$. It is more difficult to measure the transverse momentum of very high $p_{\mathrm{T}}$ muons as they do not bend significantly in the magnetic field within the detector volume.

### 2.2.6 Trigger

After particles interact with the CMS detector material, digitized signals are created via the front-end electronics associated with each detector component. These digitized signals are sent to nearby trigger racks, where they are processed on an event-by-event basis to check whether or not to keep the event to be fully reconstructed with the (time-consuming) algorithms described in Chapter 3. The CMS trigger system, discussed in great detail in [8, 9, 102, 103], is divided into two functionally different “levels” or components, the L1 trigger and the High-Level Trigger (HLT). The L1 trigger is a hardware trigger that takes the event rate of roughly 40 MHz (corresponding to 25 ns proton crossing interval) down to a maximum of 100 kHz, while the HLT is a software trigger that further reduces the rate down to the range of 100–1000 Hz.

The L1 trigger decision directly depends on two sub-units that combine digitized signals from different parts of the detector: the Global Muon Trigger (GMT) and the Global Calorimeter Trigger (GCT). The GMT makes use of the digital signals from hits in the various muon system components, which are first combined into track segments in a similar fashion to the offline reconstruction of standalone muon tracks for selected events (3.1.2). These track segments are then combined with information from the Regional Calorimeter Trigger (RCT) that looks for a minimizing-ionizing signature in the region of the calorimeter near each track segment. The GCT also makes use of the RCT, scanning $4 \times 4$ groups of trigger towers through the detector to build transverse energy sums associated with electron and photon candidates. The GCT then uses this information to find jets and primitively reconstruct the event $E_{\mathrm{T}}^{\mathrm{miss}}$ (see Section 3.4.5).

With an event selected as “interesting” by the L1 trigger due to the presence of digitized signals consistent with high event $E_{\mathrm{T}}^{\mathrm{miss}}$ or one or more muons, electrons/photons, or jets, the HLT is next run to make a final decision on whether or not to keep the event to be fully reconstructed offline. This software-level trigger is run on a computing farm containing roughly 10,000 processors, with each event taking at most 100–200 ms to be processed. The HLT is designed to allow for complex calculations making use of the entire readout of the detector, allowing for trigger decisions that are tailored to the particular topology of a signal of interest. Most often events are selected at the HLT by requiring one or more physics objects reconstructed at trigger-level to be above some $p_{\mathrm{T}}$ threshold. Frequently, additional requirements are placed on the angles between the physics objects or on the “isolation” of leptons (see Section 3.4). The particular trigger paths used in the VH($\mathrm{b\bar{b}}$) analysis are presented in Chapter 4 for both the L1 trigger and HLT.

# Chapter 3

# Physics Object Reconstruction

In this chapter, we describe the basic components of CMS physics analyses. These components are referred to as reconstructed physics objects, which are built out of more primitive detector-level information (calorimeter energy clusters, silicon tracker hits, muon system hits) and represent the kinematic information and quantum numbers of a particle hypothesis. This information can be partial, e.g. using the reconstructed transverse missing energy of an event to represent partial information of a hypothesized neutrino passing through the detector, or complete, e.g. using full tracking and calorimeter information to reconstruct the four-vector and charge of a hypothesized electron. These objects are then "cleaned" using a variety of cuts on tracking-related or calorimetry-related variables, in addition to loose cuts on kinematical variables. The entirety of the physics object reconstruction process is completed using the CMS software framework (CMSSW) [9] and the ROOT package [104], both of which make use of the C++ and Python programming languages.

Each physics analysis requires a specific combination of these reconstructed objects that corresponds to the final state of the physics signal of interest. In addition to requiring reconstructed physics objects built directly out of detector-level information, which we refer to as principal physics objects (e.g. a reconstructed electron), quite often it is necessary within an analysis to reconstruct intermediate particle resonances that decay to other particles within the detector (e.g. a Z boson that decays to two electrons). We refer to these reconstructed objects as composite physics objects, which are reconstructed with analysis-dependent strategies. The kinematic information and quantum numbers associated with these composite physics objects are often utilized to help discriminate a signal process against background processes with similar final states.

First, we discuss in Section 3.1 the detector-level information that is used to reconstruct the principal physics objects in the CMS detector. The reconstruction of primary vertices within the detector, which makes use of tracking information, is discussed in Section 3.2. In Section 3.3, we discuss the CMS Particle Flow (PF) algorithm, which is used to reconstruct individual electrons, muons, photons, charged hadrons, and neutral hadrons within the detector. A full discussion of these reconstructed physics objects and other principal physics objects built out of these particle candidates (such as jets) is presented in Section 3.4. The selection cuts and methods used specifically in the VH($b\bar{b}$) analysis are included in this discussion (as opposed to a more general selection used by analyses at CMS). The reconstructed composite physics objects used in the VH($b\bar{b}$) analysis, the massive vector bosons (W and Z) and the Higgs boson decaying to bottom quarks, are discussed in Sections **??** and 3.6, respectively. We end this chapter with a discussion of attempts at CMS to characterize these reconstructed physics objects, including mis-identification rate and efficiency measurements (Section 3.7).

## 3.1 Detector-level Objects

In this section, we discuss the low-level detector objects that are used to reconstruct higher-level physics objects in the detector. Of particular interest are tracks built solely using the silicon tracker ("tracks"), tracks built solely using the muon system ("standalone muon tracks"), and clusters of energy in the electromagnetic and hadronic calorimeters ("calorimeter clusters"). These detector-level objects, which are used by the CMS Particle Flow algorithm (see Section 3.3 below), are discussed in Sections 3.1.1, 3.1.2, and 3.1.3, respectively.

### 3.1.1 Tracks in the Silicon Tracker

The large flux of charged particles through the CMS detector volume necessitates a powerful inner tracking system in order to reconstruct primary vertices in the event (see Section 3.2) and to precisely measure the transverse momentum and impact parameter (with respect to the beam line) of individual charged particles. The physics agenda of CMS depends on measuring these quantities well, independent of the instantaneous luminosity. For instance, a high-precision measurement of charged particle transverse momentum is required to have excellent resolution for the reconstructed mass of resonant dilepton decays, enabling greater physics reach for searches of such resonances, and measuring the impact parameter of charged particles well enables physics analyses at CMS to better identify bottom-quark jets. These jets are often associated with a secondary vertex that is well-displaced from the primary vertex where the initial hard process occurred.

Tracks are produced by way of iterative tracking using the silicon pixel and strip detectors described in the previous chapter, making use of clusters of pixel or strip hits (referred to simply as "hits"). Each iteration (of which there are seven) makes use of the Combinatorial Track Finder (CTF) [10, 105] to associate tracker hits together into a complete track, removing hits from further consideration when each association is made. This reduces the combinatorial complexity of the hit-to-track assignment in successive iterations and ensures arbitration of the hits going into each track – that is, each hit is associated with only one track.

The CTF algorithm proceeds in four steps. In the first step, track seeds are formed, which are short track candidates that provide the initial helical trajectory of the track, taking into account the magnetic field in the detector but assuming it is uniform. Next, track candidates are formed using the Kalman filter method [106], which extrapolates the track outward from the track seed through the successive layers of the tracker, adding hits along the way. The next step consists of refitting the track candidate to remove any bias introduced by the seeding step, using a Runge-Kutta propagator to perform fits both inside-out and outside-in (averaging the results). Both material effects and

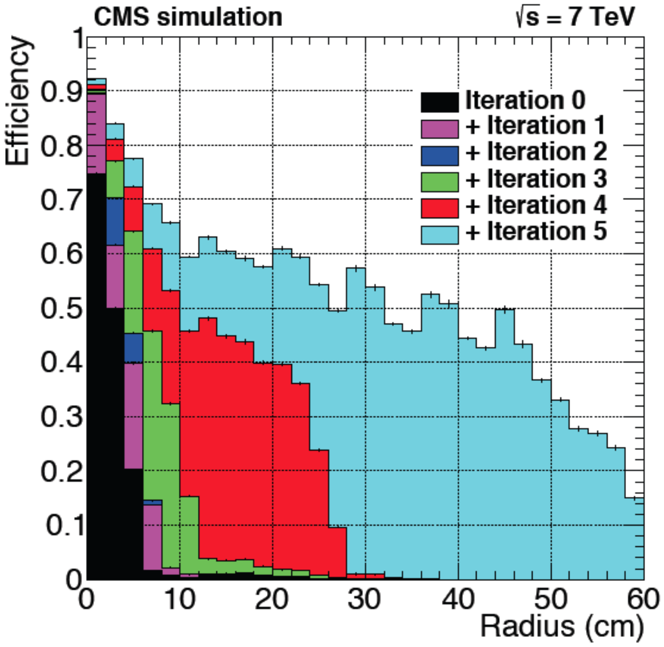

Figure 3.1: Cumulative contributions to the overall tracking performance from the six iterations in track reconstruction (before the switch to seven iterations, incorporated after 2011 data-taking) [10]. The tracking efficiency for simulated $\mathrm{t\bar{t}}$ events is shown as a function of transverse distance from the beam axis to the production point of each particle, for tracks with $p_{\mathrm{T}} > 0.9$ GeV, $|\eta| < 2.5$, and transverse (longitudinal) impact parameter $< 60$ (30) cm.

inhomogeneities of the magnetic field are considered in this step. The fourth and final step is the application of a variety of cleaning cuts on the track candidate, such as a cut on the reduced $\chi^2$ of the track.

Successive iterations have less stringent requirements on the proximity to the primary vertex but require more stringent requirements on the track quality cuts. As a result, particles produced further from the primary vertex tend to be reconstructed in later iterations, as shown in Figure 3.1. In most of the relevant kinematic range (100 MeV – 1 TeV), the rate of fake tracks is below 1%, which demonstrates excellent performance of the tracking at CMS in high-occupancy environments.

### 3.1.2 Standalone Muon Tracks

For the reconstruction of physics objects that are expected to penetrate through the CMS calorimetry, it is necessary to include detector-level information associated with activity in the muon system. Of primary interest is the reconstruction of real muons in the detector. Often it is the case that particles aside from real prompt muons contribute hits in the muon system that are combined with a real muon track (from the inner tracking system) into a “fake” reconstructed muon, and so it is important that the used reconstruction techniques try to minimize this mis-identification rate.

The detector-level object used in this effort is known as a standalone muon track, which is a track stub built strictly out of hits in the muon system. The track fitting process accounts for the detector material and magnetic field inhomogeneities. A Kalman filter is used, which takes track segments found in the innermost muon chambers as seeds for track candidates. Hits from the RPC and CSC detectors, and/or track segments from the DT detector, are added as the filter proceeds outward. The method is also applied in reverse, from the outside in. The results from both directions are combined in order to create a smooth, optimal track. Unrelated or “bad” hits caused from showers, pair production, etc. are removed with a cut on the reduced $\chi^2$ of the track. Finally, the standalone muon track candidate is refit using a constraint to the beam line to create a polished standalone muon track, to be used by the CMS PF algorithm.

### 3.1.3 Calorimeter Clusters

The reconstruction of nearly every principal physics object, whether it is charged or not, requires information about calorimeter energy deposition in the event. This includes electrons, photons, and the neutral and charged hadrons that comprise jets. Even muons need this information as an input as they should be consistent with a minimum-ionizing signature, and thus should see very little energy deposition in the calorimeter cells around the reconstructed muon candidate.

At CMS, calorimeter energy deposition is broken down into units called calorimeter clusters. These clusters of energy are built from calorimeter cells, with the clustering done separately in each sub-unit of the calorimeter (EB, EE, HB, etc.) in three steps. The first step builds cluster seeds out of local energy maxima in each sub-detector, requiring the energy to be above a detector-dependent threshold. Next, "topological clusters" are created by adding cells adjacent to the cluster seeds, again with a threshold to discriminate against electronic noise, allowing for the inclusion of multiple cluster seeds in the same topological cluster. Finally, the total energy of these topological clusters are redistributed back to the original cluster seeds to form "particle flow clusters" that are used by the PF algorithm. An iterative energy reassignment scheme is used that also adjusts the location of the particle flow clusters from the original cluster seed locations. The position is recalculated by taking the spatial average of the cells contributing energy to the particle flow cluster, weighted by the amount of energy contributed.

## 3.2 Primary Vertices

The reconstruction of the primary vertices of an event, where individual proton-proton collisions occur near the beam spot, is necessary to isolate the hard process of the triggered event. It is also essential in order to be able to account for the effects of pile-up, the softer minimum-bias proton-proton interactions that accompany the hard process of interest due to high-luminosity conditions. This effort first requires that the inner tracking has been completed, i.e. the production of a list of prompt tracks (described in Section 3.1.1). A variety of cleaning cuts are placed on the tracks (reduced $\chi^2$ of track hits fit, impact parameter significance, number of hits in the pixel and strip detectors) keeping the $p_{\mathrm{T}}$ acceptance as loose as possible to ensure high reconstruction efficiency. The remaining tracks are kept to be clustered into primary vertices [10].

The tracks are clustered according to the $z$-coordinate of each track's point of closest approach to the beam spot, requiring each track to be separated in $z$ by at least 1 cm from its closest neighbor. First, the number of primary vertices in the event and associated track list is built using the Deterministic Annealing algorithm [107]. Then the primary vertex candidates are fit using an adaptive vertex fit [108] in order to determine the spatial coordinates of the vertex. In this fit, each track is given a weight based on

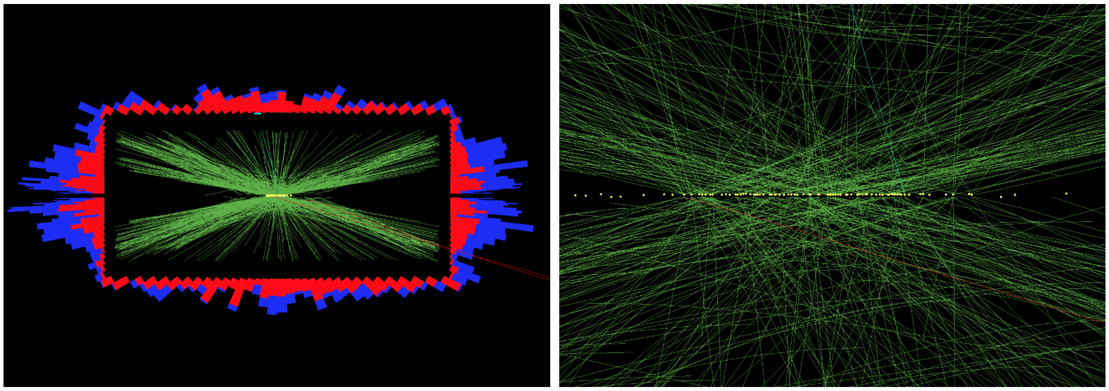

Figure 3.2: Event display illustrating the reconstruction of 78 primary vertices at CMS [11], showing both an expanded (left) and zoomed (right) view. The expanded view illustrates the calorimeter energy deposition profile of pile-up interactions that must be accounted for in physics object reconstruction. This event was observed at CMS in a data-taking run on July 10, 2012, with the LHC operating at a center-of-mass energy of 8 TeV.

its compatibility with the vertex. The "number of degrees of freedom" of the vertex is proportional to the sum of the track weights – this variable is used to identify a vertex as a real primary vertex from a proton-proton collision. The uncertainty on the of position of a primary vertex is roughly 10 $\mu$m in each direction.

Figure 3.2 illustrates the performance of primary vertex reconstruction at CMS during a high pile-up scenario (an event collected during a special exploratory run in 2012). In order to pick what we refer to as *the* primary vertex of an event, where the hard process that triggered the event occurred, we take the primary vertex with the highest value of $\sum_i [p_{\mathrm{T}i}(\text{track})]$. The efficiency of picking the correct primary vertex with this method is strongly dependent on the total amount of track momentum associated with the hard interaction. For simulated events containing a $\mathrm{Z} \to \mu^+\mu^-$ decay, this ranges from roughly 75% for a $p_{\mathrm{T}}(\mathrm{Z})$ of 20 GeV to 99% for $p_{\mathrm{T}}(\mathrm{Z}) > 100$ GeV. Similar numbers are found using simulated $\mathrm{H} \to \gamma\gamma$ decays [109].

## 3.3 Particle Flow

The CMS Particle Flow algorithm [110] is used to reconstruct all stable particles in the event. This includes individual particles – namely, electrons, muons, photons, charged hadrons, and neutral hadrons. It also includes collections of these particle candidates reconstructed into physics objects, specifically quark/gluon jets (we refer to simply as "jets"), tau jets (reconstructed hadronic decays of tau particles, which we refer to as "taus"), and missing transverse energy. The latter is inferred from the momentum imbalance in the detector once all other particles are reconstructed. Altogether, this collection of reconstructed objects accounts for all principal physics objects used by the various physics analyses at CMS, which we will describe in more detail in Section 3.4.

The PF algorithm makes use of all sub-detectors of CMS in order to distinguish different stable particle hypotheses from one another, as illustrated in Figure 3.3. In particular, this allows the reconstruction of objects like jets and taus to take advantage of tracking information to improve reconstruction and identification performance [110, 111]. The algorithm works by taking detector-level objects (described in Section 3.1) and "linking" them pair-wise, associating each link with a distance parameter that represents the quality of the link. Groups of these links are combined into "blocks" that are used in stable particle identification. The linking procedure and particle reconstruction and identification algorithm are described in Sections 3.3.1 and 3.3.2, respectively.

### 3.3.1 Linking

A single particle often results in the creation of multiple detector-level objects. In fact, particle identification at CMS depends heavily on this feature – by "linking" together multiple detector-level objects, one can create a physics object in the detector consistent with a particular particle hypothesis that is distinct from other particle hypotheses. For instance, a real muon in the detector should produce a track in the inner tracker as well as a standalone muon track in the muon system. By linking together these two detector-level objects, we can proceed to reconstruct a muon physics object in full detail.

There are three types of links that are made between different detector-level objects:

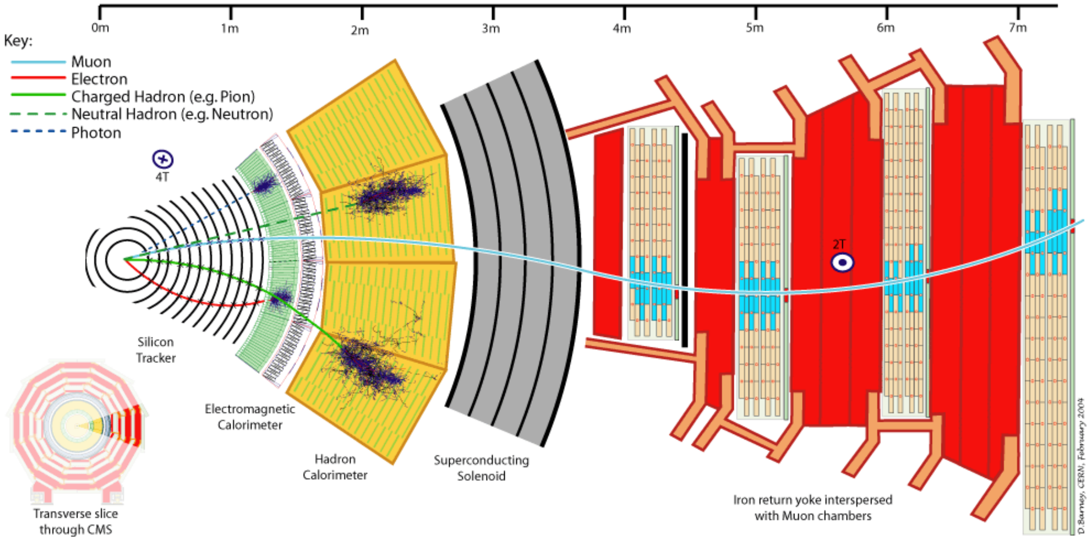
0m
1m
2m
3m
4m
5m
6m
7m
Key:
Muon
Electron
Charged Hadron (e.g. Pion)
Neutral Hadron (e.g. Neutron)
Photon
4T
2T
Silicon Tracker
Electromagnetic Calorimeter
Hadron Calorimeter
Superconducting Solenoid
Iron return yoke interspersed with Muon chambers
Transverse slice through CMS
D Barney, CERN, February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

Figure 3.3: A schematic of a slice of the CMS detector (in the $r-\phi$ plane), demonstrating the use of multiple sub-detectors in order to reconstruct stable particles in the detector.

a link between a track and a calorimeter cluster, a link between a calorimeter cluster and another calorimeter cluster, and a link between a track and a standalone muon track. In the first case, the track is extrapolated beyond the last hit in the silicon tracker into the CMS calorimetry. Any calorimeter clusters overlapping with the track (taking into account uncertainty on the shower position, gaps between cells, dead cells, and the possibility of Bremsstrahlung in the tracker for the case of electrons) are linked to the track. The distance parameter associated with the link is the distance in the $\eta - \phi$ plane between the track and the calorimeter cluster. In the second case, two calorimeter clusters are linked together with a similar distance parameter, allowing for linking between clusters in the ECAL and clusters in the HCAL (for example). The same factors discussed above, such as uncertainties due to showering, are taken into account in this case as well. Finally, a track and a standalone muon track are linked when the two can be fit into one “global” track (creating a “global muon”) with a $\chi^2$ below a maximum value. The global track $\chi^2$ is the distance parameter in this case that describes the quality of the link.

The links are grouped together into “blocks” by association. Individual detector-level objects sometimes constitute the entire block, when no link is found between the object and another detector-level object. Due to the fine granularity of the CMS detector, blocks typically contain 1-3 detector-level objects [110].

### 3.3.2 Particle Reconstruction and Identification

Once blocks are formed by the PF algorithm, reconstruction and identification of individual particles can begin. The algorithm attempts to identify individual particles within each of the blocks, treating each block independently. First, "particle-flow muons" are identified one at a time by taking a global muon (reconstructed using a block containing a track and a standalone muon track) with a momentum within three standard deviations from the momentum of the track alone [112]. Next, electrons are identified by using a block containing a track and a calorimeter cluster in the ECAL. The track is required to be consistent with an electron signature, as an electron radiates energy as it moves radially through the tracker, yielding a shorter track than other particles. The track is fit with a Gaussian-Sum Filter (GSF) [113] to determine the trajectory, and a variety of cuts related to the ECAL and tracker are used for final "particle-flow electron" identification [112]. Moving forward, tracks are required to have smaller relative $p_\mathrm{T}$ uncertainty in comparison to the relative calorimetric energy resolution expected for charged hadrons – otherwise, the track is removed from further consideration.

Photons, charged hadrons, and neutral hadrons are then identified within a block that contains a track linked with calorimeter clusters in the ECAL/HCAL [114]. First, the calibrated energy of the particle candidate is calculated by making use of calibration information obtained using simulated samples of single hadrons, along with the energy measured separately in the ECAL and HCAL. This calibrated energy obtained from the calorimetry is compared to the sum of momenta of tracks associated with the block. If it is smaller, then a search for fake tracks and additional particle-flow muons (with looser criteria) is first performed. Fake tracks are identified by a high relative uncertainty in $p_\mathrm{T}$ and removed, and extra global muons with relative $p_\mathrm{T}$ uncertainty better than 25% are reconstructed as additional particle-flow muons. Any remaining tracks after this search become “particle-flow charged hadrons” with a momentum and energy determined from assuming the mass is that of a charged pion. If the calibrated calorimeter energy is instead larger than the sum of the momenta of the associated tracks, “particle-flow photons” and “particle-flow neutral hadrons” are reconstructed with the energy excess. If the excess is greater than the total energy deposited in the ECAL, then a particle-flow photon is created with the excess ECAL energy, and a particle-flow neutral hadron is created with the excess HCAL energy. Otherwise, only a particle-flow photon is created. Finally, ECAL and HCAL calorimeter clusters never linked to a track become additional particle-flow photons, and calorimeter clusters removed from the block due to having a link too low in quality (as defined by the distance parameter) become additional particle-flow neutral hadrons (using HCAL energy only). With individual particles identified, the jets, taus, and the missing transverse energy of the event can then be reconstructed.

## 3.4 Principal Physics Objects

We refer to physics objects reconstructed directly from detector-level information (via the CMS Particle Flow algorithm) as principal physics objects. These physics objects are reconstructed for general use by the CMS collaboration, and so the identification criteria for these objects differ very little between different analysis groups (except when special techniques are necessary to reconstruct exotic physics signatures, such as for displaced jets or leptons). In Sections 3.4.1 through 3.4.5 we present the identification criteria only for physics objects used in the VH($\mathrm{b\bar{b}}$) analysis – as a result, photons are not discussed. There are some minor differences between the identification criteria used for the analysis of 2011 data versus the analysis of 2012 data – we present the identification criteria used for the latter.

### 3.4.1 Muons

Muons used in the VH($\mathrm{b\bar{b}}$) analysis are required to be reconstructed first as particle-flow muons [112, 115]. On top of this requirement, further criteria must be met:

- reduced $\chi^2 < 10$ for the global muon fit,
- muon $p_{\mathrm{T}} > 20\mathrm{GeV}$,
- tracks associated to muons must satisfy:
  - at least one pixel hit,
  - at least six tracker layers with valid hits,
  - at least one valid hit in the muon chambers,
  - at least two muon stations,
  - impact parameter in the transverse plane $d_{xy} < 2\mathrm{mm}$,
  - $|\eta| < 2.4$,
  - relative combined isolation $R < 0.12$ (computed in a cone of radius 0.4).

The relative combined isolation $R$ is calculated (for both muons and electrons) using the PF isolation equation:

$$R \equiv \frac{\sum_i \left[p_{\mathrm{T},i}(\text{PF charged had.}) + p_{\mathrm{T},i}(\text{PF neutral had.}) + p_{\mathrm{T},i}(\text{PF photon})\right]}{p_{\mathrm{T}}(\text{lepton})}, \qquad (3.1)$$

adding a term for subtraction of pile-up energy and momentum. This subtraction is based on the per-event estimated neutral energy expected to enter a cone of radius 0.3 (in $\eta - \phi$ space) around the lepton. For muons, the correction is estimated from the deposit associated to charged tracks not belonging to the primary vertex, using calibration factors obtained from data-driven CMS measurements.

### 3.4.2 Electrons

Electrons are required to be particle-flow electrons and are further identified using a multivariate (MVA) approach [112, 116, 117]. Two different cuts on the MVA ID discriminator are applied depending on the use-case, defining two different working points based on the expected selection efficiency of either 95% (“WP95”) or 80% (“WP80”). The loose WP95 working point is used for global event classification (based on vector boson type and decay), in the definition of jet cleaning, in the counting of additional leptons for the veto requirement, and in the event selection of the Z(ee)H channel. The tighter WP80 working point is used in the W(e$\nu$)H channel to suppress background with fake electrons in that final state. Electron $p_{\mathrm{T}}$ thresholds of 30 and 20 GeV are applied in the W(e$\nu$)H and Z(ee)H analyses, respectively.

For electrons the estimate of the relative combined isolation $R$ is obtained by using Equation 3.1 and subtracting an estimate of the contribution from pile-up, as for muons. This estimate is made by computing the average energy deposition $\rho$ expected in the calorimetry for the observed level of pile-up, making use of only particle-flow neutral hadrons, and then multiplying by an estimated effective area of the cone that is determined from data-driven CMS measurements. As for muons, a cut of $R < 0.12$ is used.

### 3.4.3 Taus

The most abundant final state for hadronically-decaying tau particles consists of one charged hadron (typically a charged pion) and zero or more neutral hadrons (typically neutral pions). This signature is known as a “one-prong hadronic decay” and is required for reconstructed taus in the VH($\mathrm{b\bar{b}}$) analysis. Taus are reconstructed in detail with the Hadron Plus Strips (HPS) algorithm [118], using PF objects as the input [110]. The following selection is then performed for one-prong hadronic taus:

- tau $p_{\mathrm{T}} > 40$ GeV,
- $|\eta| < 2.1$,
- leading track $p_{\mathrm{T}} > 20$ GeV,
- rejection of electrons/muons faking taus,
- loose tau-based isolation.

More details on the rejection of fakes and isolation requirements can be found in [118]. The tau reconstruction and selection has an efficiency of roughly 50% and a fake rate of less than 1%.

### 3.4.4 Jets

Jets are reconstructed in the detector in order to group together the products of fragmentation and hadronization that a gluon or quark undergoes after the initial hard process in the event. The primary goal is to provide the best measurement of jet energy and momentum (and thus mass) as possible given the final set of stable particles in the jet. This is done at CMS via a clustering of the single-particle objects output from the Particle Flow algorithm, described in Section 3.3.

Jet clustering algorithms [119] build jets by taking one detector-level object or particle-flow object at a time and piecing them together, adding the energy associated with each of these objects to the combined jet energy. Most analyses at CMS use "particle-flow jets" (or simply "PF jets") that make use of the particle-flow objects as input to the clustering, as opposed to using strictly calorimeter energy deposits (as is done for "Calo jets"). The advantage of using the objects output from the PF algorithm for jet reconstruction is better energy resolution, a result of the use of tracking to help measure roughly 65% of the jet energy that is contained in charged hadrons [110]. This comes at a cost of increased computation time, however – both in the HLT and for offline reconstruction. Jets at CMS are clustered with the anti-$k_{\mathrm{T}}$ algorithm [120] using a size parameter of 0.5 (maximal jet radius in $\eta - \phi$ space). Jet reconstruction is done using the FastJet package [121], and effects of pile-up and underlying-event are accounted for in the reconstructed jet energy [122].

Each jet is required to lie within $|\eta| < 2.5$, have at least two tracks associated to it, and have electromagnetic and hadronic energy fractions of at least 1% of the total energy to avoid mis-reconstruction derived by noise clusters. In the WH and Z($\nu\nu$)H channels, a minimum threshold of $p_{\mathrm{T}} > 30$ GeV is used, while a looser selection ($p_{\mathrm{T}} > 20$ GeV) is applied in the cleaner Z($\ell\ell$)H channels with lower di-jet boost. Standard jet energy corrections are applied for jets in both data and simulation, and the jet energy resolution is smeared for simulated jets to account for differences between data and MC, observed in studies of jet reconstruction at CMS [13, 123].

The CSV ("Combined Secondary Vertex") algorithm [12, 124] is used to identify jets that are likely to arise from the fragmentation and hadronization of b quarks (also known as “b jets”). This algorithm combines in an optimal way the information about track impact parameters and identified secondary vertices within jets, even when full vertex information is not available. Additional categories for jets where a “pseudo-vertex” is found, or no secondary vertex is identified, can be defined and combined in a likelihood discriminant to provide maximal separation of b jets from the much larger background of jets arising from charm decay, and from the fragmentation of light quarks and gluons. The algorithm provides a continuous discriminator output (between 0 and 1 when a secondary vertex is found) that can be used to select optimal working points with respect to the VH analyses, in addition to the standard Loose/Medium/Tight working points defined at CMS: CSVL ($> 0.244$), CSVM ($> 0.679$), and CSVT ($> 0.898$). Figure 3.4 shows the CSV output shapes for jets associated with different quark flavors, comparing data to simulated $\mathrm{t\bar{t}}$ events. A reshaping technique (making use of a spline fit) is used to correct the CSV output distribution seen in simulation to match the distribution using actual data – data-driven efficiency measurements made with $\mathrm{t\bar{t}}$ events are used as the the inputs for this reshaping [125]. Figure 3.5 illustrates the effect of the reshaping on the CSV output distribution.

### 3.4.5 Missing Transverse Energy

In many analyses at CMS, it is necessary to indirectly detect particles that are not expected to interact with the detector, such as neutrinos. In the $\mathrm{VH(b\bar{b})}$ analysis, the WH and $\mathrm{Z}(\nu\nu)\mathrm{H}$ channels depend on this attempt at identifying neutrinos in the final state. This is done by using the “missing transverse energy” of the event, or $E_{\mathrm{T}}^{\mathrm{miss}}$, which represents the magnitude of the transverse component of the vector sum of all neutrino momenta in an otherwise well-reconstructed signal event [126, 127]. This vector sum is estimated at CMS after the PF algorithm is run [110], computing $\vec{E}_{\mathrm{T}}^{\mathrm{miss}}$ as follows:

$$\vec{E}_{\mathrm{T}}^{\mathrm{miss}} = -\sum_{i} \vec{p}_{\mathrm{T},i}. \tag{3.2}$$

The sum in Equation 3.2 is taken over all particle-flow objects. The performance of the $E_{\mathrm{T}}^{\mathrm{miss}}$ determination, discussed in more detail in [123], depends on the performance of

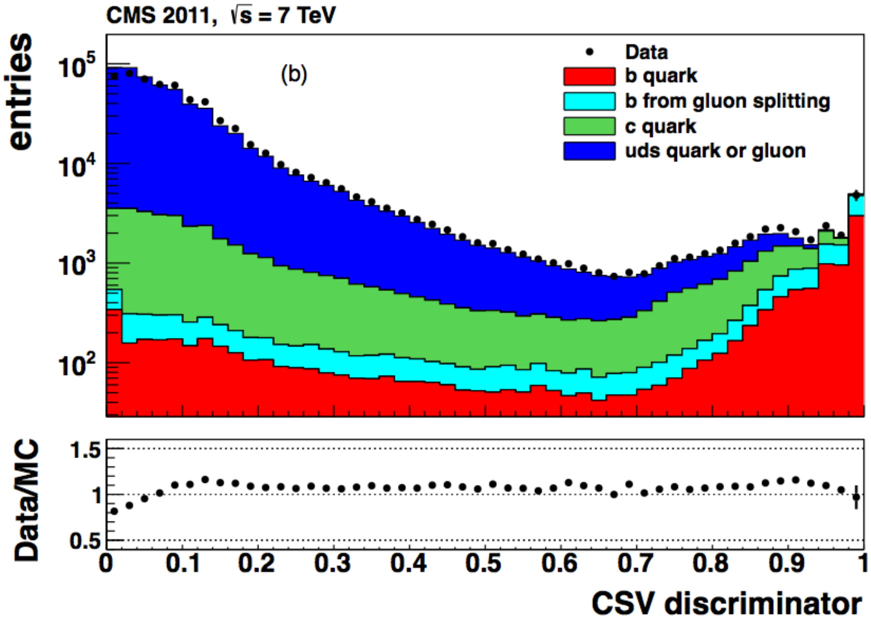

Figure 3.4: CSV output shapes for jets associated with different quark flavors [12]. Data is compared to simulated $t\bar{t}$ events.

the reconstruction of all other objects in the event. Because of this, the $E_{\mathrm{T}}^{\mathrm{miss}}$ resolution also improves from the use of the PF algorithm.

It is important to note that the $E_{\mathrm{T}}^{\mathrm{miss}}$ of an event often includes contributions from mis-reconstruction of the energy of the jets in the event, or failure for the detector to identify a lepton from the hard interaction (often due to the lepton not being within the acceptance of the tracker). As a result, even events that do not have a neutrino (or other undetectable particle) in the final state will very often have a non-zero value of $E_{\mathrm{T}}^{\mathrm{miss}}$.

## 3.5 Vector Boson Reconstruction

In the VH($b\bar{b}$) analysis, it is necessary to make use of composite physics objects in order to separate signal from background processes with a similar final state. We first discuss the reconstruction of the heavy vector boson, either W or Z, that is produced in association with a Higgs boson in our search [128]. Reconstruction of W and Z bosons begins with the identification and selection of charged leptons and $E_{\mathrm{T}}^{\mathrm{miss}}$, described in

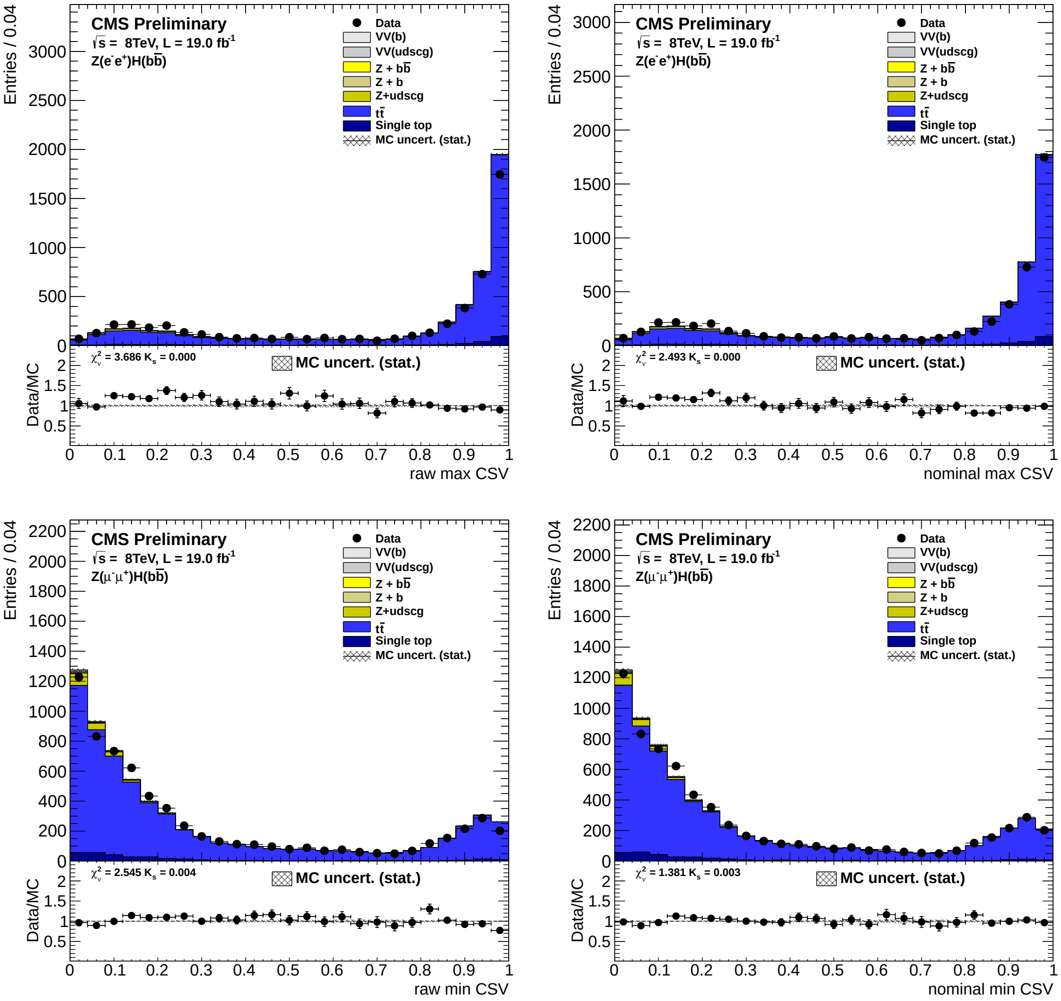

Figure 3.5: Distributions of the CSV output before reshaping (left) and after reshaping (right) in data and simulated samples in the $t\bar{t}$ control region for the $Z(\ell\ell)H$ channels. Shown are distributions of "max CSV" (the $H(b\bar{b})$ jet with the highest CSV output value) for Z(ee)H (top) and "min CSV" (the $H(b\bar{b})$ jet with the lowest CSV output value) for $Z(\mu\mu)H$ (bottom). The data/MC agreement improves after the reshaping, as seen in the data/MC ratio shown in the bottom sections of the plots. The hatched bands correspond to the statistical uncertainty on the MC simulation prediction for each bin.

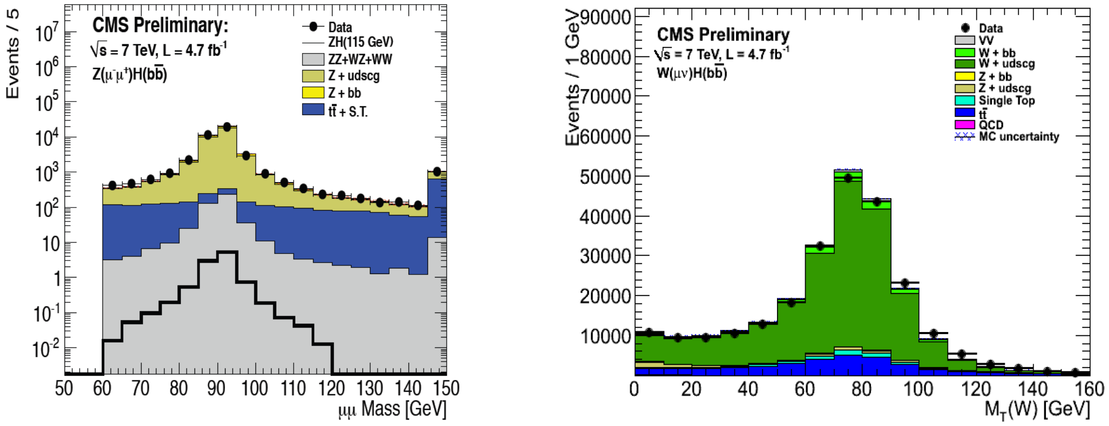

Figure 3.6: Distributions of dimuon mass (left) and W transverse mass (right) for inclusive Z($\mu\mu$)+jets and W($\mu\nu$)+jets events selected with two central jets without boosting (i.e. without tight cuts on $p_{\mathrm{T}}$(V) or $p_{\mathrm{T}}$(jj)). These plots are purely for illustration of what the data/MC agreement looks like before boosting and b-tagging.

Section 3.4. The dominant background is from real W and Z decays, given the unique signature of a highly boosted vector boson recoiling from two jets. Therefore, only a very minimal additional selection requirement is made in order to identify highly pure samples of V+jets events.

Candidate Z $\rightarrow \ell\ell$ decays are reconstructed by combining isolated electrons and muons and requiring the di-lepton invariant mass to satisfy 75 GeV $< M_{\ell\ell} <$ 105 GeV, along with a minimal requirement of $p_{\mathrm{T}}(\mathrm{Z}) >$ 50 GeV. Figure 3.6 shows the dimuon invariant mass in events selected with this loose criteria, and including two central jets with a minimum threshold of 20 GeV on the transverse momentum. It is possible to reconstruct Z $\rightarrow \nu\bar{\nu}$ decays via the selection criterion of $p_{\mathrm{T}}(\mathrm{Z}) >$ 100 GeV, where the event $E_{\mathrm{T}}^{\mathrm{miss}}$ is used to approximate the transverse momentum of the Z boson.

Candidate W $\rightarrow \ell\nu$ decays are identified primarily by the topology of a single isolated lepton (electron, muon, or tau) and additional missing transverse energy. The transverse momentum $p_{\mathrm{T}}$(W) and W candidate transverse mass $M_{\mathrm{T}}$(W) are computed as:

$$p_{\mathrm{T}}(\mathrm{W}) \quad = \quad \sqrt{(E_{\mathrm{x}}^{\mathrm{miss}} + p_x^{\ell})^2 + (E_{\mathrm{y}}^{\mathrm{miss}} + p_y^{\ell})^2}, \tag{3.3}$$

$$M_{\mathrm{T}}(\mathrm{W}) \quad = \quad \sqrt{(E_{\mathrm{T}}^{\mathrm{miss}} + p_{\mathrm{T}}^{\ell})^2 - p_{\mathrm{T}}(\mathrm{W})^2}. \tag{3.4}$$

Here $p_{\mathrm{T}}^{\ell}$ is the transverse momentum of the lepton, $p_x^{\ell}$ and $p_y^{\ell}$ are the projections of the lepton momentum onto the $x$ and $y$ axes respectively, and $E_{\mathrm{x}}^{\mathrm{miss}}$ and $E_{\mathrm{y}}^{\mathrm{miss}}$ are the projections of the event $E_{\mathrm{T}}^{\mathrm{miss}}$ onto the $x$ and $y$ axes, respectively. Additional cuts of $p_{\mathrm{T}}(\mathrm{W}) > 100$ GeV and $E_{\mathrm{T}}^{\mathrm{miss}} > 45$ GeV are required in the selection. For $\mathrm{W} \rightarrow \tau\nu$ decays, a tighter cut of $E_{\mathrm{T}}^{\mathrm{miss}} > 80$ GeV is used due to an already tight cut on this variable at trigger level. Figure 3.6 shows the distributions of $M_{\mathrm{T}}(\mathrm{W})$ in events selected with the addition of two central jets with transverse momentum above 30 GeV. It is observed that in the boosted regime, where the QCD background is much reduced, the $E_{\mathrm{T}}^{\mathrm{miss}}$ cut is sufficient to select a relatively clean sample of real W decays.

The vector boson transverse momentum is also used in the $\mathrm{VH(b\bar{b})}$ analysis to split the signal region into statistically-independent bins – up to three categories per mode. For the $\mathrm{Z}(\ell\ell)\mathrm{H}$ analysis, there are two categories:

- 50 GeV $< p_{\mathrm{T}}(\mathrm{Z}) <$ 100 GeV,

- $p_{\mathrm{T}}(\mathrm{Z}) >$ 100 GeV.

Three categories are used in the $\mathrm{Z}(\nu\nu)\mathrm{H}$ analysis:

- 100 GeV $< p_{\mathrm{T}}(\mathrm{Z}) <$ 130 GeV,

- 130 GeV $< p_{\mathrm{T}}(\mathrm{Z}) <$ 170 GeV,

- $p_{\mathrm{T}}(\mathrm{Z}) >$ 170 GeV.

Similarly, there are three categories in the $\mathrm{W}(\ell\nu)\mathrm{H}$ analysis:

- 100 GeV $< p_{\mathrm{T}}(\mathrm{W}) <$ 130 GeV,

- 130 GeV $< p_{\mathrm{T}}(\mathrm{W}) <$ 180 GeV,

- $p_{\mathrm{T}}(\mathrm{W}) >$ 180 GeV.

Due to lack of statistics, the $\mathrm{W}(\tau\nu)\mathrm{H}$ analysis only has one category, with $p_{\mathrm{T}}(\mathrm{W}) > 120$ GeV.

In an inclusive selection enriched in vector boson production, a shape difference in the vector boson transverse momentum of the simulated samples with respect to data has been observed. Due to the observed data having a softer spectrum than the simulation, a negative correction with increasing $p_{\mathrm{T}}(\mathrm{V})$ is necessary to correct for the effect. This is done with a per-event reweighting procedure, using Z + jets and W + jets control samples (described in Chapter 7) to determine the $p_{\mathrm{T}}(\mathrm{V})$-dependent weights.

## 3.6 Higgs Boson Reconstruction

The second composite physics object used in the VH($\mathrm{b\bar{b}}$) analysis is the reconstruction of the Higgs boson decaying to two bottom quarks. Reconstruction of the $\mathrm{H} \to \mathrm{b\bar{b}}$ decay is a critical element of the analysis, both in achieving the best possible mass resolution for the Higgs candidate and in maximizing the analysis sensitivity in extracting the VH($\mathrm{b\bar{b}}$) signal. In general, requiring a large boost for the di-jet system, or a high threshold on the $p_{\mathrm{T}}$ of each individual jet, improves the di-jet mass resolution. In the highly boosted regime ($p_{\mathrm{T}} > 200$ GeV), specialized jet reconstruction algorithms have been developed to optimally reconstruct the b jets from the Higgs decay, along with associated final-state radiation. This is especially useful when there are very small opening angles between the jet constituents, a case when typical jet reconstruction algorithms would lead to the merging of the two jets, and thus loss of the event. These "jet substructure" methods [82, 129–131] are expected to be very helpful in the future LHC data-taking runs at higher center-of-mass energy, but for now do not provide an improvement in sensitivity over the standard analysis. As a result, these methods are not currently used in the VH($\mathrm{b\bar{b}}$) analysis.

The candidate $\mathrm{H} \to \mathrm{b\bar{b}}$ decays are selected by identifying the di-jet combination with the largest value of di-jet $p_{\mathrm{T}}$ (also referred to as $p_{\mathrm{T}}(\mathrm{jj})$). Selection methods making use of b-tagging information to pick the candidate Higgs boson decay were ruled out due to the large $\mathrm{t\bar{t}}$ background in the W($\ell\nu$)H and Z($\nu\nu$)H analyses. A resolution of approximately 10% is achieved, with a few percent bias on the mass. This bias is largely due to energy losses characteristic of semi-leptonic B decays often occurring in jets originating from b quarks. This bias is partially corrected by using a dedicated b-jet energy regression, discussed in Chapter 6, which also leads to an improvement in the di-jet mass resolution. As an example of the performance, Figure 3.7 shows the final invariant mass shape for

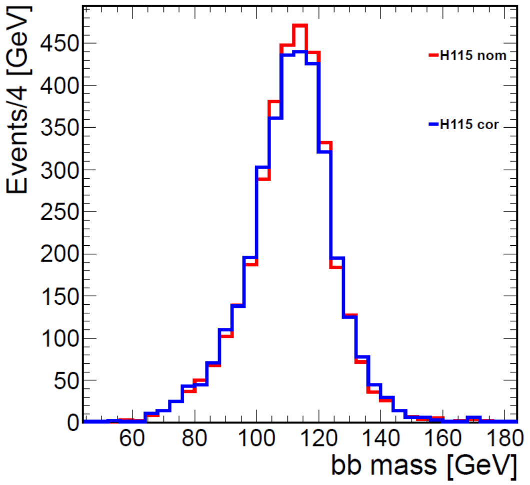

Figure 3.7: Distributions of di-jet invariant mass in simulated Z($\nu\nu$)H signal events generated with $M_H$ = 115GeV after all selection criteria described in Chapter 5 have been applied. The predicted resolution is roughly 10% for the nominal simulation (red histogram), and degrades by a negligible amount after applying $p_T$ smearing from official CMS jet corrections (blue histogram) [13], described in Section 3.4.4.

di-jets in simulated Z($\nu\nu$)H signal events ($M_H$= 115 GeV) after the selection described in Chapter 5.

## 3.7 Physics Object Characterization

Various properties of reconstructed physics objects must be characterized before they can be used in physics analyses. This includes measuring the efficiency of a physics object to be reconstructed given that there was an actual particle passing through the detector, and measuring mis-identification rates (or "fake rates") of other particles being accidentally reconstructed as a particular physics object by mistake. These measurements are carried out by CMS Physics Object Groups ("POG's") and constitute a major effort at CMS to better understand the performance of the detector. As an example, we discuss the measurement of pion-to-muon and proton-to-muon mis-identification rates at CMS, carried out by the CMS Muon POG. A more detailed discussion can be found in [132, 133].

### 3.7.1 Hadron-to-Muon Mis-identification Studies

One can obtain pure samples of pions and protons from resonant particle decays such as $K^0_\mathrm{S} \to \pi^+\pi^-$ and $\Lambda^0 \to p\pi$ (which we use to represent both $\Lambda^0 \to p\pi^-$ and $\bar{\Lambda}^0 \to \bar{p}\pi^+$), respectively. These reconstructed particles can then be checked to see if they match to a muon stub, a track reconstructed by hits in the CMS muon chambers, to form a reconstructed muon. This check looks for the overlap of reconstructed tracker-tracks of the hadron and the muon, and thus is unambiguous matching. Such a reconstructed muon is considered a "fake muon" (even if it is a real muon from pion decay-in-flight that creates the hits in the muon chambers) in the sense that the muon does not come from prompt muon production, and may constitute a background to physics analyses making use of relatively low $p_\mathrm{T}$ muons in the final-state. An example of such an analysis is the CMS $B^0_s \to \mu^+\mu^-$ (and $B^0 \to \mu^+\mu^-$) search [15].

With a pure selection of pions and protons obtained via selection cuts placed on the $K^0_\mathrm{S}$ and $\Lambda^0$ candidates (collectively called "V0's" due to the V-like topological signature of the decay), the hadron-to-muon mis-identification rates can be extracted. This is done via fits to the invariant mass distribution of the reconstructed V0 candidates. These fits are done both before and after the reconstructed V0 daughter is matched (unambiguously via overlap of reconstructed tracker hits) to a reconstructed muon, with a double Gaussian for V0 signal and a first-order polynomial for combinatorial background. This is done for two muon selections: "Loose muons" that are basically just particle-flow muons with very little extra cleaning, or "Tight muons" that include several extra cleaning cuts outlined in Section 3.4.1 (though without isolation in these studies). The hadron-to-muon mis-identification rate is then computed by taking the ratio of the signal yield after muon-matching to the signal yield before muon-matching.

The hadron-to-muon mis-identification rates for both pions and protons are presented in Figure 3.8 for Loose muons and in 3.9 for Tight muons, as a function of various different variables of importance. Jet-triggered datasets from 2011 and 2012 data-taking are combined, and a comparison to MC simulation is made using QCD events generated with PYTHIA. A cut of $L_\mathrm{xy} < 4$ cm is enforced, where $L_\mathrm{xy}$ (or "V0 radius") is the distance

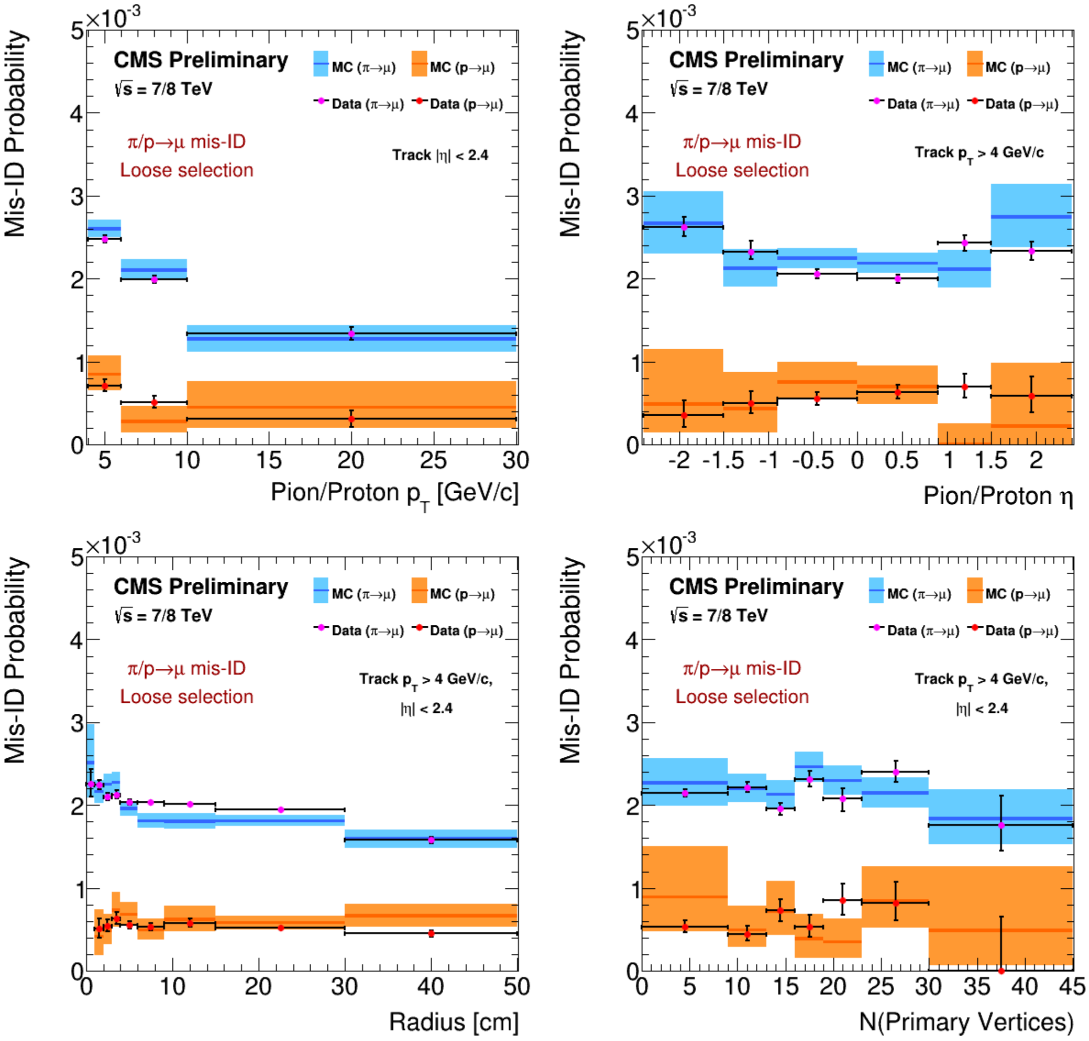

Figure 3.8: Differential hadron-to-muon mis-identification rates as a function of various different variables, for the Loose muon selection only. A comparison is made between the cases of pions and protons within each plot. An additional cut of track $p_{\mathrm{T}} > 4$ GeV is included.

in the $r - \phi$ plane between the beam line and the location of the V0 decay. This cut (not included in plots showing the hadron-to-muon mis-identification rate as a function of V0 radius) is enforced so the rates are usable in analyses where fake muons largely originate from hadrons produced within the beam pipe.

Interesting structures in these distributions, well reproduced by simulation, are observed. For pions, the mis-identification probabilities are below 0.3% even for the loosest

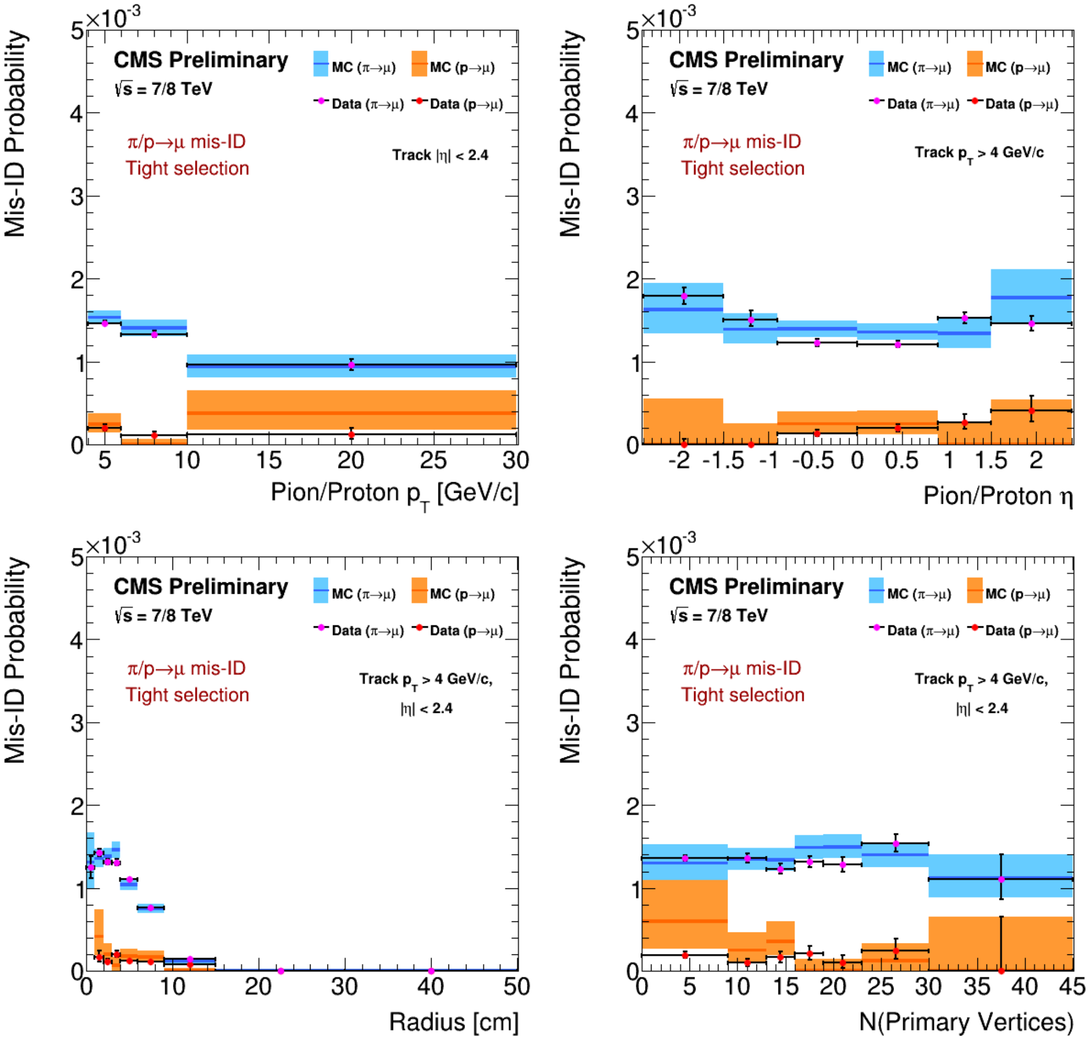

Figure 3.9: Differential hadron-to-muon mis-identification rates as a function of various different variables, for the Tight muon selection only. A comparison is made between the cases of pions and protons within each plot. An additional cut of track $p_{\mathrm{T}} > 4$ GeV is included.

Table 3.1: Table of inclusive hadron-to-muon mis-identification rates obtained using V0 decays occurring within the beam pipe, for both data and MC. An additional cut of track $p_{\mathrm{T}} > 4$ GeV is included. The rates are shown as percentages.

| **Pion-to-Muon Mis-ID Rates [%]** | | |
|---|---|---|
| Muon Selection | Data | MC |
| Loose | $0.216 \pm 0.003$ | $0.223 \pm 0.007$ |
| Tight | $0.134 \pm 0.002$ | $0.139 \pm 0.006$ |
| **Proton-to-Muon Mis-ID Rates [%]** | | |
| Muon Selection | Data | MC |
| Loose | $0.058 \pm 0.005$ | $0.058 \pm 0.013$ |
| Tight | $0.016 \pm 0.003$ | $0.018 \pm 0.007$ |

muon selection and further decrease as the $p_{\mathrm{T}}$ increases due to less of the hadrons decaying to real muons within the detector volume. For protons, the probability to be reconstructed as a muon decreases with $p_{\mathrm{T}}$ and remains much lower than the pion-to-muon fake rates in the accessible momentum range. At larger $|\eta|$, the pion-to-muon fake rate increases slightly due to the pions having more time to potentially decay to a muon within the detector volume (as the calorimetry near and in the endcaps is further away from the primary vertex). For the Tight muon selection, the rates significantly drop for decays that occur further inside the detector volume, due to the requirement of at least one hit in the pixel detector for the Tight muon selection. Finally, no significant dependence of the rates on pile-up (the number of reconstructed primary vertices in the event) is found. This is good motivation for the combination of the 2011 and 2012 datasets when determining both inclusive and differential mis-identification rates. Inclusive mis-identification rates are presented in Table 3.1. These numbers make use of a cut of track $p_{\mathrm{T}} > 4$ GeV, still requiring the V0 decay in the beam pipe.

Additional plots from these studies can be found in Appendix A, including results for a "TightMVA muon" selection (used specifically in the CMS $B_s^0 \to \mu^+\mu^-$ measurement and $B^0 \to \mu^+\mu^-$ search) that adds a cut on the output of a multivariate classifier to the Tight muon selection to further reduce the hadron-to-muon mis-identification rate.

# Chapter 4

# Data and Simulation

This chapter outlines the datasets and Monte Carlo (MC) simulation samples used in the VH($b\bar{b}$) analysis, as well as the trigger selection used to keep proton-proton collision events for further processing. Comparisons between data and MC simulation are necessary within the VH($b\bar{b}$) analysis for a variety of reasons. Examples include analysis validation (see Chapter 7) and the estimation of systematic uncertainties used in the extraction of signal (see Chapter 8.4).

We first discuss the datasets used in the VH($b\bar{b}$) analysis in Section 4.1. In Section 4.2, the MC simulation samples used in the analysis are presented, for both background and signal processes. Finally, in Section 4.3, the specific triggers used to retain signal events (and inevitably some background events in the process) are discussed, including both L1 and HLT trigger paths used to keep events during data-taking runs.

As discussed previously, we focus strictly on the final version of the VH($b\bar{b}$) analysis, which concerns data taken in 2012 only. Correspondingly, only the datasets, MC simulation samples, and trigger selection used for the analysis of 2012 data are presented in this section. For details regarding datasets, MC simulation samples, and trigger selection used for the analysis of 2011 data, please refer to [40].

## 4.1 Datasets

Events collected with the CMS detector are first selected using the L1 and HLT triggers, discussed in Section 2.2.6 and more specifically for the VH($b\bar{b}$) analysis in Section 4.3. The events are processed digitally and stored in a raw data format that is later processed offline into fully reconstructed events, containing physics objects discussed in Chapter 3. The data is then stored in Primary Datasets (PD's) for use by analysis groups at CMS. These PD's are organized by run period (e.g. `Run2012A`, `Run2012B`, etc.) and trigger paths used to retain the events. The datasets are stored on the Open Science Grid (OSG) in CMS data centers around the world for easy access by CMS analysis groups using CMS-specific grid-computing tools.

Table 4.1 summarizes the data samples used in the current analysis, the channels in which they are used, and the approximate integrated luminosity. Part of the full run range is vetoed in the analysis due to a pixel misalignment problem that impacts many observables related to b tagging. As a result, less than the full integrated luminosity collected by the CMS detector is used in the VH($b\bar{b}$) analysis. This removes about 600 $pb^{-1}$ of data and is already accounted for in Table 4.1. The distribution of the number of reconstructed primary vertices for events in these datasets is presented in Figure 4.1.

## 4.2 Monte Carlo Simulation Samples

It is important for analysis groups at CMS to understand the backgrounds that may "fake" a signal process of interest. Often there is remaining "irreducible" background in the signal region of an analysis that must be accounted for during signal extraction. Most analyses at CMS (including the VH($b\bar{b}$) analysis) make use of MC simulation to study these backgrounds in order to assess the normalization, variable shapes, and associated systematics of these processes. Equally important is the modeling of signal events using simulation, which enables analysis groups to extract signals with well-known characteristics. Some model-independent searches also use Monte Carlo simulation of signal events, as they often still make a limited set of assumptions about the signal

Table 4.1: List of 2012 data samples used for this analysis. The sum includes approximately 18.9 fb$^{-1}$ of integrated luminosity across all modes.

| Mode | Dataset | $\mathcal{L}$ (fb$^{-1}$) |
|---|---|---|
| W($\mu\nu$)H, Z($\mu\mu$)H | /SingleMu/Run2012A-13Jul2012-v1 | 0.809 |
| | /SingleMu/Run2012A-recover-06Aug2012-v1 | 0.082 |
| | /SingleMu/Run2012B-13Jul2012-v1 | 4.403 |
| | /SingleMu/Run2012C-24Aug2012-v1 | 0.405 |
| | /SingleMu/Run2012C-2012C-EcalRecover_11Dec2012-v1 | 0.090 |
| | /SingleMu/Run2012C-PromptReco-v2 | 6.445 |
| | /SingleMu/Run2012D-PromptReco-v1 | 6.803 |
| | Total Lumi | 19.04 |
| W(e$\nu$)H | /SingleElectron/Run2012A-13Jul2012-v1 | 0.809 |
| | /SingleElectron/Run2012A-recover-06Aug2012-v1 | 0.082 |
| | /SingleElectron/Run2012B-13Jul2012-v1 | 4.403 |
| | /SingleElectron/Run2012C-24Aug2012-v1 | 0.405 |
| | /SingleElectron/Run2012C-2012C-EcalRecover_11Dec2012-v1 | 0.405 |
| | /SingleElectron/Run2012C-PromptReco-v2 | 6.445 |
| | /SingleElectron/Run2012D-PromptReco-v1 | 6.803 |
| | Total Lumi | 19.04 |
| Z(ee)H | /DoubleElectron/Run2012A-13Jul2012-v1 | 0.809 |
| | /DoubleElectron/Run2012A-recover-06Aug2012-v1 | 0.082 |
| | /DoubleElectron/Run2012B-13Jul2012-v1 | 4.403 |
| | /DoubleElectron/Run2012C-24Aug2012-v1 | 0.405 |
| | /DoubleElectron/Run2012C-2012C-EcalRecover_11Dec2012-v1 | 0.090 |
| | /DoubleElectron/Run2012C-PromptReco-v2 | 6.445 |
| | /DoubleElectron/Run2012D-PromptReco-v1 | 6.803 |
| | Total Lumi | 19.04 |
| Z($\nu\nu$)H | /MET/Run2012A-13Jul2012-v1 | 0.809 |
| | /MET/Run2012A-recover-06Aug2012-v1 | 0.082 |
| | /MET/Run2012B-13Jul2012-v1 | 4.403 |
| | /MET/Run2012C-24Aug2012-v1 | 0.405 |
| | /MET/Run2012C-2012C-EcalRecover_11Dec2012-v1 | 0.090 |
| | /MET/Run2012C-PromptReco-v2 | 6.445 |
| | /MET/Run2012D-PromptReco-v1 | 6.803 |
| | Total Lumi | 19.04 |
| W($\tau\nu$)H | /Tau/Run2012A-13Jul2012-v1 | 0.751 |
| | /Tau/Run2012A-recover-06Aug2012-v1 | 0.082 |
| | /Tau/Run2012B-13Jul2012-v1 | 4.122 |
| | /Tau/Run2012C-24Aug2012-v1 | 0.405 |
| | /Tau/Run2012C-2012C-EcalRecover_11Dec2012-v1 | 0.090 |
| | /Tau/Run2012C-PromptReco-v2 | 6.190 |
| | /Tau/Run2012D-PromptReco-v1 | 6.680 |
| | Total Lumi | 18.32 |

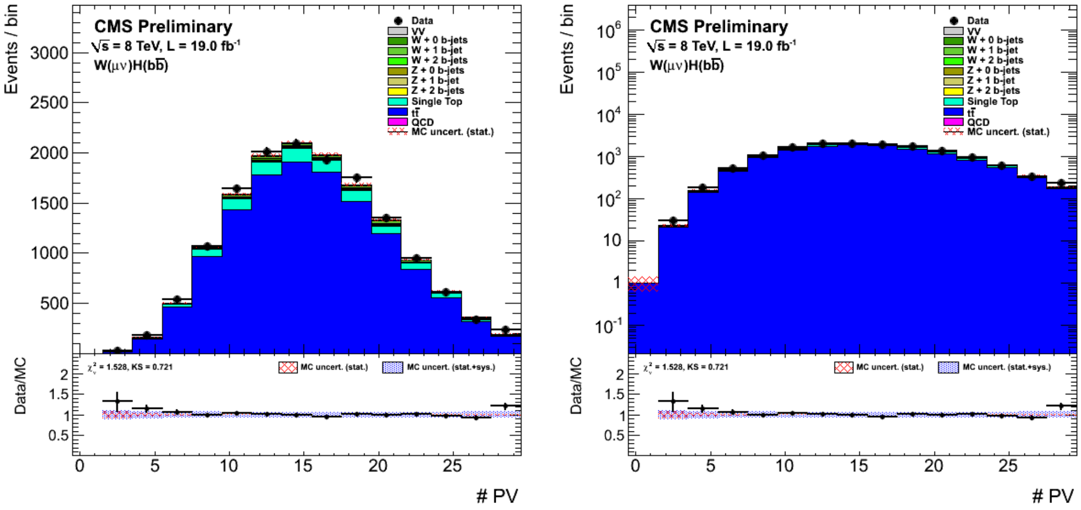

Figure 4.1: Distribution of the number of reconstructed primary vertices in data compared to simulated MC events in the t$\bar{t}$ control sample for the W($\mu\nu$)H event selection. The distribution is shown with linear (left) and logarithmic (right) scales. See Chapter 7 for the definition of this control region.

being sought that can be used in signal extraction. Such an example is the ongoing search for di-Higgs resonances at CMS [134].

Tables 4.2–4.4 summarize the simulated samples used in the VH(b$\bar{b}$) analysis, including equivalent luminosities (or cross sections) where applicable. Like the CMS datasets, Monte Carlo simulation samples are stored on the OSG and easily accessible by CMS analysis groups. Simulated events are produced by a number of different MC generators [135–139] and the particles created by the simulated proton-proton collisions are propagated through a virtual CMS detector using the Geant4 software package [140]. Appropriate pileup reweighting is applied when comparing to data, in order to represent the true primary vertex distribution in the different run ranges. This is applied in Figure 4.1 above to ensure agreement between data and MC. The samples are analyzed with CMSSW version 5.3 with the same configuration as used to analyze data events, discussed in Section 4.1. Table 4.5 summarizes the cross sections and branching fractions assumed for each signal channel and mass point. The cross sections are computed at NNLO, as described in [141].

Table 4.2: List of signal and diboson Monte Carlo samples used in the VH($b\bar{b}$) analysis, along with the integrated luminosity of the samples. Only one signal mass point is shown as an example ($M_H$ = 115 GeV).

| Mode | Dataset | $\mathcal{L}$ (fb$^{-1}$) |
|---|---|---|
| W($\ell\nu$)H | /WH_WToLNu_HToBB_M-115_8TeV-powheg-herwigpp/Summer12_DR53X-PU_S10_START53_V7A-v1 | 6005 |
| Z($\ell\ell$)H | /ZH_ZToLL_HToBB_M-115_8TeV-powheg-herwigpp/Summer12_DR53X-PU_S10_START53_V7A-v1 | 34718 |
| Z($\nu\nu$)H | /ZH_ZToNuNu_HToBB_M-115_8TeV-powheg-herwigpp/Summer12_DR53X-PU_S10_START53_V7A-v1 | 17450 |
| $ZZ$ | /WW_TuneZ2star_8TeV_pythia6_tauola/Summer12_DR53X-PU_S10_START53_V7A-v1 | 1181 |
| $WW$ | /WZ_TuneZ2star_8TeV_pythia6_tauola/Summer12_DR53X-PU_S10_START53_V7A-v1 | 176 |
| $WZ$ | /ZZ_TuneZ2star_8TeV_pythia6_tauola/Summer12_DR53X-PU_S10_START53_V7A-v1 | 295 |

Table 4.3: List of V+jets Monte Carlo samples used in the VH($b\bar{b}$) analysis. Different generator binnings and different generators have been tested to maximize the available statistics and to study systematic uncertainties (see Section 8.4). All sample names end with `/Summer12_DR53X-PU_S10_START53_V7A-v*` (not shown due to limited space).

| |
|---|
| DYJetsToLL_M-50_TuneZ2Star_8TeV-madgraph-tarball |
| DYJetsToLL_PtZ-180_TuneZ2star_8TeV-madgraph |
| DYJetsToLL_PtZ-100_TuneZ2star_8TeV-madgraph |
| DYJetsToLL_PtZ-50To70_TuneZ2star_8TeV-madgraph-tarball |
| DYJetsToLL_PtZ-70To100_TuneZ2star_8TeV-madgraph-tarball |
| WJetsToLNu_PtW-100_TuneZ2star_8TeV-madgraph |
| WJetsToLNu_PtW-100_TuneZ2star_8TeV_ext-madgraph-tarball |
| WJetsToLNu_PtW-180_TuneZ2star_8TeV-madgraph-tarball |
| WJetsToLNu_PtW-50To70_TuneZ2star_8TeV-madgraph |
| WJetsToLNu_PtW-70To100_TuneZ2star_8TeV-madgraph |
| ZJetsToNuNu_50_HT_100_TuneZ2Star_8TeV_madgraph |
| ZJetsToNuNu_100_HT_200_TuneZ2Star_8TeV_madgraph |
| ZJetsToNuNu_200_HT_400_TuneZ2Star_8TeV_madgraph |
| ZJetsToNuNu_400_HT_inf_TuneZ2Star_8TeV_madgraph |
| ZJetsToNuNu_PtZ-100_8TeV-madgraph |
| BJets_HT-100To250_8TeV-madgraph |
| DY1JetsToLL_M-50_TuneZ2Star_8TeV-madgraph |
| DY2JetsToLL_M-50_TuneZ2Star_8TeV-madgraph |
| DY4JetsToLL_M-50_TuneZ2Star_8TeV-madgraph |
| DYJetsToLL_HT-200To400_TuneZ2Star_8TeV-madgraph |
| DYJetsToLL_HT-400ToInf_TuneZ2Star_8TeV-madgraph |
| WJetsToLNu_HT-250To300_8TeV-madgraph |
| WJetsToLNu_HT-300To400_8TeV-madgraph |
| WJetsToLNu_HT-400ToInf_8TeV-madgraph |
| ZJetsToLL_Pt-100_8TeV-herwigpp |
| WJetsToLNu_PtW-100_8TeV-herwigpp |
| ZJetsToNuNu_Pt-100_8TeV-herwigpp |

Table 4.4: List of $t\bar{t}$ and single top Monte Carlo samples used in the VH($b\bar{b}$) analysis, with the corresponding assumed cross section. Different $t\bar{t}$ generators are used to study systematics (see Section 8.4).

| Mode | Dataset | $\sigma$ (pb) |
|---|---|---|
| $t\bar{t}$ | /TTJets_FullLeptMGDecays_8TeV-madgraph-part/Summer12_DR53X-PU_S10_START53_V7A-v* | 24.6 |
| $t\bar{t}$ | /TTJets_SemiLeptMGDecays_8TeV-madgraph-part/Summer12_DR53X-PU_S10_START53_V7A-v* | 103 |
| $t\bar{t}$ | /TTJets_HadronicMGDecays_8TeV-madgraph-part/Summer12_DR53X-PU_S10_START53_V7A-v* | 106 |
| $t\bar{t}$ | /TT_CT10_TuneZ2star_8TeV-powheg-tauola/Summer12_DR53X-PU_S10_START53_V7A-v* | 234 |
| Single Top (tW) | /Tbar_tW-channel-DR_TuneZ2star_8TeV-powheg-tauola/Summer12_DR53X-PU_S10_START53_V7A-v1 | 11.1 |
|  | /T_tW-channel-DR_TuneZ2star_8TeV-powheg-tauola/Summer12_DR53X-PU_S10_START53_V7A-v1 | 11.1 |
| Single Top (t-ch) | /Tbar_t-channel_TuneZ2star_8TeV-powheg-tauola/Summer12_DR53X-PU_S10_START53_V7A-v1 | 30.7 |
|  | /T_t-channel_TuneZ2star_8TeV-powheg-tauola/Summer12_DR53X-PU_S10_START53_V7A-v1 | 56.4 |
| Single Top (s-ch) | /Tbar_s-channel_TuneZ2star_8TeV-powheg-tauola/Summer12_DR53X-PU_S10_START53_V7A-v1 | 1.76 |
|  | /T_s-channel_TuneZ2star_8TeV-powheg-tauola/Summer12_DR53X-PU_S10_START53_V7A-v1 | 3.79 |

Table 4.5: Cross sections for signal events at $\sqrt{s} = 8$TeV and $H \to b\bar{b}$ branching ratios for masses from 110 to 150 GeV. Relative uncertainties are included on these estimates.

| 8TeV | WH |  | ZH |  | $H \to b\bar{b}$ |  |
|---|---|---|---|---|---|---|
| $M_H$ | $\sigma$(pb) | $\pm$ (%) | $\sigma$(pb) | $\pm$ (%) | $\mathcal{B}(H \to b\bar{b})$ | $\pm$ (%) |
| 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 |

## 4.3 Trigger Selection

A variety of different triggers are used to collect events at $\sqrt{s} = 8$ TeV consistent with the signal hypothesis in the VH(b$\bar{\text{b}}$) analysis. The analysis makes use of very simple L1 trigger paths, requiring either one muon candidate, one or two electron/photon candidates, or missing transverse energy in the detector. In addition, a dedicated set of pre-scaled L1 "utility triggers" are used to measure the efficiency of the $E_{\mathrm{T}}^{\mathrm{miss}}$ + jets paths. This information is summarized in Table 4.6 for all channels in the VH(b$\bar{\text{b}}$) analysis.

The HLT paths for the W(e$\nu$)H, W($\mu\nu$)H, and Z($\ell\ell$)H channels consist of several single lepton and di-lepton triggers with tight lepton identification. Leptons are also required to be isolated from other tracks and calorimeter energy deposits to maintain an acceptable trigger rate. For the W($\mu\nu$)H and Z($\mu\mu$)H channels, the trigger thresholds for the muon transverse momentum are in the range of 24 to 40 GeV, where the highest threshold is used without additional isolation requirements. During `Run2012A`, the muon is restricted to $|\eta| < 2.1$ – this restriction is lifted beginning in `Run2012B`. For the W(e$\nu$)H channel, a single isolated-electron trigger is used with a 27 GeV threshold on electron $p_{\mathrm{T}}$. For the Z(ee)H channel, a di-electron trigger with lower $p_{\mathrm{T}}$ thresholds (17 and 8 GeV) and tight isolation requirements is used. For the W($\tau\nu$)H channel trigger, a tau jet from a one-prong hadronically-decaying tau particle is selected for. This is similar to the offline selection, but looser. The $p_{\mathrm{T}}$ of the charged track candidate within the tau jet is required to be above 20 GeV, and the $p_{\mathrm{T}}$ of the tau jet above 35 GeV. An additional requirement of a minimum of 70 GeV is placed on the missing transverse energy.

For the Z($\nu\nu$)H channel, a combination of several HLT paths is used. A trigger requiring $E_{\mathrm{T}}^{\mathrm{miss}}$ greater than 150 GeV is utilized. Included with this trigger is a trigger path requiring two central jets satisfying $p_T > 30$ GeV (60 and 25 GeV for instantaneous luminosity above $3 \times 10^{33}$ cm$^{-2}$s$^{-1}$) and a $E_{\mathrm{T}}^{\mathrm{miss}}$ threshold of 80 GeV, with additional requirements: at least one di-jet with $p_{\mathrm{T}}$ greater than 100 GeV, and no jet in the event with $p_{\mathrm{T}}$ greater than 40 GeV closer than 0.5 in azimuthal angle to the missing transverse energy direction. In order to increase signal acceptance at lower values of missing

Table 4.6: List of L1 and HLT trigger paths used for the 2012 dataset in the VH($b\bar{b}$) analysis. The trigger paths are split according to channel.

| Channel | L1 Trigger Path | HLT Trigger Path |
|---|---|---|
| W($\mu\nu$)H | SingleMu16 | IsoMu24(_eta2p1) |
| | SingleMu16 | Mu40(_eta2p1) |
| Z($\mu\mu$)H | SingleMu16 | IsoMu24(_eta2p1) |
| | SingleMu16 | Mu40(_eta2p1) |
| W(e$\nu$)H | SingleEG20 OR SingleEG22 | Ele27_WP80 |
| Z(ee)H | DoubleEG137 | Ele17_CaloIdT_CaloIsoVL_TrkIdVL_TrkIsoVL_Ele8_CaloIdT_CaloIsoVL_TrkIdVL_TrkIsoVL |
| Z($\nu\bar{\nu}$)H | L1_ETM36 OR L1_ETM40 | HLT_PFMET150 |
| | L1_ETM36 OR L1_ETM40 | HLT_DiCentralPFJet30_PFMHT80 (Run2012A) |
| | L1_ETM36 OR L1_ETM40 | HLT_DiCentralJetSumpT100_dPhi05_DiCentralPFJet60_25_PFMET100_HBHENoiseCleaned (Run2012B-C-D) |
| | L1_ETM36 OR L1_ETM40 | DiCentralJet20_CaloMET65_BTagCSV07_PFMHT80 (Run2012A)<br>DiCentralPFJet30_PFMET80_BTagCSV07 (Run2012B-C-D) |
| W($\tau\nu$)H | L1_ETM36 OR L1_ETM40 | LooseIsoPFTau35_Trk20_Prong1_MET70 |
| Utility Triggers | L1_ETM40 | HLT_L1ETM40 |
| | L1_ETM70 | HLT_L1ETM70 |
| | L1_ETM100 | HLT_L1ETM100 |

transverse energy, triggers that require jets to be identified as coming from b quarks are used. For these triggers, two central jets with $p_{\mathrm{T}}$ above 20 or 30 GeV, depending on the luminosity conditions, are required. It is also required that at least one central jet with $p_{\mathrm{T}}$ above 20 GeV be tagged by an online version of the CSV b-tagging algorithm, discussed in Section 3.4.4. The HLT paths used for each channel are summarized in Table 4.6.

# Chapter 5

# Event Selection

With a set of fully reconstructed events (both data and MC) on hand as discussed in Chapter 4, we can begin to analyze the events by first applying a set of selection cuts on the reconstructed physics objects (principal and composite) to narrow down the phase space used for the signal search. This preliminary step is known as "event selection" and is the first step in extracting a signal consistent with the VH($\mathrm{b\bar{b}}$) final state and topology, which is discussed at length in Chapter 8 . In the end, a residual background exists that must be accounted for when extracting the signal yield. A statistical account of the contributing background processes is discussed further in Chapter 8. The background control samples used to help with the signal region background prediction is discussed in Chapter 7. These background samples are designed to be similar in phase space to the signal region we discuss in this chapter, but still orthogonal (no overlap of events).

As discussed in Section 1.3, the chief analysis strategy we employ is to require "boosted" physics objects. In particular, we require events to have high $p_{\mathrm{T}}$ for both the leptonically-decaying vector boson (W or Z) and the Higgs boson decaying to bottom quarks, which helps to reduce contributions from a number of background processes. Furthermore, we expect the vector boson and the Higgs boson to be well-separated and central (found at low values of $|\eta|$) in signal events. Requiring the Higgs jets to be consistent with bottom quarks is necessary to exclude V+$\mathrm{b\bar{b}}$ (where V generically refers to either W or Z) and QCD multi-jet events – this is done by requiring minimal values of

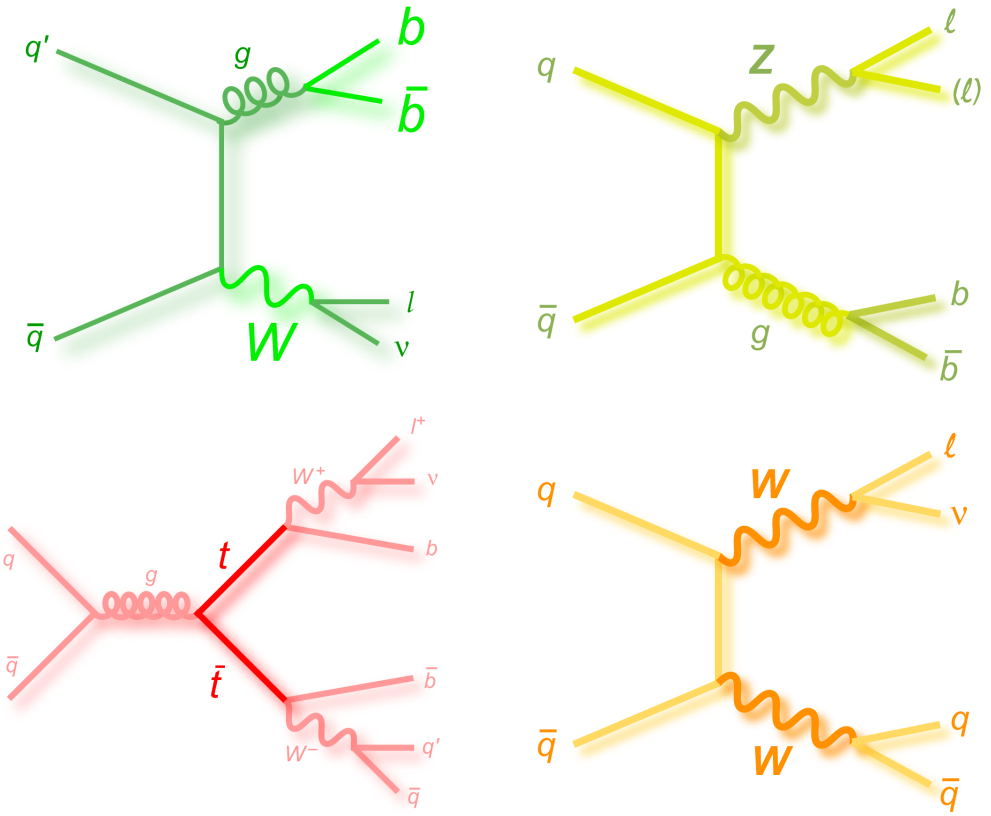

Figure 5.1: Illustration of dominant background processes in the VH(b$\bar{\text{b}}$) search [14]. Presented are the Feynman diagrams for W+b$\bar{\text{b}}$ (top left), Z+b$\bar{\text{b}}$ (top right), t$\bar{\text{t}}$ (bottom left), and di-boson (bottom right) background processes. Only Feynman diagrams associated with the quark-antiquark annihilation production mechanism are shown (for t$\bar{\text{t}}$, the gluon-gluon fusion production mechanism contributes significantly as well).

the CSV discriminant for both jets associated with the Higgs boson physics object. The additional jet activity in the event is very helpful in targeting backgrounds containing single and pair-produced top quarks, while the di-jet mass (constructed from the jets associated with the Higgs physics object) is useful in targeting the otherwise irreducible di-boson background. Finally, a set of additional cuts making use of the event $E_{\mathrm{T}}^{\mathrm{miss}}$ are used in some channels to further reduce the QCD multi-jet background levels, as this background process characteristically populates the low region of $E_{\mathrm{T}}^{\mathrm{miss}}$. The dominant background processes for the VH(b$\bar{\text{b}}$) search are illustrated in Figure 5.1.

We employ two different methods to extract a signal in the VH(b$\bar{\text{b}}$) analysis: one making use of the di-jet mass distribution M(jj) and another making use of the output discriminant of a multivariate classifier, specifically the Boosted Decision Tree (BDT) classifier. These two methods are referred to as the “M(jj) analysis” and “BDT analysis”, respectively. The extraction of signal is done using a “shape analysis” for each case, where the distribution (M(jj) or BDT output) for each simulated signal/background process is simultaneously fit to the distribution in data within the signal region. Our principal result is based solely on the BDT analysis. The M(jj) analysis serves as an important cross-check, as M(jj) is the most discriminating variable in the BDT training and has a very intuitive interpretation when it comes to looking at an excess in data. The details of the BDT training and both shape analyses are discussed further in Chapter 8.

The selection cuts used in the M(jj) and BDT analyses are somewhat different, as the BDT analysis makes use of variable correlations in a broader phase space to help increase the search sensitivity. We separately discuss each set of cuts defining the two signal regions, both of which were optimized to maximize search sensitivity for the particular method. As the BDT analysis yields our main result, we present this selection first in Section 5.1. The tighter signal region for the M(jj) cross-check analysis is presented in Section 5.2, emphasizing differences in event selection with respect to the BDT analysis.

## 5.1 BDT Signal Region

The signal region used for the BDT analysis makes use of boosting, b tagging, and topological variables as described above. The emphasis in this selection is to keep the phase space as broad as possible as to include as much signal as possible, such that the discriminating power of the correlations between the variables included in the BDT training can be fully utilized. However, making the selection too loose may dilute the power of the multi-variate classifier, as many non-interesting events would be included in the training. Therefore an optimization is necessary, taking into account both of these features, in order to settle on a set of event selection criteria. A full optimization maximizing search sensitivity was performed in order to obtain this working point.

Table 5.1 lists the selection criteria used in the event selection for the BDT analysis

Table 5.1: Signal region event selection for the BDT analysis. Entries marked with "—" indicate that the variable is not used in the given channel. If different, the entries in square brackets indicate the selection for the different $p_{\mathrm{T}}(\mathrm{V})$ regions as defined in the first row of the table. The $p_{\mathrm{T}}$ thresholds for the highest and second highest $p_{\mathrm{T}}$ Higgs jets are $p_{\mathrm{T}}(\mathrm{j}_1)$ and $p_{\mathrm{T}}(\mathrm{j}_2)$, respectively. The transverse momentum of the leading (highest $p_{\mathrm{T}}$) tau track is $p_{\mathrm{T}}(\mathrm{track})$. $\mathrm{CSV}_{\mathrm{max}}$ and $\mathrm{CSV}_{\mathrm{min}}$ refer to the CSV discriminant values associated with the Higgs jet with the highest and lowest CSV discriminant values, respectively. $N_{\mathrm{aj}}$ and $N_{\mathrm{al}}$ represent the number of additional jets and isolated leptons, respectively, that are found in addition to the number expected based on the final state of interest. $\vec{E}_{\mathrm{T}}^{\mathrm{miss}}(\mathrm{tracks})$ refers to missing transverse energy vector built for the event by making use of strictly tracking information. The values listed for kinematic variables are in units of GeV, and for angles in units of radians.

| Variable | W($\ell\nu$)H | W($\tau\nu$)H | Z($\ell\ell$)H | Z($\nu\nu$)H |
|---|---|---|---|---|
| $p_{\mathrm{T}}(\mathrm{V})$ | [100–130] [130–180] [$>$ 180] | $>$ 120 | [50–100] [$>$ 100] | [100–130] [130–170] [$>$ 170] |
| $m_{\ell\ell}$ | — | — | 75–105 | — |
| $p_{\mathrm{T}}(\mathrm{j}_1)$ | $>$ 30 | $>$ 30 | $>$ 20 | $>$ 60 |
| $p_{\mathrm{T}}(\mathrm{j}_2)$ | $>$ 30 | $>$ 30 | $>$ 20 | $>$ 30 |
| $p_{\mathrm{T}}(\mathrm{jj})$ | $>$ 100 | $>$ 120 | — | [$>$ 100] [$>$ 130] [$>$ 130] |
| $M(\mathrm{jj})$ | $<$ 250 | $<$ 250 | [40–250] [$<$ 250] | $<$ 250 |
| $E_{\mathrm{T}}^{\mathrm{miss}}$ | $>$ 45 | $>$ 80 | — | [100–130] [130–170] [$>$ 170] |
| $p_{\mathrm{T}}(\tau)$ | — | $>$ 40 | — | — |
| $p_{\mathrm{T}}(\mathrm{track})$ | — | $>$ 20 | — | — |
| $\mathrm{CSV}_{\mathrm{max}}$ | $>$ 0.40 | $>$ 0.40 | [$>$ 0.50] [$>$ 0.244] | $>$ 0.679 |
| $\mathrm{CSV}_{\mathrm{min}}$ | $>$ 0.40 | $>$ 0.40 | $>$ 0.244 | $>$ 0.244 |
| $N_{\mathrm{aj}}$ | — | — | — | [$<$ 2] [—] [—] |
| $N_{\mathrm{al}}$ | = 0 | = 0 | — | = 0 |
| $\Delta\phi(\mathrm{V},\mathrm{H})$ | — | — | — | $>$ 2.0 |
| $\Delta\phi(\vec{E}_{\mathrm{T}}^{\mathrm{miss}}, \mathrm{j}_{\mathrm{nearest}})$ | — | — | — | [$>$ 0.7] [$>$ 0.7] [$>$ 0.5] |
| $\Delta\phi(\vec{E}_{\mathrm{T}}^{\mathrm{miss}}, \vec{E}_{\mathrm{T}}^{\mathrm{miss}}(\mathrm{tracks}))$ | — | — | — | $<$ 0.5 |
| $E_{\mathrm{T}}^{\mathrm{miss}}$ significance | — | — | — | [$>$ 3] [—] [—] |
| $\Delta\phi(\vec{E}_{\mathrm{T}}^{\mathrm{miss}}, \ell)$ | $< \pi/2$ | — | — | — |

signal region. These criteria target a variety of different background processes that could imitate a signal event, described above. The selection criteria are classified according to which $p_{\mathrm{T}}(\mathrm{V})$ category the event falls into (see Section 3.5), as the different categories have differing signal-to-background ratios that influence the optimal selection criteria. Also notable is that there is no requirement placed on additional jet activity (beyond that associated with the Higgs physics object) in the event, which is the main handle against $\mathrm{t\bar{t}}$ and single top quark background. This information is instead used in the BDT training and it is found that the top background is sufficiently separated from signal in the resulting BDT discriminant output.

Another point to note about the BDT analysis event selection is that there is no cut on the di-jet mass near the Higgs mass search range. This allows us to use the same signal region definition for all mass points in the search (ranging from 110 GeV to 150 GeV). However, different BDT trainings are used when looking at a possible signal at different mass points. This is discussed in more detail in Chapter 8.

## 5.2 M(jj) Signal Region

In the VH($\mathrm{b\bar{b}}$) M(jj) analysis, a substantially tighter set of requirements is used for event selection. The selection criteria associated with the signal region for this cross-check are presented in Table 5.2. Now present is a requirement on the additional jet activity in the event, helpful in reducing $\mathrm{t\bar{t}}$ and single-top backgrounds, and tighter b tagging, used to further reduce V+$\mathrm{b\bar{b}}$ background levels. Also used in the M(jj) analysis event selection is an additional requirement on $\Delta\phi(\mathrm{V},\mathrm{H})$, utilizing the back-to-back nature of the vector boson and Higgs boson in the event. This reduces the contribution from a number of different background processes.

The di-jet mass distribution remains untouched in this selection, as this information is directly used in the shape analysis. This variable is still primarily responsible for the separation of the Higgs signal and di-boson background (most of which comes from VZ($\mathrm{b\bar{b}}$) events). It is therefore important to improve the di-jet mass resolution as much as possible, separating both di-boson and VH($\mathrm{b\bar{b}}$) peaks from a relatively smooth di-jet mass spectrum arising from other background processes. We address this concern in Chapter 6.

Table 5.2: Signal region event selection for the M(jj) analysis. Entries marked with "—" indicate that the variable is not used in the given channel. If different, the entries in square brackets indicate the selection for the different boost regions as defined in the first row of the table. Note that the W(e$\nu$)H channel only has two $p_{\mathrm{T}}(\mathrm{V})$ categories in the M(jj) analysis due to limited statistics with the tighter selection criteria. The values listed for kinematic variables are in units of GeV, and for angles in units of radians. For a description of previously undefined variables, please refer to Table 5.1.

| Variable | W($\ell\nu$)H | W($\tau\nu$)H | Z($\ell\ell$)H | Z($\nu\nu$)H |
|---|---|---|---|---|
| $p_{\mathrm{T}}(\mathrm{V})$ | [100–150] [$>$ 150] (e)<br>[100–130] [130–180] [$>$ 180] ($\mu$) | $>$ 120 | [50–100] [100–150] [$>$ 150] | [100–130] [130–170] [$>$ 170] |
| $m_{\ell\ell}$ | — | — | 75–105 | — |
| $p_{\mathrm{T}}(j_1)$ | $>$ 30 | $>$ 30 | $>$ 20 | [$>$ 60] [$>$ 60] [$>$ 80] |
| $p_{\mathrm{T}}(j_2)$ | $>$ 30 | $>$ 30 | $>$ 20 | $>$ 30 |
| $p_{\mathrm{T}}(\mathrm{jj})$ | $>$ 100 | $>$ 120 | — | [$>$ 110] [$>$ 140] [$>$ 190] |
| $N_{\mathrm{aj}}$ | = 0 | = 0 | — | = 0 |
| $N_{\mathrm{al}}$ | = 0 | = 0 | — | = 0 |
| $E_{\mathrm{T}}^{\mathrm{miss}}$ | $>$ 45 | $>$ 80 | $<$ 60 | — |
| $p_{\mathrm{T}}(\tau)$ | — | $>$ 40 | — | — |
| $p_{\mathrm{T}}(\mathrm{track})$ | — | $>$ 20 | — | — |
| $\mathrm{CSV}_{\mathrm{max}}$ | $>$ 0.898 | $>$ 0.898 | $>$ 0.679 | $>$ 0.898 |
| $\mathrm{CSV}_{\mathrm{min}}$ | $>$ 0.5 | $>$ 0.4 | $>$ 0.5 | $>$ 0.5 |
| $\Delta\phi(\mathrm{V},\mathrm{H})$ | $>$ 2.95 | $>$ 2.95 | — | $>$ 2.95 |
| $\Delta R(\mathrm{jj})$ | — | — | [—] [—] [$<$ 1.6] | — |
| $\Delta\phi(\vec{E}_{\mathrm{T}}^{\mathrm{miss}}, \mathrm{j}_{\mathrm{nearest}})$ | — | — | — | [$>$ 0.7] [$>$ 0.7] [$>$ 0.5] |
| $\Delta\phi(\vec{E}_{\mathrm{T}}^{\mathrm{miss}}, \vec{E}_{\mathrm{T}}^{\mathrm{miss}}(\mathrm{tracks}))$ | — | — | — | $<$ 0.5 |
| $\Delta\phi(\vec{E}_{\mathrm{T}}^{\mathrm{miss}}, \ell)$ | $< \pi/2$ | — | — | — |

# Chapter 6

# Jet Energy Regression

A principal concern in the VH($\mathrm{b\bar{b}}$) analysis is ensuring the di-jet mass resolution is as good as possible in order to separate $\mathrm{H} \to \mathrm{b\bar{b}}$ decays from the $\mathrm{Z} \to \mathrm{b\bar{b}}$ resonance. This is essential in order to show that any significant excess above background prediction is coming from a new particle and not due to mis-modeling of the $\mathrm{Z} \to \mathrm{b\bar{b}}$ mass peak. Additionally, the di-jet mass (constructed using the reconstructed b jets formed from the $\mathrm{H} \to \mathrm{b\bar{b}}$ decay) is the most discriminating variable when it comes to separating signal from background in the signal regions defined in Chapter 5. This is true for even for the BDT analysis, which includes the $M(\mathrm{jj})$ variable in the BDT training. Improving the di-jet mass resolution results in less background under the $\mathrm{H} \to \mathrm{b\bar{b}}$ signal mass peak, which translates to increased sensitivity of the VH($\mathrm{b\bar{b}}$) signal search.

The di-jet mass resolution suffers both from detector effects, i.e. the finite resolution of the CMS detector hardware in reconstructing the energy of the jet, as well as smearing from fundamental physics processes such as hadronization. One handle analysts have in improving the resolution is to better reconstruct the energy of b jets in the event, primarily those from the Higgs boson decay. Such a correction can be made on top of the nominal CMS jet energy corrections, discussed in Section 3.4.4, which do not specifically target characteristics of b jets. Presented in this chapter is a dedicated b-jet energy regression and based on regression techniques previously developed at the Tevatron [142]. This regression makes use of boosted decision trees [143] as does the multivariate classifier used for signal extraction, discussed in Chapter 8. The regression, which attempts to correct the $p_{\mathrm{T}}$ of each individual reconstructed jet back to the true $p_{\mathrm{T}}$ of the jet as seen in MC simulation (using the same jet clustering algorithm and nominal jet corrections), improves the resolution and reduces the jet energy scale bias on a per-jet basis. This is done by targeting a variety of variables correlated with the energy scale of b jets, including missing transverse energy and the presence of soft, non-isolated leptons that are characteristic of semi-leptonic B decays that frequently occur in b jets. The resulting improvement in di-jet mass resolution is approximately 15–20% depending on the channel.

We present the details of the b-jet energy regression in two separate sections. First, we discuss aspects of the regression BDT training in Section 6.1, additionally covering the performance of the regression upon application. In Section 6.2 the validation of the tool using data-driven techniques is discussed.

## 6.1 Regression Training

The b-jet energy regression is a multivariate tool that uses simulated events (in our case, VH($\mathrm{b\bar{b}}$) signal events) to learn the correlations of different b-jet observables with the true energy scale of the jet, outputting a set of weights. These weights can then be applied to individual b jets in events from either CMS-collected data or simulation (both background and signal, ensuring no overlap of events with the training sample) to correct event observables (most importantly the di-jet mass) that are input to the final shape fit for signal extraction. We make use of the TMVA package [144] to perform the regression, using the BDT algorithm. This multivariate algorithm was picked for both the regression and classifier (discussed in Chapter 8) due to its high-performance and widespread use within the CMS collaboration.

Simulated events from an ensemble of Higgs signal mass points, every 5 GeV between 110 GeV and 150 GeV (inclusive), are used in the training to ensure that there is no bias related to the mass scale of the di-jet system. On the order of 50,000 events from each signal MC sample are used in the training. Separate sets of events are used to train and test the samples, the latter step used to check against over-training of the regression, where the regression picks up on statistical features of the samples instead of true variable correlations. The three main topologies (Z($\ell\ell$)H, Z($\nu\nu$)H, and W($\ell\nu$)H) are trained separately, using a common set of training variables for all regression trainings. The target of the regression is the generator-level $p_{\mathrm{T}}$ of the jets associated with the Higgs decay (looking at one jet at a time), including neutrinos – in other words, the generator-level jets associated with each reconstructed b jet from the H $\rightarrow$ b$\bar{\mathrm{b}}$ decay.

The variables used in the BDT regression training can be classified into a few different categories. First of all, kinematical variables associated with the jet, such as jet $p_{\mathrm{T}}$, jet $\eta$, etc. are used in the training. Other jet-related properties are used as well, such as the total number of constituents in the jet and the $p_{\mathrm{T}}$ of the leading (highest $p_{\mathrm{T}}$) track in the jet. Additionally, if a secondary vertex is reconstructed for the jet, then a variety of vertex-related variables are used in the training, such as the distance between the primary vertex and the location of the secondary vertex, as well as the relative uncertainty on that length. If a soft lepton is present in the jet, targeting semi-leptonic B decays in the jet, then various properties of the lepton are used in the training, including the $p_{\mathrm{T}}$ of the lepton, the momentum of the lepton transverse to the jet axis, and the separation of the lepton from the jet center in the $\eta - \phi$ plane. Finally, for the Z($\ell\ell$)H modes only, the $E_{\mathrm{T}}^{\mathrm{miss}}$ of the event (and related variables, such as the azimuthal separation of $\vec{E}_{\mathrm{T}}^{\mathrm{miss}}$ and the transverse momentum of the nearest jet) is used in the training, as there should be no "real" $E_{\mathrm{T}}^{\mathrm{miss}}$ coming from neutrinos in these final states, attributing the observed $E_{\mathrm{T}}^{\mathrm{miss}}$ to jet energy mis-measurement. Agreement between data and MC for each of these variables is first ensured before being used in the regression training. An example of the typical agreement between data and MC for the training variables is presented in Figure 6.1 for a small subset of the variables used. Data/MC comparison for more of these variables is presented in Chapter 7.

The regression is trained on jets in signal MC events satisfying $p_{\mathrm{T}} > 20$ GeV, $|\eta| < 2.4$, and CSV output greater than zero. The results are evaluated on independent single MC samples and shown in Figure 6.2, comparing di-jet mass resolution before and after appli-

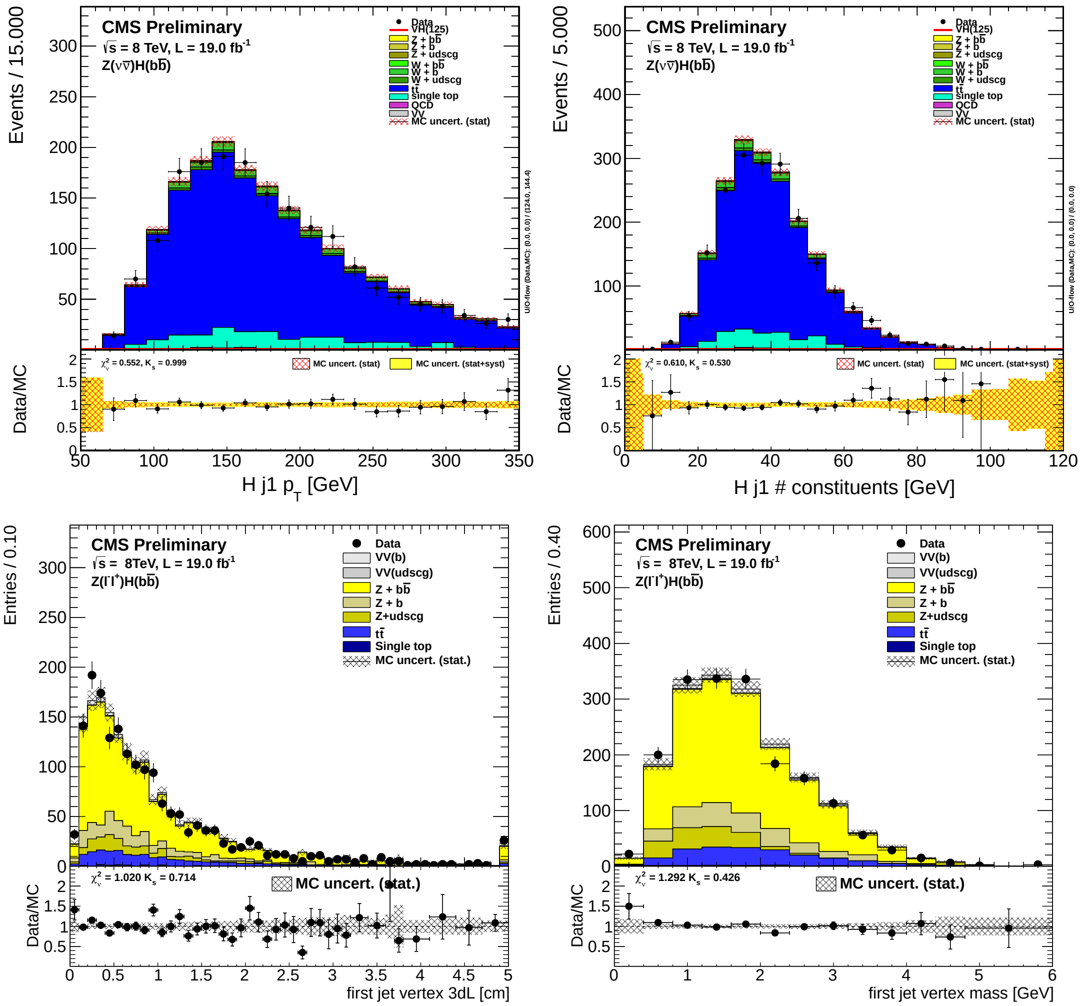

Figure 6.1: Comparison of input variables to the b-jet energy regression in the Z($\nu\nu$)H $t\bar{t}$ control sample (top row) and Z($\ell\ell$)H Z+$b\bar{b}$ control sample (bottom row). Definitions of these control samples can be found in Chapter 7. The variables shown (all for the highest $p_{\mathrm{T}}$ Higgs jet in the event) are the jet $p_{\mathrm{T}}$ (top left), the number of constituents in the jet (top right), the three-dimensional distance between the primary vertex and the secondary vertex of the jet (bottom left), and the mass of the secondary vertex of the jet (bottom right). Good agreement between data and MC is found for all variables.

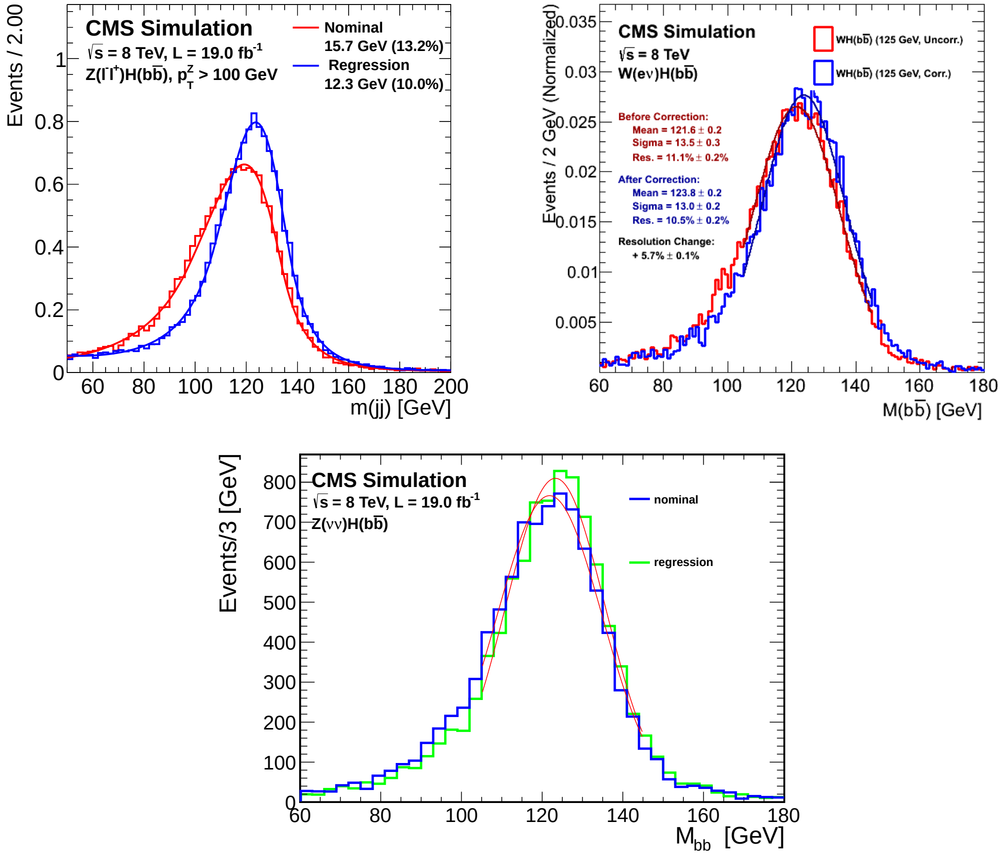

Figure 6.2: Distributions of di-jet invariant mass in MC signal events for the Z($\ell\ell$)H (top left), W($\ell\nu$)H (top right), and Z($\nu\nu$)H (bottom) channels before and after the b-jet energy regression is applied.

cation of the regression. Resolution gains of 15–20% are found for the Z($\ell\ell$)H channels, depending on $p_{\mathrm{T}}$(V) category, while gains of 5–10% are observed for the W($\ell\nu$)H and Z($\nu\nu$)H channels. The neutrinos present in semi-leptonic B decays within the b jet degrade the jet momentum measurement significantly, leading to a bias in the reconstructed momentum and a worsening of the momentum resolution. Most of the power of the regression comes from targeting the effects of the neutrinos, leading to more gains in the Z($\ell\ell$)H channel where the event $E_{\mathrm{T}}^{\mathrm{miss}}$ can be used in the regression training. The separation between VH(b$\bar{\mathrm{b}}$) and VZ(b$\bar{\mathrm{b}}$) di-jet invariant mass resonances also improves considerably as shown in Figure 6.3 for the Z($\ell\ell$)H channel. This translates to increased analysis sensitivity, with an approximate gain of 15% in sensitivity across all channels.

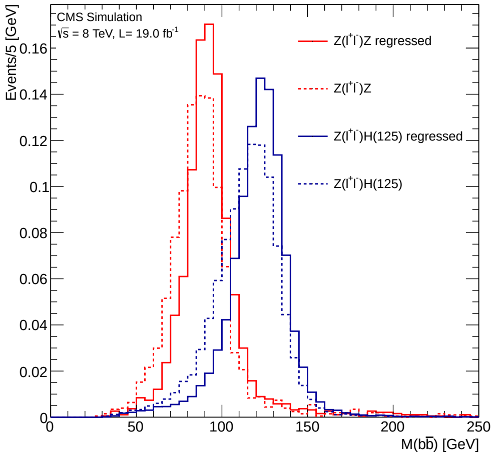

Figure 6.3: ZZ($\mathrm{b\bar{b}}$) and ZH($\mathrm{b\bar{b}}$) di-jet invariant mass resonances in the Z($\ell\ell$)H channel before and after the regression is applied. The regression significantly improves the separation.

## 6.2 Regression Validation

It is important to validate the b-jet energy regression with data to make sure that the algorithm is functioning as expected, leading to no bias in extracting a hypothesized signal. The first check we make is to look for an explicit bias in the di-jet mass after the application of the regression. In order to do this, we make use of the background control samples since the signal region remains blinded until the techniques of the analysis are fully validated. Figure 6.4 illustrates comparisons of data and MC simulation for different control samples in the W($\ell\nu$)H channel (as an example), demonstrating that backgrounds are not biased by the regression technique.

Two further data-driven checks are done in order to validate the b-jet energy regression methodology. The first check makes use of a sample of data events from a control sample defined very similarly to the Z($\ell\ell$)H Z+$\mathrm{b\bar{b}}$ control sample discussed in Chapter 7. This control sample requires basic jet and lepton identification cuts, a di-lepton mass cut consistent with the Z boson mass [145], no additional jets in the event (beyond the two associated with the Higgs boson candidate), CSV output values greater than 0.5 for both Higgs jets, and $p_{\mathrm{T}}(\mathrm{Z}) > 50$ GeV. The electron and muon modes are combined

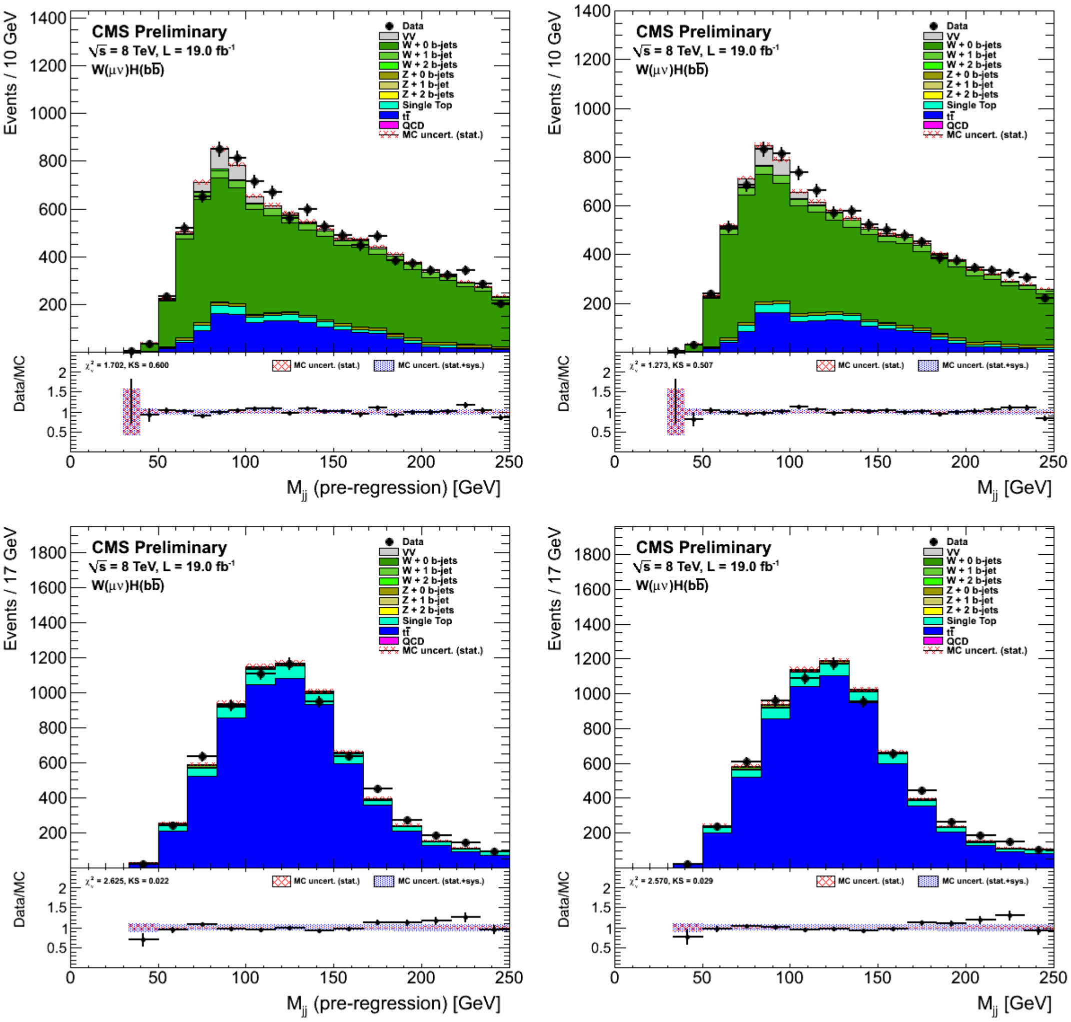

Figure 6.4: Distributions of di-jet invariant mass in background control samples for the W($\ell\nu$)H channel enhanced in W+udscg (top row) and t$\bar{\text{t}}$ (bottom row) backgrounds, both before (left) and after (right) the b-jet energy regression is applied. "W+udscg" refers to a W boson produced with one or more jets that are not associated with a b quark (or antiquark).

Table 6.1: Fit parameters from Gaussian fits to the core of the $p_{\mathrm{T}\mathrm{Balance}}$ distribution, for both data and MC simulation. Events are selected using a Z+b$\bar{\mathrm{b}}$ control sample in the Z($\ell\ell$)H channel.

| **Data** | Before Regression | After Regression |
|---|---|---|
| Constant | $517.1 \pm 13.3$ | $579.5 \pm 14.8$ |
| Mean | $0.929 \pm 0.008$ | $0.985 \pm 0.005$ |
| Sigma | $0.226 \pm 0.010$ | $0.192 \pm 0.007$ |
| **MC** | Before Regression | After Regression |
| Constant | $535.5 \pm 8.2$ | $586.4 \pm 8.8$ |
| Mean | $0.938 \pm 0.004$ | $0.981 \pm 0.003$ |
| Sigma | $0.205 \pm 0.005$ | $0.189 \pm 0.004$ |

together in this sample. In these events, with a Z+b$\bar{\mathrm{b}}$ purity of approximately 70%, we use the distribution of the ratio between the $p_{\mathrm{T}}$(jj) and the $p_{\mathrm{T}}$ of the di-lepton system,

$$p_{\mathrm{T}\mathrm{Balance}} = \frac{p_{\mathrm{T}}(\mathrm{jj})}{p_{\mathrm{T}}(\ell\ell)}, \tag{6.1}$$

to quantify the improvement coming from the regression. The distribution of this variable, for both data and MC simulation, before and after applying the regression is presented in Figure 6.5. The mean of the distribution is sensitive to the jet energy scale, while the width is sensitive to the jet energy resolution. The agreement between data and MC, already good before the application of the regression, further improves with the correction. Also presented in Figure 6.5 is a Gaussian fit the core of the distribution, using events in data, before and after the application of the regression. The fit parameters are shown in Table 6.1, for both data and simulated events.. The scale bias is improved with the regression as the mean shifts toward unity for both data and simulation, as expected from the topology of Z+b$\bar{\mathrm{b}}$ events, while the resolution after the application of the regression is similar for both data and simulated events.

The second data-driven check makes use of a W($\ell\nu$)H control sample enriched with single top events, containing a large amount of t$\bar{\mathrm{t}}$ background as well. The sample is constructed by requiring additional jet activity in the forward regions of the detector (characteristic of single top events) and less-boosted physics objects, and is otherwise very similar to the W($\ell\nu$)H W+b$\bar{\mathrm{b}}$ control sample defined in Chapter 7. We reconstruct

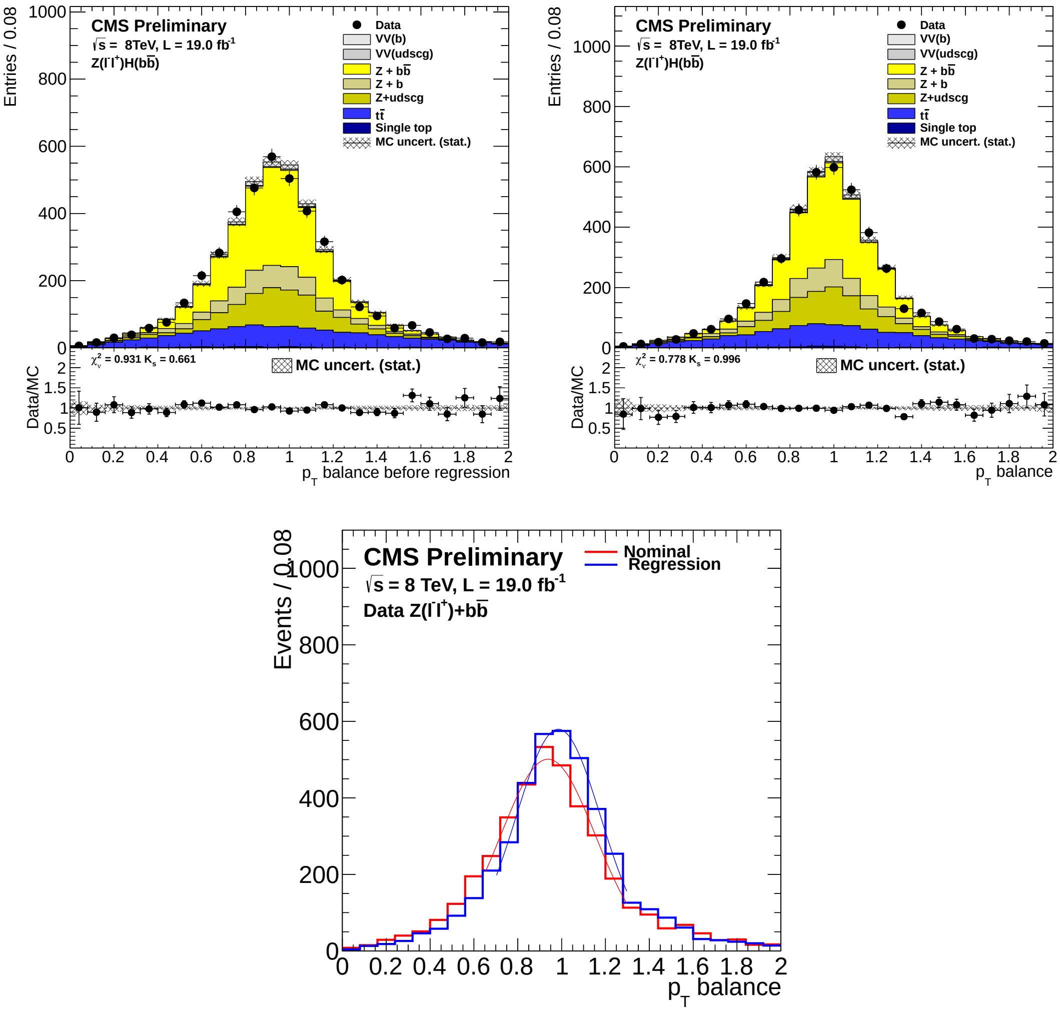

Figure 6.5: Distribution of $p_{\mathrm{T}\,\mathrm{Balance}}$ in both data and simulated events before (top left) and after (top right) the application of the b-jet energy regression, for a Z($\ell\ell$)H Z+b$\bar{\mathrm{b}}$ control sample. Also illustrated is a Gaussian fit to the core of this distribution using events in data (bottom), before (red) and after (blue) the regression application. Both mean and resolution of the fitted Gaussian improve with the use of the regression.

a top quark candidate by taking the b jet with the highest CSV output in the event and combining it with the leptonically-decaying W candidate, imposing a W boson mass [145] constraint to reconstruct the full W boson four-vector. The mass of the top quark candidate is then used as a candle to test the performance of the b-jet energy regression. In Figure 6.6 the reconstructed top mass is shown both before and after the application of the regression, comparing data to MC simulation. Also illustrated are Gaussian fits to the core of the top mass distribution for both data and MC simulation (combining all backgrounds together), again before and after the regression. The top quark mass moves closer to the expected value [145] after the regression is applied, demonstrating a decrease in bias with respect to the jet energy scale. Additionally, the top quark mass resolution improves with the regression, in line with expectations.

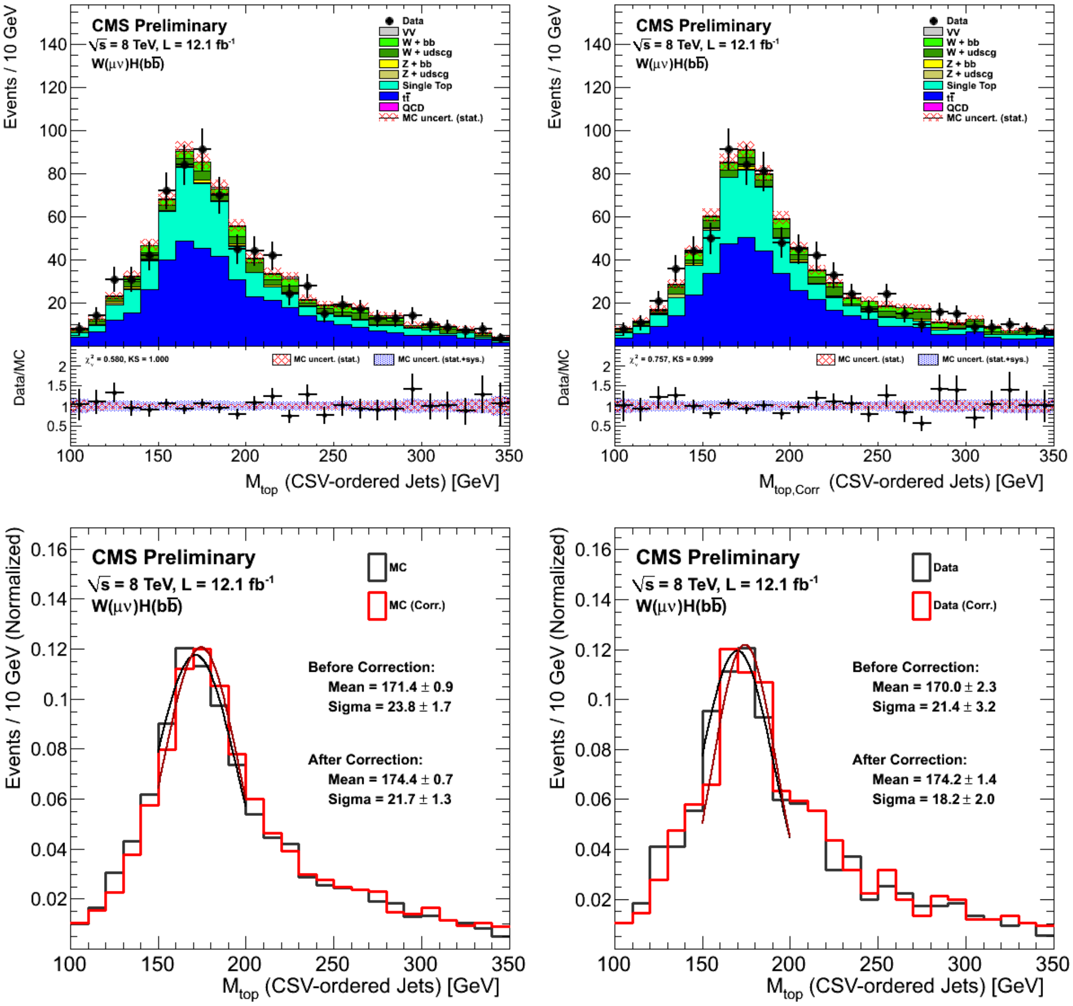

Figure 6.6: Reconstructed top quark mass distribution before (top left) and after (top right) the application of the b-jet energy regression, comparing data and simulated events. Also shown are Gaussian fits to the core the top mass distribution for both data (bottom left) and MC simulation (bottom right), comparing before and after the regression is applied. For both data and simulation there is less bias on the reconstructed top quark mass after the regression, which also considerably improves the top mass resolution.

# Chapter 7

# Background Control Samples

The VH($\mathrm{b\bar{b}}$) analysis employs a thorough blinding strategy in order to ensure that there is no bias in signal extraction due to mis-understood behavior of background processes in the signal region (defined in Chapter 5). This effort requires the construction of background control samples, or control regions, where individual background processes are isolated with relatively high purity so that they may be studied in detail. Each control sample (CS) is defined to be as close as possible to the signal region in terms of phase space, while ensuring orthogonality to the signal region (i.e. no shared events). It is also sometimes necessary to loosen certain selection criteria in these control regions in order to gain enough events to effectively compare data and MC simulation of a particular background process. This comparison of data and MC simulation is essential as the analysis makes use of simulated events in the signal region for the prediction of the contribution from background events.

Control samples in the VH($\mathrm{b\bar{b}}$) analysis are established separately for each channel (i.e. Z($\ell\ell$)H, Z($\nu\nu$)H, W($\ell\nu$)H, and W($\tau\nu$)H), combining together the electron and muon modes where applicable. Only backgrounds that contribute significantly in the signal region are targeted with this approach. For W($\ell\nu$)H and W($\tau\nu$)H, three different background control samples are required: W+udscg, W+$\mathrm{b\bar{b}}$, and $\mathrm{t\bar{t}}$. Here, "W+udscg" refers to an enrichment of W+jets events where the jets originate from light quarks or gluons only, while "W+$\mathrm{b\bar{b}}$" refers to W+jets events where very often at least one of the two jets associated with the Higgs boson candidate originates from a bottom quark or bottom antiquark (usually due to a gluon splitting into a $b\bar{b}$ pair within the event). The control samples for the $Z(\ell\ell)H$ channel are similar to the ones for $W(\ell\nu)H$ and $W(\tau\nu)H$: Z+udscg, $Z+b\bar{b}$, and $t\bar{t}$, with the two Z+jets control samples defined similarly as for W+jets, described above. The $Z(\nu\nu)H$ channel signal region has more contributing background processes in comparison to the other channels, and so requires five different control samples: W+udscg, $W+b\bar{b}$, Z+udscg, $Z+b\bar{b}$, and $t\bar{t}$.

A variety of different background processes are looked at in each control sample by way of MC simulation (the associated MC samples are discussed in Chapter 4). Some of the more important background processes are discussed briefly in Chapter 5. The following background processes are studied: W+0b, W+1b, W+2b, Z+0b, Z+1b, Z+2b, $t\bar{t}$, single top backgrounds (tW, t-channel, and s-channel production mechanisms), diboson backgrounds (WW, WZ, and ZZ), and QCD multi-jet events. Here, "V+Nb" refers to an event with a leptonically-decaying W or Z boson and N of the jets associated with the Higgs boson candidate (up to two) originating from a bottom quark or bottom antiquark. In the case of W+jets and Z+jets events, one must be careful not to confuse the background processes with the control samples of similar name. For example, the $W+b\bar{b}$ control sample contains W+0b, W+1b, and W+2b events, but is particularly enriched in W+2b events (and somewhat enriched in W+1b events). Likewise, the W+udscg control sample contains all three processes but is more enriched with W+0b events.

For each channel, a set of simultaneous fits to data are performed using all control samples, floating background normalization for each of the simulated V+Nb and $t\bar{t}$ processes. The shapes used in these fits vary by channel and control sample: either the CSV discriminant outputs associated with the Higgs candidate jets ($\mathrm{CSV}_{\mathrm{max}}$ or $\mathrm{CSV}_{\mathrm{min}}$) or $M(\mathrm{jj})$, the former due to the ability to discriminate between the different W+jets and Z+jets processes (e.g. Z+1b vs. Z+2b) and the latter for historical reasons. The output from each simultaneous fit is a set of data/MC scale factors for the V+Nb and $t\bar{t}$ background processes that are then applied to the MC background prediction in the signal region. In this way we control the background level in the signal region with a data-driven method. Associated with these scale factors are a set of systematic uncertainties, described in Section 8.4, to be used in extracting the hypothesized VH($\mathrm{b\bar{b}}$) signal. After the data/MC scale factors are computed, they are applied to the background processes in each control sample. Agreement between data and MC in the control samples is then sought for all variables used in the VH($\mathrm{b\bar{b}}$) analysis, serving as a method of validating the analysis before proceeding with unblinding the data in the signal region and attempting the extraction of a hypothesized signal.

Presented in Sections 7.1, 7.2, 7.3, and 7.4 are the definitions of the background control samples for the Z($\ell\ell$)H, Z($\nu\nu$)H, W($\ell\nu$)H, and W($\tau\nu$)H channels, respectively. In each section, a comparison of data and MC is shown for several variables used in the analysis (after the application of the b jet energy regression discussed in Chapter 6, where applicable). This represents only a small number of plots scrutinized in the VH($\mathrm{b\bar{b}}$) analysis before unblinding – the full set is not listed here due to space limitations. Finally, in Section 7.5 the data/MC scale factors obtained in the simultaneous fits to the control samples are presented and discussed for all channels.

## 7.1 Z($\ell\ell$)H Control Samples

The various control samples utilized by the Z($\ell\ell$)H channel analysis are defined in Table 7.1. The control samples are split by lepton type (electron or muon) only – the same control samples are used to validate both $p_{\mathrm{T}}$(V) categories for the Z($\ell\ell$)H analysis. A sampling of the control plots formed using these control regions are presented in Figure 7.1. Here the di-jet $p_{\mathrm{T}}$ and vector boson $p_{\mathrm{T}}$ are highlighted.

It is important to note that in the case of Z($\ell\ell$)H, the control samples used to obtain the data/MC scale factors are different from the ones used to validate the MC simulation variable shapes used in the analysis. This is because the fit control samples are closer to the signal region, but less pure in the backgrounds of interest (as opposed to those defined above) and so less helpful in visually comparing data and MC in terms of variable shapes. These alternative control samples used in the simultaneous fit for Z($\ell\ell$)H are

Table 7.1: Definition of the three control samples for the Z(ee)H and Z($\mu\mu$)H channels. The values listed for kinematic variables are in units of GeV, and for angles in units of radians. Note that there is only one control sample for each background, used for both $p_{\mathrm{T}}(\mathrm{V})$ categories, in the Z(ee)H and Z($\mu\mu$)H analyses.

| Variable | Z+udscg | $\mathrm{t\bar{t}}$ | Z+$\mathrm{b\bar{b}}$ |
|---|---|---|---|
| $p_{\mathrm{T}}(\mathrm{j}_1)$ | $> 20$ | $> 20$ | $> 20$ |
| $p_{\mathrm{T}}(\mathrm{j}_2)$ | $> 20$ | $> 20$ | $> 20$ |
| $p_{\mathrm{T}}(\mathrm{jj})$ | $> 100$ | $> 100$ | — |
| $p_{\mathrm{T}}(\mathrm{Z})$ | $> 100$ | — | — |
| $\mathrm{CSV}_{\mathrm{max}}$ | 0.0–0.898 | $> 0.898$ | $> 0.898$ |
| $\mathrm{CSV}_{\mathrm{min}}$ | $> 0.0$ | $> 0.5$ | $> 0.5$ |
| $N_{\mathrm{aj}}$ | $< 2$ | — | $< 2$ |
| $N_{\mathrm{al}}$ | $= 0$ | $= 0$ | $= 0$ |
| $\Delta\phi(\mathrm{V}, \mathrm{H})$ | $> 2.9$ | — | $> 2.9$ |
| $m_{\ell\ell}$ | — | $\notin$[75–120] | — |
| $M(\mathrm{jj})$ | — | — | $\notin$[90–145] |

presented in Section 7.5.

## 7.2 Z($\nu\nu$)H Control Samples

The five control samples made use of in the Z($\nu\nu$)H analysis are defined in Table 7.2. There is one control sample per background for each $p_{\mathrm{T}}(\mathrm{V})$ category, defined in terms of the event $E_{\mathrm{T}}^{\mathrm{miss}}$, as opposed to the case of Z($\ell\ell$)H. Various variable distributions are illustrated in Figure 7.2, sampling from the various Z($\nu\nu$)H control samples (for the high $p_{\mathrm{T}}(\mathrm{V})$ category only).

The prediction of the contribution from QCD multi-jet events using MC simulation is difficult for the Z($\nu\nu$)H channel due to statistical issues – a negligible number of simulated events from high cross section processes remain after all selection criteria have been applied. Instead, data-driven methods are used to check the contribution in the signal region and control samples. These methods employ an extrapolated yield prediction using the yield in data after inverting anti-QCD selection requirements, i.e. the requirements placed on $\Delta\phi(\vec{E}_{\mathrm{T}}^{\mathrm{miss}}, \mathrm{j}_{\mathrm{nearest}})$ and $\Delta\phi(\vec{E}_{\mathrm{T}}^{\mathrm{miss}}, \vec{E}_{\mathrm{T}}^{\mathrm{miss}}(\mathrm{tracks}))$. It is found

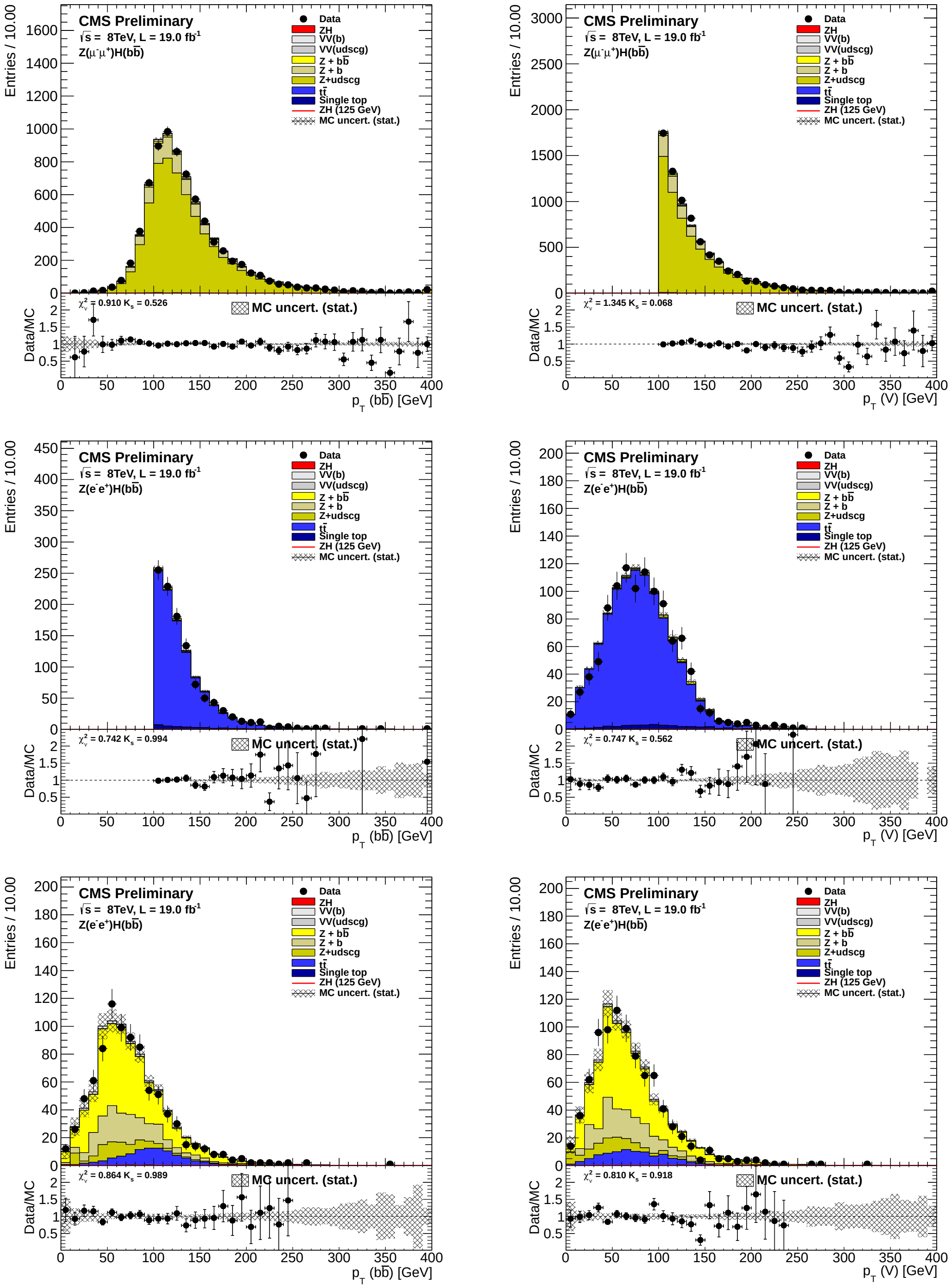

Figure 7.1: Distributions of variables in data and simulated samples in the Z($\ell\ell$)H channels for the Z+udscg (top row), t$\bar{\text{t}}$ (middle row), and Z+b$\bar{\text{b}}$ (bottom row) control samples. The $p_{\text{T}}$(jj) (left column) and $p_{\text{T}}$(Z) (right column) variables are highlighted here. The plots are normalized using the scale factors to facilitate shape comparison. The Z+udscg control sample shown is for the Z($\mu\mu$)H channel while the others are for the Z(ee)H channel.

Table 7.2: Definition of the five control regions for the Z($\nu\nu$)H channel. As different $p_{\mathrm{T}}$(V) categories of the analysis have their own control samples, the entries in square brackets indicate differences in selection for the different categories, labeled near the top of the table (in terms of event $E_{\mathrm{T}}^{\mathrm{miss}}$). The values listed for kinematic variables are in units of GeV, and for angles in units of radians.

| Variable | Z+udscg | Z+b$\bar{\mathrm{b}}$ | t$\bar{\mathrm{t}}$ | W+udscg | W+b$\bar{\mathrm{b}}$ |
|---|---|---|---|---|---|
| $E_{\mathrm{T}}^{\mathrm{miss}}$ | [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 | $> 0.679$ |
| $\mathrm{CSV}_{\mathrm{min}}$ | — | $> 0.244$ | — | — | $> 0.244$ |
| $N_{\mathrm{aj}}$ | [< 2] [—] [—] | [< 2] [—] [—] | $\geq 1$ | $= 0$ | $= 0$ |
| $N_{\mathrm{al}}$ | $= 0$ | $= 0$ | $= 1$ | $= 1$ | $= 1$ |
| $\Delta\phi(\mathrm{V,H})$ | — | $> 2.0$ | — | — | $> 2.0$ |
| $\Delta\phi(\vec{E}_{\mathrm{T}}^{\mathrm{miss}}, \mathrm{j}_{\mathrm{nearest}})$ | [> 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] |
| $\Delta\phi(\vec{E}_{\mathrm{T}}^{\mathrm{miss}}, \vec{E}_{\mathrm{T}}^{\mathrm{miss}}(\mathrm{tracks}))$ | $< 0.5$ | $< 0.5$ | — | — | — |
| $E_{\mathrm{T}}^{\mathrm{miss}}$ significance | [> 3] [—] [—] | [> 3] [—] [—] | [> 3] [—] [—] | [> 3] [—] [—] | [> 3] [—] [—] |

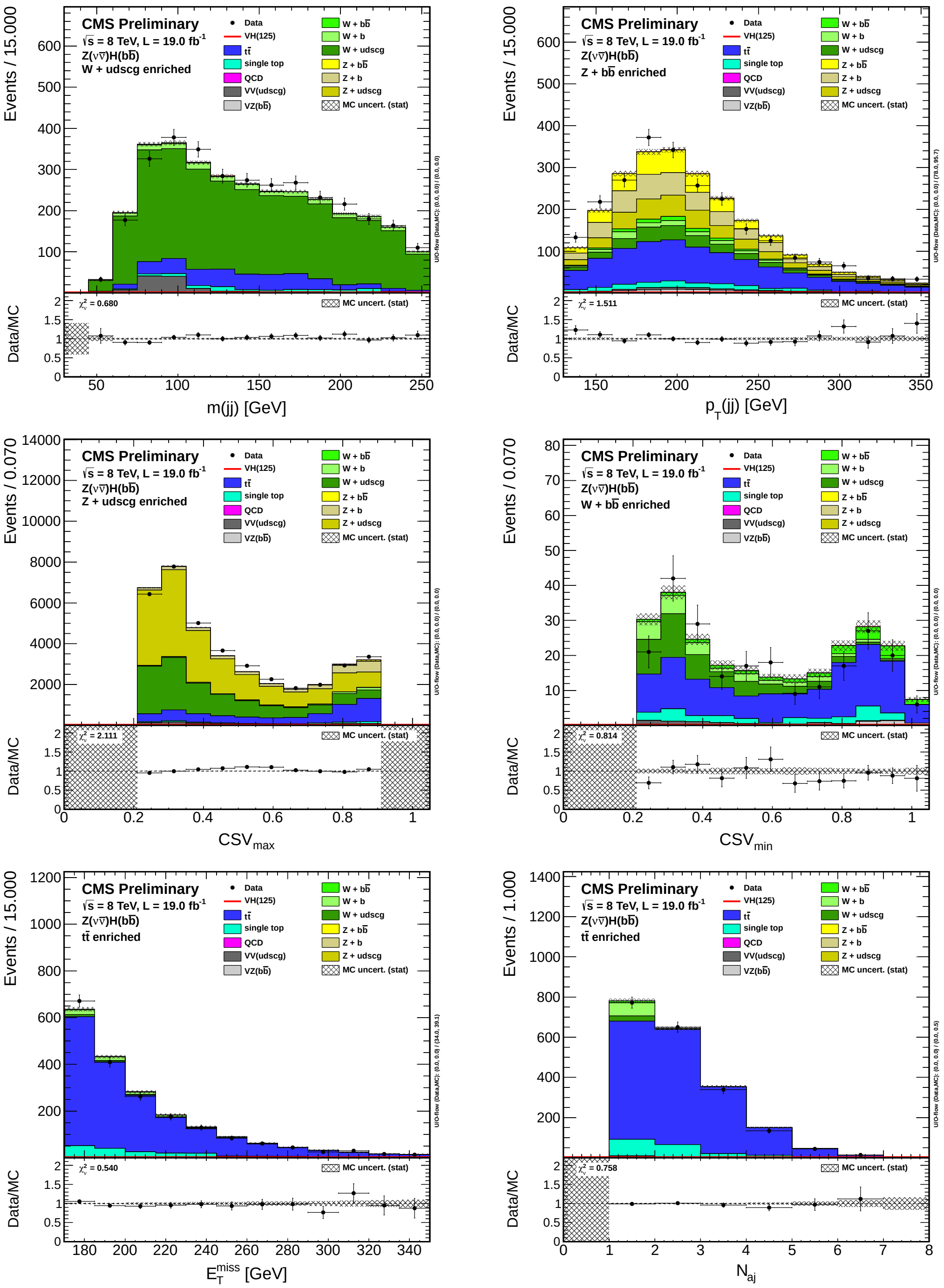

Figure 7.2: Various control plots from the Z($\nu\nu$)H analysis control samples, normalizing MC to data. The variables shown include the $M$(jj) distribution in the W+udscg control sample (upper left), the $p_{\mathrm{T}}$(jj) distribution in the Z+b$\bar{\mathrm{b}}$ control sample (upper right), the CSV$_{\mathrm{max}}$ distribution in the Z+udscg control sample (middle left), the CSV$_{\mathrm{min}}$ distribution in the W+b$\bar{\mathrm{b}}$ control sample (middle right), and both the $E_{\mathrm{T}}^{\mathrm{miss}}$ and $N_{\mathrm{aj}}$ distributions in the t$\bar{\mathrm{t}}$ control sample (bottom row). Only the high $p_{\mathrm{T}}$(V) category is shown here.

Table 7.3: Definition of the control samples for the W($\mu\nu$)H and W(e$\nu$)H channels. The same selection is used for all $p_{\mathrm{T}}$(V) categories (aside from the requirements on $p_{\mathrm{T}}$(W)). The values listed for kinematic variables are in units of GeV.

| Variable | W+udscg | $\mathrm{t\bar{t}}$ | W+$\mathrm{b\bar{b}}$ |
|---|---|---|---|
| $p_{\mathrm{T}}(\mathrm{j}_1)$ | $> 30$ | $> 30$ | $> 30$ |
| $p_{\mathrm{T}}(\mathrm{j}_2)$ | $> 30$ | $> 30$ | $> 30$ |
| $p_{\mathrm{T}}(\mathrm{jj})$ | $> 100$ | $> 100$ | $> 100$ |
| $M(\mathrm{jj})$ | $< 250$ | $< 250$ | $< 250$, $\notin$[90–150] |
| $\mathrm{CSV}_{\mathrm{max}}$ | 0.244–0.898 | $> 0.898$ | $> 0.898$ |
| $N_{\mathrm{aj}}$ | $< 2$ | $> 1$ | $= 0$ |
| $N_{\mathrm{al}}$ | $= 0$ | $= 0$ | $= 0$ |
| $E_{\mathrm{T}}^{\mathrm{miss}}$ | $> 45$ | $> 45$ | $> 45$ |
| $E_{\mathrm{T}}^{\mathrm{miss}}$ significance | $> 2.0$ ($\mu$), $> 3.0$ (e) | — | — |

that the QCD multi-jet contribution in the high $p_{\mathrm{T}}$(V) and intermediate $p_{\mathrm{T}}$(V) categories is negligible, but at the few percent level for the low $p_{\mathrm{T}}$(V) category. The extrapolation is used to create QCD multi-jet variable shapes to be used in the signal region during signal extraction.

## 7.3 W($\ell\nu$)H Control Samples

The selection criteria for the W($\ell\nu$)H control samples are presented in Table 7.3, as for the other channels. The $E_{\mathrm{T}}^{\mathrm{miss}}$ requirements used are to define the control samples as closely as possible to the signal region, which makes use of the transverse missing energy in the event. For the W+$\mathrm{b\bar{b}}$ control sample, the second jet is not required to be b-tagged in order to increase the statistics of the sample. Several variable distributions in each of the W($\ell\nu$)H control samples are presented in Figure 7.3, highlighting some of the more important variables in the analysis.

As for the Z($\nu\nu$)H channel, special methods are used to check for contribution from QCD multi-jet events in the W($\ell\nu$)H channels. A yield extrapolation is done using excess data events (above the MC prediction from other background processes) after inverting the $E_{\mathrm{T}}^{\mathrm{miss}}$ and lepton isolation requirements, nominally used to suppress QCD multi-jet

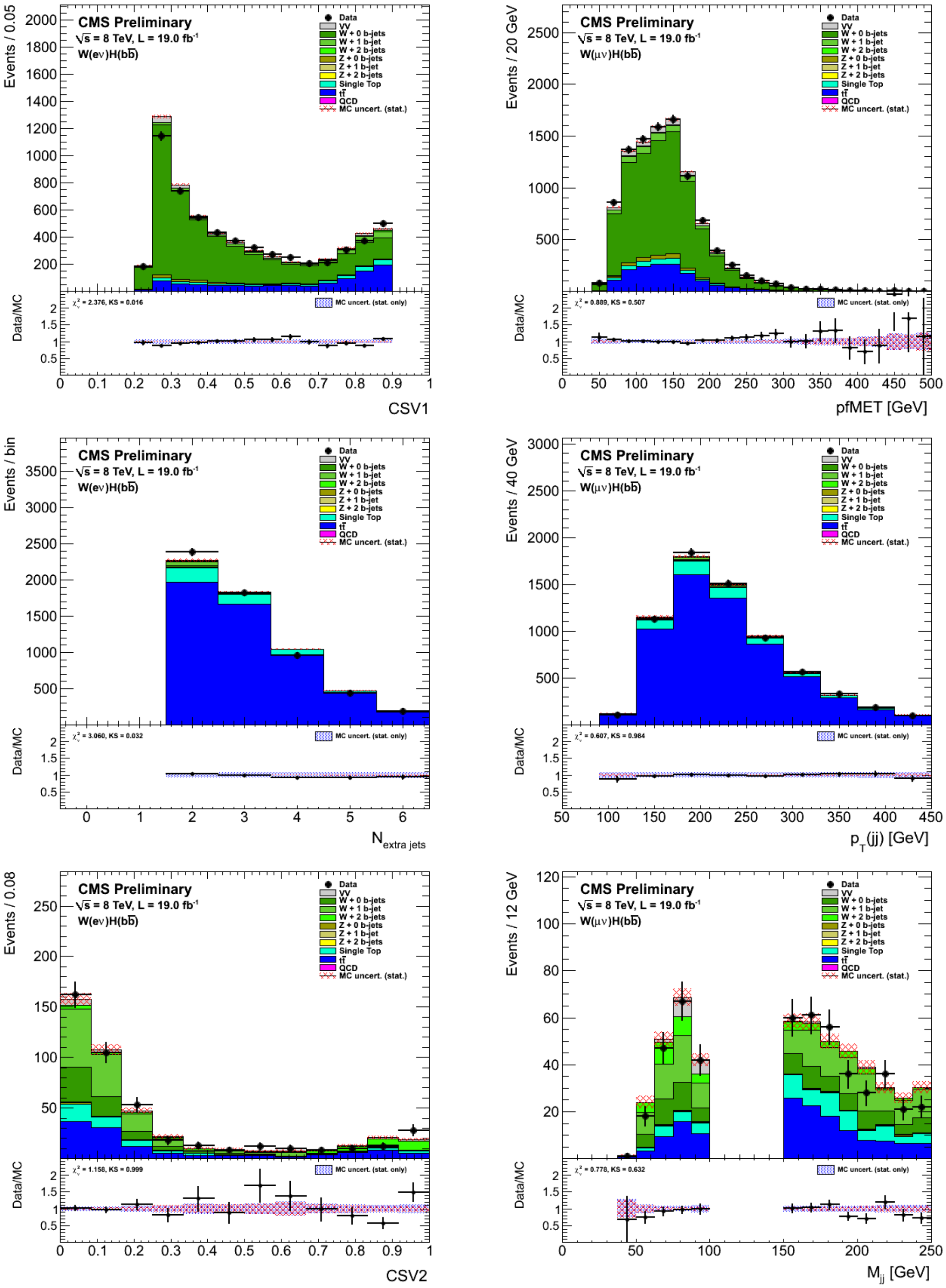

Figure 7.3: Comparison of data and MC simulation in the various high $p_{\mathrm{T}}$(V) control samples from the W(e$\nu$)H and W($\mu\nu$)H analyses, normalizing MC to data. Included distributions: CSV$_{\mathrm{max}}$ and $E_{\mathrm{T}}^{\mathrm{miss}}$ in the W+udscg control sample (top row), $N_{\mathrm{aj}}$ and $p_{\mathrm{T}}$(jj) in the t$\bar{\mathrm{t}}$ control sample (middle row), and CSV$_{\mathrm{min}}$ and $M$(jj) in the W+b$\bar{\mathrm{b}}$ control sample (bottom row). Control samples are shown for both the W(e$\nu$)H (left column) and W($\mu\nu$)H (right column) channels.

Table 7.4: Definition of control samples for the W($\tau\nu$)H channel. Here $p_{\mathrm{T}}$(track) refers to the transverse momentum of the leading tau track. The values listed for kinematical variables are in units of GeV.

| Variable | W+udscg | $\mathrm{t\bar{t}}$ | W+$\mathrm{b\bar{b}}$ |
|---|---|---|---|
| $p_{\mathrm{T}}(\mathrm{j}_1)$ | $> 30$ | $> 30$ | $> 30$ |
| $p_{\mathrm{T}}(\mathrm{j}_2)$ | $> 30$ | $> 30$ | $> 30$ |
| $p_{\mathrm{T}}(\mathrm{jj})$ | $> 120$ | $> 120$ | $> 120$ |
| $p_{\mathrm{T}}(\tau)$ | $> 40$ | $> 40$ | $> 40$ |
| $p_{\mathrm{T}}(\mathrm{W})$ | $> 120$ | $> 120$ | $> 120$ |
| $p_{\mathrm{T}}(\mathrm{track})$ | $> 20$ | $> 20$ | $> 20$ |
| $M(\mathrm{jj})$ | $< 250$ | $< 250$ | $\notin$[90–150] |
| $\mathrm{CSV}_{\mathrm{max}}$ | 0.244–0.898 | $> 0.898$ | $> 0.898$ |
| $N_{\mathrm{aj}}$ | $< 2$ | $> 1$ | $= 0$ |
| $N_{\mathrm{al}}$ | $= 0$ | $= 0$ | $= 0$ |
| $E_{\mathrm{T}}^{\mathrm{miss}}$ | $> 80$ | $> 80$ | $> 80$ |
| $E_{\mathrm{T}}^{\mathrm{miss}}$ significance | $> 2.0$ | – | – |

background. The extrapolated QCD multi-jet yields for all control samples and the signal region are found to be consistent with zero for all $p_{\mathrm{T}}$(V) categories.

## 7.4 W($\tau\nu$)H Control Samples

The control sample definitions for the W($\tau\nu$)H analysis are presented in Table 7.4. The control samples are similar to those for the W($\ell\nu$)H channels, with a few important differences. First of all, requirements on tau-specific variables are included. Also, the event $E_{\mathrm{T}}^{\mathrm{miss}}$ requirement is much higher due to trigger constraints. Finally, there is only one $p_{\mathrm{T}}$(V) category for the W($\tau\nu$)H analysis, so an inclusive selection requirement of $p_{\mathrm{T}}(\mathrm{W}) > 120$ GeV is used. The distributions of several variables important to the W($\tau\nu$)H analysis are illustrated in Figure 7.4, sampling from various control samples. As for all other channels, good agreement is found between data and MC simulation.

An important note for this channel is that the control samples defined in this section are not used for extracting data/MC scale factors for the mode – instead the W($\ell\nu$)H control samples are used. This is described in more detail in Section 7.5.

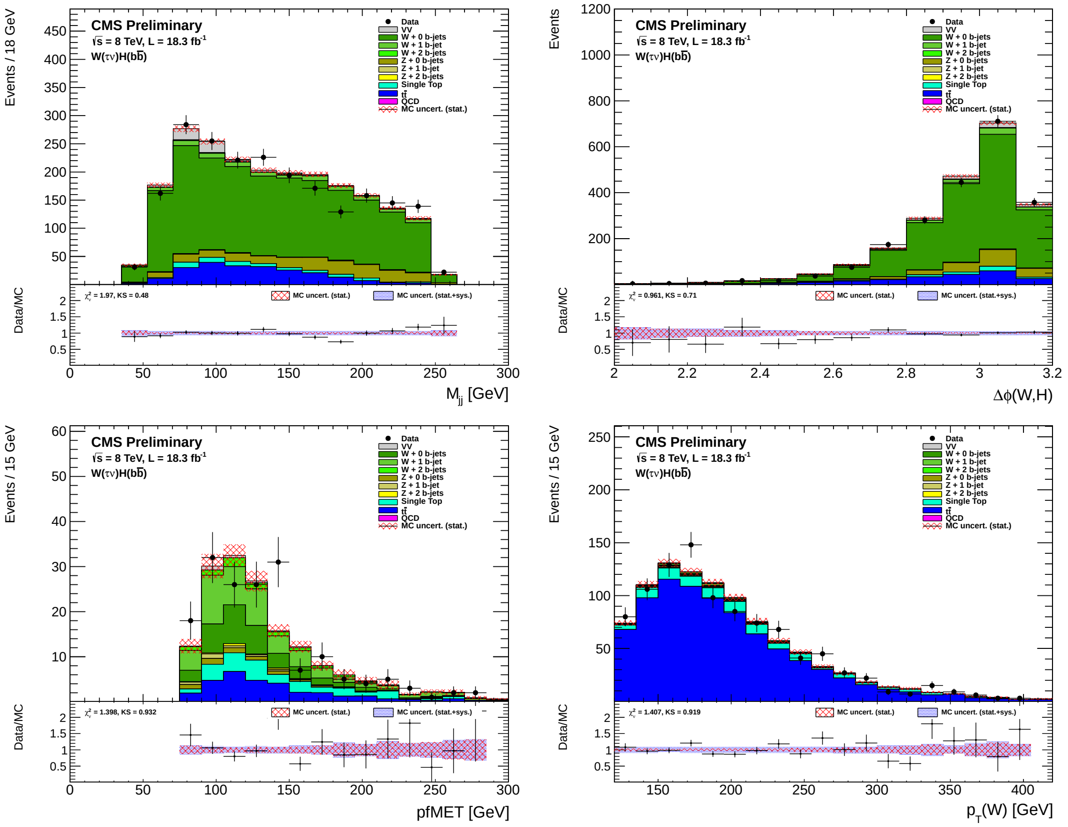

Figure 7.4: Control plots associated with the W($\tau\nu$)H channel control samples, with data/MC scale factors applied to the simulation to facilitate shape comparison. The $M$(jj) and $\Delta\phi(\mathrm{V},\mathrm{H})$ distributions are shown in the W+udscg control sample (top row), the $E_{\mathrm{T}}^{\mathrm{miss}}$ distribution in the W+b$\bar{\mathrm{b}}$ control sample (bottom left), and the $p_{\mathrm{T}}$(W) distribution in the t$\bar{\mathrm{t}}$ control sample (bottom right).

Table 7.5: Definition of the two alternative control samples used in the scale factor fit for the Z($\ell\ell$)H channels. The values listed for kinematic variables are in units of GeV.

| Variable | Z+jets | $\mathrm{t\bar{t}}$ |
|---|---|---|
| $m_{\ell\ell}$ | 75–105 | $\notin$[75–105] |
| $p_{\mathrm{T}}(\mathrm{j}_1)$ | $> 20$ | $> 20$ |
| $p_{\mathrm{T}}(\mathrm{j}_2)$ | $> 20$ | $> 20$ |
| $p_{\mathrm{T}}(\mathrm{V})$ | $> 50$ | 50–100 |
| $M(\mathrm{jj})$ | $< 250$, $\notin$[80–150] | $< 250$, $\notin$[80–150] |
| $\mathrm{CSV}_{\mathrm{max}}$ | $> 0.244$ | $> 0.244$ |
| $\mathrm{CSV}_{\mathrm{min}}$ | $> 0.244$ | $> 0.244$ |

## 7.5 Scale Factors

As discussed earlier in the chapter, the various control samples associated with each channel in the analysis are used to extract data/MC scale factors used to correct the background yield prediction in the signal region. This is done via a binned maximum likelihood fit, using one variable from each control sample (within a particular channel, combining electron and muon modes where applicable). Both CSV output and $M$(jj) variables are used in these fits, depending on the channel in question. As described in Section 7.1, the control samples used for the Z($\ell\ell$)H fit differ from the nominal ones used to compare data and MC simulation in terms of variable shapes. These alternative control samples, defined in Table 7.5, make use of the $m_{\ell\ell}$ and $M$(jj) sidebands to be as close to the signal region as possible.

The data/MC scale factors for all channels in the VH($\mathrm{b\bar{b}}$) analysis are summarized in Table 7.6. Listed are scale factors for each individual $p_{\mathrm{T}}$(V) category – note however that the two Z($\ell\ell$)H $p_{\mathrm{T}}$(V) categories make use of the same set of scale factors. Also, the W($\tau\nu$)H scale factors are obtained from the W($\ell\nu$)H fits to data, instead of making use of the control samples defined for the W($\tau\nu$)H channel. This is due to a limited number of MC simulation events in some of the W($\tau\nu$)H control samples, particularly the W+$\mathrm{b\bar{b}}$ control sample, which would lead to much less precise estimates of the scale factors if used. The uncertainties on the scale factors are used as normalization systematics in the final signal extraction (described in Chapter 8) and are properly correlated with the shape systematics used in obtaining the final results, as discussed in Section 8.4.

The uncertainties (both statistical and systematic) are largest for the W+1b and W+2b processes due to limited MC simulation statistics and a relatively low purity of the backgrounds in the W+b$\bar{\text{b}}$ control samples for the W($\ell\nu$)H and Z($\nu\nu$)H channels. This fact is reflected in the control plots shown earlier in the chapter. The scale factors are found to be close to unity for all processes except for W+1b and Z+1b, for which the scale factors are consistently found to be larger than two. This discrepancy is interpreted as arising mainly from mis-modeling in the generator parton shower of the process of gluon-splitting to b$\bar{\text{b}}$ pairs, as most of the excess occurs in the region of phase space where the secondary vertices associated with the b jets are found relatively close to each other ($\Delta R < 0.5$). This observation is consistent with similar observations in other studies of vector bosons in association with heavy-flavor quarks by the ATLAS and CMS experiments [146–148].

Table 7.6: Data/MC scale factors for the leading background processes in all channels of the VH($b\bar{b}$) analysis, combining electron and muon modes where applicable. Included with the scale factors are both statistical (listed first) and systematic (listed second) uncertainties. The scale factors for the W($\tau\nu$)H channel are obtained from the W($\ell\nu$)H fit, as there are not enough simulated events in the W($\tau\nu$)H control samples to estimate scale factors with comparable precision. Scale factors are listed separately for each $p_{\mathrm{T}}$(V) category. Note that there are only two $p_{\mathrm{T}}$(V) categories for the Z($\ell\ell$)H analysis, which uses the same scale factors for both categories.

| Category/Process | W($\ell\nu$)H/W($\tau\nu$)H | Z($\ell\ell$)H | Z($\nu\nu$)H |
|---|---|---|---|
| Low $p_{\mathrm{T}}$(V) | | | |
| W+0b | $1.03 \pm 0.01 \pm 0.05$ | — | $0.83 \pm 0.02 \pm 0.04$ |
| W+1b | $2.22 \pm 0.25 \pm 0.20$ | — | $2.30 \pm 0.21 \pm 0.11$ |
| W+2b | $1.58 \pm 0.26 \pm 0.24$ | — | $0.85 \pm 0.24 \pm 0.14$ |
| Z+0b | — | $1.11 \pm 0.04 \pm 0.06$ | $1.24 \pm 0.03 \pm 0.09$ |
| Z+1b | — | $1.59 \pm 0.07 \pm 0.08$ | $2.06 \pm 0.06 \pm 0.09$ |
| Z+2b | — | $0.98 \pm 0.10 \pm 0.08$ | $1.25 \pm 0.05 \pm 0.11$ |
| $t\bar{t}$ | $1.03 \pm 0.01 \pm 0.04$ | $1.10 \pm 0.05 \pm 0.06$ | $1.01 \pm 0.02 \pm 0.04$ |
| Intermediate $p_{\mathrm{T}}$(V) | | | |
| W+0b | $1.02 \pm 0.01 \pm 0.07$ | — | $0.93 \pm 0.02 \pm 0.04$ |
| W+1b | $2.90 \pm 0.26 \pm 0.20$ | — | $2.08 \pm 0.20 \pm 0.12$ |
| W+2b | $1.30 \pm 0.23 \pm 0.14$ | — | $0.75 \pm 0.26 \pm 0.11$ |
| Z+0b | — | — | $1.19 \pm 0.03 \pm 0.07$ |
| Z+1b | — | — | $2.30 \pm 0.07 \pm 0.08$ |
| Z+2b | — | — | $1.11 \pm 0.06 \pm 0.12$ |
| $t\bar{t}$ | $1.02 \pm 0.01 \pm 0.15$ | — | $0.99 \pm 0.02 \pm 0.03$ |
| High $p_{\mathrm{T}}$(V) | | | |
| W+0b | $1.04 \pm 0.01 \pm 0.07$ | — | $0.93 \pm 0.02 \pm 0.03$ |
| W+1b | $2.46 \pm 0.33 \pm 0.22$ | — | $2.12 \pm 0.22 \pm 0.10$ |
| W+2b | $0.77 \pm 0.25 \pm 0.08$ | — | $0.71 \pm 0.25 \pm 0.15$ |
| Z+0b | — | $1.11 \pm 0.04 \pm 0.06$ | $1.17 \pm 0.02 \pm 0.08$ |
| Z+1b | — | $1.59 \pm 0.07 \pm 0.08$ | $2.13 \pm 0.05 \pm 0.07$ |
| Z+2b | — | $0.98 \pm 0.10 \pm 0.08$ | $1.12 \pm 0.04 \pm 0.10$ |
| $t\bar{t}$ | $1.00 \pm 0.01 \pm 0.11$ | $1.10 \pm 0.05 \pm 0.06$ | $0.99 \pm 0.02 \pm 0.03$ |

# Chapter 8

# Signal Extraction

This chapter discusses the methodology employed to isolate and potentially extract (if an excess in data is present) a VH($b\bar{b}$) signal. As discussed in previous chapters, the nominal result of the VH($b\bar{b}$) analysis makes use of multi-variate techniques to separate hypothetical signal from background, more specifically a BDT classifier [143] using the TMVA package [144]. This is not to be confused with the BDT regression used to correct the b-jet energy scale, discussed in Chapter 6. Simulated events are used to train (and test) the classifier, yielding weights that are then applied to an orthogonal set of simulated events and the actual data in the signal region. The application of the weights creates an output discriminant for each event, which we refer to as "BDT output" below. This distribution is then fit to extract signal using the Standard Model VH($b\bar{b}$) signal hypothesis – details of this fit are discussed both in this chapter and in Chapter 9.

The BDT classifier separates signal from background in such a way that the background is pushed more toward the left of the distribution, towards values of zero, and the signal is pushed more toward the right of the distribution, toward values of unity. This separation power draws from a variety of different variables used in the BDT training, making use of the full topology of the event. Electron and muon modes are combined in the BDT training, where applicable. A single BDT classifier is used for the Z($\ell\ell$)H analysis, providing an excellent improvement in sensitivity of roughly 35–40% (measured in terms of exclusion limit of SM VH($b\bar{b}$) signal at the 95% confidence level) in comparison to an analysis making use of a fit to the $M$(jj) distribution. In the W($\ell\nu$)H and Z($\nu\nu$)H channels, enough SM signal is expected at the current integrated luminosity such that further gains can be obtained via the use of multiple BDT classifiers, used together in what we refer to as a "multi-BDT" analysis. The additional BDT classifiers used for these channels make use of BDT trainings that include only one background process at a time, and help to sort the events into subsets organized by signal-to-background ratio. The overall gain from using the multi-BDT approach, relative to an optimized analysis employing a single BDT classifier, is roughly 5–10% in sensitivity for the relevant channels. The W($\tau\nu$)H channel does not make use of this technique due to a limited number of training events, and so uses a single BDT training as in the case of the Z($\ell\ell$)H analysis.

We first present the BDT analysis signal extraction in Section 8.1, covering the training and validation of the BDT classifier as well as the multi-BDT technique used for the W($\ell\nu$)H and Z($\nu\nu$)H channels. In Section 8.2, the M(jj) cross-check analysis is discussed. We discuss in Section 8.3 a further cross-check that attempts to measure the cross section of the VZ($\mathrm{b\bar{b}}$) background process in order to validate both the BDT technique and the handling of systematic uncertainties in the analysis. Finally, in Section 8.4, the systematic uncertainties used by the VH($\mathrm{b\bar{b}}$) analysis for the purpose of signal extraction are discussed.

## 8.1 BDT Analysis

The VH($\mathrm{b\bar{b}}$) BDT analysis makes use of variable shape information provided by MC simulation samples for both SM VH($\mathrm{b\bar{b}}$) signal and a variety of different background processes, discussed in Section 4.2. Half of the events in these samples are used in the BDT training, while the other half is used to test for over-training and for the final signal extraction. Both the training and final signal extraction is performed in a signal region defined in Section 5.1.

The BDT training is discussed at length in Section 8.1.1. In Section 8.1.2, the validation of the BDT methodology is presented, making use of the background control samples defined in Chapter 7. Finally, in Section 8.1.3 the multi-BDT technique is

Table 8.1: Variables used in the training of the event BDT classifier discriminant. Note that some variables are used only for a subset of the channels.

| Variable |
|---|
| $p_{\mathrm{T}}(\mathrm{j}_1), p_{\mathrm{T}}(\mathrm{j}_2)$: transverse momentum of each jet associated with the Higgs boson candidate (ordered by $p_{\mathrm{T}}$) |
| $M(\mathrm{jj})$: di-jet invariant mass |
| $p_{\mathrm{T}}(\mathrm{jj})$: di-jet transverse momentum |
| $p_{\mathrm{T}}(\mathrm{V})$: vector boson transverse momentum (or $E_{\mathrm{T}}^{\mathrm{miss}}$) |
| $N_{\mathrm{aj}}$: number of additional jets (beyond those associated with Higgs boson candidate) |
| $\mathrm{CSV}_{\mathrm{max}}$: value of CSV output for the Higgs boson daughter with largest CSV output value |
| $\mathrm{CSV}_{\mathrm{min}}$: value of CSV output for the Higgs boson daughter with second largest CSV output value |
| $\Delta\phi(\mathrm{V},\mathrm{H})$: azimuthal angle between vector boson and di-jet |
| $\Delta\eta(\mathrm{jj})$: difference in $\eta$ between Higgs boson daughters |
| $\Delta R(\mathrm{jj})$: distance in $\eta$–$\phi$ between Higgs boson daughters |
| $\Delta\theta_{\mathrm{pull}}$: color pull angle |
| $\Delta\phi(\vec{E}_{\mathrm{T}}^{\mathrm{miss}}, \mathrm{j}_{\mathrm{nearest}})$: azimuthal angle between $E_{\mathrm{T}}^{\mathrm{miss}}$ and the closest jet (only for $\mathrm{Z}(\nu\nu)\mathrm{H}$) |
| $\mathrm{maxCSV}_{\mathrm{aj}}$: maximum CSV of the additional jets in an event (only for $\mathrm{Z}(\nu\nu)\mathrm{H}$ and $\mathrm{W}(\ell\nu)\mathrm{H}$) |
| $\min\Delta R(\mathrm{H},\mathrm{aj})$: minimum distance between an additional jet and the Higgs boson candidate (only for $\mathrm{Z}(\nu\nu)\mathrm{H}$ and $\mathrm{W}(\ell\nu)\mathrm{H}$) |
| Invariant mass of the VH system (only for $\mathrm{Z}(\ell\ell)\mathrm{H}$) |
| Cosine of the angle between the direction of the vector boson in the rest frame of the VH system and the direction of the VH system in the laboratory frame (only for $\mathrm{Z}(\ell\ell)\mathrm{H}$) |
| Cosine of the angle between the direction of one of the leptons in the rest frame of the Z boson and the direction of the Z boson in the laboratory frame (only for $\mathrm{Z}(\ell\ell)\mathrm{H}$) |
| Cosine of the angle between the direction of one of the jets in the rest frame of the reconstructed Higgs boson and the direction of the reconstructed Higgs boson in the laboratory frame (only for $\mathrm{Z}(\ell\ell)\mathrm{H}$) |

discussed in more detail.

### 8.1.1 Training

The BDT training is carried out separately for each channel and Higgs mass point (every 5 GeV in the range 110-150 GeV, inclusive), combining the electron and muon decay modes in the case of $\mathrm{Z}(\ell\ell)\mathrm{H}$ and $\mathrm{W}(\ell\nu)\mathrm{H}$. All backgrounds are combined together in this training, weighting each background by the expected yield contribution in the signal region. A variety of different variables are used in the trainings to maximize the sensitivity of the analysis in extracting a hypothesized $\mathrm{VH}(\mathrm{b\bar{b}})$ signal. The full list of variables used is presented in Table 8.1, including a summary of the variable definitions. Jets (not associated with the Higgs boson candidate) are counted as additional jets if they satisfy the following: $p_{\mathrm{T}} > 20$ GeV and $|\eta| < 4.5$ for $\mathrm{W}(\ell\nu)\mathrm{H}$, $p_{\mathrm{T}} > 20$ GeV and $|\eta| < 2.5$ for $\mathrm{Z}(\ell\ell)\mathrm{H}$, and $p_{\mathrm{T}} > 25$ GeV and $|\eta| < 4.5$ for $\mathrm{Z}(\nu\nu)\mathrm{H}$. Also used here is the "color pull angle" [149] which discriminates between color-octet background and the color singlet $\mathrm{H} \to \mathrm{b\bar{b}}$ decay.

Before the final training is done, each variable is checked to see if it is well modeled by simulation. This is done chiefly using the control samples discussed in Chapter 7. In addition to checking control plots for each variable in all background control samples, looking for good agreement between data and MC, the modeling of the variable correlations must be checked. Figure 8.1 illustrates a small set of the variable correlations looked at before finalizing the BDT training, shown the $\mathrm{t\bar{t}}$ control sample of the $\mathrm{W(e\nu)H}$ channel as an example. Excellent agreement between data and MC is observed.

The variable list represented in Table 8.1 reflects considerable effort in maximizing the separation power of the BDT classifiers while using as few variables as possible. The latter effort is based on a desire to keep the analysis as simple as possible once maximal separation power has been achieved. The variables that are included provide separation power that is at least somewhat orthogonal to the rest of the variable set, improving the analysis sensitivity by at least 3–5% individually (some variables, such as $M(\mathrm{jj})$, provide much more separation power). The separation between signal and background in the final BDT shape for the $\mathrm{Z(\ell\ell)H}$ channels is presented in Figure 8.2, where the signal and total background levels are both normalized to unity. Note that a re-binning of the BDT output (included in the distributions shown in Figure 8.2) is applied before signal extraction, in order to prevent too few simulated events in any one bin. This is necessary for the statistical treatment of the simulated events in each bin during signal extraction, as discussed further in Section 8.4.

### 8.1.2 Validation

An important effort in any analysis that makes use of multi-variate techniques is the validation of the output discriminant. In the case of the $\mathrm{VH(b\bar{b})}$ analysis at CMS, this is done in three separate ways. The first validation method makes use of cross-check analyses that validate various aspects of the signal extraction, such as the BDT training, the treatment of systematic uncertainties, the fitting machinery employed, etc. The $\mathrm{VH(b\bar{b})}$ analysis makes use of two separate cross-check analyses, a $\mathrm{VH(b\bar{b})}$ M(jj) analysis and a SM di-boson analysis that attempts to measure the cross section of the

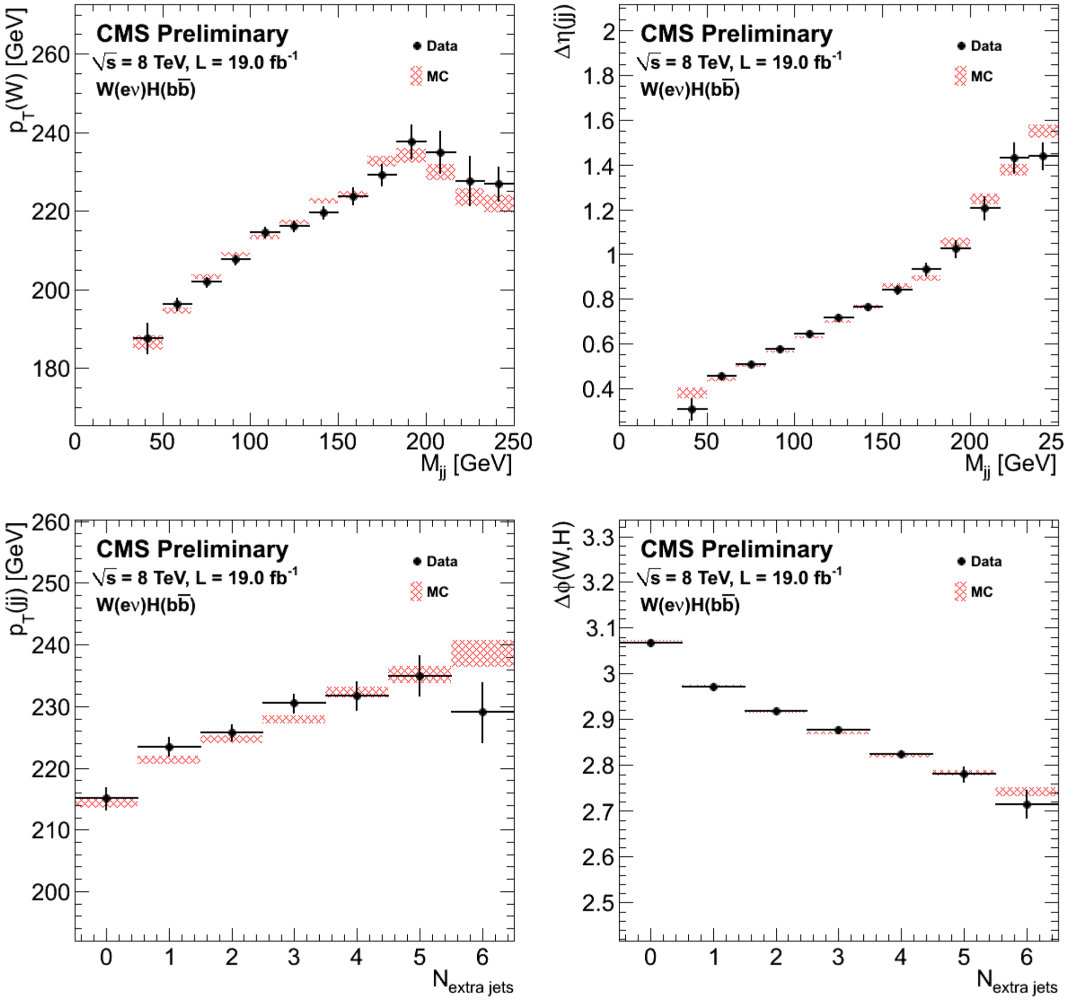

Figure 8.1: Correlations between BDT input variables as evaluated in the W(e$\nu$)H t$\bar{\text{t}}$ control sample with data overlayed on simulated t$\bar{\text{t}}$ events. Included comparisons: $p_{\mathrm{T}}$(W) vs. $M$(jj) (top left), $\Delta\eta$(jj) vs. $M$(jj) (top right), $p_{\mathrm{T}}$(jj) vs. $N_{\mathrm{aj}}$ (bottom left), and $\Delta\phi$(W, H) vs. $N_{\mathrm{aj}}$ (bottom right). Good agreement is found between data and MC simulation in all cases (including the other correlation plots not shown here due to space limitations).

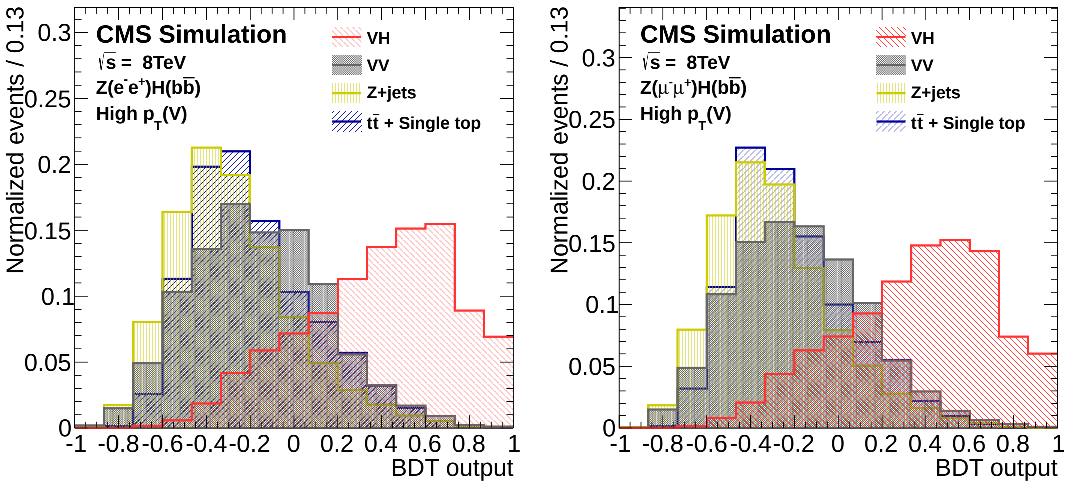

Figure 8.2: BDT output distributions for the Z(ee)H (left) and Z($\mu\mu$)H (right) channels (after re-binning) in the high $p_{\mathrm{T}}$(V) category BDT signal region. The total background and signal prediction (for SM VH(b$\bar{\mathrm{b}}$) signal) are both normalized to unity.

VZ(b$\bar{\mathrm{b}}$) background process. These cross-checks are discussed in Sections 8.2 and 8.3, respectively.

The second validation method makes use of the BDT sideband. The sideband (left-hand side) of the BDT discriminant output is looked at before the signal-rich (right-hand side) part of the distribution is unblinded and signal extraction is completed. Agreement between data and MC is sought before continuing with the signal extraction. Good agreement is found in the sidebands of all BDT distributions, as shown in Chapter 9.

The final validation method made use of in the VH(b$\bar{\mathrm{b}}$) analysis is a visual inspection of the BDT output distribution in the various control samples of the different analysis channels. A sampling of these distributions is illustrated in Figure 8.3. The excellent agreement between data and MC simulation is suggestive that the BDT output distribution is modeled well by simulation, and is a requirement for proceeding with the unblinding of all data in the signal region and the extraction of a hypothesized VH(b$\bar{\mathrm{b}}$) signal.

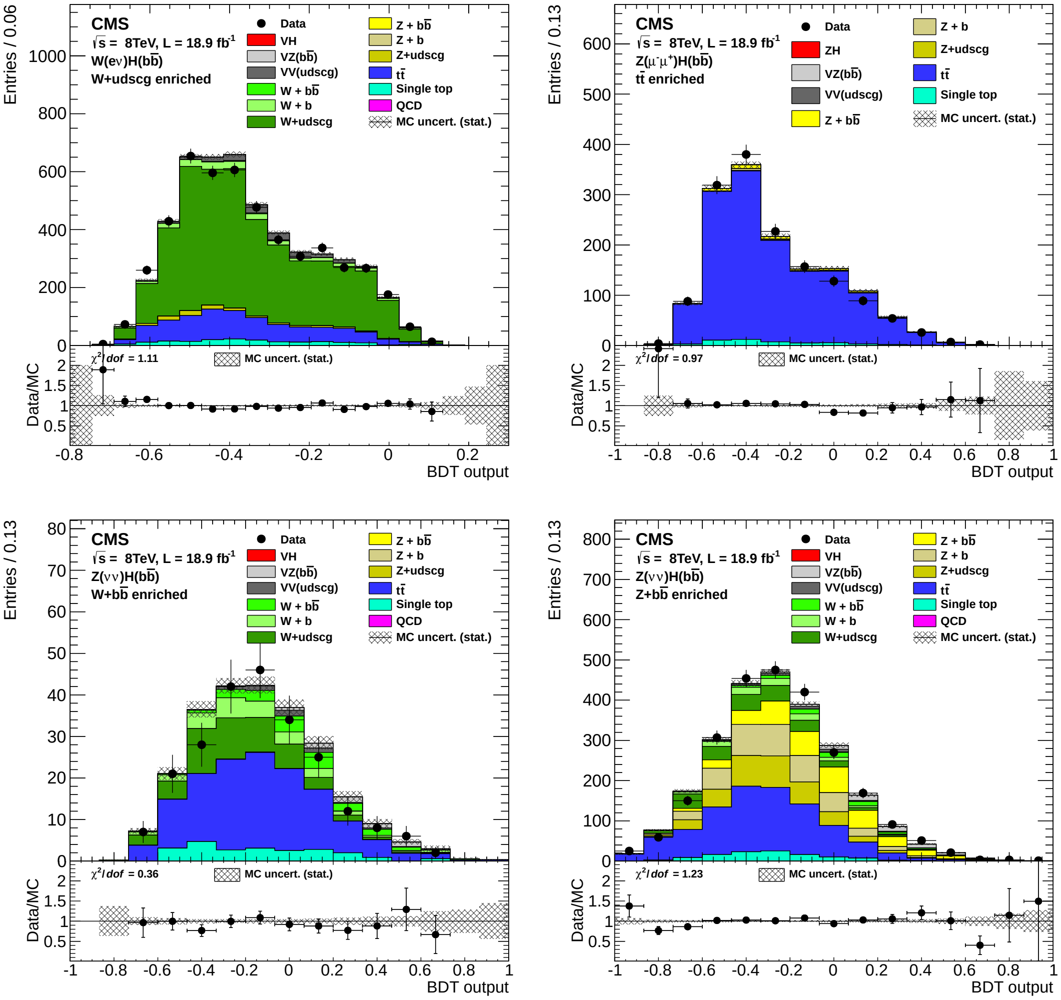

Figure 8.3: BDT output distributions in various control samples for some of the search channels. Shown is the BDT output in the W(e$\nu$)H W+udscg control sample (top left), the Z($\mu\mu$)H t$\bar{\text{t}}$ control sample (top right), the Z($\nu\nu$)H W+b$\bar{\text{b}}$ control sample (bottom left), and the Z($\nu\nu$)H Z+b$\bar{\text{b}}$ control sample (bottom right). Very good agreement between data and MC is observed in all cases.

### 8.1.3 Multi-BDT Technique

While the Z($\ell\ell$)H and W($\tau\nu$)H channels each make use of a single BDT training in order to classify events as either signal-like or background-like, the W($\ell\nu$)H and Z($\nu\nu$)H channels make use of more advanced methods to further enhance the sensitivity of the analysis, as described at the beginning of the chapter. These channels make use of a "multi-BDT" technique similar to the one used by the CDF collaboration [150], which divides the events in the signal region into four distinct subsets that are enriched in $\mathrm{t\bar{t}}$, V+jets, di-boson, and VH signal events (in order of increasing signal-to-background ratio). The technique uses three additional BDT classifiers that are trained on signal and only one background process at a time: $\mathrm{t\bar{t}}$, V+jets, or di-boson events. This results in three additional BDT output values per event, which are cut on to place the events into one of the four subsets described above. The set of variables used in the training of the background-specific BDT's is the same as for the nominal BDT training. As previously mentioned, the Z($\ell\ell$)H channels do not use this technique because of a relatively small final expected signal yield at the present level of integrated luminosity, in addition to less types of background being relevant in the signal region (primarily Z+jets background events). The W($\tau\nu$)H channel instead lacks enough simulated events after the signal region cuts to make use of the multi-BDT technique.

The event classification using the multi-BDT approach works as follows: if an event fails the cut on the $\mathrm{t\bar{t}}$ BDT output, then the event is $\mathrm{t\bar{t}}$-like and is placed in the first subset. Otherwise, the event is checked to see if it passes a cut on the V+jets BDT output. If not, the event is placed in the second subset. Finally, if it passes this cut, the di-boson BDT output is looked at. Failing this cut places the event in the third subset, and passing it places it in the final subset that is most signal-like. The nominal BDT distribution for each subset is separately re-binned, as for the single BDT case, and all four subsets are "pasted" together to form one continuous distribution (in the same order as discussed), mapping the new distribution to values from zero to unity. This distribution (one per $p_{\mathrm{T}}$(V) category) is then fit during signal extraction for the W($\ell\nu$)H and Z($\nu\nu$)H channels. The separation between signal and background in the

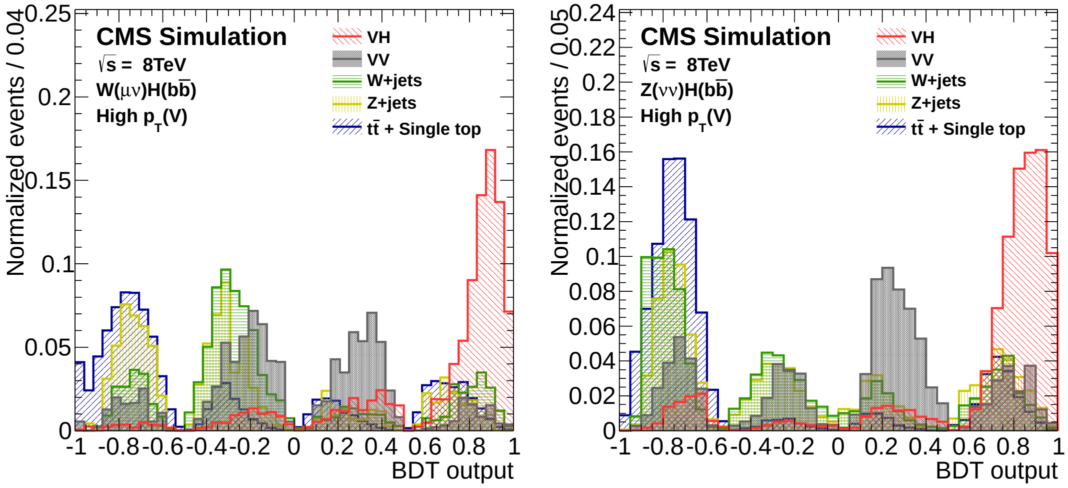

Figure 8.4: Multi-BDT output distributions for the W($\mu\nu$)H (left) and Z($\nu\nu$)H (right) channels in the high $p_{\mathrm{T}}$(V) category BDT signal region. The total background and signal prediction (for SM VH(b$\bar{\mathrm{b}}$) signal) are both normalized to unity.

final multi-BDT shape for the W($\mu\nu$)H and Z($\nu\nu$)H channels is presented in Figure 8.4, where the signal and total background levels are both normalized to unity.

Note that these background-specific BDT's serve as an optimized way to further sort events into categories with differing signal-to-background ratios, and do not add more separation power to the nominal BDT training. This is seen when using only the fourth subset of events in the signal extraction, as it is found to lead to similar sensitivity when compared to a single BDT training without splitting the events into subsets. The added sensitivity comes from separating out very background-like signal events and fitting them in the first, second, and third event subsets with relatively smaller total background levels.

## 8.2 M(jj) Analysis

The M(jj) analysis serves as a cross-check to the nominal VH(b$\bar{\mathrm{b}}$) result obtained with the BDT analysis that is described in Section 8.1. Signal extraction for the M(jj) analysis proceeds in a similar fashion as for the BDT analysis, except that the $M$(jj) distribution is fit instead of the BDT output. An alternative signal region is used as well, presented in Section 5.2, which makes use of cuts on several variables that are nominally used in the BDT training for increased signal/background separation power. If an excess is present in data above the background prediction made using MC simulation, the M(jj) analysis provides an especially clear way to interpret the excess as a resonant signal. The $M$(jj) distribution from the different VH(b$\bar{\text{b}}$) channels can be combined together to present a visually compelling picture in the case of an excess in data arising from a true resonant signal process.

In addition, the M(jj) analysis provides a way to validate the fitting machinery used to extract signal, as well as the background prediction methodology employed in the VH(b$\bar{\text{b}}$) analysis. Aside from the differences listed above, the background prediction and signal extraction proceeds identically for the M(jj) analysis. Results for the M(jj) analysis are presented in Chapter 9 along with the results of the BDT analysis.

## 8.3 Di-boson Cross-check

An additional cross-check to the nominal VH(b$\bar{\text{b}}$) BDT analysis is an attempt to measure the cross section of the resonant di-boson background that normally serves as an irreducible background in extracting a hypothesized VH(b$\bar{\text{b}}$) signal. In particular, the VZ(b$\bar{\text{b}}$) background is targeted, including both WZ and ZZ semi-leptonic processes, due to the great similarity to the VH(b$\bar{\text{b}}$) event topology. The primary differences between the two processes are the spin and mass of the boson decaying to two b jets, as the processes are similar in most other respects. This provides a useful opportunity to validate the BDT technique (and multi-BDT technique, where applicable) that is used in the nominal VH(b$\bar{\text{b}}$) result.

The di-boson cross-check mimics the VH(b$\bar{\text{b}}$) BDT and M(jj) analyses, making use of the same analysis methodology and signal region definitions, except treating the VZ(b$\bar{\text{b}}$) process as a signal process and the VH(b$\bar{\text{b}}$) process as background (using the SM prediction for the expected yield, with large uncertainty). This requires retraining the BDT classifiers used to extract signal and construct the multi-BDT distributions. Also, the W($\tau\nu$)H channel and the low $p_{\mathrm{T}}$(V) categories of the Z($\ell\ell$)H channels are removed as they contribute minimally in this cross-check analysis. The nominal BDT and background-specific BDT trainings for the VZ($\mathrm{b\bar{b}}$) analysis make use of the same set of variables as for the VH($\mathrm{b\bar{b}}$) analysis. The only difference (aside from using VZ($\mathrm{b\bar{b}}$) as signal instead of VH($\mathrm{b\bar{b}}$)) is that there are only three subsets of events in the multi-BDT distribution, which is due to the di-boson processes no longer serving as a background in the cross-check. A separate subset of the multi-BDT distribution is not formed for "VH($\mathrm{b\bar{b}}$)-like background" as the expected yield contribution is relatively small compared to VZ($\mathrm{b\bar{b}}$).

Results of the di-boson cross-check analysis, including VZ($\mathrm{b\bar{b}}$) versions of both BDT and M(jj) analyses, are presented in Chapter 9 alongside the nominal VH($\mathrm{b\bar{b}}$) search results. For a more in-depth discussion of the VZ($\mathrm{b\bar{b}}$) analysis, please refer to [151], which presents a stand-alone VZ($\mathrm{b\bar{b}}$) cross section measurement that includes further improvements beyond those used in this cross-check.

## 8.4 Systematic Uncertainties

In addition to the Poissonian statistical uncertainties on the data points in the ensemble of binned maximum likelihood fits used during signal extraction, a number of systematic effects are accounted for in these fits. The systematic uncertainties are applied to the MC simulation events and come in two types: normalization and shape. The normalization uncertainties are simply uncertainties on the predicted background and signal yields in the signal region, and are flat with respect to variable shape. The shape uncertainties allow for smooth deformations in the variable shape (either $M$(jj) or BDT output, the latter including correlated shape variations across multi-BDT bins) defined by two limits, an "up" shape and a "down" shape that serve as the "one-sigma" bounds on shape variations due to a particular systematic. In both cases, the amount of variation in normalization or shape are controlled by "nuisance parameters" that signify the degree of impact of the systematic on the final signal extraction. The various systematic uncertainties that play a role in the VH($\mathrm{b\bar{b}}$) analysis and their impact on both the total yield uncertainty and final fitted signal strength (for a 125 GeV SM Higgs boson) are presented in Table 8.2 and discussed in detail below. The numbers presented in this table

Table 8.2: Summary of systematic uncertainties in the VH($b\bar{b}$) analysis. Listed information includes whether each systematic affects the shape or normalization of the BDT output, the associated uncertainty on signal or background event yields, and the relative contribution to the expected uncertainty on the extracted signal strength, $\mu$, defined as the ratio of the best-fit value for the production cross section for a 125 GeV Higgs boson, relative to the SM cross section. MC modeling refers to the use of different MC generators in producing certain backgrounds, allowing a variation in BDT output shape. Concerning the uncertainty on $\mu$, the second-to-last column shows the individual contribution of the systematic with no other systematics included in the signal extraction, while the last column shows the impact of removing strictly that systematic with all other systematics included in the signal extraction. These two numbers are different due to correlations between the different systematics. Ranges indicate a variance in impact over different channels or event categories in the analysis.

| | | Event yield uncertainty | Individual contribution | Effect of removal |
|---|---|---|---|---|
| Source | Type | range [%] | to $\mu$ uncertainty [%] | on $\mu$ uncertainty [%] |
| Luminosity | norm. | 3 | $< 2$ | $< 0.1$ |
| Lepton efficiency and trigger (per lepton) | norm. | 3 | $< 2$ | $< 0.1$ |
| Z($\nu\nu$)H triggers | 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 | norm. | 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 ($p_{\mathrm{T}}$ spectrum, EW/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 quark (simulation estimate) | norm. | 15 | 5.0 | 0.5 |
| Di-boson (simulation estimate) | norm. | 15 | 5.0 | 0.5 |
| MC modeling (V+jets and $t\bar{t}$) | shape | 10 | 7.4 | 1.1 |

cover the VH($b\bar{b}$) BDT analysis only, though the systematics have a similar impact on the M(jj) analysis results.

The uncertainty on the CMS luminosity measurement [152, 153], the uncertainties on lepton reconstruction/trigger efficiencies (obtained from studies using leptonically-decaying Z bosons), the uncertainty on the Z($\nu\nu$)H trigger efficiency (obtained from varying the parameters describing the efficiency turn-on curve, yielding a shape systematic), and uncertainties on the event $E_{\mathrm{T}}^{\mathrm{miss}}$ estimate each contribute very minimally to the VH($b\bar{b}$) analysis in terms of yield estimates and sensitivity. The normalization systematics on the VH($b\bar{b}$) signal yield estimate contribute minimally as well in the most recent round of the analysis. The total VH($b\bar{b}$) cross section is calculated to NNLO accuracy [154] with NNLO QCD corrections [155] and NLO electroweak (EW) corrections [156–158] to account for differences in the V/H $p_{\mathrm{T}}$ spectrum between data and MC, and the systematics applied after these corrections are relatively small (5% and 2%, respectively). The impact on the analysis sensitivity due to scale variations and PDF uncertainties [159–163] are similarly small.

Jet-related uncertainties serve as a principal source of systematic uncertainty in the analysis. The jet energy scale is varied within its uncertainty as a function of $p_{\mathrm{T}}$ and $\eta$, while the jet energy resolution is evaluated by smearing the jet energies according to the measured uncertainty. The b-tagging systematics are evaluated by varying the b-tagging data/MC scale factors, measured using $\mathrm{t\bar{t}}$ events and used to re-shape the CSV output distribution as discussed in Section 3.4.4, within the associated uncertainties. These uncertainties are 3% per b-quark tag, 6% per c-quark tag, and 15% per mis-tagged jet (originating from gluons or u/d/s quarks) [12]. All of these jet-related systematics are independently varied up/down within uncertainty as discussed, then propagated through the analysis to obtain altered BDT output shapes (or $M(\mathrm{jj})$ shapes) that serve as up/down shape systematics in the signal extraction.

Several systematics are related to the background yield estimation in the signal region. The data-driven background estimation using the data/MC scale factors measured with the background control samples and discussed in Section 7.5 has associated systematics from the uncertainties on the scale factors. This includes the statistical uncertainty on the scale factors, which is represented in Table 8.2 as the “data estimate” of backgrounds, resulting in one nuisance parameter per background (V+jets and $\mathrm{t\bar{t}}$ processes only) per channel. The systematics on the data/MC scale factors due to jet energy scale, jet energy resolution, and b-tagging variations are folded into the jet-related shape systematics described above. The di-boson and single-top background processes make use of MC simulation for the yield prediction in the signal region, and have associated systematic uncertainties that are derived from CMS cross section measurements [164–166]. For the V+jets and $\mathrm{t\bar{t}}$ backgrounds, additional shape systematics are used that cover differences arising from different MC generators. Finally, for all signal and background processes, we add a bin-by-bin up/down shape systematic that reflects the statistical uncertainty of the background prediction in each bin, treated independently, using simulated events ("Monte Carlo statistics" in Table 8.2).

As is evident from Table 8.2, b-tagging, data-driven background estimation, and MC statistical uncertainties have the greatest impact on the sensitivity of the analysis when it comes to systematic effects. However, the impact of all systematics on the relative uncertainty of the signal strength (at a Higgs mass of 125 GeV) is about 20%, compared to the roughly 50% total relative uncertainty on the signal strength as shown in Chapter 9. Correspondingly, the impact of the systematics is roughly 15% on the expected upper limit on the Higgs boson production cross section and on the expected significance of an observation when the Higgs boson is present in the data at the predicted SM rate. Thus, the VH($b\bar{b}$) analysis is not currently a systematics-limited analysis, as the Poissonian uncertainty on the events in data primarily drives the sensitivity of the analysis.

# Chapter 9

# Results

The results of the VH($b\bar{b}$) search are presented in detail in this chapter, covering results specific to the final analysis of 2012 data and additional results associated with the analysis of the full LHC dataset from both 2011 and 2012 [40, 41]. In addition to the findings of the BDT analysis, which comprise the nominal results for the VH($b\bar{b}$) search, the results of the M(jj) analysis and di-boson cross-checks, discussed in Chapter 8, are presented.

We start with a brief discussion of the methodology employed in calculating exclusion limits and evaluating the significance of excesses in data, which is generic to the Higgs searches at both ATLAS and CMS [167], in Section 9.1, before presenting results for the VH($b\bar{b}$) BDT analysis in Section 9.2. The results of the associated cross-checks, the VH($b\bar{b}$) M(jj) analysis and di-boson cross-check, are presented in Sections 9.3 and 9.4, respectively.

## 9.1 Statistical Treatment

Three principal measurements are carried out for each analysis, making use of the full dataset: the setting of an exclusion limit on the cross section times $H \rightarrow b\bar{b}$ branching fraction of a Standard Model Higgs boson (of different mass hypotheses), the calculation of $p$-values (and associated significance) of an excess in data above predicted background levels, and, if an excess is present, the measurement of the best-fit signal strength ($\mu$, the ratio of the measured cross section times $\mathrm{H} \to \mathrm{b\bar{b}}$ branching fraction to that of the Standard Model prediction) associated with the excess for the applicable mass points. The methodology employed for each of these measurements is identical for both BDT and M(jj) analyses, and in treating either VH($\mathrm{b\bar{b}}$) or VZ($\mathrm{b\bar{b}}$) events as signal. The treatment of the nuisance parameters, $\theta$, arising from systematic effects in the analysis (discussed in Section 8.4) is also the same for the different analyses.

In order to set upper exclusion limits on the cross section of a Standard Model Higgs boson, the modified frequentist method $\mathrm{CL_s}$ is employed [168, 169]. This prescription makes use of a profile likelihood test statistic $\tilde{q}_\mu$:

$$\tilde{q}_\mu = -2\ln\frac{\mathcal{L}(\mathrm{data}|\mu,\hat{\theta}_\mu)}{\mathcal{L}(\mathrm{data}|\hat{\mu},\hat{\theta})}, \quad 0 \leq \hat{\mu} \leq \mu, \tag{9.1}$$

with the likelihood product (over all bins $i$ of all channels) $\mathcal{L}$ given by

$$\mathcal{L}(\mathrm{data}|\mu,\theta) = \prod_i \Big[\mathrm{Poisson}\left(N_i|\, \mu \cdot s_i(\theta) + b_i(\theta)\right) \cdot p(\tilde{\theta}|\theta)\Big]. \tag{9.2}$$

Here, $s(\theta)$ and $b(\theta)$ refer to the expected signal and background yields (as a function of the nuisance parameters), $N$ refers to the observed yield, $\tilde{\theta}$ refers to the best estimate of the nuisance parameters prior to the measurement, $p(\tilde{\theta}|\theta)$ encodes information about the systematic uncertainties, and "data" refers to either the actual experimental *observation* or *pseudo-data* sampled from simulated events in order to construct a probability distribution function $f(\tilde{q}_\mu|\mu,\theta)$ for the test statistic. Furthermore, the quantities $\hat{\mu}$ and $\hat{\theta}$ refer to the maximum likelihood estimates or best-fit values for the signal strength and nuisance parameters, respectively, while $\hat{\theta}_\mu$ denotes the conditional maximum likelihood estimate of the nuisance parameters with $\mu$ fixed.

With the construction of $f(\tilde{q}_\mu|\mu,\theta)$ using pseudo-data sampled from MC simulation, one can proceed with calculating the $\mathrm{CL_s}$ value from a ratio of two probabilities:

$$\mathrm{CL_s}(\mu) = \frac{P(\tilde{q}_\mu \geq \tilde{q}_\mu^{\mathrm{obs}}|\mu,\hat{\theta}_\mu^{\mathrm{obs}})}{P(\tilde{q}_\mu \geq \tilde{q}_\mu^{\mathrm{obs}}|0,\hat{\theta}_0^{\mathrm{obs}})}. \tag{9.3}$$

Here, the quantities labeled with "obs" refer to the variables defined above evaluated using the experimentally observed data. The probabilities in Equation 9.3 are calculated as follows:

$$P(\tilde{q}_\mu \geq \tilde{q}_\mu^{\rm obs}|\mu, \theta) = \int_{\tilde{q}_\mu^{\rm obs}}^{\infty} f(\tilde{q}_\mu|\mu, \theta)\, d\tilde{q}_\mu. \tag{9.4}$$

In the modified frequentist approach, one says the signal is excluded at the 95% confidence level (the standard for most analyses in particles physics), or C.L., if $\mathrm{CL_s} = 0.05$. The $\mathrm{CL_s}$ prescription provides one-sided exclusion limits and prevents downward fluctuations of background from giving senseless results, as can happen for classical frequentist methods.

The above discussion concerns the computation of *observed* limits, which make use of the experimental observation (the data collected by CMS) and result in one measurement per Higgs mass point. For *expected* limits, which signify the expected range of values that the observed limit should fall within given the background-only hypothesis, the procedure is slightly modified. Instead, one generates pseudo-data using strictly background MC and runs the limit calculation discussed above for each of a large ensemble of generated pseudo-datasets, treating each one as if it were the observation. This creates a distribution of 95% confidence limits per Higgs mass point. One then takes the median of this distribution as the expected limit, the crossings of the 16% and 84% quantiles as the $\pm\, 1\sigma$ band, and the crossings of the 2.5% and 97.5% quantiles as the $\pm\, 2\sigma$ band. The bands are colored in green and yellow, respectively, within the plots in the later sections of this chapter.

In order to quantify an excess of events, one must first calculate the associated $p$-values that characterize the significance of the excess (as a function of Higgs mass) beyond the predicted background levels. Each $p$-value is the probability that an excess of magnitude equal to or greater than the observed excess, at a particular Higgs mass point, is a result of an upward fluctuation in background with no actual signal present. A particular result for a $p$-value corresponds to a unique level of significance, quantified in terms of "sigmas" or number of standard deviations corresponding to a Gaussian distribution. The methodology used to measure the $p$-values in the VH($\mathrm{b\bar{b}}$) analysis is slightly simpler than that used for the setting of 95% upper exclusion limits on the cross section times $\mathrm{H \to b\bar{b}}$ branching fraction of the SM Higgs boson. The test statistic from Equation 9.1 is used, except now $\mu = 0$ is enforced. The $p$-value is equivalent to $P(\tilde{q}_0 \geq \tilde{q}_0^{\mathrm{obs}} | 0, \hat{\theta}_0^{\mathrm{obs}})$. This is translated into an observed significance via

$$p = \int_Z^\infty \frac{1}{\sqrt{2\pi}} \exp\left(-x^2/2\right) dx, \tag{9.5}$$

where $Z$ is the significance of the observed excess in number of standard deviations.

The last measurement made in the VH($\mathrm{b\bar{b}}$) analysis (and associated cross-checks) is the computation of the best-fit signal strength $\mu$ at a given Higgs mass point, assuming that the SM Higgs signal hypothesis is the source of an excess in observed data, if any. This measurement is simpler than the others, and is done via a binned maximum likelihood fit to either the BDT or $M(\mathrm{jj})$ distribution (combining all channels in the fit), allowing $\mu$ to vary freely and the nuisance parameters $\theta$ within their uncertainties. We focus on the 125 GeV Higgs mass point for this measurement, as we observe an excess in data around this mass and because of the previous observation at ATLAS and CMS of a SM Higgs boson near this mass point.

## 9.2 BDT Analysis Results

The measurements discussed in technical detail within Section 9.1 are carried out for the VH($\mathrm{b\bar{b}}$) BDT analysis, standing as the main results for the search as discussed in Section 8.1. For the 8 TeV results, a total of 14 separate BDT (and multi-BDT) distributions are included in the fits, sharing several common nuisance parameters where applicable. A sampling of these BDT/multi-BDT distributions, after the binned maximum likelihood fit is performed to extract signal strength, is shown in Figure 9.1 for the 125 GeV Higgs mass point. Included here is the high $p_{\mathrm{T}}(\mathrm{V})$ category only. More plots of the BDT and multi-BDT distributions after the fit are illustrated in Appendix B.

The expected and observed exclusion limits at the 95% confidence level on the cross section times $\mathrm{H} \rightarrow \mathrm{b\bar{b}}$ branching fraction of a Standard Model Higgs boson (for the VH($\mathrm{b\bar{b}}$) BDT analysis) are presented in Figures 9.2 and 9.3. The limits are presented in two separate plots, one for the entire LHC dataset (2011 and 2012 data included) in Figure 9.2, and one for only the 2012 data in Figure 9.3, for which a larger search

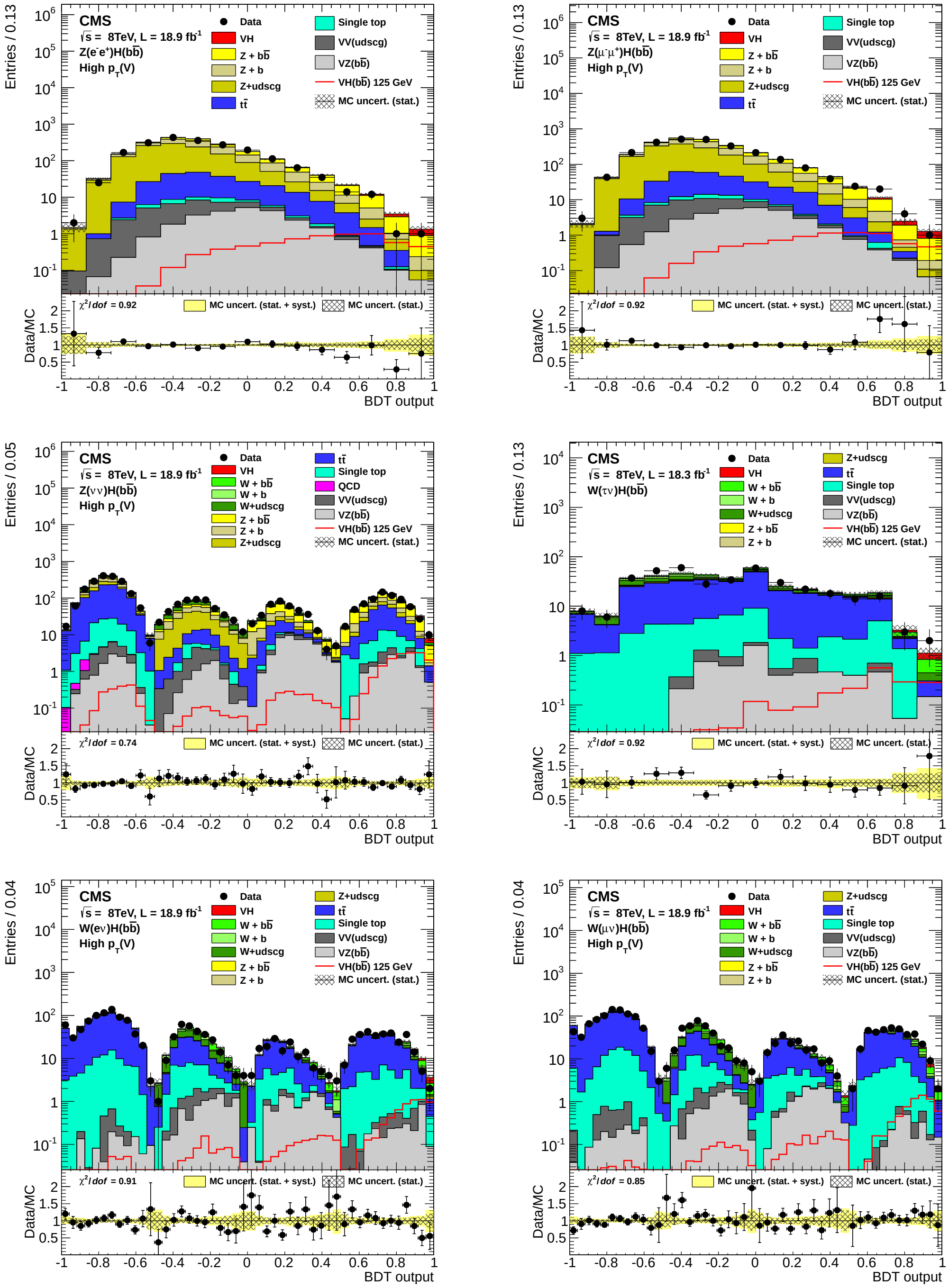

Figure 9.1: BDT and multi-BDT output distributions, after the binned maximum likelihood fit performed to extract signal strength (at a Higgs mass of 125 GeV), for the high $p_{\mathrm{T}}$(V) category of the Z(ee)H channel (top left), the Z($\mu\mu$)H channel (top right), the Z($\nu\nu$)H channel (middle left), the W($\tau\nu$)H channel (middle right), the W(e$\nu$)H channel (bottom left), and the W($\mu\nu$)H channel (bottom right). Also presented in these plots is the predicted signal yield for a 125 GeV SM Higgs boson.

range is used in terms of mass of the Higgs boson (up to 150 GeV). For a Higgs boson mass of 125 GeV the expected limit is 0.95 and the observed limit is 1.89, using the full LHC dataset – as the observed limit is above one, we cannot exclude a SM Higgs boson decaying to bottom quarks at the 95% confidence level for this mass point. In fact, the observed limit distribution (as a function of Higgs mass) is quite consistent with a SM Higgs boson at that mass, as is seen in a comparison to the expected limits obtained when 125 GeV VH($b\bar{b}$) signal (at the SM rate) is added to the background prediction.

The calculation of $p$-values (and correspondingly, significance) is also done and presented as a function of Higgs mass in Figure 9.2 for the analysis of the entire LHC dataset and in Figure 9.3 for the analysis of 2012 data only, alongside the limits discussed above. There is a clear excess of observed data throughout the search range – an excess above background levels with a significance of 2.1 standard deviations is observed for the 125 GeV mass point, compared to an expected value of 2.1 standard deviations for the presence of a SM Higgs boson at that mass point. The exclusion limits and significance of the excess at the 125 GeV Higgs mass point are summarized in Table 9.1 for the VH($b\bar{b}$) BDT analysis, separating the results by topologically distinct groups of channels. Note that all significances presented here are "local" in that they do not take into account the look-elsewhere effect [170], which impacts the significances only very minimally due to the resolution of the search with respect to di-jet mass being of similar size to the relevant Higgs mass search range in the analysis. We do not need the correction at any rate since we are now testing the specific hypothesis that the Higgs boson recently found at a mass of 125 GeV decays (or doesn't decay) to $b\bar{b}$ pairs.

Assuming that the excess quantified above can be explained by a SM Higgs boson (or a similar boson of different cross section times branching fraction of decays to $b\bar{b}$), the signal strength $\mu$ is extracted as a function of signal mass, which is presented in Figure 9.4 for the combined LHC dataset. Also presented in Figure 9.4 is the breakdown of the observed signal strength by channel for a 125 GeV Higgs boson. The compatibility between the different channels is consistent with a SM Higgs boson at 125 GeV, with a total signal strength of 1.0 ± 0.5 at this mass point. While consistent with a SM Higgs

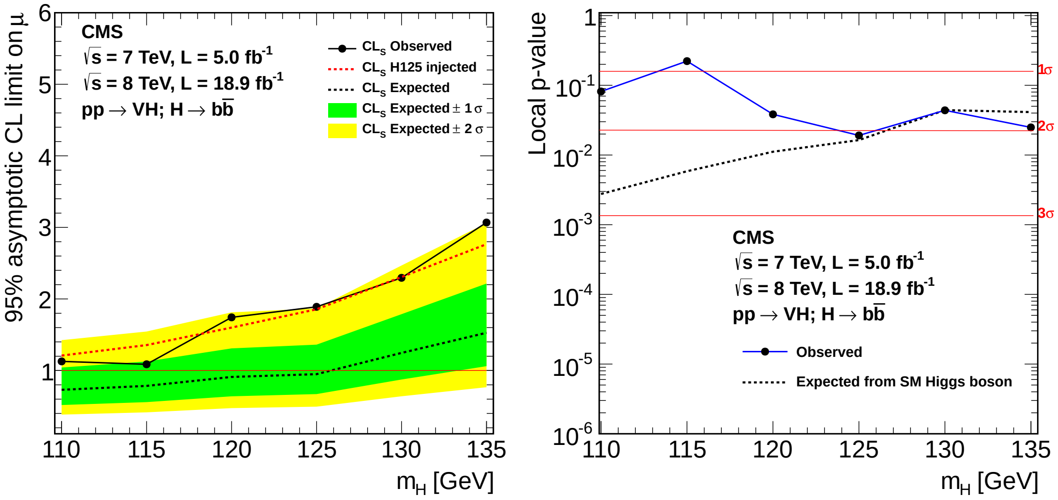

Figure 9.2: Observed/expected exclusion limits on the product of the cross section and the H → b$\bar{b}$ branching fraction for a SM Higgs boson at the 95% confidence level (left) and *p*-values/significances of an excess in observed data (right) for the VH(b$\bar{b}$) BDT analysis, making use of the full LHC dataset (2011 and 2012 datasets).

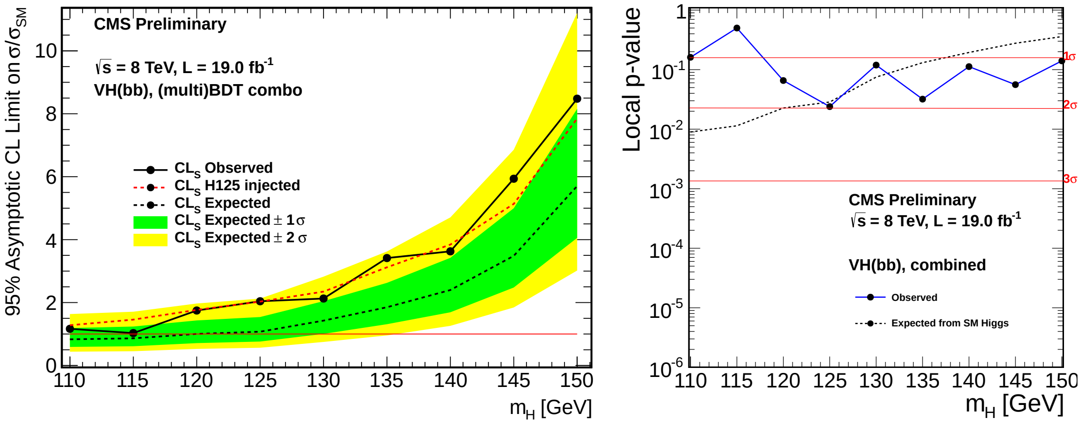

Figure 9.3: Observed/expected exclusion limits on the product of the cross section and the H → b$\bar{b}$ branching fraction for a SM Higgs boson at the 95% confidence level (left) and *p*-values/significances of an excess in observed data (right) for the VH(b$\bar{b}$) BDT analysis, from an analysis of 2012 data only.

Table 9.1: The expected and observed 95% C.L. upper limits on the product of the cross section and H → b$\bar{\text{b}}$ branching fraction, with respect to the expectations for a 125 GeV SM Higgs boson, for partial combinations of channels and for all channels combined within the VH(b$\bar{\text{b}}$) BDT analysis. Also shown are the expected and observed significances.

| **($\mathbf{m_H}$ = 125 GeV)** | $\sigma/\sigma_{\text{SM}}$ (95% C.L.) | $\sigma/\sigma_{\text{SM}}$ (95% C.L.) | Signif. | Signif. |
|---|---|---|---|---|
| | median expected | observed | expected | observed |
| W($\ell\nu, \tau\nu$)H | 1.6 | 2.3 | 1.3 | 1.4 |
| Z($\ell\ell$)H | 1.9 | 2.8 | 1.1 | 0.8 |
| Z($\nu\nu$)H | 1.6 | 2.6 | 1.3 | 1.3 |
| All channels | 0.95 | 1.89 | 2.1 | 2.1 |

boson at 125 GeV (as seen in other Higgs searches at ATLAS and CMS for different decays of the Higgs boson), this result cannot be definitively labeled as a discovery or even evidence for a Higgs boson decaying to bottom quarks – a much larger dataset is needed to make claims of this nature.

One very important feature desired in a physics analysis is the ability to visually observe an excess of events, consistent with a signal hypothesis, in the final fitted distributions. Looking at the distributions in Figure 9.1, there are slight hints of an excess on the right-hand sides of the BDT and multi-BDT distributions, where signal would be expected with the existence of a SM Higgs boson decaying to bottom quarks, but because many different channels are combined in extracting the significance of the excess, this is far from obvious. As an effort to further visualize the excess that is present in the VH(b$\bar{\text{b}}$) BDT analysis, we combine the events of all channels into one distribution, illustrated in Figure 9.5, for the 125 GeV Higgs mass point only. In this plot we combine the BDT outputs of all channels where the events are gathered into bins of similar expected signal-to-background ratio, as given by the value of the output of their corresponding BDT discriminant. The excess is clearer in this distribution, and is compatible with a SM Higgs boson at 125 GeV above expected background levels over a wide range of values of the signal-to-background ratio.

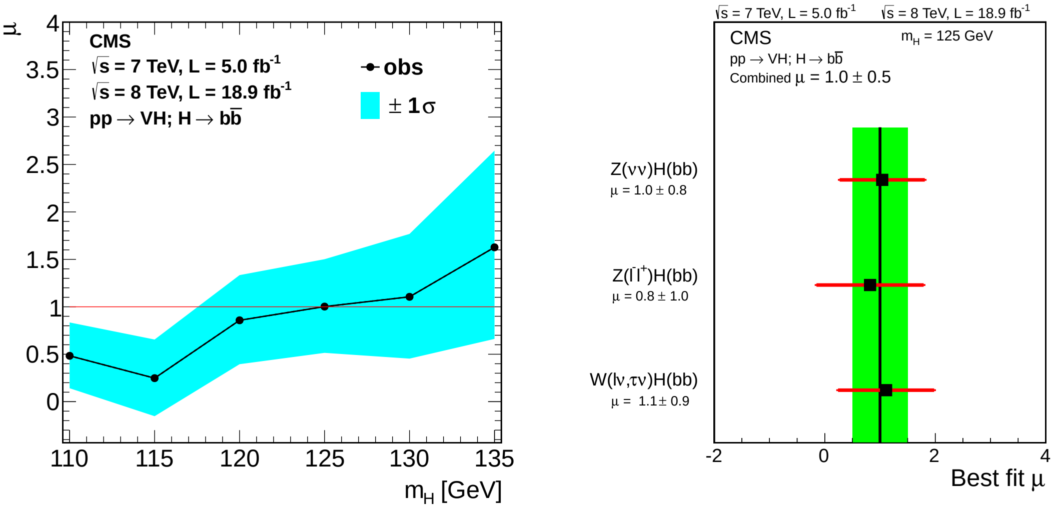

Figure 9.4: Extracted signal strength $\mu$ as a function of Higgs mass point (left) and the extracted signal strength for a 125 GeV Higgs boson for partial combinations of channels and for all channels combined (right), using the full LHC dataset. Results are compatible with the presence of a 125 GeV SM Higgs boson, while the background-only hypothesis is somewhat disfavored.

## 9.3 M(jj) Analysis Results

For the VH(b$\bar{\text{b}}$) M(jj) analysis cross-check (described in Section 8.2) we perform the same measurements as for the VH(b$\bar{\text{b}}$) BDT analysis, presenting results here for the combined analysis of the entire LHC dataset (data collected in both 2011 and 2012) strictly. In Figure 9.6, the combined $M$(jj) distribution is presented, summing over all channels and $p_{\mathrm{T}}$(V) categories with events weighted by the ratio of the expected number of signal events (from a SM Higgs boson with a mass of 125 GeV) to the sum of expected signal and background events in a window of $M$(jj) values between 105 GeV and 150 GeV. Also present in Figure 9.6 is the weighted and summed $M$(jj) distribution after all backgrounds except for di-boson processes are subtracted. It is clear from this background-subtracted distribution that there is a mild excess around 125 GeV, slightly more consistent with a SM Higgs boson than the background-only hypothesis.

At the 125 GeV Higgs boson mass point, a signal strength of 0.8 ± 0.7 is found in the VH(b$\bar{\text{b}}$) M(jj) analysis. An observed significance of 1.1 standard deviations associated with the excess in observed data is found. A comparison of the expected and observed

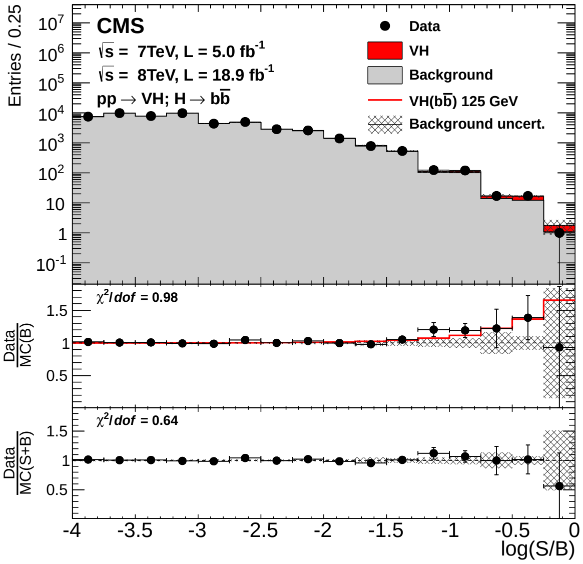

Figure 9.5: Combination of all channels/categories in the VH(b$\bar{b}$) BDT analysis into one distribution, allowing for visualization of the excess of events above background levels observed in many different channels of the analysis. Events are sorted in bins of similar expected signal-to-background 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 two bottom insets show the ratio of the data to the background-only prediction (above) and to the predicted sum of background and SM Higgs boson signal with a mass of 125 GeV (below).

exclusion limits on the cross section times H → b$\bar{b}$ branching fraction of a SM Higgs boson, at the 95% confidence level, is made between the BDT analysis and M(jj) analysis in Table 9.2. Here we can see the substantial improvement in analysis sensitivity with the use of the BDT method across all mass points, with a similar trend between expected and observed limit for both the BDT and M(jj) analyses.

## 9.4 Di-boson Cross-check Results

We also present results for the di-boson cross-check to the VH(b$\bar{b}$) analysis, described in Section 8.3, providing cross-checks for both the BDT and M(jj) analyses. Specifically, we attempt to extract the VZ(b$\bar{b}$) process as signal, making use of the methodology of both the VH(b$\bar{b}$) BDT analysis and the VH(b$\bar{b}$) M(jj) analysis (a cross-check in itself).

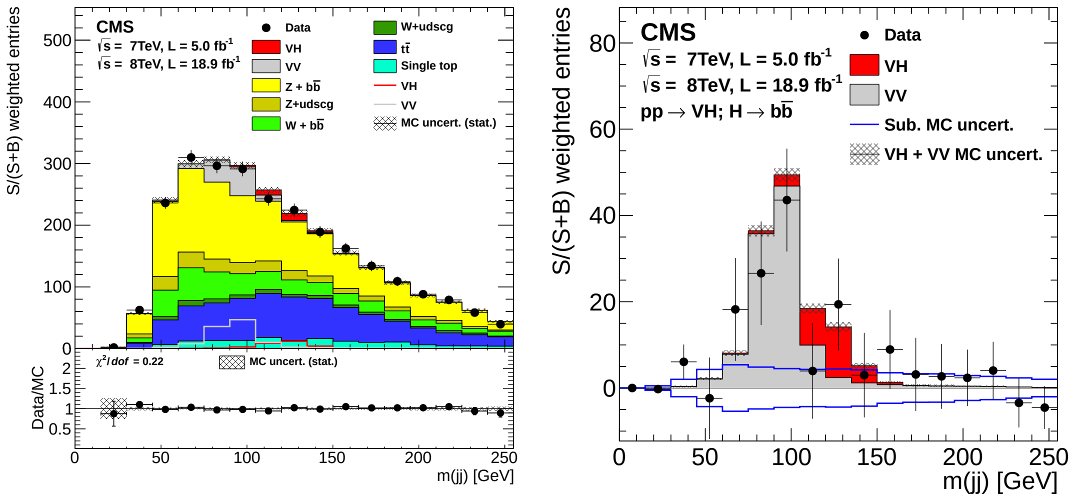

Figure 9.6: Di-jet mass distribution in the signal region of the M(jj) analysis, summing over all channels and $p_{\mathrm{T}}$(V) categories with events weighted by the ratio of the expected number of signal events from a 125 GeV SM Higgs boson to the sum of expected signal and background events in a window of $M$(jj) values between 105 GeV and 150 GeV, before (left) and after (right) subtraction of all backgrounds except for the di-boson processes. The weight for the high $p_{\mathrm{T}}$(V) category is set to 1.0 and all other weights are adjusted proportionally. The background-subtracted plot demonstrates our ability to reconstruct the VZ(b$\bar{\text{b}}$) peak, and is also suggestive of the presence of another resonance around 125 GeV.

In Figure 9.6, after the background-subtraction is performed, one can see a well-defined VZ(b$\bar{\text{b}}$) peak in the $M$(jj) spectrum that is well reconstructed with the observed data. This provides good validation of our ability to reconstruct b jets with the correct energy, without bias with respect to MC simulation.

In Table 9.3 we show the results of the measurements discussed in Section 9.1 completed for the di-boson cross-check. Shown are results for the M(jj) and BDT VZ(b$\bar{\text{b}}$) analyses, this time explicitly separating out the impact of using the multi-BDT methodology (which is implicitly included in the presentation of the VH(b$\bar{\text{b}}$) BDT analysis results in Section 9.2). We measure the signal strength of the VZ(b$\bar{\text{b}}$) process to be 1.19 ± 0.25, with the full multi-BDT methodology employed as it is for the VH(b$\bar{\text{b}}$) BDT analysis. The fact that this measurement is in agreement with the SM VZ(b$\bar{\text{b}}$) cross section prediction ($\mu = 1$) is excellent validation of the BDT methods that we make use of in the VH(b$\bar{\text{b}}$) analysis.

Table 9.2: Expected and observed 95% confidence level upper limits on cross section times H → b$\bar{\text{b}}$ branching fraction, with respect to the expectations for a Standard Model Higgs boson, in the M(jj) and BDT analyses using the full LHC dataset. Results are presented for the relevant Higgs mass range.

| $m_{\text{H}}$ [GeV ] | 110 | 115 | 120 | 125 | 130 | 135 |
|---|---|---|---|---|---|---|
| M(jj) Exp. limit | 1.00 | 1.07 | 1.21 | 1.36 | 1.77 | 2.26 |
| M(jj) Obs. limit | 1.09 | 1.23 | 1.56 | 2.00 | 2.54 | 3.21 |
| BDT Exp. limit | 0.73 | 0.79 | 0.91 | 0.95 | 1.25 | 1.53 |
| BDT Obs. limit | 1.13 | 1.09 | 1.74 | 1.89 | 2.30 | 3.07 |

Table 9.3: Expected/observed significances of the excess above predicted non-resonant background due to the VZ(b$\bar{\text{b}}$) process along with the extracted VZ(b$\bar{\text{b}}$) signal strength for the various analyses covered in the di-boson cross-check. Here the BDT results are separated into two further categories, one with and one without use of the multi-BDT technique.

| Analysis | M(jj) | BDT | multi-BDT |
|---|---|---|---|
| Expected significance | 4.3 | 5.6 | 6.3 |
| Observed significance | 3.7 | 6.3 | 7.5 |
| Signal strength $\mu$ | $0.78 \pm 0.27$ | $1.10 \pm 0.28$ | $1.19 \pm 0.25$ |

# Chapter 10

# Conclusions

This work presents a search for the Standard Model Higgs boson when produced in association with a vector boson (W or Z) and decaying to $\mathrm{b\bar{b}}$, making use of six distinct final-states: Z(ee)H, Z($\mu\mu$)H, Z($\nu\nu$)H, W(e$\nu$)H, W($\mu\nu$)H, and W($\tau\nu$)H. While the analysis methodology discussed here is most relevant for the final analysis of the 2012 dataset, the results presented in this work cover the analysis of the full LHC dataset, including 5.0 $\mathrm{fb}^{-1}$ of data collected at $\sqrt{s} = 7$ TeV in 2011 and 18.9 $\mathrm{fb}^{-1}$ of data collected at $\sqrt{s} = 8$ TeV in 2012. The analysis utilizes a wide array of tools and validation techniques, using background control samples to study the backgrounds that contribute to the signal region, a b-jet energy regression to improve the di-jet mass resolution of signal events, BDT classifiers to better separate signal and background shapes in the signal region, and cross-check analyses to help validate the main results of the VH($\mathrm{b\bar{b}}$) analysis.

We find expected and observed exclusion limits, at the 95% confidence level, on the cross section times $\mathrm{H \rightarrow b\bar{b}}$ branching fraction of the SM Higgs boson over a Higgs mass search range of 110-150 GeV (110-135 GeV for the full LHC dataset). At 125 GeV, roughly the mass of the SM Higgs boson discovery made at ATLAS and CMS using other Higgs decay channels, we find an observed exclusion limit of 1.89 compared to an expected limit of 0.95, signifying the presence of an excess in data above predicted background levels. This excess of observed events corresponds to a significance of 2.1 standard deviations, compared to an expected significance of 2.1 standard deviations from the presence of a SM Higgs boson decaying to bottom quarks with a mass of 125 GeV. The best-fit signal strength associated with this excess, under the assumption that the excess is explained by SM VH($b\bar{b}$) decays, is 1.0 ± 0.5, suggesting that the excess is somewhat more in line with the existence of a $b\bar{b}$ resonance at 125 GeV than the background-only hypothesis. In order to make a more definitive statement about whether or not the Higgs boson discovered recently at 125 GeV couples to the bottom sector, or any flavor of quark, a much larger dataset must be analyzed. However, the sensitivity of the CMS VH($b\bar{b}$) search presented in this work, as represented by the expected significance for a 125 GeV Higgs boson, is the highest for a single experiment thus far and also higher in comparison to the combined results of the CDF and D0 experiments. Given the decades-long effort of hunting for $H \rightarrow b\bar{b}$ decays, this is a considerable feat – not only for the analysts, but also for the designers and builders of the LHC and CMS detector.

As the center-of-mass energy of the LHC is raised to 13–14 TeV and the instantaneous luminosity of proton-proton collisions is increased, a number of possible issues with the VH($b\bar{b}$) analysis must be tackled in order to push the expected sensitivity towards evidence and eventual discovery of the coupling of the SM Higgs boson to bottom quarks. The decreasing separation of the Higgs jets in the $\eta - \phi$ plane on average will result in lower signal efficiency due to resulting problems in jet reconstruction, i.e. jet merging. One possible way to tackle this problem is the use of jet substructure methods [82, 129–131], which aim to improve jet reconstruction for boosted di-jet and tri-jet systems. Additionally, the increased instantaneous luminosity may lead to a degradation in the di-jet mass resolution due to effects related to the increased pile-up in events. It may be necessary to do significant “jet cleaning” [171] in order to enjoy the same di-jet resolution as seen in the present round of the VH($b\bar{b}$) analysis. There are considerable efforts at CMS in addressing these problems, and it is expected that they will be solved, as countless other problems at the energy frontier have been when faced with the prospect of failure. As they say, necessity is the mother of creativity.

# Appendix A

# Additional Hadron-to-Muon Mis-identification Plots

Presented in Figures A.1–A.4 are the fits to the inclusive V0 invariant mass distributions used in computing the hadron-to-muon mis-identification rates that are discussed in Section 3.7.1. Also included in Figure A.5 are differential hadron-to-muon mis-identification rates for the TightMVA muon selection.

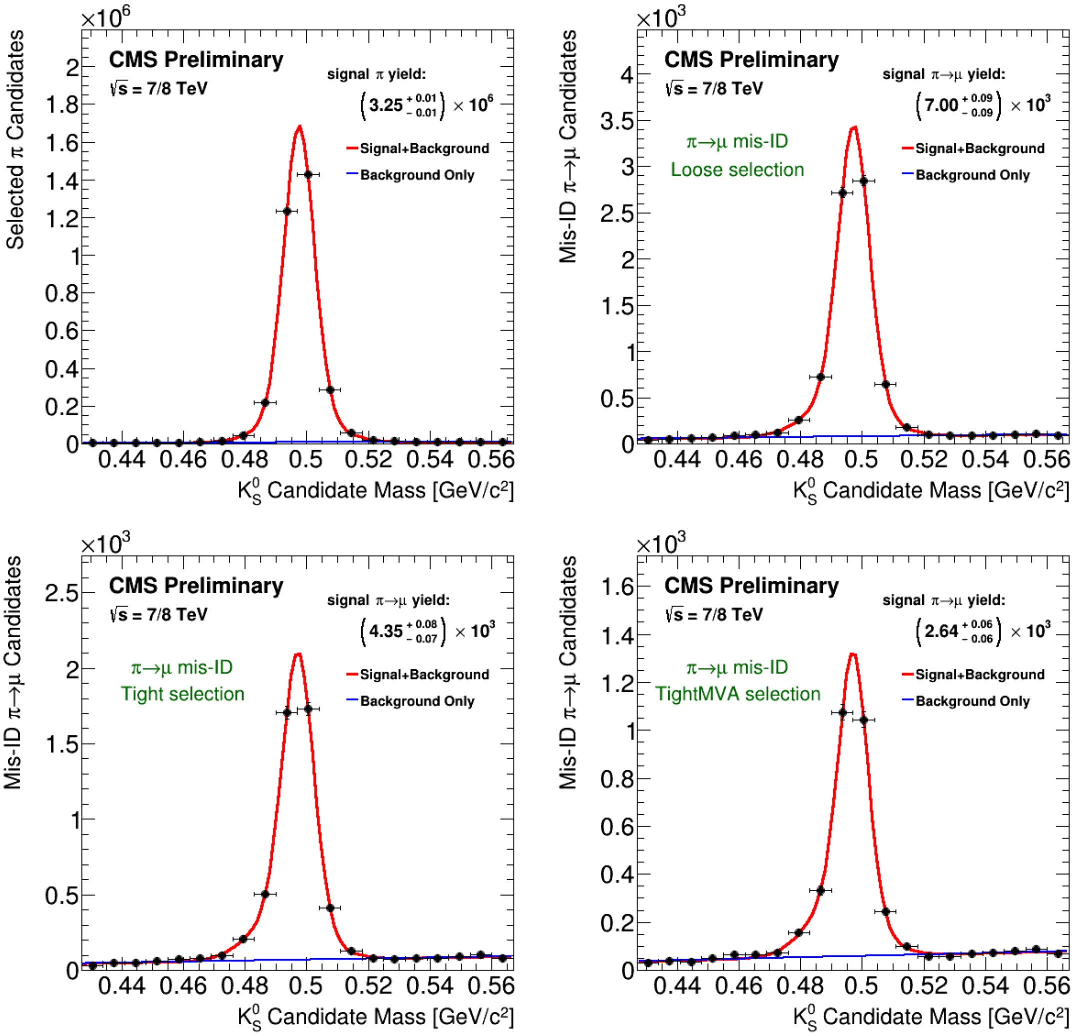

Figure A.1: Fits to the inclusive $K^0_\mathrm{S}$ invariant mass distribution in 2011/2012 data before and after muon-matching, with additional cuts of track $p_\mathrm{T} > 4$ GeV and $K^0_\mathrm{S}$ $L_\mathrm{xy} < 4$ cm. Shown is the distribution before matching (top left), after matching to Loose muons (top right), after matching to Tight muons (bottom left), and after matching to TightMVA muons (bottom right).

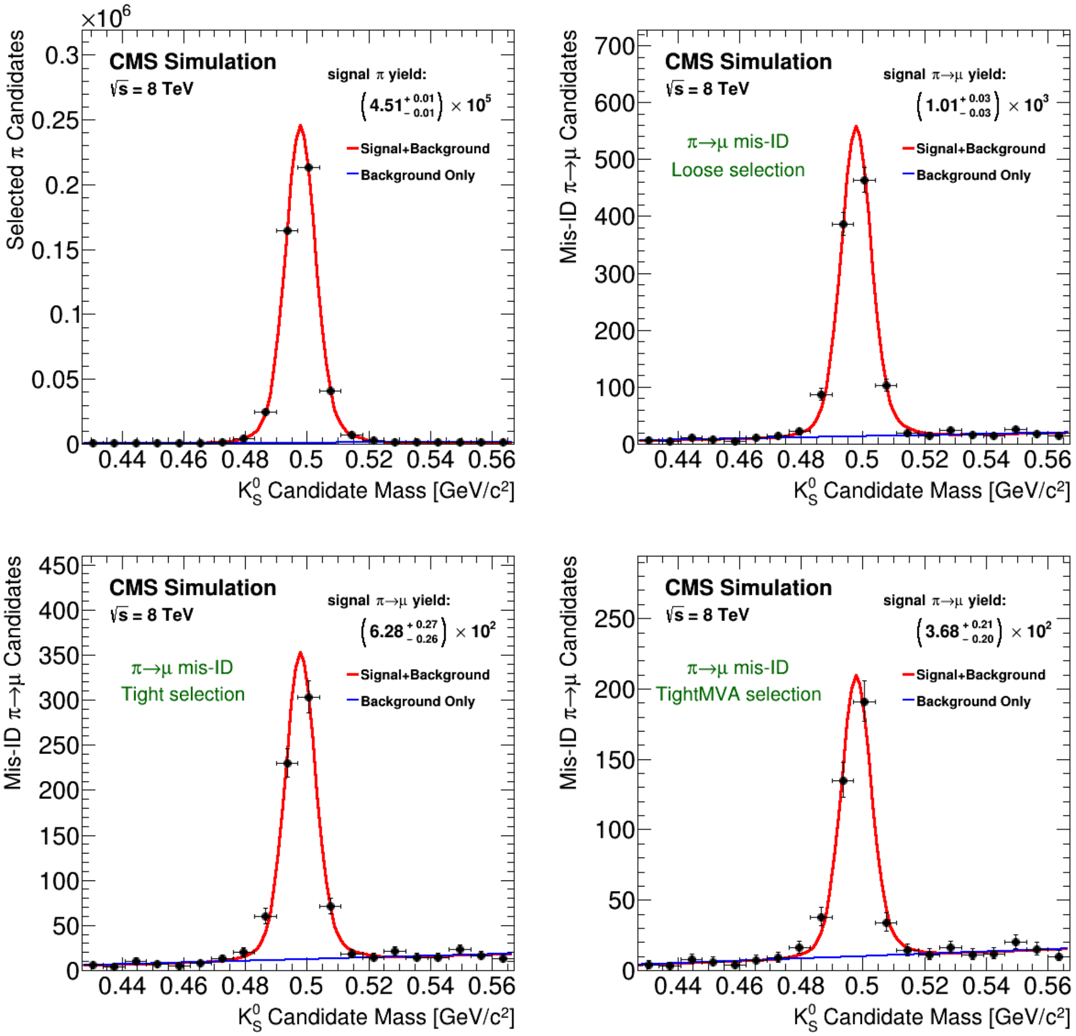

Figure A.2: Fits to the inclusive $K_S^0$ invariant mass distribution in QCD MC (2012 pile-up scenario) before and after muon-matching, with additional cuts of track $p_T > 4$ GeV and $K_S^0$ $L_{xy} < 4$ cm. Shown is the distribution before matching (top left), after matching to Loose muons (top right), after matching to Tight muons (bottom left), and after matching to TightMVA muons (bottom right).

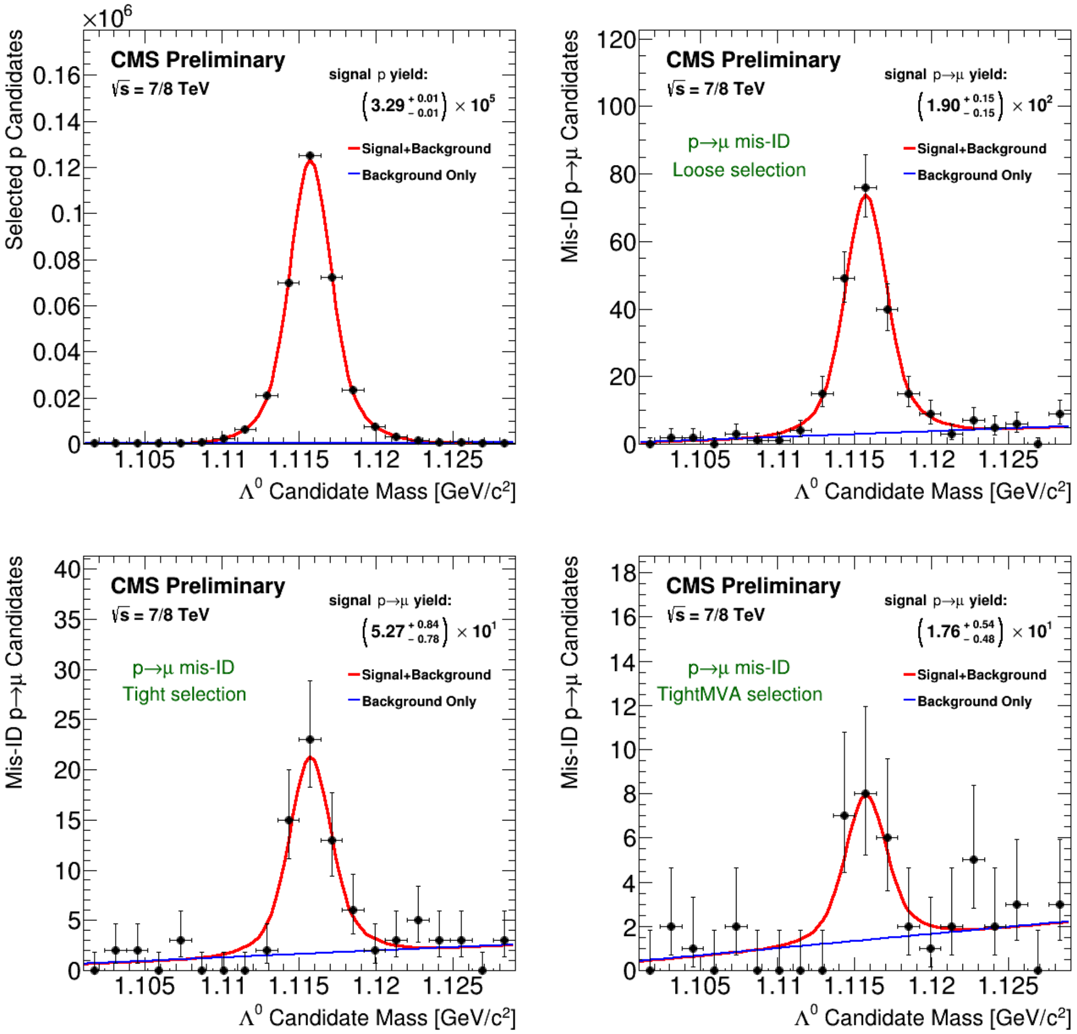

Figure A.3: Fits to the inclusive $\Lambda^0$ invariant mass distribution in 2011/2012 data before and after muon-matching, with additional cuts of track $p_{\mathrm{T}} > 4$ GeV and $\Lambda^0$ $L_{\mathrm{xy}} < 4$ cm. Shown is the distribution before matching (top left), after matching to Loose muons (top right), after matching to Tight muons (bottom left), and after matching to TightMVA muons (bottom right).

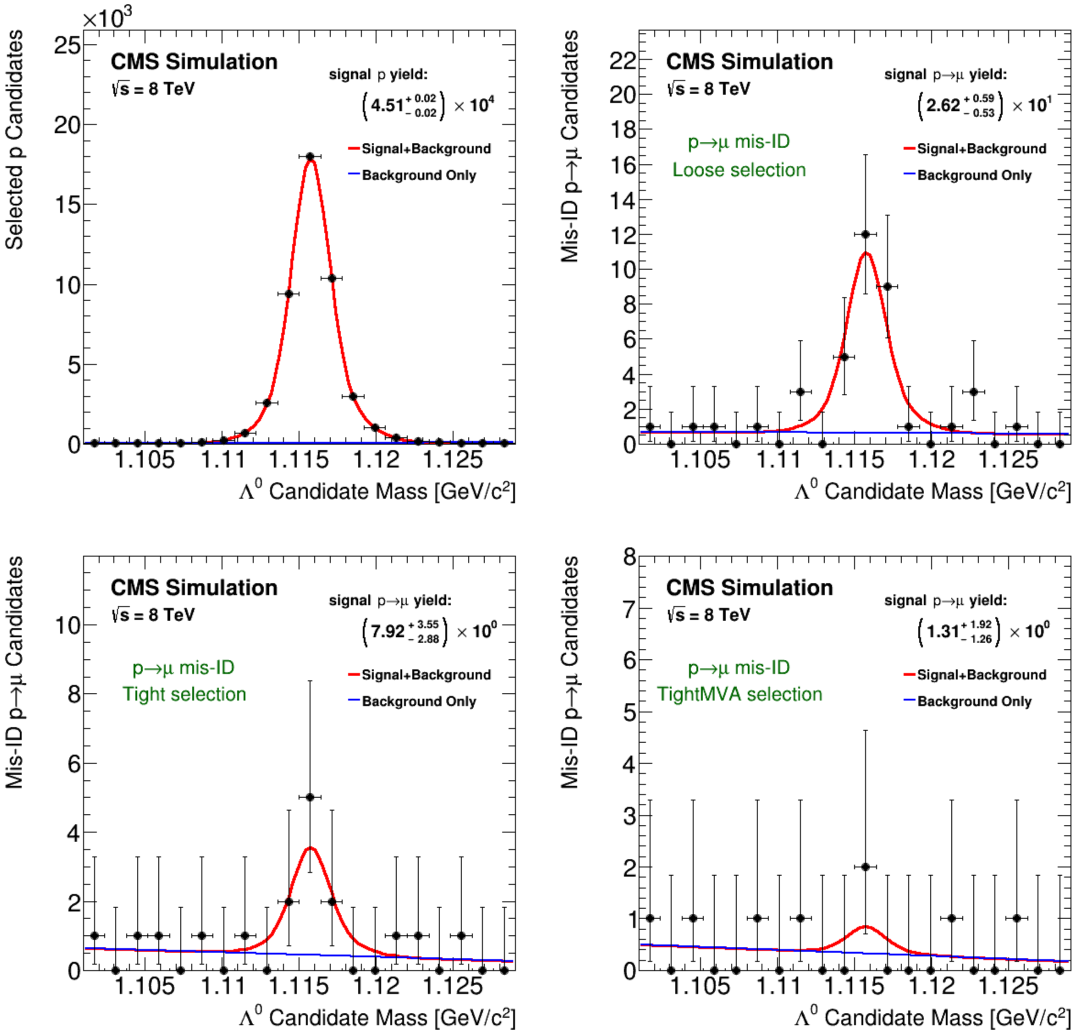

Figure A.4: Fits to the inclusive $\Lambda^0$ invariant mass distribution in QCD MC (2012 pile-up scenario) before and after muon-matching, with additional cuts of track $p_{\mathrm{T}} > 4$ GeV and $\Lambda^0$ $L_{\mathrm{xy}} < 4$ cm. Shown is the distribution before matching (top left), after matching to Loose muons (top right), after matching to Tight muons (bottom left), and after matching to TightMVA muons (bottom right).

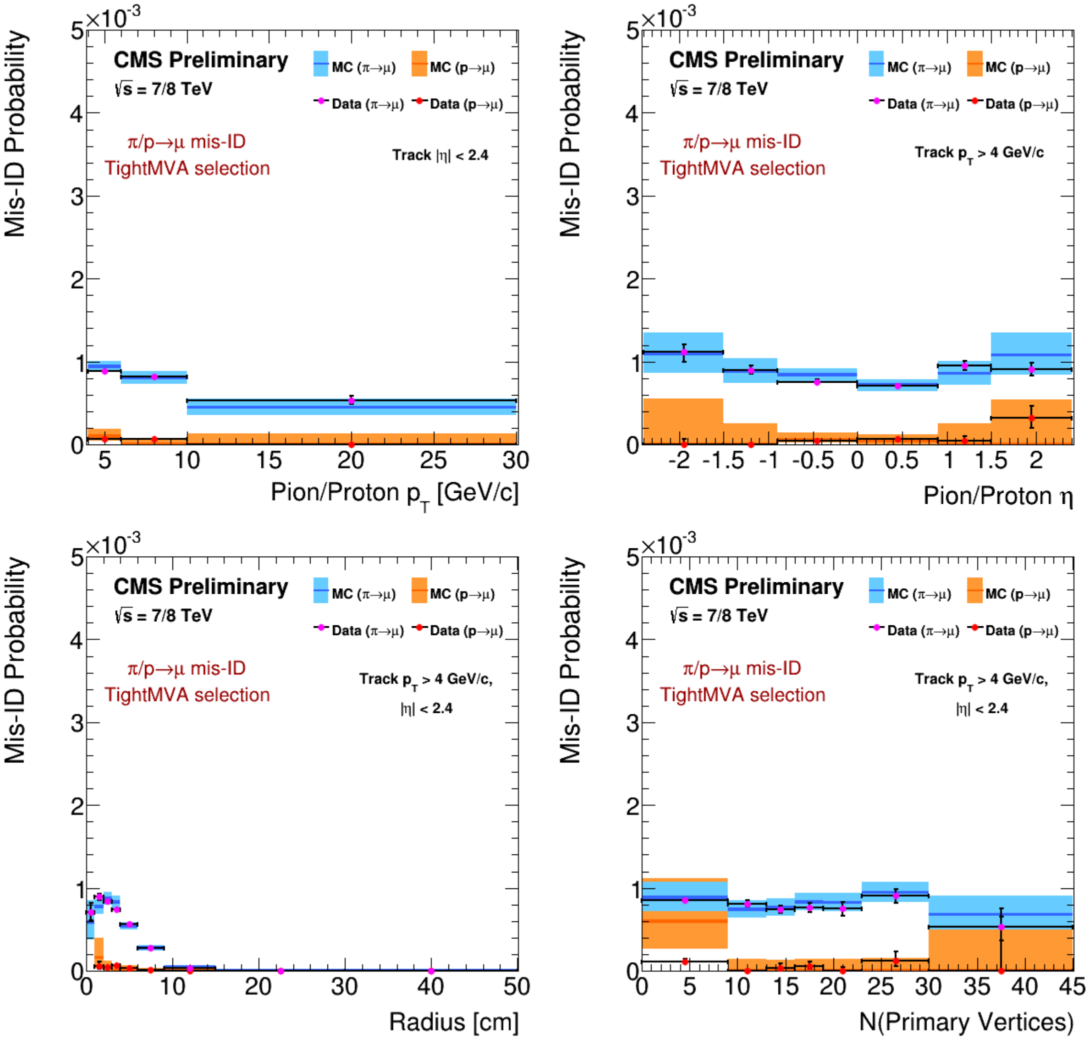

Figure A.5: Differential hadron-to-muon mis-identification rates as a function of various different variables, for the TightMVA muon selection only. This muon selection is used exclusively in the CMS $B_s^0 \to \mu^+\mu^-$ measurement and $B^0 \to \mu^+\mu^-$ search [15]. A comparison is made between the cases of pions and protons within each plot. An additional cut of track $p_{\mathrm{T}} > 4$ GeV is included.

# Appendix B

# Additional VH($b\bar{b}$) Plots

Presented in Figures B.1 and B.2 are BDT and multi-BDT output distributions after the binned maximum likelihood fit discussed in Section 9.2 is performed. These plots complement those presented in Figure 9.1 that show post-fit BDT and multi-BDT output distributions for only the high $p_{\mathrm{T}}(\mathrm{V})$ category of each channel.

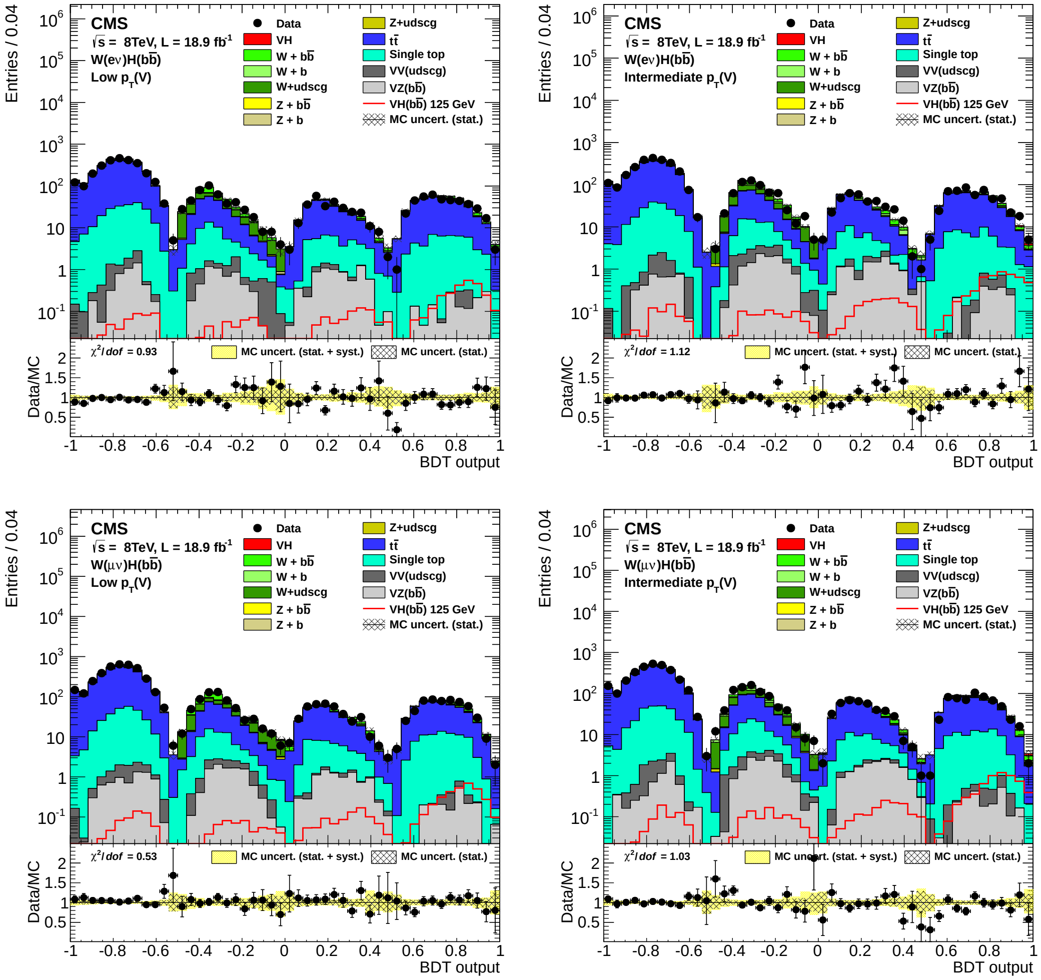

Figure B.1: Multi-BDT output distributions, after the binned maximum likelihood fit performed to extract signal strength (at a Higgs mass of 125 GeV), for the low $p_{\mathrm{T}}$(W) category (left column) and medium $p_{\mathrm{T}}$(W) category (right column) of the W(e$\nu$)H channel (top row) and the W($\mu\nu$)H channel (bottom row). Also presented in these plots is the predicted signal yield for a 125 GeV SM Higgs boson.

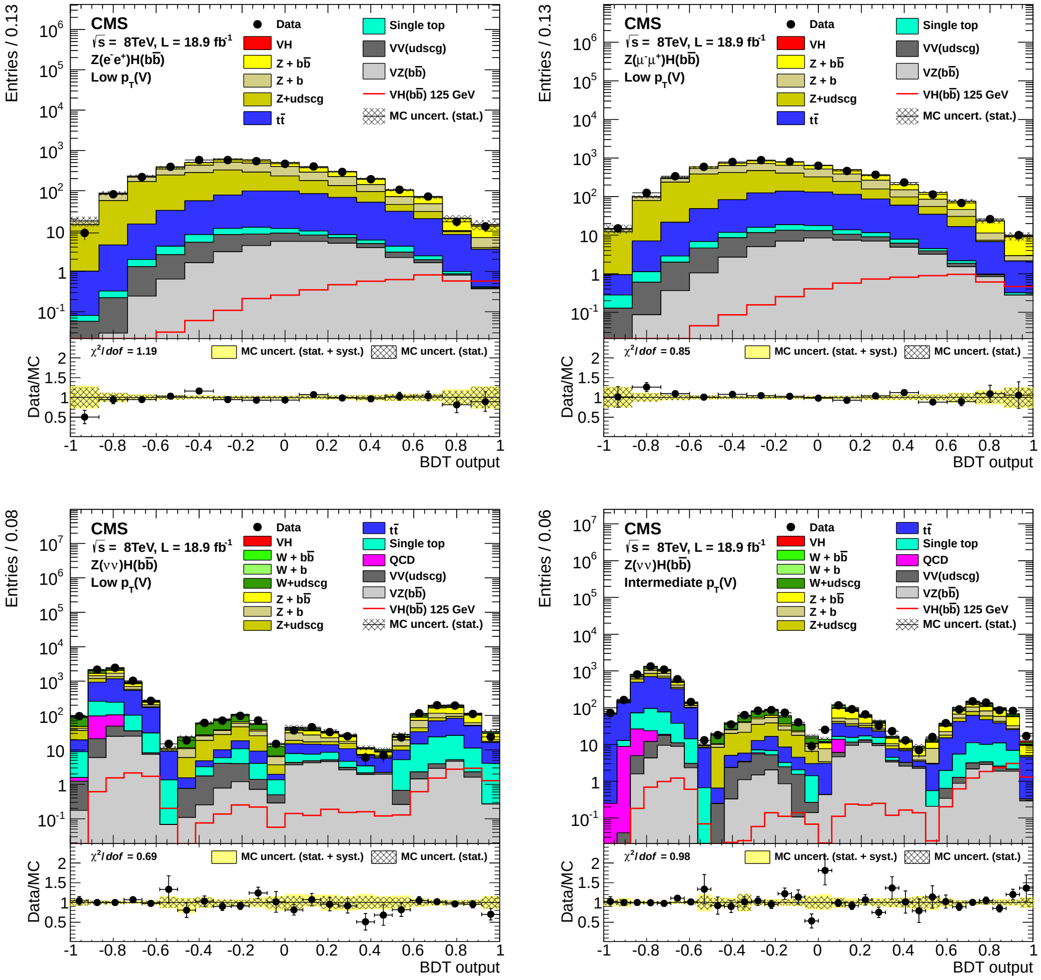

Figure B.2: BDT output distributions, after the binned maximum likelihood fit performed to extract signal strength (at a Higgs mass of 125 GeV), for the low $p_T$(Z) category of the Z(ee)H channel (top left) and the Z($\mu\mu$)H channel (top right), as well as multi-BDT output distributions after the same fit for the low $p_T$(Z) category (bottom left) and medium $p_T$(Z) category (bottom right) of the Z($\nu\nu$)H channel. Also presented in these plots is the predicted signal yield for a 125 GeV SM Higgs boson.

# References

[1] 2012 CERN Webfest, "Standard Model Infographic," http://www.isgtw.org/spotlight/go-particle-quest-first-cern-hackfest/, 2012.

[2] L. Alvarez-Gaume and J. Ellis, "Eyes on a prize particle," *Nat.Phys.*, vol. 7, pp. 2–3, 2011.

[3] WISE Simulation of LHC Optics, "Some Feynman diagrams of anticipated Higgs events," http://wise.web.cern.ch/wise/, 2012.

[4] LHC Higgs Cross Section Working Group, "Standard Model Higgs boson production cross sections and decay branching ratios," https://twiki.cern.ch/twiki/bin/view/lhcphysics/crosssections/, 2012.

[5] CMS Luminosity Public Results, "CMS Integrated Luminosity, p$\bar{\text{p}}$, 2011 and 2012," https://twiki.cern.ch/twiki/bin/view/cmspublic/lumipublicresults/, 2014.

[6] J.-L. Caron, "CERN Accelerator Complex." LHC Project Illustrations, Jun 1991.

[7] J.-L. Caron, "LHC Layout." LHC Project Illustrations, Sep 1997.

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

[9] G. Bayatian *et al.*, "CMS physics: Technical design report," CMS Technical Design Report, CERN-LHCC-2006-001, CMS-TDR-008-1, 2006.

[10] S. Chatrchyan *et al.*, "Description and performance of track and primary-vertex reconstruction with the CMS tracker." Unpublished (submitted to *JINST*), 2014.

[11] A. Holzner, “78 reconstructed vertices in event from high-pileup run 198609.” CMS Collection, Sep 2012.

[12] S. Chatrchyan *et al.*, “Identification of b-quark jets with the CMS experiment,” *JINST*, vol. 8, p. P04013, 2013.

[13] S. Chatrchyan *et al.*, “Determination of Jet Energy Calibration and Transverse Momentum Resolution in CMS,” *JINST*, vol. 6, p. P11002, 2011.

[14] D0 Collaboration, “Useful Diagrams of Top Signals and Backgrounds,” http://www-d0.fnal.gov/Run2Physics/top/top_public_web_pages/top_feynman_diagrams.html, 2011.

[15] S. Chatrchyan *et al.*, “Measurement of the B(s) to mu+ mu- branching fraction and search for B0 to mu+ mu- with the CMS Experiment,” *Phys.Rev.Lett.*, vol. 111, p. 101804, 2013.

[16] G. Arnison *et al.*, “Experimental Observation of Isolated Large Transverse Energy Electrons with Associated Missing Energy at s**(1/2) = 540-GeV,” *Phys.Lett.*, vol. B122, pp. 103–116, 1983.

[17] M. Banner *et al.*, “Observation of Single Isolated Electrons of High Transverse Momentum in Events with Missing Transverse Energy at the CERN anti-p p Collider,” *Phys.Lett.*, vol. B122, pp. 476–485, 1983.

[18] G. Arnison *et al.*, “Experimental Observation of Lepton Pairs of Invariant Mass Around 95-GeV/c**2 at the CERN SPS Collider,” *Phys.Lett.*, vol. B126, pp. 398–410, 1983.

[19] P. Bagnaia *et al.*, “Evidence for Z0 → e+ e- at the CERN anti-p p Collider,” *Phys.Lett.*, vol. B129, pp. 130–140, 1983.

[20] F. Abe *et al.*, “Observation of top quark production in $\bar{p}p$ collisions,” *Phys.Rev.Lett.*, vol. 74, pp. 2626–2631, 1995.

[21] S. Abachi *et al.*, "Observation of the top quark," *Phys.Rev.Lett.*, vol. 74, pp. 2632–2637, 1995.

[22] Y. Fukuda *et al.*, "Evidence for oscillation of atmospheric neutrinos," *Phys.Rev.Lett.*, vol. 81, pp. 1562–1567, 1998.

[23] G. Aad *et al.*, "Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC," *Phys.Lett.*, vol. B716, pp. 1–29, 2012.

[24] S. Chatrchyan *et al.*, "Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC," *Phys.Lett.*, vol. B716, pp. 30–61, 2012.

[25] G. Aad *et al.*, "Combined search for the Standard Model Higgs boson using up to 4.9 $\mathrm{fb}^{-1}$ of $pp$ collision data at $\sqrt{s} = 7$ TeV with the ATLAS detector at the LHC," *Phys.Lett.*, vol. B710, pp. 49–66, 2012.

[26] S. Chatrchyan *et al.*, "Combined results of searches for the standard model Higgs boson in $pp$ collisions at $\sqrt{s} = 7$ TeV," *Phys.Lett.*, vol. B710, pp. 26–48, 2012.

[27] T. Aaltonen *et al.*, "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.*, vol. 109, p. 071804, 2012.

[28] CDF and D0 Collaborations, "Updated Combination of CDF and D0 Searches for Standard Model Higgs Boson Production with up to 10.0 $\mathrm{fb}^{-1}$ of Data," CDF/D0 Conference Note, FERMILAB-CONF-12-318-E, CDF-NOTE-10884, D0-NOTE-6348, 2012.

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

[30] V. Khachatryan *et al.*, "Constraints on the Higgs boson width from off-shell pro-

duction and decay to $Z$-boson pairs." Unpublished (submitted to *Phys.Lett. B*), 2014.

[31] V. Khachatryan *et al.*, "Observation of the diphoton decay of the Higgs boson and measurement of its properties." Unpublished (submitted to *Eur.Phys.J. C*), 2014.

[32] CMS Collaboration, "Constraints on anomalous HVV interactions using H to 4l decays," CMS Physics Analysis Summary, CMS-PAS-HIG-14-014, 2014.

[33] CMS Collaboration, "Constraints on Anomalous HWW Interactions using Higgs boson decays to W+W- in the fully leptonic final state," CMS Physics Analysis Summary, CMS-PAS-HIG-14-012, 2014.

[34] CMS Collaboration, "Precise determination of the mass of the Higgs boson and studies of the compatibility of its couplings with the standard model," CMS Physics Analysis Summary, CMS-PAS-HIG-14-009, 2014.

[35] G. Aad *et al.*, "Evidence for the spin-0 nature of the Higgs boson using ATLAS data," *Phys.Lett.*, vol. B726, pp. 120–144, 2013.

[36] G. Aad *et al.*, "Measurements of Higgs boson production and couplings in diboson final states with the ATLAS detector at the LHC," *Phys.Lett.*, vol. B726, pp. 88–119, 2013.

[37] ATLAS Collaboration, "Updated coupling measurements of the Higgs boson with the ATLAS detector using up to 25 $\text{fb}^{-1}$ of proton-proton collision data," ATLAS Conference Note, ATLAS-CONF-2014-009, 2014.

[38] G. Aad *et al.*, "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 $\text{fb}^{-1}$ of $pp$ collision data." Unpublished (submitted to *Phys.Rev. D*), 2014.

[39] ATLAS Collaboration, "Determination of the off-shell Higgs boson signal strength in the high-mass ZZ final state with the ATLAS detector," ATLAS Conference Note, ATLAS-CONF-2014-042, 2014.

[40] S. Chatrchyan *et al.*, "Search for the standard model Higgs boson decaying to bottom quarks in *pp* collisions at $\sqrt{s} = 7$ TeV," *Phys.Lett.*, vol. B710, pp. 284–306, 2012.

[41] S. Chatrchyan *et al.*, "Search for the standard model Higgs boson produced in association with a W or a Z boson and decaying to bottom quarks," *Phys.Rev.*, vol. D89, p. 012003, 2014.

[42] S. Chatrchyan *et al.*, "Evidence for the direct decay of the 125 GeV Higgs boson to fermions," *Nature Phys.*, vol. 10, 2014.

[43] S. Chatrchyan *et al.*, "Evidence for the 125 GeV Higgs boson decaying to a pair of $\tau$ leptons," *JHEP*, vol. 1405, p. 104, 2014.

[44] J. J. Thomson, "Cathode Rays," *Phil.Mag.*, vol. 44, p. 293, 1897.

[45] J. Schechter and J. W. F. Valle, "Neutrino Masses in SU(2) x U(1) Theories," *Phys.Rev.*, vol. D22, p. 2227, 1980.

[46] W. Pauli, "Relativistic Field Theories of Elementary Particles," *Rev.Mod.Phys.*, vol. 13, p. 203, 1941.

[47] S. Tomonaga, "On a Relativistically Invariant Formulation of the Quantum Theory of Wave Fields," *Progr.Theoret.Phys.*, vol. 1, no. 2, pp. 27–42, 1946.

[48] J. Schwinger, "On Quantum-Electrodynamics and the Magnetic Moment of the Electron," *Phys.Rev.*, vol. 73, p. 416, 1948.

[49] J. Schwinger, "Quantum Electrodynamics. I. A Covariant Formulation," *Phys.Rev.*, vol. 74, p. 1439, 1948.

[50] R. P. Feynman, "The Theory of Positrons," *Phys.Rev.*, vol. 76, p. 749, 1949.

[51] R. P. Feynman, "Space-Time Approach to Quantum Electrodynamics," *Phys.Rev.*, vol. 76, p. 769, 1949.

[52] D. J. Gross and F. Wilczek, "Ultraviolet Behavior of Nonabelian Gauge Theories," *Phys.Rev.Lett.*, vol. 30, pp. 1343–1346, 1973.

[53] H. D. Politzer, "Reliable Perturbative Results for Strong Interactions?," *Phys.Rev.Lett.*, vol. 30, pp. 1346–1349, 1973.

[54] H. Fritzsch, M. Gell-Mann, and H. Leutwyler, "Advantages of the Color Octet Gluon Picture," *Phys.Lett.*, vol. B47, pp. 365–368, 1973.

[55] C.-N. Yang and R. L. Mills, "Conservation of Isotopic Spin and Isotopic Gauge Invariance," *Phys.Rev.*, vol. 96, pp. 191–195, 1954.

[56] J. Bjorken and E. A. Paschos, "Inelastic Electron Proton and gamma Proton Scattering, and the Structure of the Nucleon," *Phys.Rev.*, vol. 185, pp. 1975–1982, 1969.

[57] G. Zweig, "An SU(3) model for strong interaction symmetry and its breaking. Version 2," pp. 22–101, 1964.

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

[59] F. Englert and R. Brout, "Broken Symmetry and the Mass of Gauge Vector Mesons," *Phys.Rev.Lett.*, vol. 13, pp. 321–323, 1964.

[60] P. W. Higgs, "Broken symmetries, massless particles and gauge fields," *Phys.Lett.*, vol. 12, pp. 132–133, 1964.

[61] G. Guralnik, C. Hagen, and T. Kibble, "Global Conservation Laws and Massless Particles," *Phys.Rev.Lett.*, vol. 13, pp. 585–587, 1964.

[62] S. Glashow, "Partial Symmetries of Weak Interactions," *Nucl.Phys.*, vol. 22, pp. 579–588, 1961.

[63] S. Weinberg, "A Model of Leptons," *Phys.Rev.Lett.*, vol. 19, pp. 1264–1266, 1967.

[64] A. Salam, “Weak and Electromagnetic Interactions,” *Conf.Proc.*, vol. C680519, pp. 367–377, 1968.

[65] M. Green, J. Schwarz, and E. Witten, *Superstring Theory: Volume 1, Introduction.* Cambridge University Press, 1988.

[66] S. Weinberg, “The Cosmological Constant Problem,” *Rev.Mod.Phys.*, vol. 61, pp. 1–23, 1989.

[67] A. G. Riess *et al.*, “Observational evidence from supernovae for an accelerating universe and a cosmological constant,” *Astron.J.*, vol. 116, pp. 1009–1038, 1998.

[68] S. P. Martin, “A Supersymmetry primer,” *Adv.Ser.Direct.High Energy Phys.*, vol. 21, pp. 1–153, 2010.

[69] N. Arkani-Hamed, S. Dimopoulos, and G. Dvali, “The Hierarchy problem and new dimensions at a millimeter,” *Phys.Lett.*, vol. B429, pp. 263–272, 1998.

[70] L. Randall and R. Sundrum, “A Large mass hierarchy from a small extra dimension,” *Phys.Rev.Lett.*, vol. 83, pp. 3370–3373, 1999.

[71] H. Georgi and S. Glashow, “Unity of All Elementary Particle Forces,” *Phys.Rev.Lett.*, vol. 32, pp. 438–441, 1974.

[72] A. Buras, J. R. Ellis, M. Gaillard, and D. V. Nanopoulos, “Aspects of the Grand Unification of Strong, Weak and Electromagnetic Interactions,” *Nucl.Phys.*, vol. B135, pp. 66–92, 1978.

[73] G. Bertone, D. Hooper, and J. Silk, “Particle dark matter: Evidence, candidates and constraints,” *Phys.Rept.*, vol. 405, pp. 279–390, 2005.

[74] P. Peebles and B. Ratra, “The Cosmological constant and dark energy,” *Rev.Mod.Phys.*, vol. 75, pp. 559–606, 2003.

[75] E. J. Copeland, M. Sami, and S. Tsujikawa, “Dynamics of dark energy,” *Int.J.Mod.Phys.*, vol. D15, pp. 1753–1936, 2006.

[76] P. Ade *et al.*, “Planck 2013 results. I. Overview of products and scientific results,” 2013.

[77] M. Fukugita and T. Yanagida, “Baryogenesis Without Grand Unification,” *Phys.Lett.*, vol. B174, p. 45, 1986.

[78] D. Decamp *et al.*, “Determination of the Number of Light Neutrino Species,” *Phys.Lett.*, vol. B231, p. 519, 1989.

[79] I. J. R. Aitchison and A. J. G. Hey, *Gauge Theories in Particle Physics, Vol. 2: Non-Abelian Gauge Theories: QCD and the Electroweak Theory.* CRC Press, 2003.

[80] J. F. Gunion, H. E. Haber, G. Kane, and S. Dawson, *The Higgs Hunter’s Guide.* Westview Press, 2000.

[81] R. Barate *et al.*, “Search for the standard model Higgs boson at LEP,” *Phys.Lett.*, vol. B565, pp. 61–75, 2003.

[82] 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.*, vol. 100, p. 242001, 2008.

[83] L. Evans and P. Bryant, “LHC Machine,” *JINST*, vol. 3, p. S08001, 2008.

[84] K. Aamodt *et al.*, “The ALICE experiment at the CERN LHC,” *JINST*, vol. 3, p. S08002, 2008.

[85] J. Alves, A. Augusto *et al.*, “The LHCb Detector at the LHC,” *JINST*, vol. 3, p. S08005, 2008.

[86] G. Aad *et al.*, “The ATLAS Experiment at the CERN Large Hadron Collider,” *JINST*, vol. 3, p. S08003, 2008.

[87] D. Decamp *et al.*, “ALEPH: A detector for electron-positron annnihilations at LEP,” *Nucl.Instrum.Meth.*, vol. A294, pp. 121–178, 1990.

[88] P. Aarnio *et al.*, “The DELPHI detector at LEP,” *Nucl.Instrum.Meth.*, vol. A303, pp. 233–276, 1991.

[89] L3 Collaboration, “The Construction of the L3 Experiment,” *Nucl.Instrum.Meth.*, vol. A289, pp. 35–102, 1990.

[90] K. Ahmet *et al.*, “The OPAL detector at LEP,” *Nucl.Instrum.Meth.*, vol. A305, pp. 275–319, 1991.

[91] F. Abe *et al.*, “The CDF Detector: An Overview,” *Nucl.Instrum.Meth.*, vol. A271, pp. 387–403, 1988.

[92] S. Abachi *et al.*, “The D0 Detector,” *Nucl.Instrum.Meth.*, vol. A338, pp. 185–253, 1994.

[93] V. Abazov *et al.*, “The Upgraded D0 detector,” *Nucl.Instrum.Meth.*, vol. A565, pp. 463–537, 2006.

[94] J. C. Collins and M. Perry, “Superdense Matter: Neutrons Or Asymptotically Free Quarks?,” *Phys.Rev.Lett.*, vol. 34, p. 1353, 1975.

[95] N. Cabibbo and G. Parisi, “Exponential Hadronic Spectrum and Quark Liberation,” *Phys.Lett.*, vol. B59, pp. 67–69, 1975.

[96] T. Matsui and H. Satz, “$J/\psi$ Suppression by Quark-Gluon Plasma Formation,” *Phys.Lett.*, vol. B178, p. 416, 1986.

[97] CMS Collaboration, “The CMS tracker system project: Technical Design Report,” Technical Design Report, CERN-LHCC-98-006, CMS-TDR-5, 1997.

[98] CMS Collaboration, “The CMS electromagnetic calorimeter project: Technical Design Report,” Technical Design Report, CERN-LHCC-97-033, CMS-TDR-4, 1997.

[99] CMS Collaboration, “The CMS hadron calorimeter project: Technical Design Report,” Technical Design Report, CERN-LHCC-97-031, CMS-TDR-2, 1997.

[100] CMS Collaboration, "The CMS magnet project: Technical Design Report," Technical Design Report, CERN-LHCC-97-010, CMS-TDR-1, 1997.

[101] CMS Collaboration, "The CMS muon project: Technical Design Report," Technical Design Report, CERN-LHCC-97-032, CMS-TDR-3, 1997.

[102] CMS Collaboration, "CMS TriDAS project: Technical Design Report, Volume 1: The Trigger Systems," Technical Design Report, CERN-LHCC-2000-038, CMS-TDR-6-1, 2000.

[103] G. Bayatian *et al.*, "CMS technical design report, volume II: Physics performance," *J.Phys.*, vol. G34, pp. 995–1579, 2007.

[104] R. Brun and F. Rademakers, "ROOT: An object oriented data analysis framework," *Nucl.Instrum.Meth.*, vol. A389, pp. 81–86, 1997.

[105] V. Khachatryan *et al.*, "CMS Tracking Performance Results from early LHC Operation," *Eur.Phys.J.*, vol. C70, pp. 1165–1192, 2010.

[106] R. Fruhwirth, "Application of Kalman filtering to track and vertex fitting," *Nucl.Instrum.Meth.*, vol. A262, pp. 444–450, 1987.

[107] K. Rose, "Deterministic annealing for clustering, compression, classification, regression, and related optimization problems," *Proceedings of the IEEE*, vol. 86, pp. 2210–2239, Nov 1998.

[108] R. Fruhwirth, W. Waltenberger, and P. Vanlaer, "Adaptive vertex fitting," *J.Phys.*, vol. G34, p. N343, 2007.

[109] CMS Collaboration, "Search for a Higgs boson decaying into two photons in the CMS detector," CMS Physics Analysis Summary, CMS-PAS-HIG-11-010, 2011.

[110] CMS Collaboration, "Particle-Flow Event Reconstruction in CMS and Performance for Jets, Taus, and MET," CMS Physics Analysis Summary, CMS-PAS-PFT-09-001, 2009.

[111] 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.

[112] CMS Collaboration, “Particle-flow commissioning with muons and electrons from J/Psi and W events at 7 TeV,” CMS Physics Analysis Summary, CMS-PAS-PFT-10-003, 2010.

[113] W. Adam, R. Fruhwirth, A. Strandlie, and T. Todorov, “Reconstruction of electrons with the Gaussian sum filter in the CMS tracker at LHC,” *eConf*, vol. C0303241, p. TULT009, 2003.

[114] CMS Collaboration, “Commissioning of the Particle-Flow reconstruction in Minimum-Bias and Jet Events from pp Collisions at 7 TeV,” CMS Physics Analysis Summary, CMS-PAS-PFT-10-002, 2010.

[115] CMS Collaboration, “Performance of muon identification in pp collisions at sqrt(s) = 7 TeV,” CMS Physics Analysis Summary, CMS-PAS-MUO-10-002, 2010.

[116] CMS Collaboration, “Electron reconstruction and identification at sqrt(s) = 7 TeV,” CMS Physics Analysis Summary, CMS-PAS-EGM-10-004, 2010.

[117] S. Chatrchyan *et al.*, “Energy calibration and resolution of the CMS electromagnetic calorimeter in pp collisions at s = 7 TeV,” *JINST*, vol. 8, p. P09009, 2013.

[118] CMS Collaboration, “Tau identification in CMS,” CMS Physics Analysis Summary, CMS-PAS-TAU-11-001, 2011.

[119] G. P. Salam, “Towards Jetography,” *Eur.Phys.J.*, vol. C67, pp. 637–686, 2010.

[120] M. Cacciari, G. P. Salam, and G. Soyez, “The Anti-k(t) jet clustering algorithm,” *JHEP*, vol. 0804, p. 063, 2008.

[121] M. Cacciari, G. P. Salam, and G. Soyez, “FastJet User Manual,” *Eur.Phys.J.*, vol. C72, p. 1896, 2012.

[122] M. Cacciari and G. P. Salam, “Pileup subtraction using jet areas,” *Phys.Lett.*, vol. B659, pp. 119–126, 2008.

[123] CMS Collaboration, “Jet Performance in pp Collisions at 7 TeV,” CMS Physics Analysis Summary, CMS-PAS-JME-10-003, 2010.

[124] C. Weiser, “A Combined Secondary Vertex Based B-Tagging Algorithm in CMS,” CMS Note, CMS-NOTE-2006-014, 2006.

[125] CMS Collaboration, “Performance of b tagging at sqrt(s) = 8 TeV in multijet, ttbar and boosted topology events,” CMS Physics Analysis Summary, CMS-PAS-BTV-13-001, 2013.

[126] S. Chatrchyan *et al.*, “Missing transverse energy performance of the CMS detector,” *JINST*, vol. 6, p. P09001, 2011.

[127] 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.

[128] CMS Collaboration, “Measurement of the W and Z inclusive production cross sections at sqrt(s) = 7 TeV with the CMS experiment at the LHC,” CMS Physics Analysis Summary, CMS-PAS-EWK-10-002, 2010.

[129] Y. L. Dokshitzer, G. Leder, S. Moretti, and B. Webber, “Better jet clustering algorithms,” *JHEP*, vol. 9708, p. 001, 1997.

[130] CMS Collaboration, “Jet Substructure Algorithms,” CMS Physics Analysis Summary, CMS-PAS-JME-10-013, 2011.

[131] S. Chatrchyan *et al.*, “Studies of jet mass in dijet and W/Z + jet events,” *JHEP*, vol. 1305, p. 090, 2013.

[132] S. Chatrchyan *et al.*, “Performance of CMS muon reconstruction in $pp$ collision events at $\sqrt{s} = 7$ TeV,” *JINST*, vol. 7, p. P10002, 2012.

[133] CMS Collaboration, “Muon identification performance: hadron mis-identification measurements and RPC muon selections,” CMS Detector Performance Summary, CMS-DP-2014-018, 2014.

[134] CMS Collaboration, “Search for the resonant production of two Higgs bosons in the final state with two photons and two bottom quarks,” CMS Physics Analysis Summary, CMS-PAS-HIG-13-032, 2014.

[135] G. Corcella, I. Knowles, G. Marchesini, S. Moretti, K. Odagiri, *et al.*, “HERWIG 6: An Event generator for hadron emission reactions with interfering gluons (including supersymmetric processes),” *JHEP*, vol. 0101, p. 010, 2001.

[136] T. Sjostrand, S. Mrenna, and P. Z. Skands, “PYTHIA 6.4 Physics and Manual,” *JHEP*, vol. 0605, p. 026, 2006.

[137] S. Frixione, P. Nason, and C. Oleari, “Matching NLO QCD computations with Parton Shower simulations: the POWHEG method,” *JHEP*, vol. 0711, p. 070, 2007.

[138] J. Alwall, P. Demin, S. de Visscher, R. Frederix, M. Herquet, *et al.*, “MadGraph/MadEvent v4: The New Web Generation,” *JHEP*, vol. 0709, p. 028, 2007.

[139] J. Alwall, M. Herquet, F. Maltoni, O. Mattelaer, and T. Stelzer, “MadGraph 5 : Going Beyond,” *JHEP*, vol. 1106, p. 128, 2011.

[140] S. Agostinelli *et al.*, “GEANT4: A Simulation toolkit,” *Nucl.Instrum.Meth.*, vol. A506, pp. 250–303, 2003.

[141] LHC Higgs Cross Section Working Group, S. Dittmaier, C. Mariotti, G. Passarino, and R. Tanaka (Eds.), “Handbook of LHC Higgs Cross Sections: 1. Inclusive Observables,” CERN-2011-002, 2011.

[142] T. Aaltonen, A. Buzatu, B. Kilminster, Y. Nagai, and W. Yao, “Improved $b$-jet Energy Correction for $H \to b\bar{b}$ Searches at CDF,” Fermilab PPD Technical Memo, FERMILAB-TM-2513-PPD, 2011.

[143] B. P. Roe, H.-J. Yang, J. Zhu, Y. Liu, I. Stancu, *et al.*, "Boosted decision trees, an alternative to artificial neural networks," *Nucl.Instrum.Meth.*, vol. A543, pp. 577–584, 2005.

[144] A. Hocker, J. Stelzer, F. Tegenfeldt, H. Voss, K. Voss, *et al.*, "TMVA - Toolkit for Multivariate Data Analysis," *PoS*, vol. ACAT, p. 040, 2007.

[145] J. Beringer *et al.*, "Review of Particle Physics (RPP)," *Phys.Rev.*, vol. D86, p. 010001, 2012.

[146] G. Aad *et al.*, "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*, vol. 1306, p. 084, 2013.

[147] S. Chatrchyan *et al.*, "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*, vol. 1312, p. 039, 2013.

[148] S. Chatrchyan *et al.*, "Measurement of the Z/gamma*+b-jet cross section in pp collisions at 7 TeV," *JHEP*, vol. 1206, p. 126, 2012.

[149] J. Gallicchio and M. D. Schwartz, "Seeing in Color: Jet Superstructure," *Phys.Rev.Lett.*, vol. 105, p. 022001, 2010.

[150] T. Aaltonen *et al.*, "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.*, vol. 109, p. 111803, 2012.

[151] CMS Collaboration, "Measurement of VZ production cross sections in VZ to Vbb-bar decay channels in pp collisions at 8 TeV," CMS Physics Analysis Summary, CMS-PAS-SMP-13-011, 2013.

[152] CMS Collaboration, "Absolute calibration of the luminosity measurement at CMS: Winter 2012 update," CMS Physics Analysis Summary CMS-PAS-SMP-12-008, 2012.

[153] CMS Collaboration, “CMS luminosity measurement based on pixel cluster counting: Summer 2012 update,” CMS Physics Analysis Summary, CMS-PAS-LUM-12-001, 2012.

[154] LHC Higgs Cross Section Working Group, S. Dittmaier, C. Mariotti, G. Passarino, and R. Tanaka (Eds.), “Handbook of LHC Higgs Cross Sections: 2. Differential Distributions,” CERN-2012-002, 2012.

[155] G. Ferrera, M. Grazzini, and F. Tramontano, “Associated WH production at hadron colliders: a fully exclusive QCD calculation at NNLO,” *Phys. Rev. Lett.*, vol. 107, p. 152003, 2011.

[156] 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.*, vol. 99, p. 161803, 2007.

[157] M. Ciccolini, A. Denner, and S. Dittmaier, “Electroweak and QCD corrections to Higgs production via vector-boson fusion at the LHC,” *Phys. Rev. D*, vol. 77, p. 013002, 2008.

[158] 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*, vol. 03, p. 075, 2012.

[159] A. D. Martin, W. J. Stirling, R. S. Thorne, and G. Watt, “Parton distributions for the LHC,” *Eur. Phys. J. C*, vol. 63, p. 189, 2009.

[160] M. Botje *et al.*, “The PDF4LHC Working Group interim recommendations.” 2011.

[161] S. Alekhin *et al.*, “The PDF4LHC Working Group interim report.” 2011.

[162] H.-L. Lai, M. Guzzi, J. Huston, Z. Li, P. M. Nadolsky, J. Pumplin, and C.-P. Yuan, “New parton distributions for collider physics,” *Phys. Rev. D*, vol. 82, p. 074024, 2010.

[163] R. D. Ball, V. Bertone, F. Cerutti, L. Del Debbio, S. Forte, A. Guffanti, J. I. Latorre, J. Rojo, and M. Ubiali, “Impact of Heavy Quark Masses on Parton Distributions and LHC Phenomenology,” *Nucl. Phys. B*, vol. 849, p. 296, 2011.

[164] S. Chatrchyan *et al.*, “Measurement of W+W- and ZZ production cross sections in pp collisions at sqrt(s) = 8 TeV,” *Phys.Lett.*, vol. B721, pp. 190–211, 2013.

[165] CMS Collaboration, “Measurement of the single-top t-channel cross section in pp collisions at centre-of-mass energy of 8 TeV,” CMS Physics Analysis Summary, CMS-PAS-TOP-12-011, 2012.

[166] CMS Collaboration, “Single Top associated tW production at 8 TeV in the two lepton final state,” CMS Physics Analysis Summary, CMS-PAS-TOP-12-040, 2013.

[167] ATLAS and CMS Collaborations, “Procedure for the LHC Higgs boson search combination in summer 2011,” ATLAS/CMS Note, ATL-PHYS-PUB-2011-011/CMS-NOTE-2011-005, 2011.

[168] T. Junk, “Confidence level computation for combining searches with small statistics,” *Nucl.Instrum.Meth.*, vol. A434, pp. 435–443, 1999.

[169] A. L. Read, “Presentation of search results: The CL(s) technique,” *J.Phys.*, vol. G28, pp. 2693–2704, 2002.

[170] E. Gross and O. Vitells, “Trial factors or the look elsewhere effect in high energy physics,” *Eur.Phys.J.*, vol. C70, pp. 525–530, 2010.

[171] D. Krohn, M. D. Schwartz, M. Low, and L.-T. Wang, “Jet Cleansing: Pileup Removal at High Luminosity.” 2013.