# EVIDENCE FOR A STANDARD MODEL HIGGS BOSON PRODUCED IN ASSOCIATION WITH A TOP QUARK PAIR AND DECAYING TO LEPTONS

A Dissertation

Submitted to the Graduate School
of the University of Notre Dame
in Partial Fulfillment of the Requirements
for the Degree of

Doctor of Philosophy

by

Charles Mueller

---

Kevin Lannon, Director

Graduate Program in Physics
Notre Dame, Indiana
November 2017

EVIDENCE FOR A STANDARD MODEL HIGGS BOSON PRODUCED IN ASSOCIATION WITH A TOP QUARK PAIR AND DECAYING TO LEPTONS

Abstract

by

Charles Mueller

A search for the standard model Higgs boson produced in association with a top quark pair is presented, using the full pp collision dataset corresponding to an integrated luminosity of 35.9 $\text{fb}^{-1}$ collected by the CMS experiment at a center of mass energy of $\sqrt{s} = 13$ TeV. MVA-based event reconstruction techniques are used to identify final states where the Higgs boson decays to either a W, Z or tau pair by selecting events with two isolated same-sign leptons, and b-jets. The observed best-fit $\text{t}\bar{\text{t}}\text{H}$ signal strength is $1.7^{+0.6}_{-0.5}$ times the Standard Model prediction, corresponding to a significance of 3.3 standard deviations above the background-only hypothesis. The observed 95% CL upper limit on the signal strength is 2.9 times the Standard Model prediction, compared to the expected upper limit of $1.0^{+0.5}_{-0.3}$.

To my parents, Charles and Toni.

CONTENTS

FIGURES

TABLES

# ACKNOWLEDGMENTS

I would first like to acknowledge my advisor Kevin Lannon, whose support and guidance helped me through my graduate school experience. Thanks to Kevin's advising, I was regularly engaged in useful, interesting and visible projects. Kevin taught me the principles of scientific investigation, and I am confident that any further success I enjoy will be in part due to the advising I received during my time at Notre Dame.

I would also like to acknowledge the faculty and staff in the High Energy Physics group at Notre Dame. In addition to my advisor, Mike Hildreth, Colin Jessop, and other CMS faculty members helped create an effective and impactful research effort on CMS. Thanks to their generous support, I was able to spend several years at CERN in an intellectually stimulating environment that facilitated my transition into a researcher. I must also express my gratitude to the Notre Dame community, especially the Physics Department staff: Sherry Herman, Shelly Goethals, and Susan Baxmeyer, who made Notre Dame and South Bend feel like home.

The measurements presented in this dissertation would not be possible without the efforts of thousands of dedicated scientists, engineers and students working on the LHC and CMS experiment. I am grateful for the time I spent working in the CMS Trigger Studies Group, where I learned both the nuances of creating and operating sophisticated software, and, more importantly, the nuances of working with other people. A special thanks is due to Andrea Bocci, Tulika Bose, Aram Avestiyan, and Roberta Acridiacono. More directly, this work is the result of the collaboration of many talented scientists working on $t\bar{t}H$: Wuming Luo, Christopher Neu, Matthias

Wolf, Jason Slaunwhite, Marco Peruzzi, Francesco Romeo, Binghuan Li, and Giovanni Petrucciani. Special credit is due to Geoffrey Smith, the friend, officemate and postdoc whom I worked with most.

During my time at CERN, I was fortunate enough to form friendships with special people who set examples in kindness and scientific aptitude that I still do my best to follow. In addition to those mentioned above, this includes Justin Pilot, Christine McClean, Ted Kolberg, Rachel Yohay, and Sean Flowers. A special acknowledgement is due to my camrades de chambre and now close friends at Boulevard des Philosophes: Andrea Tognina, Charlie Goodlake, Benjamin Tannenwald, and Johannes Fexer. I will not soon forget our days on the lake, nights in Geneva, or adventures in the Alps. This circle of friends made work and play far more enjoyable than I could have anticipated. The same is true of my friends/classmates at Notre Dame: Andrew Brinkerhoff, Joseph Hagmann, Nil Valls, Doug and Tessa Berry, Anna Woodard, Nabarun Dev, Fanbo Meng, and Anthony Ruth. A special thanks is in order to my long-time officemate and good friend Michael Planer, whom I enjoyed many long conversations, only some of which pertained to physics. I am especially grateful for Diane Polydoris, who was patient with me when I was stubborn, kind to me when I was rude, and understanding when I was upset. I am still learning to appreciate the extent to which her support has grounded and helped me over the years.

Finally, I thank my family, who supported my ambitions when success was uncertain.

# CHAPTER 1

# INTRODUCTION

The recent observation of the Higgs boson confirmed the mechanism by which matter acquires mass. While this discovery made significant progress towards completing the Standard Model's overall successful description of nature, many important questions regarding the origins of mass remain unanswered. Do the observed properties of the Higgs boson match SM expectations? Why do particles have the specific masses we observe? Is the top quark's large mass coming solely from its interaction with the Higgs? Answering these questions will provide crucial insight into the underlying principles that govern our universe. The research and analysis presented here attempts to address these questions.

This analysis aims to discover processes from proton-proton collisions at the LHC where a Higgs boson is produced in association with a top-antitop quark pair (denoted as $t\bar{t}H$) and decays to final states with two or more charged leptons in the CMS detector at a center-of-mass collision energy of 13 TeV. Referred to as $t\bar{t}H$ multilepton processes, these provide an efficient probe with which to test the Standard Model.

The purpose of $t\bar{t}H$ searches is to measure the top-Higgs Yuakawa coupling. While this coupling is already indirectly measured via tight constraints on the gluon-gluon fusion production mode of the Higgs (where it is assumed a top quark dominates the loop), a direct measurement of this coupling at tree-level is the ultimate motivation for $t\bar{t}H$.

There are significant experimental challenges to measuring $t\bar{t}H$ processes. Besides the large backgrounds and small signal, $t\bar{t}H$ events are themselves complicated. There

are 10 final state particles in the signal diagram in Figure 1.1, this translates to at most 9 final state objects reconstructed in the detector since the neutrinos are only detected due to a single missing transverse energy quantity. While having as many of the final state objects as possible is desirable, it also makes the task of object assignment more complicated, since many objects are indistinguishable due to finite detector resolution and uncertainties. This is also what makes reconstructing the visible Higgs mass nearly impossible. On the other hand, events passing the signal region selection of this analysis rarely contain the full 9 objects due to a number of reasons. The missing objects could be mis-reconstructed and not identified properly by the detector, or they fail the selection quality requirements, or they are outside the fiducial region of the detector we restrict our search to. A full event reconstruction is also nearly impossible with missing objects also, and additionally the backgrounds look more similar to partially reconstructed $t\bar{t}H$ signal events.

There are several reasons for considering $t\bar{t}H$ events with multiple lepton final states. The primary experimentally-driven reason is due to the efficiency with which CMS identifies and reconstructs charged leptons. Reconstructing electrons and muons accurately is far easier compared to reconstructing jets, which will be discussed in the following chapters. From a theoretical perspective, the bb decay mode of the Higgs is the most desirable having the largest branching fraction, however this is balanced by the large uncertainties and experimental difficulty of accurately identifying and reconstructing jets with CMS. Finally, there tend to be smaller backgrounds capable of producing multiple leptons consistent with $t\bar{t}H$ with respect to other Higgs decay modes.

The results presented here build on previous measurements of the top-Higgs Yukawa coupling that were performed in $t\bar{t}H$ analyses by both ATLAS and CMS. The initial searches analyzed 20 $fb^{-1}$ of 8 TeV pp collisions during Run I of the LHC. Combining the results of Higgs decay channels of bb, $\gamma\gamma$, and leptonic final states

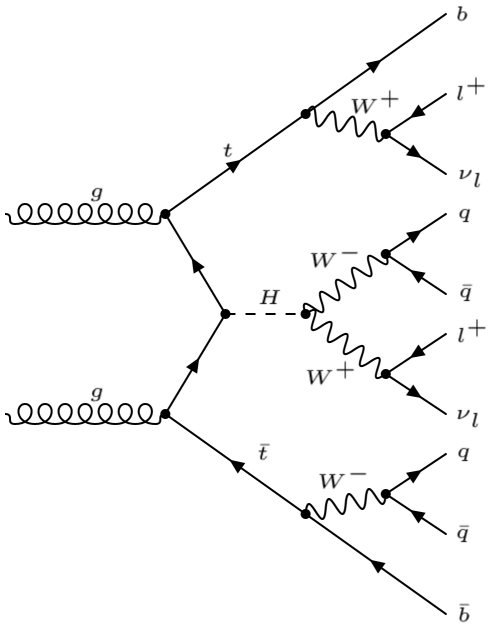



Figure 1.1. A feynman diagram of the primary signal process.

produced a signal strength of $\mu = 2.8 \pm 1.0$, where $\mu = \sigma/\sigma_{SM}$. The multilepton channel alone observed $\mu = 3.7^{1.6}_{1.4}$ [**?** ]. Subsequent iterations of the multilepton analysis were performed again on 13 TeV pp collisions from Run II of the LHC, the most recent signal strength observations are $\mu = 2.0^{0.8}_{0.7}$, corresponding to 12.9 fb$^{-1}$, and the subsequent analysis on the full 2016 dataset on which this dissertation is based, and $\mu = 1.5 \pm 0.5$ ,corresponding to 35.9 fb$^{-1}$.

# CHAPTER 2

# THEORY

## 2.1 The Standard Model

The Standard Model (SM) of particle physics provides the context in which this analysis is performed and the results interpreted. It explains three of four fundamental forces via gauge symmetries, while characterizing unknown matter into separate generations of particles called quarks and leptons. Since its inception in the early 1960s, the SM has predicted the existence of nearly every fundamental particle that has been discovered to-date. The SM distills the real-world observables of matter and energy into discrete elementary particles and their kinematics. The SM is the theory on which all of the following research is based, and also the hypothesis being tested in this analysis.

### 2.1.1 Particle Content

The particles in the SM are first characterized by their intrinisic angular momentum, more commonly referred to as spin. Particles with half-integer spin, quantized in units proportional to Planck's constant $\hbar$, are fermions, while particles with integer spin are bosons. This distinction is important because the spin values govern behavior and interactions of collections of particles.

The fermions in the SM are the most fundamental examples of matter in nature. Fermions behave and interact according to Fermi-Dirac statistics, and obey the Pauli exclusion principle. Fermions are further categorized, based on their primary interaction mechanism, into quarks and leptons. There are six different flavors of quarks

in the SM: the up and down, the charm and strange, and the top and bottom quarks are organized into three generations of doublets below.

$$\begin{pmatrix} u \\ d \end{pmatrix} \quad \begin{pmatrix} c \\ s \end{pmatrix} \quad \begin{pmatrix} t \\ b \end{pmatrix} \tag{2.1}$$

An increasing mass (from left to right) distinguishes each generation, while the upper and lower elements of each doublet are distinguished by an electric charge of +2/3 and -1/3 respectively in each generation. Quarks interact via the strong and electroweak forces. Quarks also carry a color charge, which can assume one of three values (red, blue, green) as a result of the strong interaction described by Quantum Chromodynamics (QCD). The leptons in the SM can also be arranged into three increasingly massive generations of doublets.

$$\begin{pmatrix} e^- \\ \nu_e \end{pmatrix} \quad \begin{pmatrix} \mu^- \\ \nu_\mu \end{pmatrix} \quad \begin{pmatrix} \tau^- \\ \nu_\tau \end{pmatrix} \tag{2.2}$$

The upper elements in each lepton doublet are the familiar electron, and the less familiar but much heavier, muon and tau. Due to their increased mass, the muon and tau have short lifetimes which causes them to decay rapidly to lighter, more stable particles. In the context of CMS however, the muon is stable. The taus decay inside CMS and the subsequent reconstruction will be discussed in the following chapters. The electron, muon, and tau all have the same electric charge of -1.

The lower elements in each doublet are the lightweight and electrically neutral counterparts called neutrinos, which also come in three flavors; the electron-neutrino, the muon-neutrino, and the tau-neutrino. Neutrinos interact primarily through the weak force. In an experimental context such as CMS, neutrinos are characterized by how weakly they interact. They effectively don't interact at all for what concerns CMS. They interact so weakly that they can pass through the earth without a single interaction. This property makes it impossible to directly detect their presence at

CMS. For every electrically charged example of matter described above, there exists a nearly identical anti-matter version. Antimatter is identical to matter, except that the signs on all charges and spins are opposite that of ordinary matter. Matter and anti-matter interact via the same forces/gauge bosons.

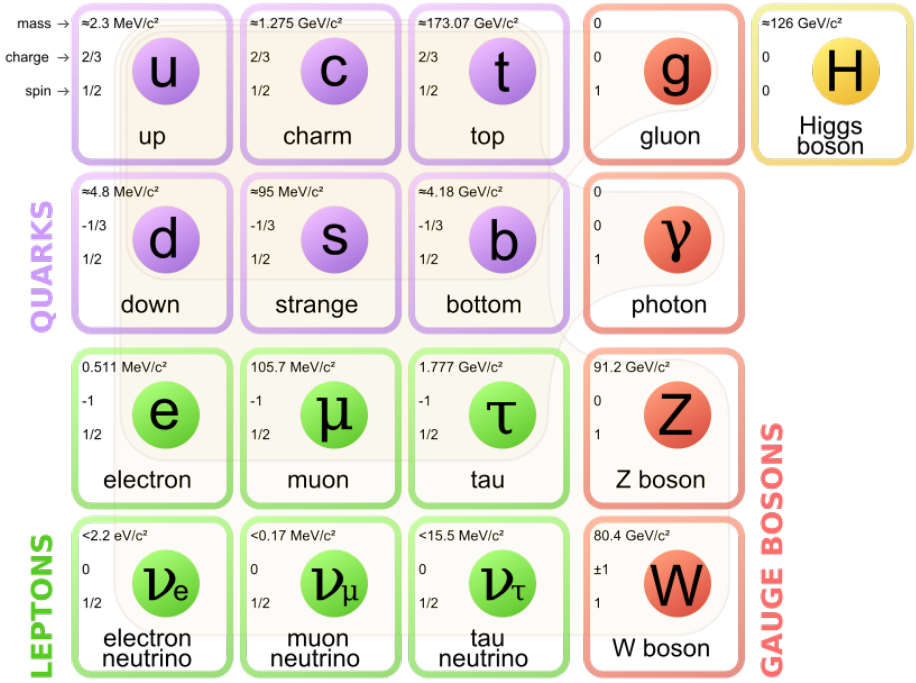



Figure 2.1. A summary of all elementary particles and their interactions in the Standard Model [**?** ].

The bosons in the SM are also fundamental, but are not examples of matter. Bosons are characterized by their integer-quantized (in units proportional to $\hbar$) angular momentum and behave according to Bose-Einstein statistics. There are five elementary SM bosons, the four force-carrying gauge bosons, and one scalar (spin-0)

boson that was recently discovered in 2012, known as the Higgs boson. Three of the four forces (strong, weak, electromagnetic) through which particles interact are all carried by corresponding gauge bosons. The SM currently does not explain or incorporate gravity and there is no SM particle that carries its force. The hypothetical gauge boson believed to be responsible for gravity is the graviton, and has yet to be discovered because its coupling to SM particles is too weak to be probed by the current center-of-mass energies that today's colliders are capable of producing. The strongest of the four forces, the appropriately-named strong force is carried by the gluon. Gluons are spin-1, electrically neutral, massless, and carry a color charge. Gluons mediate the strong force through which quarks interact. Due to the nature of color charge and confinement, gluons keep the quarks *glued* together, confined inside hadrons. Additionally, the strong force also binds protons and neutrons together to form nuclei of atoms. Any particle carrying color-charge is capable of strong interactions. The photon is the gauge boson that mediates the next strongest interaction, the electromagnetic force. The photon is massless, spin-1, electrically neutral, and travels at the speed of light. Aside from gravity, the electromagnetic force is the most familiar, responsible for keeping electron orbitals bound to nuclei, forming atoms, and it is also responsible for the attractive and repulsive forces that bond atoms together into molecules. Any particle carrying electric charge is capable of interacting electromagnetically. The weakest force explained by the SM, the appropriately named weak force is mediated by the massive W and Z gauge bosons. There are two types of weak interactions, charged and neutral. There are two W bosons, $W^+$,$W^-$ which are identical except for their electric charges of +1 and -1 respectively. The spin-1 W boson has a mass of 80.4 GeV [**?** ], and mediates the weak charged interaction. The electrically neutral, spin-1 Z boson has a mass of 91.2 GeV and mediates the weak neutral interaction [**?** ]. The decay of unstable atoms, which is harnessed for nuclear power, is possible thanks to the weak interaction. The final elementary SM boson

is the Higgs. The Higgs is a massive spin-0, electrically neutral boson that interacts with both fermions and other bosons and has a mass of approximately 125 GeV [**?** ]. While the Higgs doesn't mediate a force, it does represent the Higgs field and the mechanism by which elementary particles (matter) obtain masses. The Higgs boson and the origins of mass will be explained the coming sections.

TABLE 2.1

Relative dimensionless strengths of the fundamental forces

| Force | Relative Strength |
|---|---|
| Gravity | 1 |
| Weak | $10^{25}$ |
| Electromagnetic | $10^{36}$ |
| Strong | $10^{38}$ |

### 2.1.2 Fields and Symmetries

The notion of symmetry is of central importance to the SM. The SM describes three of the fundamental forces with gauge field theories. A gauge theory is a field theory whose Lagrangian is invariant under a specific type of transformation, called a gauge transformation. These transformations form a symmetry (group), because the Lagrangian is invariant under the transformation. The word symmetry is used to describe these transformations because the underlying dynamics of the theory are left unchanged or *symmetric* with respect to the transformation. Thanks to Noether's

Theorem, these symmetries are critical to the SM because they describe and invoke the conservation laws that SM particles obey. Each force within the SM corresponds to its own gauge field theory and symmetry that describe it. In each gauge theory, the corresponding field or force-carrying gauge boson is represented by the generators of the symmetry group.

Originally, the gauge theory that described the electromagnetic force, Quantum Electrodynamics (QED), was invariant under a $U(1)$ symmetry that described the photon as the force-carrying gauge boson and the symmetry required electric charge to be conserved. QED was first developed by Sin-Itiro Tomonaga, Julian Schwinger, and Richard Feynman. This laid the foundation on which the rest of the SM was built and won them the Nobel Prize in 1965 [**?** ].

Not long after the development of QED, C.N. Yang and Robert Mills formalized the gauge field theory techniques that were further refined by Sheldon Glashow, Abdus Salam, and Steven Weinberg in the 1960s to combine QED with the description of the weak force, producing a unified electroweak theory that described electromagnetism and the weak force [**?** ]. With this unification came new quantum numbers and conservation laws associated with the electroweak force. Among them are the conserved quantities of isospin and hypercharge[1]. These conserved quantites are actually used to redefine the already conserved and familiar quantity of electric charge

$$Q = T_3 + \frac{Y_W}{2} \tag{2.3}$$

where $Q$ is the familiar electric charge, $T_3$ is the third component of weak isospin and $Y_W$ is the hypercharge. With these newly introduced quantum numbers, the notion of chirality or handedness is now very relevant in describing the electroweak force. Chirality is defined as the orientation of a particle's spin vector with respect to its

[1]Also commonly referred to as weak isospin and weak hypercharge

linear momentum vector. Particles are said to be lefthanded when their spin vector is antialigned with their linear momentum vector, and have $T_3 = \pm 1/2$, and righthanded when the two vectors are aligned, and have $T_3 = 0$. This unified electroweak theory is a $U_Y(1) \otimes SU_L(2)$ gauge theory, where Y represents the weak hypercharge, and L means that only left-handed fermions participate in the weak interaction. The corresponding gauge bosons are the massless photon, and the massive weak force-carrying W and Z bosons. The initial shortcoming of electroweak theory was that it lacked an explanation for the broken symmetry: it didn't describe why the W and Z gauge bosons had mass in contrast to the massless gauge boson, the photon, from QED. The explanation for and mechanism behind this electroweak symmetry breaking is described in the following subsection.

The remaining gauge theory necessary for the SM is QCD. The theoretical underpinnings of QCD were developed in 1965 by Moo-Young Han, Yoichiro Nambu, and Oscar Greenberg, after significant work from Murray Gell-Mann and others describing the interactions of quarks [**?** ]. Unlike the QED, QCD is a non-abelian gauge field theory. The consequence of this is that the force-carrying gauge boson of QCD, the gluon, can interact with itself (and other gluons). This non-abelian gauge theory is denoted as $SU_c(3)$, with the conserved property of QCD being color charge. Bare color-charged particles cannot exist alone, rather they are confined to color-neutral states. Everyday examples of this include baryons (which include protons and neutrons) where each constituent quark carries a unique color charge. At very short distances and very high energies however, it is possible to observe an approximately unconfined quark, however the strong confining force increases with increasing distance between color charges until it becomes energetically favorable to generate a new color-neutral quark-antiquark pair from the vacuum. This process, known as fragmentation[2], has very important consequences in the context of experimental

[2]Also referred to as hadronization.

physics at CMS, as the hard-scatter LHC collisions often produce bare quarks which immediately hadronize into jets, or collimated sprays of energetic particles. The experimental techniques used to detect and reconstruct these jets are described in the Section 3.2.

Combining electroweak theory with QCD, we have the SM, an $U_Y(1) \otimes SU_L(2) \otimes SU_c(3)$ gauge field theory that describes all interactions and particles in Figure 2.1. The remaining unexplained portions of electroweak theory, namely the Higgs mechanism and its implications on particle masses are explained next.

### 2.1.3 Electroweak Symmetry Breaking

In electroweak $U_Y(1) \otimes SU_L(2)$ gauge theory, gauge invariance requires all of the force-carrying gauge bosons to be massless. For everything to make sense, the weak force-carrying W and Z gauge bosons would need to be massless, just like the photon. However observations indicate that the W and Z are massive, along with the fermions. So why are the W and Z massive and the photon massless? Where do fermions get their mass? Enter electroweak spontaneous symmetry breaking (EWSB). Thanks to the identification of EWSB, in 1964 three groups almost simultaneously explained the origins of particle masses using EWSB and gauge invariance in what became known as the 1964 PRL symmetry breaking papers [**?** ][**?** ][**?** ]. The explanation by which particles acquire masses is the Higgs Mechanism. For the particles in electroweak theory to have mass, they must be coupled[3] to a field. Adding the Higgs field to the electroweak Lagrangian and imposing gauge invariance gives masses to the W and Z gauge bosons as well as the fermions, but it does not interact or give mass to the photon. Giving mass to the W and Z but not the photon spontaneously breaks the $U_Y(1) \otimes SU_L(2)$ symmetry. Adding the Higgs field to the electroweak Lagrangian

[3]The phrases "coupling to" and "interacting with" have the same meaning and are used interchangeably.

must be done in such a way as to minimize the new potential. The addition of the Higgs field means that new potential is minimized at a non-zero value. This non-zero value is referred to as the vacuum expectation value (V.E.V.) and is the ground state of the SM. The fields in the SM are considered fluctuations around the V.E.V. The addition of the Higgs field and minimization of the potential to the resulting ground state is said to spontaneously break the symmetry because no external impetus for it exists. That is, there are many ways to break the symmetry and one is chosen by nature at random. A famous and more intuitive example of spontaneous symmetry breaking can be illustrated by imagining a plastic ruler held vertically between your hands with the skinny edge (not the face) facing you and then compressing your hands so that the center of the ruler bows to the right or the left and spontaneously breaks the right left symmetry of the ruler-hand system [**?** ]. This Higgs field, similar to the gauge fields, is represented by a particle called the Higgs boson. For the field to give all electroweak particles mass, it interacts with bosons and fermions. The interaction of the Higgs boson (or any boson) with fermions is known as the Yukawa interaction or coupling. Measuring the strength of the Yukawa coupling of the Higgs boson with the top quark is the primary focus of this dissertation.

### 2.1.4 The Higgs Boson

The particle manifestation of the Higgs field is the Higgs boson. After the Higgs mechanism was first introduced in 1964 to explain the origins of particle mass, the race was on to experimentally confirm the massive gauge bosons predicted by the unified electroweak gauge theory including the Higgs mechanism [**?** ]. The W and Z were discovered and their masses confirmed in 1983 by the UA1 and UA2 experiments at the Super Proton Synchrotron (SPS) at CERN [**?** ][**?** ][**?** ].

The missing piece of the puzzle was the Higgs boson. The Large Electron Positron Collider (LEP) at CERN was the next place to look for the Higgs. Beginning in 1989,

experiments at LEP searched for the Higgs at center-of-mass energies ranging from 45 GeV early on, to over 200 GeV in 2000 [**?** ]. Meanwhile, Higgs searches were also being conducted at the Tevatron collider at Fermilab. The Tevatron reached higher collision energies than LEP, with its second run lasting from 2001 to 2011, colliding protons and anti-protons at a center-of-mass energy 10x greater than that of LEP, reaching nearly 2 TeV [**?** ]. The next collider, the Large Hadron Collider (LHC) at CERN came online in 2008. The search for the Higgs came to an end when the ATLAS and CMS experiments announced the discovery of the Higgs in 2012 at the LHC [**?** ][**?** ].

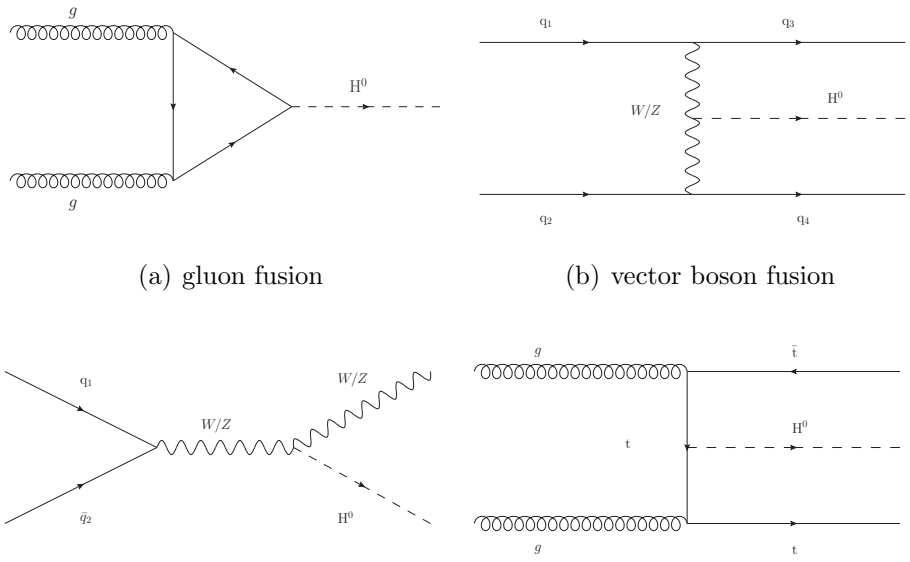
Almost 48 years after being theorized, the Higgs discovery proved the existence of the Higgs field, and validated the Higgs mechanism as well as the SM. The 2013 Nobel Prize was awarded to Peter Higgs and Francois Englert[4] for their explanation of the origins of mass[5] [**?** ].

While this Higgs discovery is important, it is more relevant now to address the production mechanisms that made it possible. The LHC collides beams of protons together; however, it is the quarks and gluons inside the protons that are actually colliding. The most common processes that produce Higgs bosons at the LHC are described in the Feynman diagrams in Figure 2.2. Of these processes, gluon fusion is the most common and also the mode targeted and used by the analyses that discovered the Higgs. Associated (Higgs) production is the mechanism that produces the Higgs processes studied in this dissertation, and occurs much less frequently compared to the other production modes. These cross sections of each process are shown in Figure 2.3.

The decay modes of the Higgs are important in the context of an experimental

[4]Robert Brout contributed equally to this work, but died in 2011.

[5]This is technically known and published as the BEH mechanism after the initial authors [**?** ] [**?** ].




(a) gluon fusion

(b) vector boson fusion

(c) Higgs-strahlung

(d) associated production

Figure 2.2. Higgs boson production modes at the LHC: gluon fusion $gg \to H$ (a), vector boson fusion $qq \to qqH$ (b), Higgs-strahlung $q\bar{q} \to W(Z)H$ (c), and associated production $gg \to t\bar{t}H$ (d).

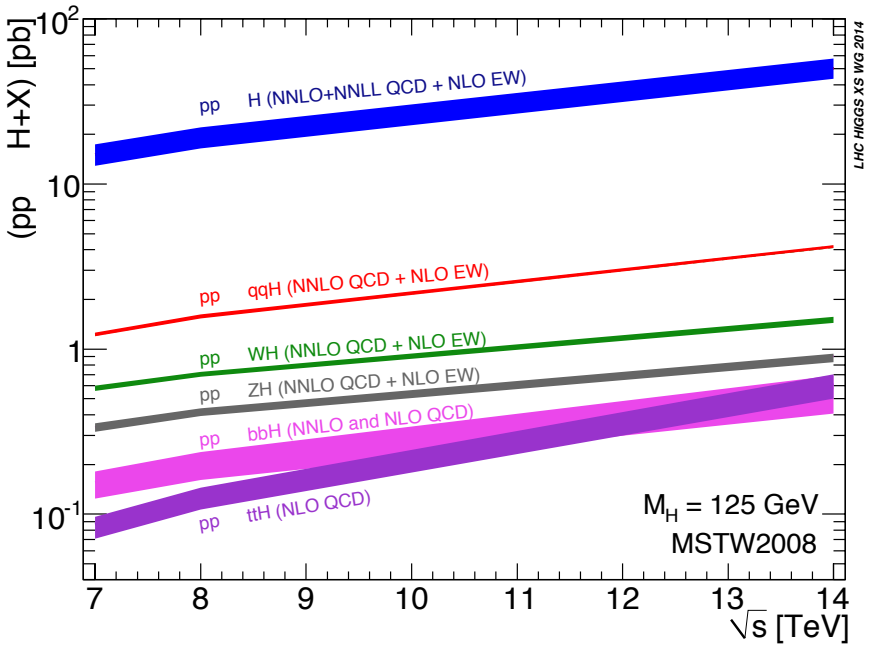



Figure 2.3. Higgs production process cross section as a function of center-of-mass LHC collision energy [**?** ]. This dissertation analyzes events produced at 13 TeV.

collider search. The Higgs decay mode and to some extent, the production mode determine which decay channels are most relevant for an experimental search. After being produced, the Higgs decays almost instantly[6] to pairs of identical SM particles. The fraction of total Higgs decays that produce a given pair of SM particles, is referred to as the branching fraction. This branching fraction value is unique to each set of final state particles the Higgs decays to. The Higgs couples more strongly to massive particles and less strongly to lighter particles. This means a decay to heavy particles is more likely than a decay to light particles. This is true with the caveat that this effect is balanced by the fact that the Higgs itself has a mass of approximately 125 GeV, and decaying to a particle pair with mass greater than the Higgs mass is strongly suppressed. Decays to heavier states, such as $H \to WW$ are allowed, but at least one W is produced off-shell, that is with a lighter mass. It is the decay to off-shell particles that suppresses the branching fraction. The greater the off-shell particle's mass differs from the pole mass, the larger the suppression. This explains some of the decay mode behavior as a function of Higgs mass in Figure 2.4.

While the SM as described here paints a picture of a complete theory, many important questions remain, such as those posed in the introduction. A thorough study and understanding of $t\bar{t}H$ processes can help address these questions.

## 2.2 ttH

### 2.2.1 Description

In a broader theoretical context, $t\bar{t}H$ searches are in essence a probe of the SM, *directly* testing the Yukawa coupling strength of the Higgs boson to top quarks. The fact that a measurement of $t\bar{t}H$ is a direct probe is an important distinction between an indirect measurement. As mentioned previously, the Higgs was discovered and

[6]The Higgs has a lifetime of $10^{-22}s$ [**?** ]

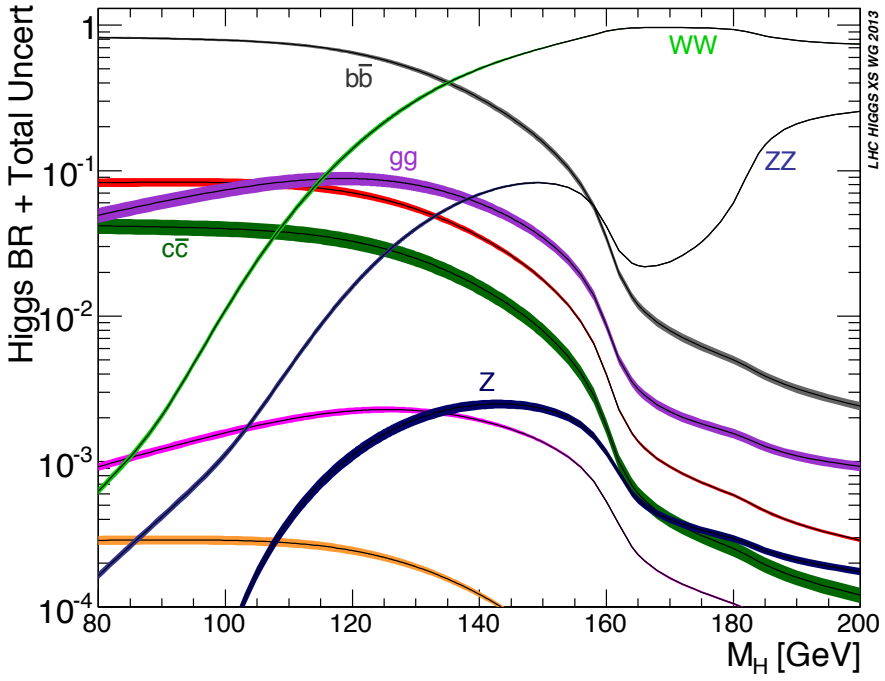



Figure 2.4. Higgs branching fractions as a function of Higgs mass. While the Higgs mass has now been measured to approximately 125 GeV, this illustrates how the branching fractions are affected by varying the Higgs mass. [**?** ].

proven consistent with the SM via the gluon fusion production mode in Figure 2.2. The gluon fusion diagram includes a fermion loop (triangle) that produces the Higgs. Because the Higgs couples more strongly to massive particles, the top quark contribution dominates in this loop. And because the Higgs was already confirmed and proven consistent with SM predictions being produced this way, the top-Higgs Yukawa coupling has already been measured, albeit *indirectly*. Beacuse we don't directly observe the fermions in this loop, it is possible that other particles beyond the SM contribute. Enter $t\bar{t}H$. While the primary goal of searching for $t\bar{t}H$ has been motivated, other important questions about the SM can be addressed simultaneously. The top quark is unique with respect to all other quarks. It is the heaviest, but curiously it is approximately 40x heavier than the next heaviest quark. The top quark mass being so much greater than any other quark inspires questions about the pattern of the quark masses. Does the top quark's mass come only from the Higgs? Or could it also come from something beyond the SM? Direct $t\bar{t}H$ searches will help answer these questions.

In an experimental context, $t\bar{t}H$ is produced by two gluons, each connecting a top-antitop quark pair, where a top and an antitop from each gluon connect, and the Higgs can be produced off of any (anti)top quark line, represented by the feynman diagram in Figure 2.5. The remaining top and antitop quark decay to a $W$ boson and b quark with nearly 100% probability[7]. The $W$ boson produced in the top decays instantly to pairs of SM particles. These pairs include a charged lepton and a neutrino of the same flavor approximately 1/3 of the time, while the rest of the time [**?** ] the $W$ decays to a quark-antiquark where the quark and antiquark have different flavors. Like the $W$, the Higgs is free to decay to numerous states according to Figure 2.4. Because of the various decay modes of the Higgs and the $W$s there are many final states possible in $t\bar{t}H$. These numerous possible final states dictate the experimental

[7]While the top can in principle decay to quark flavors other than bottoms, these decays are so heavily CKM suppressed that they are neglected [**?** ].

techniques employed in searches, with completely separate analysis efforts dedicated to a single $t\bar{t}H$ Higgs final state, or a closely related collection of Higgs final states.

### 2.2.2 Multilepton Final States

The specific signal targeted by this analysis is $t\bar{t}H$ decaying to final states with two or more charged leptons. Examples of this signal are in Figure 2.5. There are multiple Higgs decays included in this definition, specifically $WW$, $ZZ$ and $\tau\tau$. The fractions of the Higgs decays is in Figure 2.6.

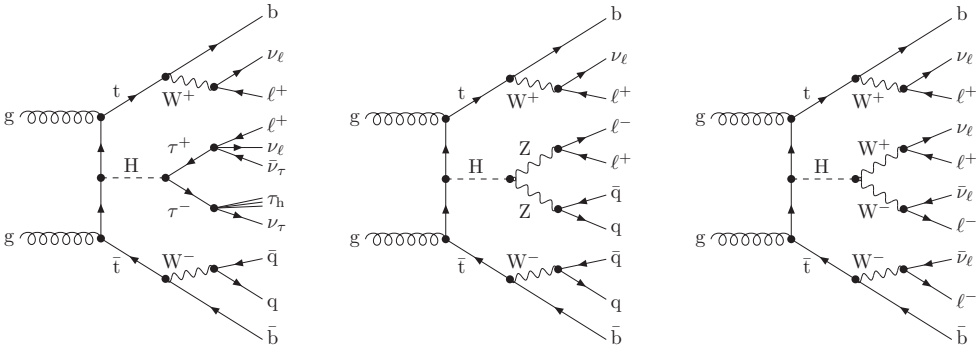



Figure 2.5. Examples of leading order Feynman diagrams for $t\bar{t}H$ production at pp colliders, with the Higgs boson decaying to $\tau\tau$, $ZZ^*$, and $WW^*$ (from left to right). The first, second, and third diagrams are examples of the two same-sign lepton signature, the three lepton signature, and the four lepton signature, respectively.

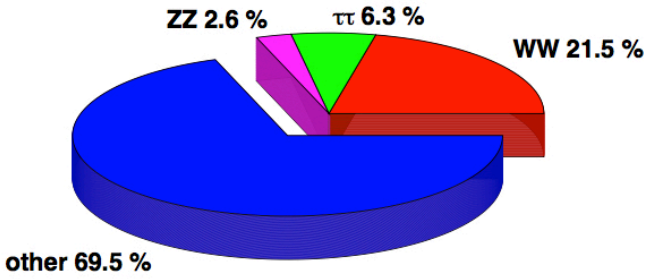



Figure 2.6. The Higgs branching fractions assuming a mass of 125 GeV. The slices removed are the decays targeted in this analysis.

# CHAPTER 3

# EXPERIMENTAL APPARATUS

## 3.1 The Large Hadron Collider

The most powerful machine of its kind, the Large Hadron Collider (LHC) accelerates and collides particles at high center-of-mass energies producing rare particles and interactions that would be otherwise unobservable in the laboratory, making it the best tool available for producing $t\bar{t}H$ processes. The LHC is a circular particle accelerator/collider. Measuring 27 km in circumference, it collides beams of protons in head-on collisions at a center-of-mass energy of 13 TeV.

Originally conceived in the early 1980s and approved in 1994, the LHC was designed to replace the then-operating Large Electron Positron Collider (LEP), re-using the same underground tunnel. Located just outside Geneva, Switzerland, at the Center for European Nuclear Research (CERN), the LHC stretches across the border into France. There are four detectors positioned around the LHC: ALICE, ATLAS, CMS, and LHCb. The two general purpose, functionally-equivalent detectors are ATLAS and CMS, while ALICE studies heavy ion collisions, and LHCb focuses on flavor physics. The motivation for having two equivalent detectors is to provide cross checks on results, as each result is produced with separate analysis teams, studying separate collisions recorded with a separate detector. Each detector is centered on an interaction point, where the beams are steered into each other to produce collisions. The LHC itself is technically the final element in a series of accelerators that bring particle beams from rest to successively higher energies. This system of of accelerators, referred to as the LHC Accelerator Complex is depicted below in Figure 3.2.

The acceleration is accomplished with radio frequency (RF) cavities. RF cavities are a linear series of cylindrical conductors, which sustain a resonant electromagnetic field produced by a generator. As charged particles pass through the cavities, they experience a force (acceleration) from the resonant alternating field in each cavity. This acceleration process begins with hydrogen gas in the linear accelerator (LINAC 2). The hydrogen atoms in the gas are stripped of electrons in an applied electric field, leaving only protons, which are then accelerated along the linear beam pipe with RF cavities to an energy of 50 MeV. After the LINAC, the beams of protons enter the Proton Synchrotron Booster rings where they are accelerated to 1.4 GeV, before reaching the Proton Synchrotron (PS). The circular PS, measuring more than 600 m in circumference, accelerates the beams to 25 GeV before injection into the larger, Super Proton Synchrotron (SPS). The SPS at over 7 km around, provides the final acceleration before the beams reach the LHC at an energy 450 GeV. The SPS injects the beams into the LHC in opposite directions to facilitate collisions later. After the beams are fully injected, the LHC ramps the beam energy to 6.5 TeV per beam, providing 13 TeV center-of-mass proton collisions [**?** ].

The beam pipes of the entire accelerator complex are kept at an ultra high vacuum to avoid detrimental beam interactions before the collisions. Since the beams are made up of protons which have electric charge, they can be focused and steered around the LHC with 392 quadrupole and 1232 dipole superconducting magnets. To accomplish this, the magnets produce a field of over 8 Tesla. This is possible thanks to the superconducting niobium-titanium (NbTi) coils which are cooled with superfluid helium-4. These magnets operate at 1.9 K allowing them to carry a current of over 11,000 amperes. In addition to the LHC, similar superconducting magnets are used throughout the accelerator complex [**?** ].

Inside the LHC, the beams travel in opposite directions in separate but adjacent pipes inside of the superconducting magnets. As a result of the RF cavity accelera-

tion, the beams are comprised of individual 'bunches' of protons. There are over 2800 bunches in each beam, with each bunch spaced 25 ns apart. This spacing is choosen to produce as many bunch crossings as possible, without overloading the detector instrumentation and data acquisition. Extra space is placed at certain intervals of bunches both for injection purposes and to allow for additional time to activate special kicker magnets that dump t/e beams. The extra space, up to 3 $\mu$s, is called the abort gap. Of the over 2800 bunches in each beam, there are approximately $10^{11}$ protons in each bunch, but due to the small cross section of the protons, only approximately 20 collisions occur in each bunch crossing. The beams travel around the LHC over 11,246 times per second, over 99.99% the speed of light. This translates to around 600 million collisions per second [**?** ].

The LHC's ability to collide protons at the energies and intensities described above make it an excellent tool for producing interesting and rare physics processes like $t\bar{t}H$. A sufficient center-of-mass collision energy is needed to produce new, heavy particles. With the current center-of-mass collision energy at 13 TeV, any particle with mass less than or equal to this, is technically within reach of the LHC. While this is technically the upper bound, the practical bound on a new particle mass is substantially smaller, since the proton collisions are actually collisions of the individual partons[1] inside the protons, and the partons carry fractions of the total proton momentum[2]. The momentum fractions carried by each of the partons, known as Parton Distribution Functions (PDFs) are determined based on inelastic scattering experiments, and theoretical calculations, since they vary with momentum transfer. Examples of the PDFs used in this analysis are in Figure 3.1.

[1]Partons refer here to any quark or gluon. The phrase was originally coined by Richard Feynman, as an individual parton makes up 'part' of a meson or baryon.

[2]The LHC is, on occasion, referred to as a gluon collider. This is due to gluons being the most likely parton to carry a non-negligible fraction of the proton momentum, thus being the most likely parton to participate in a hard-scatter collision. This is why many physics processes studied at the LHC, including $t\bar{t}H$, have incoming gluons in the initial state.

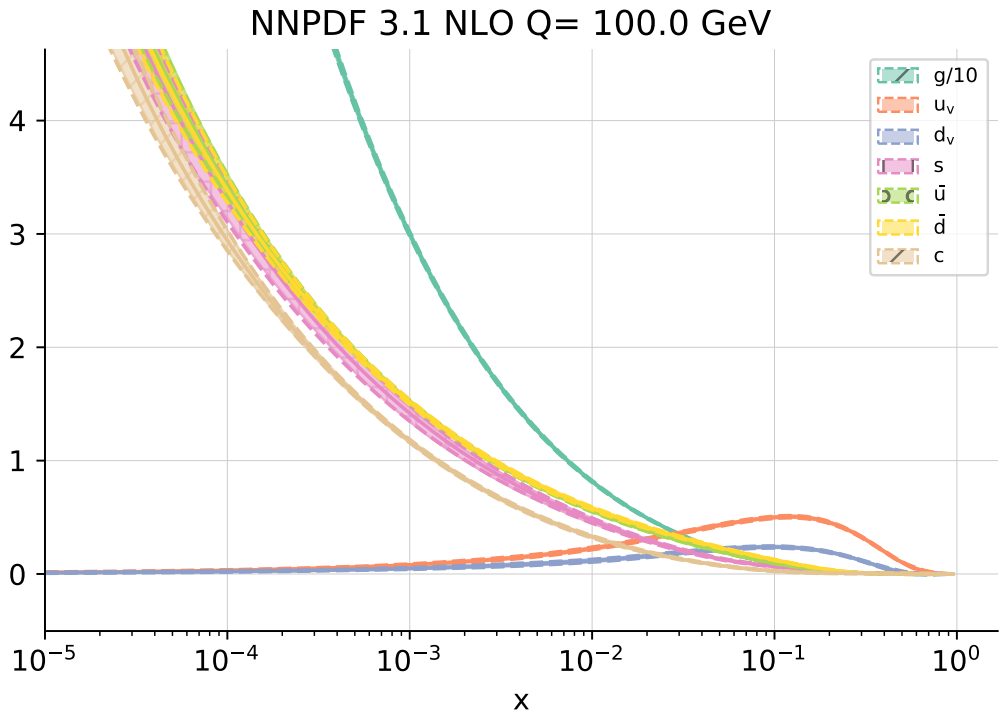



Figure 3.1. NLO parton distribution function for the proton at momentum transfer of 100 GeV. The x-axis represents the fraction of the total proton momentum carried by a given parton type while the y-axis corresponds to relative frequency[**?** ].

Because many interesting physics processes are exceedingly rare (small cross section), many collisions are needed. The quantity used to describe the rate of collisions is luminosity. Described in equation 3.1 below, instantaneous luminosity represents the number of collisions (events) occuring per unit time [**?** ].

$$\mathcal{L}_{inst} = \frac{N_b^2 n_b f_{\text{rev}} \gamma}{4\pi \epsilon_n \beta^*} F \tag{3.1}$$

Where $N_b$ is the number of protons per bunch, $n_b$ is the total number of bunches, $f_{\text{rev}}$ is the number of beam revolutions around the LHC per second, $\gamma$ is a relativistic factor, $\epsilon_n$ is factor related to beam emittance, $\beta^*$ is the beta function, which is related to the transverse beam size at the interaction point, and $F$ describes the factor related to the crossing angle of the beams. Because not every collision can be recorded, knowing the total number of collisions is critically important to every analysis at an LHC collider experiment. The integrated luminosity is the time integral of the instantaneous luminosity. Combining this with the proton-proton inelastic cross section, $\sigma_{pp}$, the number of collisions occurring in time interval $t$ is calculated in equation 3.2 below.

$$N_{pp} = \int \sigma_{pp} \mathcal{L}_{inst} dt \tag{3.2}$$

## 3.2 The Compact Muon Solenoid (CMS) Detector

100 m beneath the town of Cessy, France, sits the Compact Muon Solenoid (CMS) detector, a multi-purpose particle detector designed to record and identify particles from collisions produced by the LHC. CMS is comprised of layers of subdetectors which are connected to readout electronics and the trigger and data acquisition systems. The various subdetectors record information about the collision, while the trigger and data acquisition systems record and save the collisions. The acronym

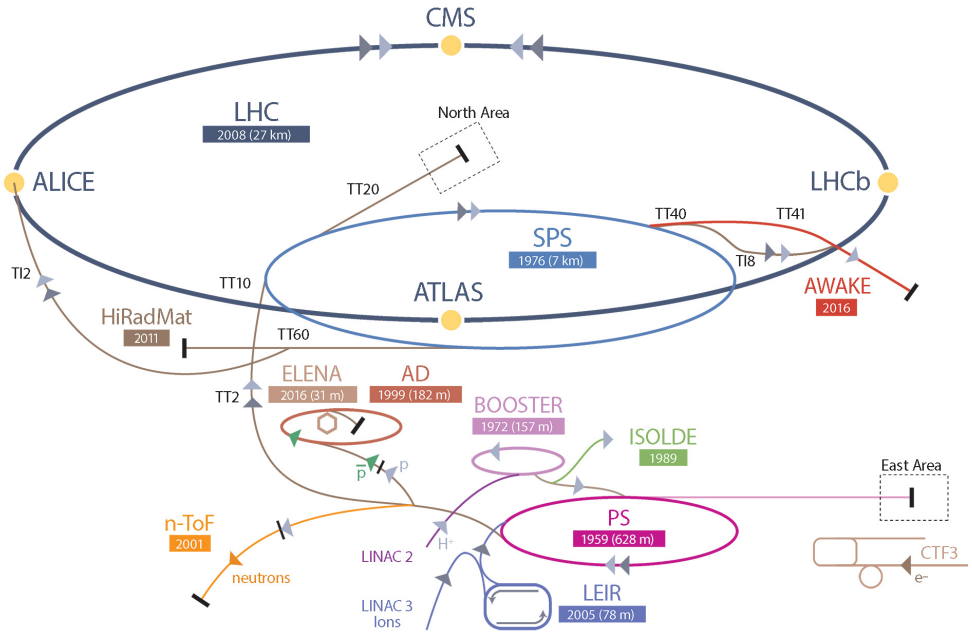



Figure 3.2. An overview of the LHC accelerator complex [**?** ].

CMS comes from being more compact than its sister detector ATLAS, at 15 m in diameter and over 21 m long. The muon in the acronym comes from the fact that detecting muons is among the most important tasks of the experiment. The solenoid part of the name is due the powerful solenoid magnet which is crucial for particle tracking and identification. CMS is cylindrical in shape, as pictured in Figure 3.3. The various subdetectors and components are described in more detail below.

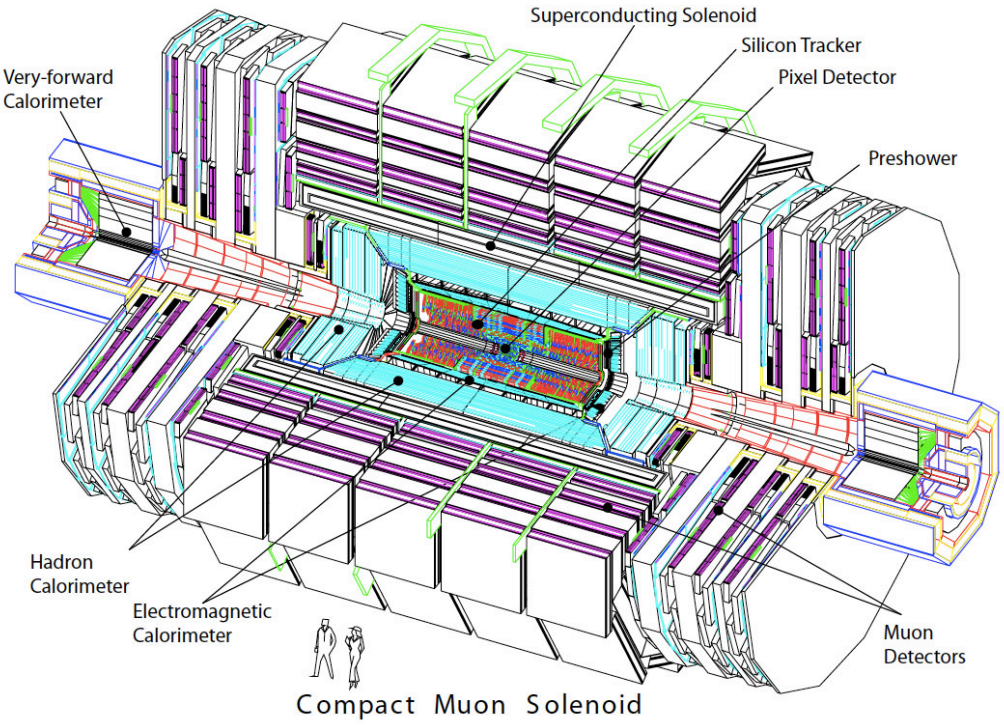



Figure 3.3. A qualitative overview of the CMS detector [**?** ].

### 3.2.1 Coordinate System

Due to the geometry and symmetry of the collisions produced inside CMS, a special coordinate system is used for simplicity. With the origin at the interaction point, the y-axis in the vertical direction, the z-axis parallel to/along the beam direction, and the x-axis in the horizontal direction perpendicular to both the y and z axes. In the transverse x-y plane, the azimuthal angle formed with respect to the x-axis is $\phi$. In the y-z plane, the angle formed with respect to the z-axis is $\theta$. The number of particles traveling through a given area, known as flux, increases towards the beam line, due to many more glancing collisons occurring than head-on collisions. The particles from the collisions are moving relativistically. Because of this, a special coordinate $\eta = -\ln\tan(\theta/2)$, is used to describe the angle with respect to the z-axis, known as *pseudorapidity*. Finally, the transverse plane is of special importance for momentum and energy measurements due to energy/momentum conservation, since the head-on inelastic pp collisions have only momentum in the z-direction.

### 3.2.2 Tracker

The CMS tracking system sits closest to the interaction point at the core of CMS and provides precise tracking information from its millions of silicon sensors. The tracking system is comprised of two concentric cylindrical subdetectors, the pixel and strip detectors, which are named after the geometry of the sensors they use. The pixels provide the highest granularity tracking information, essential for distinguishing between collision vertices, and sit nearest to the interaction point, inside of the strip detector. The strips provide slightly less resolution compared with the pixels, since they sit further from the interaction point. There is approximately one radiation length[3] of tracker material (pixels + strips) between the interaction

[3]A radiation length is defined as the distance an energetic electron travels in material before losing all but 1/e of its initial energy due to bremsstrahlung emission [**?** ].

point and the beginning of the next subdetector, the ECAL. The total detector area of the tracking system (pixels+strips) sums to more than a tennis court of silicon.

As a charged particle travels through a silicon sensor (pixel or strip), it ionizes a small fraction of the silicon atoms, releasing electrons from the valence band of n-type or p-type semiconductor, creating electron-hole pairs. A voltage is applied to the sensor, attracting the now free electrons, creating a small current. This current is measured and interpreted as a hit. A small silicon pn junction is attached to the back of each sensor to read the voltage and send the hit information to the data acquisition system, so the hit patterns can be reconstructed into particle tracks later. As the particles travel through successive layers of silicon sensors, they leave collections of hits. The hits are measured with a precision of 10 $\mu$m, and the particle track is reconstructed from this collections of hits. Because the information from the tracker has such a high granularity, track reconstruction algorithms need more time than is feasable for trigger decisions and are mostly performed offline.

#### 3.2.2.1 Pixel Detector

The pixel detector is comprised of 65 million silicon pixels, allowing it to record precise trajectories of the charged particles resulting from collisions. Precise tracking information is critical to differentiating nearby particles, and identifying exactly where an interesting collision took place.

The pixel detector is about the size of a shoe box and contains 3 layers at 4 cm, 7 cm, and 11 cm from the beam. Silicon is used in the tracker over other materials, as it strikes the best balance between performance, radiation resistance, and affordability. Radiation resistance, also known as radiation hardness, is very important for the tracker as it sits only a few centimeters from the interaction point receiving around 10 million particles per square centimeter per second. Each silicon pixel measures 100 $\mu$m by 150 $\mu$m and is readout by an electronic chip bump-bonded to the back of

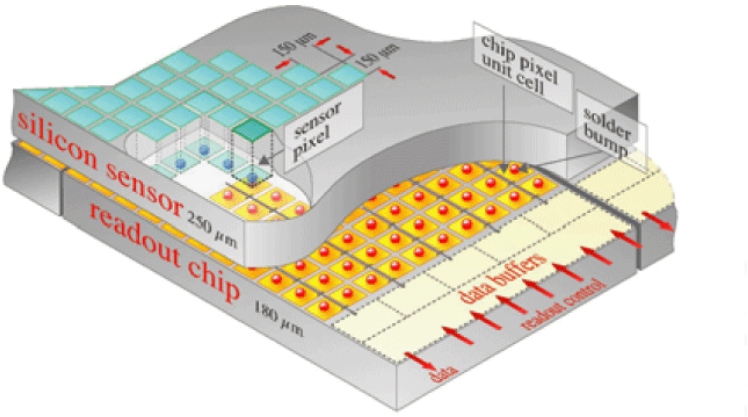



Figure 3.4. An overview of a CMS pixel element [**?** ].

each pixel. The layout of a pixel sensor is depicted in Figure 3.4.

#### 3.2.2.2 Strip Detector

Outside the pixel detector is the silicon strip detector. Here, silicon strips are favored over pixels as they are less costly, and the resolution provided by the pixels is not necessary at greater distances from the interaction point, allowing for larger and fewer silicon modules. The silicon strip detector consists of about 10 million strips in 10 layers, extending to 130 cm from the beam. The strip detector is comprised of 4 distinct sections of silicon strips, depicted in Figure 3.5. The outer most sections are the tracker outer barrel (TOB) and tracker end cap (TEC+,-) sections. Between the TOB and the pixels, sits the tracker inner barrel (TIB) section, and the tracker inner endcaps (TID+,-). The silicon strips in each section are different, specifically suited to that section. Each silicon strip measures between 10-25 cm by 180 $\mu$m, depending on the position. Precise tracking information is essential to measuring particle momenta and identifying particles in the strong magnetic field.

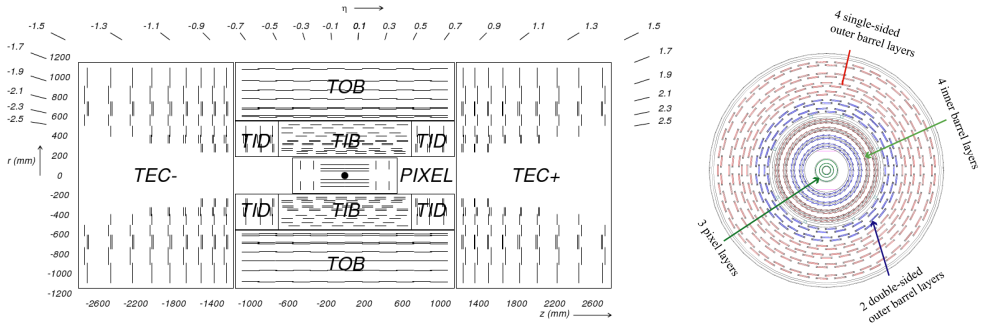



Figure 3.5. The CMS silicon tracking system including both the pixel and strip detectors in y-z plane (left), and transverse x-y plane (right) [**?** ].

### 3.2.3 ECAL

The Electromagnetic Calorimeter (ECAL) measures the energy of charged particles[4]. The ECAL is situated immediately outside the tracker and is made up of over 75,000 of lead tungstate ($PbWO_4$) crystals in three distinct sections. The ECAL provides good energy resolution, fast readouts, and radiation hardness which make it ideal for recording the frequent collisions produced by the LHC.

After particles pass through the silicon tracking system, they enter the ECAL. The ECAL measures the energy of electromagnetically-interacting particles from collisions, namely electrons and photons. The choice of the lead tungstate material is motivated by needing to stop the particles, and also scintillate light effectively to allow an accurate energy measurement. The stopping action is accomplished with the lead atoms in the crystal, while the scintillation is accomplished with the crystalline oxygen. These combined properties produce photons (light) in proportion to the amount of energy deposited by the stopped particles.

The ECAL is made up of three sections; the barrel, endcaps, and preshower, depicted in Figure 3.6. The barrel section is cylindrical and covers the full $\phi$ range, and extends between $|\eta| < 1.479$ in y-z. The barrel consists of 61200 crystals, each measuring approximately 2.2 cm x 2.2cm x 23 cm. The length of each barrel crystal translates to approximately 26 radiation lengths. The two endcaps sit on each end of the barrel, extending between $1.479 < |\eta| < 3.0$, 315 cm on either side of the interaction point. There 14648 identical crystals in the endcaps, each measuring 3 cm x 3 cm x 22 cm. Like the crystals in the barrel, these crystals have a small taper with the smaller face pointing towards the interaction point, mimicking the conical shapes of electromagnetic particle showers in the $PbWO_4$ material. The preshower disc sits on the ends of the barrel and in front of the endcaps. The preshower measures

[4]The ECAL gets its name because it detects charged particles, which interact via the electromagnetic force

2.5 m in circumference and is 20 cm thick. The ECAL preshower consists of two layers of lead, followed by silicon sensors measuring 2mm x 2mm. The preshower allows the ECAL to resolve nearby photon pairs, and discriminates against those coming from in-flight decays of neutral pions. The scintillation light from each crystal is detected by Avalanche Photo Diodes (APDs) in the barrel, and Vacuum Photo Triodes (VPTs) in the endcaps. These readouts convert the scintillation light to a voltage pulse that is passed further downstream to the trigger and DAQ.

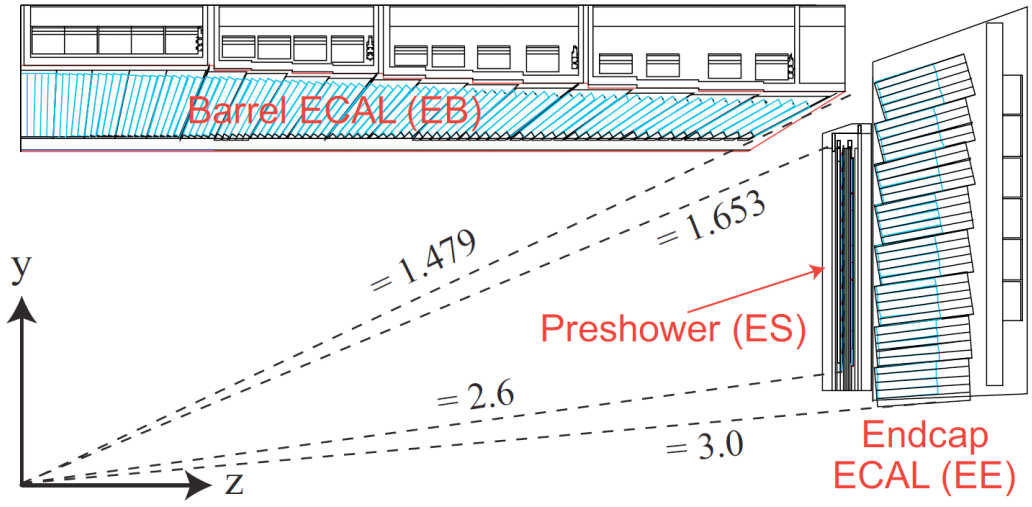



Figure 3.6. Longitudinal view of one quarter of the ECAL [**?** ].

### 3.2.4 HCAL

The CMS Hadronic Calorimeter (HCAL) measures the energy of hadrons via their strong force interaction with the detector material. The HCAL is situated outside

the ECAL. The HCAL is a sampling calorimeter made up of alternating layers of brass absorber and plastic scintillator material in four separate sections. The HCAL provides energy resolution that helps reconstruct and tag hadronic particle decays, as well as reconstructing the *missing transverse energy* or $E_T^{miss}$, which can be interpreted as the presence of neutrinos[5]. Within each plastic scintillator tile are optical wavelength-shifting fibers. Measuring less than 1 mm in diameter, these fibers connect to other optical fibers which send the light signals to Hybrid Photodiodes (HPDs) for readout. The fast and radiation-resistant HPDs amplify the light and convert it into an electronic signal via the photoelectric effect.

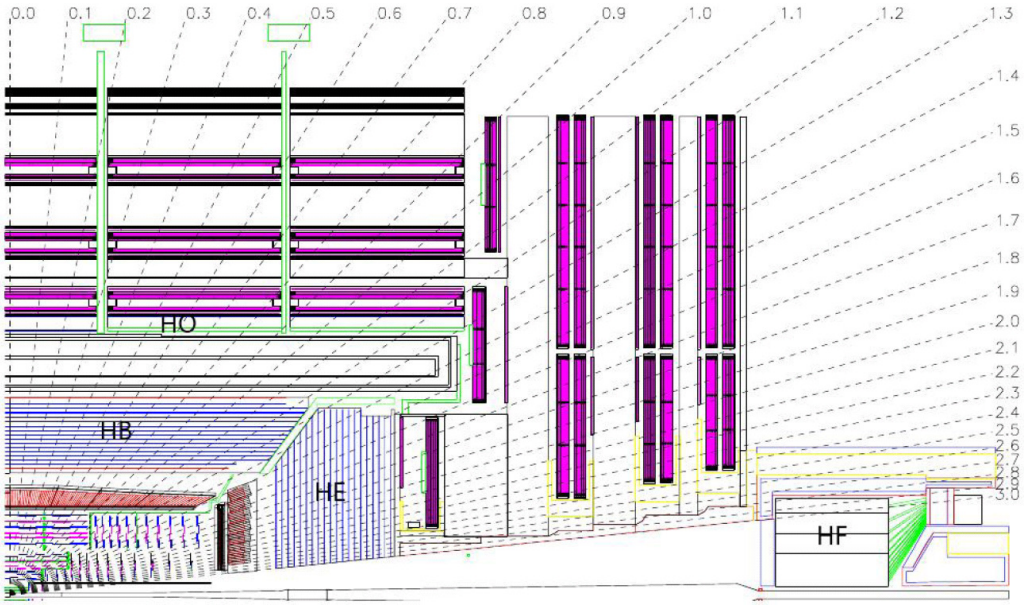



Figure 3.7. Longitudinal view of one quarter of the HCAL [**?** ].

[5]The $E_T^{miss}$ reconstruction actually occurs later, but the HCAL information provides some of the necessary quantities

As strongly-interacting particles travel through the HCAL, they interact with the dense brass absorber material decaying and producing showers of secondary particles which further cascade and travel through the scintillator and absorber material. As the shower particles pass through the scintillator material, they emit a blue-violet light. This blue-violet light is then converted to green light via the wavelength shifting fibers and sent to the HPDs for readout. Like the ECAL, the amount of light emitted corresponds to the amount of energy deposited.

The four sections of the HCAL are the barrel (HB), the endcaps (HE), the outer barrel (HO) and the forward HCAL (HF) in Figure 3.7. The barrel extends from 1.77 m to 2.95 m from the interaction point and sits between the outside of the ECAL barrel and the inside of the magnet coil in the transverse plane. The HB covers a pseudorapidity range of $|\eta| < 1.3$. The endcaps cover a pseudorapidity of $1.3 < |\eta| < 3$. The HO sits outside and around the solenoid coil covering the same pseudorapidity range as the HB. The HO utilizes the magnet coil as an additional absorbing layer ensuring detection of any particle exiting the HB. The HF sits outside of all other subdetectors at $\pm 11$ m from the interaction point in the z-direction, covering a pseudorapidity range of $3.0 < |\eta| < 5.0$. The HF was specially designed to resist the intense radiation deposited in this forward region. In total, the endcap calorimeters cover roughly 10 radiation lengths.

### 3.2.5 Solenoid

Central to the design and name of CMS, the CMS magnet is one of the largest superconducting solenoids in the world. At 12 m in length, it produces a field of nearly 3.8 T inside the 6 m diameter free bore, with the flux returned through a 10000 ton iron yoke. The magnetic field outside the free bore in the muon chambers is approximately 2 T. Similar to the LHC magnets, the solenoid uses NbTi coils cooled to 1.8 K but carries a current of over 19000 amps. When fully energized, the

magnet stores 2.6 GJ of energy, enough to power 24 American homes for 1 day[**?** ]. The central barrel of the solenoid sits between the HCAL and the muon chambers, with the return yoke interwoven with the muon chambers.

A strong magnetic field is essential for measuring the momentum of charged particles. Charged particles moving relativistically in a magnetic field are subject to the Lorentz force, described in equation 3.3

$$\vec{F} = \gamma q \vec{v} \times \vec{B} \tag{3.3}$$

The solenoid produces a magnetic field along the z-direction and by the right-hand-rule from the cross product in equation 3.3, this means the particles experience a centripetal force, which curves or deflects their trajectories. By setting the Lorentz force equal to the relativistic centripetal force in equation 3.4, the momentum and charge of the particle can be found by measuring the radius of the track. Thus the field produced by the CMS solenoid together with accurate track reconstruction, allows for precise momentum measurements of the particles.

$$\gamma q v B = \frac{m \gamma^2 v^2}{r} \tag{3.4}$$

### 3.2.6 Muon Chambers

A central feature of CMS, the muon chambers are solely dedicated to detecting and measuring muons. Muons are not stopped by the ECAL due to their large mass, and travel through the HCAL and solenoid yoke mostly unimpeded since they don't interact via the strong force. The muon chambers are the outermost subdetector for this reason. The muon chambers are comprised of 3 unique gas detection systems consisting of Drift Tubes (DTs), Cathode Strip Chambers (CSCs), and Resistive Plate Chambers (RPCs), which each serve a specific purpose. Each of the gas chambers

produce voltage pulses as muons pass through and leave hits. Similar to the tracker, the hits are reconstructed into muon tracks, which can then be used to determine the momentum of the muon in the magnetic field.

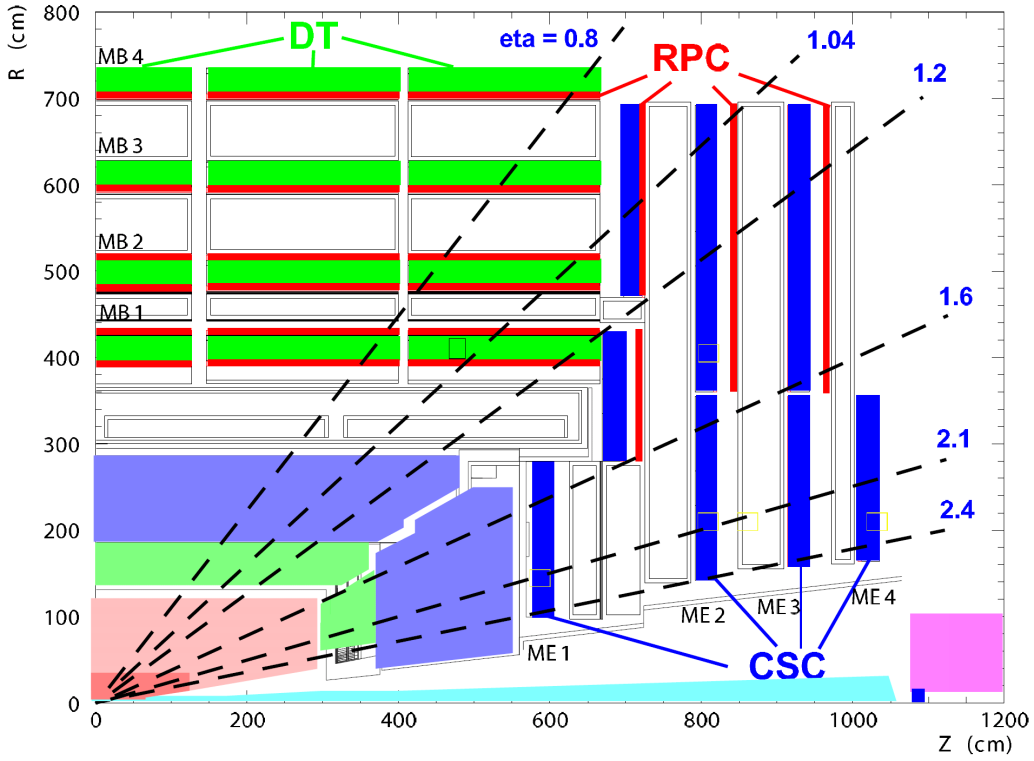



Figure 3.8. Longitudinal view of one quarter of the CMS muon chambers [**?** ].

The DTs cover the barrel region of the detector. There are three concentric cylindrical layers each containing 60 drift tube chambers and one additional outer layer with 70. Each drift tube is constructed from light weight aluminum and measures 4 cm wide by 2.4 m long, filled with a mixture of 85% Ar and 15% $CO_2$ gas. Cathode

strips held at -1.2 kV run the length of each cell with a gold-plated steel anode wire held at 3.6 kV runs down the center (see Figure 3.9).

As an incident muon ionizes the gas atoms, electrons are attracted to the anode wire while positively charged ions are attracted to the cathodes. The ions induce a current fluctuation in the electrodes which is readout and interpreted as a hit. By knowing the drift velocities of ions in the gas and measuring the timing of the current pulses precisely, the position of the muon from the anode wire is inferred. By alternating the orientation of these chambers between wires parallel to the beamline and wires perpendicular to the beamline, the DTs can resolve muon positions to 100 $\mu$m. The DTs are used in the barrel only, where there are lower particle fluxes and intensities.

The CSCs are used only in the endcaps, where there is a greater particle intensity and because the large, non-uniform magnetic field would affect the ion drift of DTs - making them unsuitable in this region. 468 CSCs in 7 layers are in the endcaps, where each chamber is trapezoidal in shape. Each chamber consists of positively charged anode wires which cross negatively charged cathode strips that are enclosed in a chamber filled with gas. As muons enter a CSC, they ionize the gas causing freed electrons to travel to the positive wires and positive charges migrate towards the cathode strips. The charge pulses are detected in both the strips and the wires, which are perpendicular to each other, providing two coordinates for each muon hit. The CSCs provide good position information quickly, making them suitable for use in the endcaps.

The third type of muon detector used is the RPC. The RPCs are used in both the barrel and the endcaps covering out to $|\eta| < 1.6$, providing redundant coverage. The RPCs are characterized by their fast response time which makes them suitable for triggering, unlike the DTs and CSCs, which are limited by their relatively longer drift times. Due to this very fast response time ( 1 ns), the RPCs easily identify which

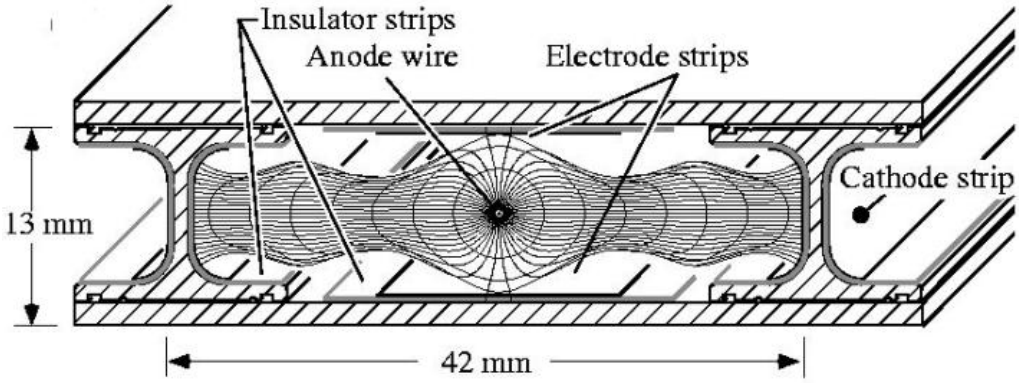



Figure 3.9. A map of the electric field of the drift tubes in absence of magnetic field [**?** ].

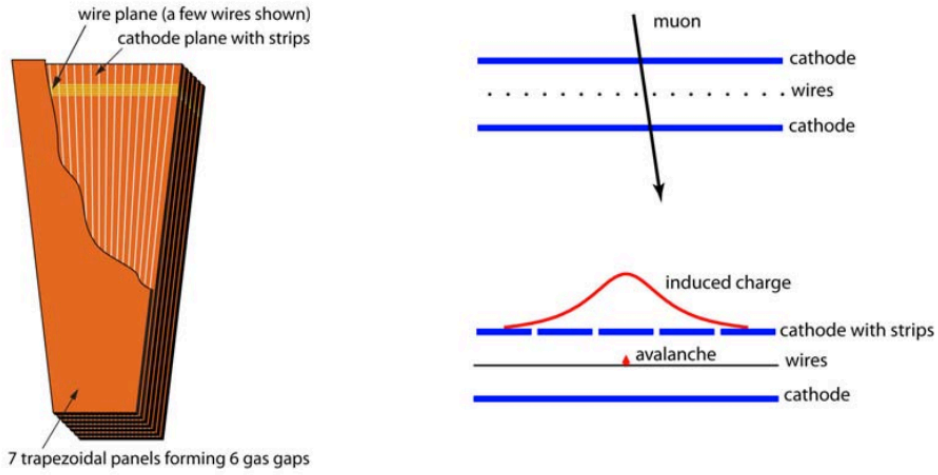



Figure 3.10. CSC module (left) and a depiction of the CSC operation (right). The strips provide precise position information [**?** ].

bunch crossing a detected muon is associated with, since the bunch time spacing is 25ns. The RPCs are gaseous parallel plate detectors. Each RPC measures between 2.1-2.5 m in length and 1.5-2.5 m wide, with a thickness of 2mm filled with gas separating two resistive plastic parallel plates, an anode and cathode. As an incident muon ionizes the gas atoms releasing electrons, the electrons in turn ionize nearby gas atoms resulting in an avalanche of electrons which drift towards the anode and are picked up by metallic detecting strips. The hit pattern on the strips is used to determine the muon momentum for fast trigger decisions.

### 3.2.7 Trigger and Data Acquisition

The LHC collides proton bunches at a frequency of almost 35 MHz, making efficient data collection and filtering necessary. This task is accomplished with the trigger and Data Acquisition system (DAQ). Because not all collisions can be recorded, and most collisons don't produce events that warrant further analysis, the trigger "fires"

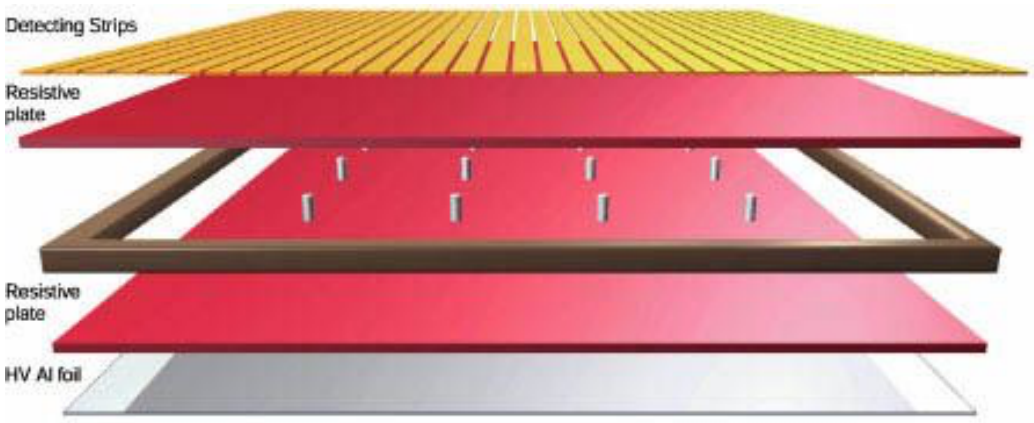



Figure 3.11. A qualitative depiction of an RPC [**?** ].

on interesting events only, and sends the information readout in each subdetector to the DAQ. The CMS trigger is comprised of two parts or levels. The Level 1 (L1) is the frontline of the CMS trigger system, comprised of firmware and basic software that is directly connected to each subdetector's readout electronics. The second part of the trigger is called the High Level Trigger (HLT) and is entirely software-based, running a filter farm of high-performance CPUs.

The L1 is the first place where an event can be discarded or selected based on predefined conditions. From an event rate of 40 MHz produced by the LHC, the L1 filters and reduces this to between 80-100 kHz. Each subdetector readout, with the exception of the silicon trackers, is directly connected to the L1. Since the trackers have very high occupancies, and millions of pixels and strips, sophisticated track reconstruction takes much longer than is acceptable for a L1 trigger decision. Therefore when the L1 fires based on other subdetector information, the tracker is immediately readout and the information is saved but not reconstructed. The software portion of the L1 consists of 128 algorithms called triggers or bits, that are constantly processing detector readout information. Any one trigger accepting an event, will pass all the detector information downstream to the HLT.

The HLT receives events that pass the L1, and is the final filter that determines which events are saved or discarded. This means that all events passing the HLT are saved and reconstructed for later analysis. The HLT reduces the L1 input rate from 100 kHz to around 1 kHz. The HLT software is the foundation on which the rest of the offline CMS analysis software, CMSSW, is built. The HLT software consists of over 400 trigger algorithms that make accept/reject decisions based on quantities such as particle momenta, multiplicity, energy, position, and other more sophisticated variables that are available thanks to the advanced event reconstruction that takes place at the HLT. The more than 400 HLT algorithms (triggers) are gouped into categories based on the specific types of objects that fire the trigger such as electron/photons, muons, Jets/MET etc. These groups are called primary datasets and only events which passed a trigger allocated to that primary dataset will be found in the associated dataset. Common examples of primary datasets include SingleEG, DoubleEG, SingleMu, DoubleMu, MuonEG etc. The events are then sent further downstream grouped together in these primary datasets. The collection of 400 algorithms comprise the HLT menu. Because the HLT, DAQ and downstream hardware can only handle a maximum trigger rate, the trigger menu uses prescales to control the rates of triggers that would otherwise fire too often. This allows lower-priority triggers to collect data without consuming large rates. The prescale controls the frequency of trigger accepts. For example a trigger with a prescale of 2 means that data is recorded every other time the trigger fires. The HLT and L1 algorithms are thoroughly tested and highly efficient at passing the objects they are designed to accept. Typical HLT trigger efficiencies are well above 90%. See the muon efficiency at the HLT in Figure 3.12 below.

The CMS DAQ handles everything from the detector readout to interfacing the various parts of the L1 and HLT, as well as running the 1800+ CPU filter farm that the HLT software runs on. The DAQ also handles all data recorded by CMS,

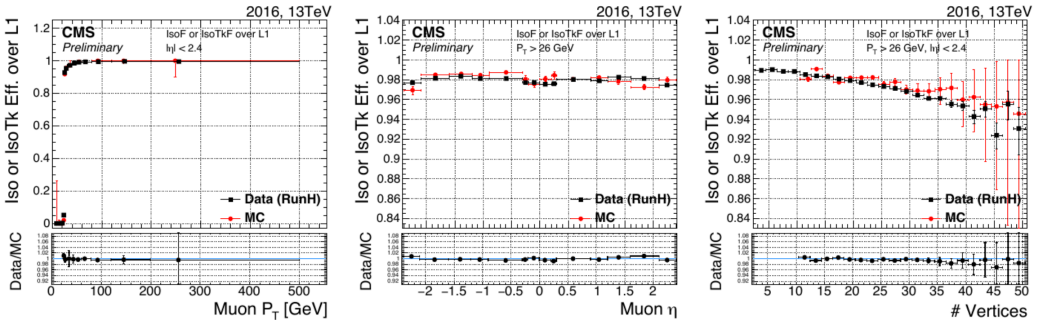



Figure 3.12. Muon trigger efficiency at the HLT as a function of muon transverse momenta, pseudorapidity, and number of reconstructed vertices.

sending it from the HLT to an offsite computing facility for additonal reconstruction and distribution for analysis. The DAQ compiles information from the L1 and all subdetector readouts and synchronizes and combines this information together to build a complete picture of an event before it even reaches the HLT.

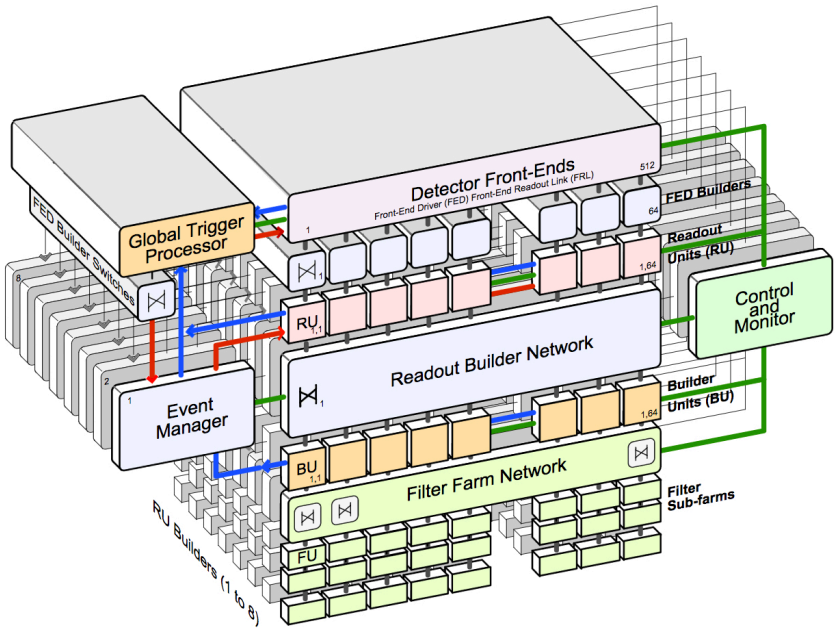



Figure 3.13. A schematic of the DAQ system. After passing the L1, the events are built by combining all subdetector information into one coherent picture for the HLT (labeled Filter Farm Network) to make more sophisticated accept decisions [**?** ].

# CHAPTER 4

# PHYSICS OBJECTS

Each of the CMS subdetectors (neglecting the trigger system) technically only record and detect hits and energy deposits. While these hits and energy deposits are almost always due to passing particles, the detectors and readouts themselves only produce information about the position, multiplicity, and value of these hits and energy deposits respectively. Sophisticated algorithms are used to reconstruct these hits and deposits into particles. These reconstructed particles are referred to as physics *objects*. The reconstruction techniques vary greatly with different objects and suit the subdetectors used to detect and record their hits and energy deposits.

## 4.1 Object Reconstruction and Particle Flow

The particle flow algorithm is used by CMS to reconstruct physics objects from hits and energy deposits. Particle flow and CMS are unique in the sense that nearly all physics analyses performed on data collected by CMS use objects reconstructed with this approach. Particle flow coordinates the many reconstruction algorithms across all CMS subdetectors into a single global picture of each event. The primary advantage of this strategy is uniform and consistent object definitions across nearly all papers published on behalf of CMS. The purpose of particle flow is to identify all final-state stable particles in an event recorded by CMS, specifically electrons, muons, taus, jets, and photons. Particle flow optimally combines building-block information (hits and energy clusters) from all subdetectors to reconstruct objects and determine particle type, position, and momentum. To accomplish this, two different primary

reconstruction techniques are used. One for reconstructing tracks from tracker hits, and one for clustering energy deposits from individual calorimeter cells.

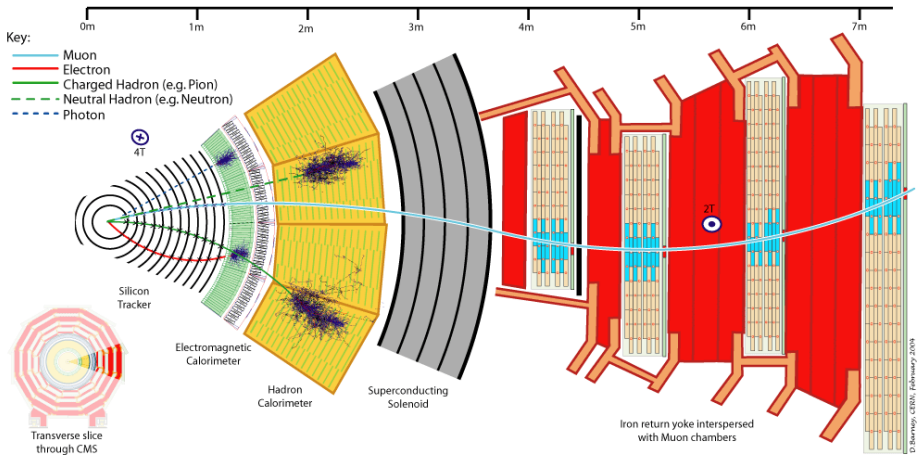



Figure 4.1. An overview of how CMS detects different types of particles. The slice of CMS in in the x-y plane. [**?** ].

Object reconstruction begins with grouping collections of hits into tracks in an iterative process [**?** ]. In the first iteration, tracks are seeded with initial hits and subject to very tight criteria, sacrificing efficiency for a low fake rate. In the following iterations, hits assigned to tracks in the previous iteration are removed from further consideration, and the criteria for candidate tracks is gradually relaxed with each iteration. In the final iterations, the constraints on the track seed are relaxed to account for secondary decays from photon conversions and nuclear interactions with the silicon tracker material. This technique reconstructs tracks with as few as three hits and $p_T$ as small as 150 MeV with a fake rate in the single digits [**?** ]. A similar but separate track reconstruction is performed in the muon chambers, however the

minimum muon $p_T$ is around 4 GeV.

Object reconstruction continues in the calorimeters, where it relies on a clustering algorithm to identify individual energy deposits and associate them to an object. This algorithm is designed to yield a high efficiency even for low energy, and nearby objects. The clustering process is performed separately in the ECAL and HCAL, and furthermore in the barrel, and endcaps. The ECAL preshower also uses separate clusterings in each of its two layers. No clustering is performed in the HF, where each module is an independent cluster. The clustering algorithm begins by identifying cells (cluster seeds) with an energy above a "seed" energy threshold. Clusters are then increased by adding adjacent[1] cells that have an energy above a given threshold. The calorimeter granularity is used to optimize the determination of cluster energies and positions [**?** ].

After the tracking and clustering is complete, Particle Flow then matches tracks in the tracker to energy clusters in the calorimeters, and to tracks in the muon chambers. In a given event, there can be many different objects and Particle Flow utilizes a process-of-elimination strategy. Because they are the easiest to identify unambiguously, muons are identified first by matching tracks in the inner tracker to the tracks in the muon chambers. The matching criteria is based on a $\chi^2$ fit threshold. When multiple sets of tracks are matched, the set with the lowest $\chi^2$ is selected as the muon object. Muons are first reconstructed this way and called global muons [**?** ]. Particle flow muons are identifed from global muons when the $p_T$ measurements in the tracker and the muon chambers agree to within three standard deviations. The tracks corresponding to the muon object are then removed from consideration for the remaining object reconstruction.

Electrons are reconstructed next. Electrons emit characteristic Bremsstrahlung radiation due to their small mass as they are deflected in the magnetic field. This

[1]sharing at least one edge with the seed

Bremsstrahlung radiation is emitted tangentially to the deflected electron's track, and is subsequenctly detected in the ECAL. Tracks are flagged based on these characteristics as potential electron candidates and refitted with a Gaussian-Sum Filter (GSF) [**?** ] and the resulting tracks are then matched to ECAL energy clusters. The ECAL clustering and track matching accounts for the Bremsstrahlung radiation in electron object reconstruction. These matches are then subject to additional quality criteria before they are considered particle flow electrons, and their corresponding tracks and energy deposits removed from further consideration for the remaining object reconstruction.

At this point in the event/object reconstruction, the so-called "low hanging fruit" has been picked, and the more difficult objects are all that remain. These objects are charged and neutral hadrons (jets), and photons. The neutral particles are difficult to reconstruct because they don't leave hits in the tracker, so only calorimeter information is available. Photons interact electromagnetically and are stopped by the ECAL, while neutral hadrons are stopped and deposit their energy in the HCAL. The remaining tracks are subject to quality cuts aimed at reducing the fake-rate. The high-quality tracks passing these thresholds are matched to ECAL and HCAL deposits, and give rise to particle flow charge hadrons. The momentum of these objects is measured from the track radius and compared to the corresponding energy deposit in the calorimeters assuming the object is a charged pion. If the two measurements are compatible, the momenta is refined with additional fits to the tracks and energy deposit(s) [**?** ].

The charged and neutral hadron objects with tracks and matched energy deposits are only considered particle flow candidates at this stage. An additional clustering step is necessary to reconstruct jet objects. As mentioned previously, when free quarks produced in the pp collision decay in the fragmentation/hadronization process, they produce energetic and collimated sprays of charged and neutral hadrons, called

jets. To accurately determine the direction and momentum of the initial quark, as many of the corresponding charged and neutral hadrons as possible must not only be reconstructed correctly and identified, but also clustered together (matched) correctly. These jets are then interpreted experimentally to be the free quark. A depiction of the hadronization and jet detection process is below in Figure 4.2.

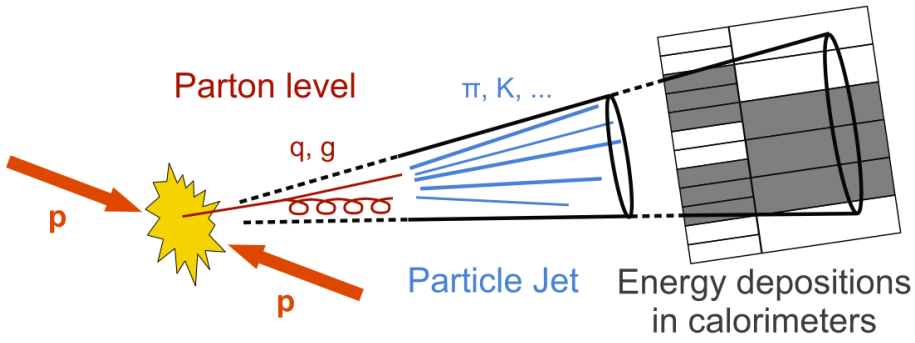



Figure 4.2. An example of quark hadronization and the resulting jet. [**?** ].

Because the hadronization process is challenging to reconstruct, jet clustering is accomplished with specialized algorithms. These algorithms exploit information from the detector with theoretical knowledge of the hadronization process to cluster the jets in a sensible and reproducible manner. While many clustering algorithms exist, CMS uses the anti-$k_T$ algorithm [**?** ] for jet clustering. Like other sequential clustering algorithms, anti-$k_T$ begins by finding the highest $p_T$ candidate or seed, and

calculating the distance measures in equations 4.1.

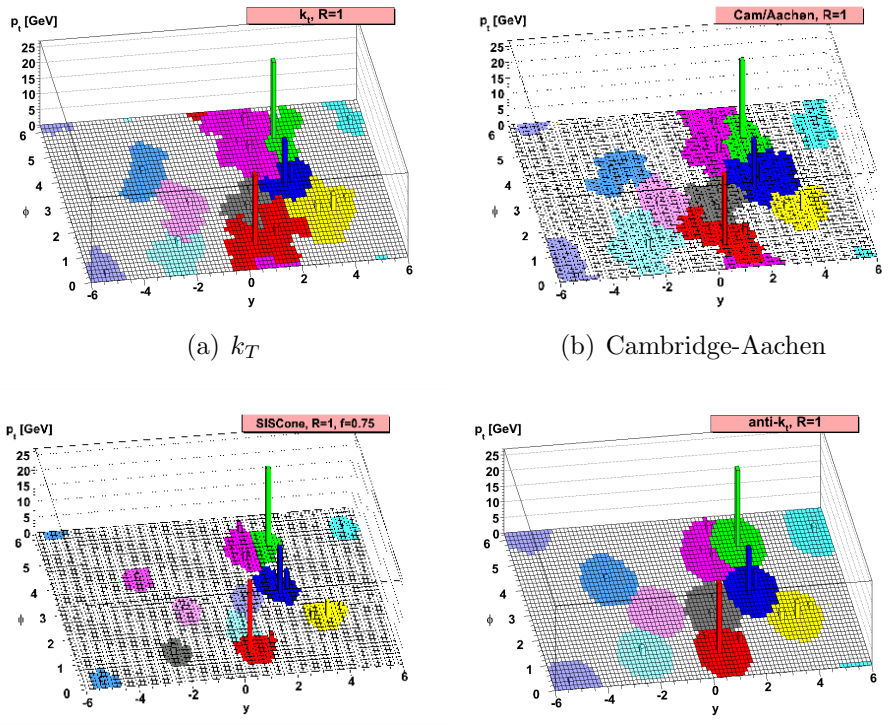
$$d_{ij} = min(k_{Ti}^{2p}, k_{Tj}^{2p})\frac{\Delta_{ij}^2}{R^2}$$
$$d_{iB} = k_{Ti}^{2p} \tag{4.1}$$

where $\Delta_{ij}^2 = (y_i - y_j)^2 + (\phi_i - \phi_j)^2$ and $y_i$, $k_{Ti}$ are the rapidity and transverse momenta of candidate i respectively. After the distance measures have been calculated for each candidate in the event, the smallest $d_{ij}$ are merged into one object by summing the 4-momenta of particles i and j, the distance measures are updated and the algorithm moves onto the next smallest $d_{ij}$. If a particle has the smallest $d_{iB}$, it is removed and called a jet. This iterative process continues until all PF candidates are clustered into jets.

Conceptually, the exponents with $p$ dictate how transverse momenta varies with distance parameter. Negative values of $p$ merge higher transverse momenta, nearby candidates first. What differentiates anti-$k_T$ from Cambridge-Aachen and other similar algorithms is the choice of $p$ = -1 in equation 4.1. This negative value explains the name of the algorithm, and the tendency for it to produce circular[2] jets, centered on the highest $p_T$ candidate of the jet. The effect of these parameter values with respect to other available algorithms is below in Figure 4.3.

The final parameter in equation 4.1 is defined as $R = \sqrt{\Delta\eta^2 + \Delta\phi^2}$ and is the conesize of the jet being clustered. The cone size used for clustering in this analysis is R = 0.4. Versions of this analysis on 7 TeV, and 8 TeV datasets used a wider cone of 0.5. The move to smaller cone size was motivated by the fact that jets tend to be more energetic and thus narrower and more collimated as the center-of-mass collision energy increases to 13 TeV. All techniques mentioned above provide CMS analyzers the basic physics objects needed to perform analysis, and also standardizes

[2]circular in the $\eta$-$\phi$ plane




(a) $k_T$ (b) Cambridge-Aachen

(c) SIScone (d) anti-$k_T$

Figure 4.3. Jets in a sample MC event clustered with various algorithms.

the object reconstruction across the experiment.

## 4.2 Primary Vertex Identification and Pile-up

Due to the way the LHC collides bunches, there are multiple pp collisions in each LHC bunch crossing at CMS. Unfortunately, many of these collisions produce multiple objects that are reconstructed, but at most one of these collisions is a hard scatter head-on collision capable of producing a $t\bar{t}H$ event. These additional collisions that don't include the process of interest, as well as their resulting reconstructed objects are referred to as pileup, because they are said to *pile up* on top of the objects from the collision of interest. The location of the collision of interest is called the primary vertex. In this analysis, the primary vertex is defined as the vertex with the highest $p_T$ sum of all constituent tracks. Pileup is problematic because it can make matching objects to collisions difficult. Fortunately, CMS has an excellent tracker which makes the process of matching tracks to vertices (collisions) fairly straightforward. Aside from the tracking, pileup is also problematic because the energy deposits in the calorimeters from the pilup objects can distort the energy measurements and clustering of objects originating from the primary vertex. There are numerous techniques to account for these effects. The technique employed in this analysis is to calculate the energy due to pileup in the detector, rho, and subtract it off from each reconstructed jet energy measurement. This is called the rho correction. An example of the reconstructed tracks and vertices in an LHC bunch crossing is below in Figure 4.4.

The peak pileup values in the data analyzed vary from 30 to 45, meaning there are, on average, between 30 and 45 collisions (including the primary vertex) in each bunch crossing recorded by CMS[3].

[3]For a given bunch crossing, the actual number of pileup collisions varies according to a Poisson distribution, where the average is taken.

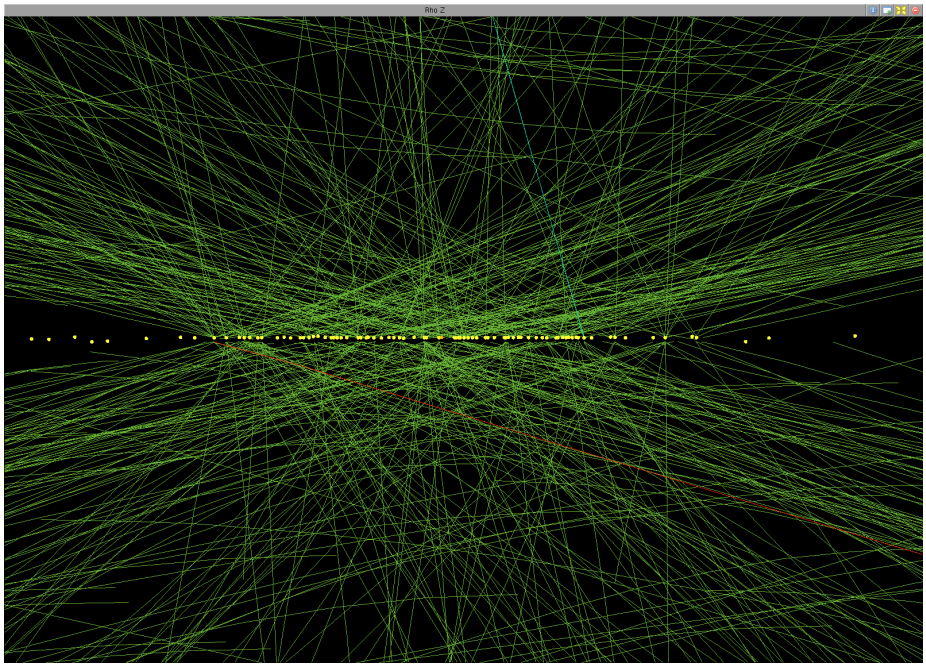


Figure 4.4. A side view of the CMS tracker's reconstructed vertices in a special high pileup LHC run. The pileup here is 78, meaning there are 78 collisions in a single bunch crossing [**?** ].

## 4.3 Object Selection

After the basic object reconstruction is performed, each event is ready for the first steps of analysis. The analysis begins with object selection where the objects in each event are subjected to tighter criteria to ensure quality objects in the analysis that are consistent with signal expectations and reject background. This selection is tailored to the type of objects in the desired signal event. Because the $t\bar{t}H$ multilepton final state is so complicated, almost[4] all objects available from the particle flow reconstruction are needed to identify events that are consistent with the signal, namely jets, missing energy, charged leptons, and hadronic taus.

### 4.3.1 Jets

Jets are clustered with anti-kt and reconstructed from PF candidates as described previously using the FASTJET algorithm [**?** ][**?** ]. Charged hadrons not originating from the primary vertex are subtracted from the jet clustering. Because the detector performance varies with time the measured jet energies must be updated to correct for performance degradation in the calorimeters. The jets are corrected in bins of jet $E_T$ (transverse energy) and $\eta$ [**?** ].

The jet selection for this analysis requires jets to have a $p_T > 25$ GeV and $|\eta| < 2.4$. Jets are removed when they overlap with a fakeable lepton (described later) within a $\Delta R$ cone of $< 0.4$. Finally, the jets must pass the working point of an MVA used to discriminate against jets from pileup vertices. This discriminator uses inputs characterizing the shape of the jet, the relative amounts of charged and neutral candidates within the jet, and the $p_T$ ratio between the candidates.

[4] all objects except the photon, which are used in the $t\bar{t}H, H \to \gamma\gamma$ search

### 4.3.2 b-jet Identification

Identifying jets from b-quark hadronization (b-jets) is crucial to selecting $t\bar{t}H$ events since the tops almost always produce b-quarks. The properties of heavy flavor quarks, specifically the bottom quark, have unique characteristics which make it possible to tag and identify their hadronizations in the detector. This identification is called b-tagging and is used throughout this and many other analyses. These heavy flavor quarks are characterized and distinguished from light flavor by large[5] masses, long lifetimes, high $p_T$ decay products, among which there is an occasional charged lepton. The long lifetime provides a key handle for identifying b-decays, because the b quark is further displaced from the primary vertex before hadronizing, resulting in a secondary vertex. Identifying this secondary vertex plays a central role in tagging jets from b-decays.

The algorithm used to identify b-jets in this analysis is the Combined Secondary Vertex (CSVv2) tagger [**?** ]. The CSVv2 tagger is an MVA discriminator that combines secondary vertex information with impact parameter variables that discriminate heavy flavor quarks from light flavor quarks and gluons. The tagging variable is assigned for each jet, and can range from -1 to +1, where the higher values correspond to higher likelihood that the jet originated from a b-quark. The jets described in the previous subsection are *tagged* when they pass specific working points of the tagger. The specific working point is choosen to balance keeping the fake rate low, and the efficiency high. The working points used in this analysis were selected after studying the tagger performance, as a function of jet $p_T$ and $\eta$ in data samples with b-jets. Two working points are used for b-tagging in this analysis. The medium working point (CSVv2 $> 0.8484$) corresponds to 70% efficiency for tagging b-jets with a 1.5% mistag rate, while the loose working point (CSVv2 $> 0.5426$) corresponds to an 85%

[5]large relative to light quarks and gluons

efficiency and a 10% mistag rate. The mistag rate is the probability to tag jets originating from light flavor quarks or gluons. The motivation for two working points is explained in Chapter 5.

#### 4.3.2.1 b-jet Scale Factors

The shape of the CSVv2 distribution of b-jets in data differs somewhat between that of MC. Because MC is being used to predict both the signal and many of the backgrounds, we apply scale factors (SFs) to correct the shape in the MC to match that of data. These scale factors are derived separately for heavy (b,c quark) flavor and light flavor as a function of the jet CSVv2, $p_T$, $|\eta|$. The SF is calculated according to equation 4.2 below.

$$SF(CSVv2, p_T, \eta) = \frac{DATA - MC_A}{MC_B} \tag{4.2}$$

The scale factor for heavy flavor has $MC_A$, $MC_B$ = light, heavy flavor MC respectively, while the scale factor applied to light flavor has $MC_A$, $MC_B$ = heavy, light flavor MC respectively. The scale factors are derived from a tag-and-probe technique. First, a control region is defined by selecting opposite sign $e - \mu$ events with exactly two jets. When this selection is applied to data, it yields events enriched with dileptonic $t\bar{t}$ decays where the two jets are most likely b-jets. Then one of the jets (the *tag*) is required to pass the medium CSVv2 working point (CSVv2 $> 0.8484$) to increase the $t\bar{t}$ purity. In real $t\bar{t}$ events, the remaining jet (the *probe*) is very likely a b-jet. Because control region applied to data is only *enriched* in $t\bar{t}$ events and not pure $t\bar{t}$ events, we subtract off the contribution of light flavor jets as estimated in MC, which explains the second term in the numerator of equation 4.2. Now we can directly apply this same control region to pure $t\bar{t}$ MC, and the ratio of the corrected yields in data to the pure $t\bar{t}$ MC gives the scale factor described in equation 4.2. The

same process is used for the light flavor correction, but the control region definition is modified to instead select dileptonic events with two leptons that have the same flavor and opposite sign, with exactly two jets where one of the jets *fails* the medium CSVv2 WP. This control region applied on data yields events enriched in Z+jets and the tag-and-probe process is repeated. The full details of this technique are described in [**?** ] and [**?** ].

### 4.3.3 Missing Energy

Missing energy, or more accurately missing *transverse* energy is calculated as the negative vector sum of $p_T$ of all PF candidates in the event[6], denoted as $E_T^{miss}$. It is called missing because the gluons (or quarks) colliding in the hard scatter that produce events such as $t\bar{t}H$ have no initial transverse momentum. By momentum conservation, the vector sum of all particles produced in the collision should match the sum before the collision. Any time the $E_T^{miss}$ is non-zero, it is assumed to be either from mismeasurements, or carried off by undetected "missing" particles. In the context of this and many other analyses, the $E_T^{miss}$ is interpreted as the presence of neutrinos. The two same-sign lepton requirement on the signal ensures that there will always be two neutrinos in the event, making it impossible to fully reconstruct the individual neutrino momenta. The presence of $E_T^{miss}$ can also occur due to mismeasurements/incorrect object reconstructions. To discriminate between the different sources of $E_T^{miss}$ we use another variable called , which is the negative vector sum of all selected jets and leptons in the event. While has lower resolution than the $E_T^{miss}$, it relies only on the higher $p_T$ selected objects and not the softer objects failing the selection. To exploit the fact that and $E_T^{miss}$ are less correlated in events with incidental missing energy, and more correlated in events with real missing energy, a

[6]The word "Energy" is used here in place of momentum for historical reasons. This quantity used to be calculated from the energy measured in the calorimeters of older experiments, and still is for neutral objects where momentum information is unavailable due to the lack of a curved track.

linear discriminant of both varibles is defined in equation 4.3 below.

$$\mathrm{E_T^{miss}}LD = \mathrm{E_T^{miss}} \times 0.00397 + \times 0.00265. \tag{4.3}$$

The coefficients are tuned to scale the $\mathrm{E_T^{miss}}$ term by 60% and the term by 40%, which were empirically found to deliver the best signal efficiency and background rejection. The $\mathrm{E_T^{miss}}$ LD threshold is described in Chapter 5.

### 4.3.4 Leptons

The flagship objects that characterize this analysis are leptons. The lepton selection is the foundation of this $\mathrm{t\bar{t}H}$ search and drives an important component of the background estimation. The ultimate goal of the lepton selection is to identify and select prompt leptons and reject non-prompt leptons, also known as fake leptons. In this analysis, leptons are defined as being electrons or muons. Taus are excluded from this defnition because they are unstable and decay in the detector. In this context, a prompt lepton is a lepton that originates directly (promptly) from a W, Z or $\tau$ decay, while non-prompt leptons predominantly originate from b-hadron decays, but also in-flight decays of pions, and photon conversions[7] . The term *fake* is used to describe non-prompt leptons because non-prompt leptons that pass the criteria designed to select prompt leptons are faking prompt leptons. These fakes are one of the largest backgrounds in this analysis and the lepton selection is designed to reduce both the quantity of fakes entering the signal regions, and the systematic uncertainties on that quantity.

[7]Prompt leptons produced in $\tau$ decays are actually coming from an offshell W produced by the decaying $\tau$

#### 4.3.4.1 Electron Identification

Electrons are reconstructed from tracker hits (GSF tracks) and ECAL clusters via particle flow as described previously. While many other requirements are placed on electrons later, they first must have $|\eta|$ <2.5 to guarantee they passed through the tracker, and minimum $p_T$ > 7 GeV. Electrons must pass the working point of an MVA designed to identify electrons using shower-shape variables, track-cluster consistency variables, and track quality variables [**?** ]. We apply loose cuts as a function of $|\eta|$ on this MVA value. The value of this MVA discriminator is used as an input to the Lepton MVA, which will be described in the following sections. For the tightest selections, we also require the charge measurements in the pixels and strips to agree and be well-matched to ECAL clusters. Additionally, electrons must not have missing hits in the inner tracker, and pass additional veto which suppresses contributions from photon conversions.

#### 4.3.4.2 Muon Identification

Muons are reconstructed by combining tracks from the silicon tracker with tracks in the muon chambers and are required to be global muons. Muons are first required to have $p_T > 5$ GeV and $|\eta|$ <2.4. We use the selectrion criteria developed by the CMS Muon Physics Object Group (POG) and require all muons to pass the Loose ID [**?** ] and in some cases the Medium ID [**?** ] depending on the selection. Additionally, we require the sign of the muon electric charge to be well measured by appling a cut related to the significance of the muon $p_T$ measurement called *tight charge* defined as: $\Delta\ p_T/p_T < 0.2$, where $\Delta\ p_T$ is the uncertainty on the muon $p_T$. This cut ensures a high confidence on the charge sign of the muons. This cut is only required on the tightest selection of muons.

#### 4.3.4.3 Isolation

Lepton isolation is a measure of how spatially separated the lepton is from other physics objects in an event. Prompt leptons are typically isolated from hadronic activity (jets) in the event, while non-prompt leptons, which are often produced in hadronic decays, are significantly less isolated and close to or overlapping with jets. The standard isolation considers all charged and neutral hadrons as well as photons within a fixed cone size in R around the lepton. Then if there is too much energy relative to the lepton energy from the other objects inside the fixed cone radius, the lepton fails the isolation criteria. This analysis uses a variation of this technique that varies the cone size with the $p_T$ of the lepton, since boosted (very high momentum) objects tend to be more collimated and thus should have a more collimated cone definition. This isolation technique is called mini isolation ($I_{\text{mini}}$) and the cone size is varied with lepton $p_T$ according to figure 4.5 below.

The amount of activity inside this cone is adjusted to correct for pileup using the $\rho$-correction, which uses a specific pileup-density event-by-event in different $|\eta|$ ranges in the detector. The pileup density $\rho$, is multiplied by effective areas as a function of $|\eta|$. The $I_{\text{mini}}$ quantity is defined in equation 4.4 below

$$I_{mini} = \frac{\sum_R p_T(h^{\pm}) - max(0, \sum_R p_T(h^0) + p_T(\gamma) - \rho A(\frac{R}{0.3})^2)}{p_T(l)} \tag{4.4}$$

where $\sum_R p_T(h^{\pm})$ is the energy of the charged hadrons inside the cone of radius R, $\sum_R p_T(h^0)$ is the energy of the neutral hadrons inside the cone, $p_T(\gamma)$ is the energy from photons inside the cone, and $-\rho A(\frac{R}{0.3})^2$ is the correction for pileup. The effective areas, A, are defined separately for muons and electrons and are based on a fixed cone size of R = 0.3 and listed in tables 4.14.2, which explains the denominator in the pileup correction term. Both muons and electrons are subject to the same isolation criteria, which requires $I_{\text{mini}} < 0.4$.

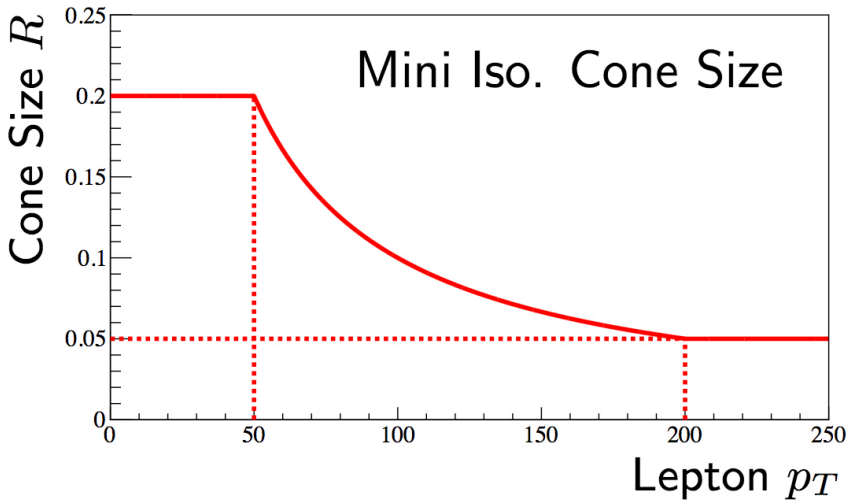



Figure 4.5. Mini isolation cone size vs lepton $p_T$. For $p_T < 20$ GeV, R = 0.2, for $p_T > 200$ GeV, R = 0.05. [**?** ].

TABLE 4.1

Electron effective areas for the pileup correction.

| $\|\eta\|$ range | effective area (A) |
|---|---|
| $\|\eta\| < 1.0$ | 0.1752 |
| $1.0 < \|\eta\| < 1.479$ | 0.1862 |
| $1.479 < \|\eta\| < 2.0$ | 0.1411 |
| $2.0 < \|\eta\| < 2.2$ | 0.1534 |
| $2.2 < \|\eta\| < 2.3$ | 0.1903 |
| $2.3 < \|\eta\| < 2.4$ | 0.2243 |
| $2.4 < \|\eta\|$ | 0.2687 |

TABLE 4.2

Muon effective areas for the pileup correction.

| $\|\eta\|$ range | effective area (A) |
|---|---|
| $\|\eta\| < 0.8$ | 0.0735 |
| $0.8 < \|\eta\| < 1.3$ | 0.0619 |
| $1.3 < \|\eta\| < 2.0$ | 0.0465 |
| $2.0 < \|\eta\| < 2.2$ | 0.0433 |
| $2.2 < \|\eta\|$ | 0.0577 |

#### 4.3.4.4 Vertexing

Vertexing requirements are placed on the leptons to help to ensure they are coming from (associated with) the primary vertex and to remove leptons from misreconstructed tracks and hadron decays from further consideration. The requirements placed on the leptons include the impact parameter in both the transverse plane ($d_{xy}$), the z-direction ($d_z$) and the significance on the three-dimensional impact parameter ($\text{SIP}_{3D}$). The impact parameter is the distance of closest approach of the lepton track to the primary vertex. The values of thes cuts are detailed in tables 4.3.4.7 and 4.3.4.7. These cuts are designed to reduce the contribution of pileup and misreconstructed tracks. All of these variables are useful in distinguishing prompt from non-prompt leptons and are inputs to the lepton MVA.

#### 4.3.4.5 Jet-related Variables

In order to reduce the contribution of fake leptons from hadron decays, specifically b-decays, we construct several variables that incorporate the characteristics of the nearest jet to the selected lepton. Using the jet selection described previously, but lowering the $p_T$ threshold to 15 GeV, we consider the closest jet to each lepton within an $R < 0.5$ of the lepton. The variables include the ratio of the jet to the ratio of the lepton, the jet $p_T$, the jet CSVv2 value, the number of charged tracks inside the jet, and pt of the lepton relative to the pt of the jet ($p_T^{rel}$), which is defined in equation 4.5 below.

$$p_T^{rel} = \frac{(\vec{p}(jet) - \vec{p}(l)) \cdot \vec{p}(l)}{||\vec{p}(jet) - \vec{p}(l)||} \tag{4.5}$$

We apply a modified version of the jet energy corrections described previously to avoid biasing the prompt lepton selection by over-correcting the jets near the lepton.

#### 4.3.4.6 Lepton MVA

In addition to the above selection requirements, we use a multivariate discriminator designed to select prompt leptons and reject fakes. This MVA is called the lepton MVA and is based on a boosted decision tree classifier. We use two versions of this MVA, one specifically for muons and one for electrons. The lepton MVA is trained on prompt leptons in $t\bar{t}H$ MC as signal, and the background sample consists of fake leptons from $t\bar{t}$ MC sample. The input variables are:

- lepton $p_T$
- lepton $|\eta|$
- number of charged tracks
- charged component of mini isolation
- neutral component of mini isolation ($\rho$ corrected)

- jet $p_T^{rel}$
- min(jet $\mathrm{p_T}$ ratio, 1.5)
- jet CSVv2
- $\mathrm{SIP}_{3D}$
- $\log|d_{xy}|$
- $\log|d_z|$
- segment compatibility (muons only)
- electron MVA ID (electrons only)

Passing the tight working point of the lepton MVA (MVA $> 0.9$) is the defining characteristic of the tight leptons used in the signal regions.

#### 4.3.4.7 Lepton Selection

The lepton selection consists of three increasingly selective classes: loose, fakeable, and tight. The loose is a preselection, while the fakeable object selection, which is a tighter subset of the loose, is used to define control regions for estimating the background due to fake leptons. The tight leptons, which are a subset of the fakeable, define the leptons used in the signal regions of this analysis. The details of these selections are described in tables 4.3.4.7 and 4.3.4.7.

#### 4.3.4.8 Lepton Effiency Scale Factors

The rate at which objects pass the described lepton selections (known as efficiency) differs slightly between data and MC. Because we use MC predictions and compare directly to data, the lepton efficiency in MC is corrected to match the lepton efficiency in data. Because we utilize two lepton selections (loose, tight) for the MC predictions, we apply scale factors derived uniquely for each selection. Because the tight selection

Requirements on each of the three muon selections. A few extra requirements are applied for fakeable objects that fail the lepton MVA requirement, to better control the extrapolation in fragmentation and flavor composition and are marked with a †.

| **Cut** | **Loose** | **Fakeable** | **Tight** |
|---|---|---|---|
| $\lvert\eta\rvert < 2.4$ | ✓ | ✓ | ✓ |
| $\mathrm{p_T}$ | $> 5$ | $> 15$ | $> 15$ |
| $\lvert d_{xy}\rvert < 0.05$ (cm) | ✓ | ✓ | ✓ |
| $\lvert d_z\rvert < 0.1$ (cm) | ✓ | ✓ | ✓ |
| $\mathrm{SIP}_{3D} < 8$ | ✓ | ✓ | ✓ |
| $I_{\mathrm{mini}} < 0.4$ | ✓ | ✓ | ✓ |
| is Loose Muon | ✓ | ✓ | ✓ |
| $\mathrm{p_T^{ratio}}$ | – | $> 0.5$† / – | – |
| segmentCompatibility | – | $> 0.3$† / – | – |
| jet CSV | – | $< 0.3$† / $< 0.8484$ | $< 0.8484$ |
| is Medium Muon | – | – | ✓ |
| tight-charge | – | – | ✓ |
| lepMVA $> 0.90$ | – | – | ✓ |

Requirements on each of the three electron selections. In some cases, the cut values change for different $\eta$ ranges. These ranges are $0 < |\eta| < 0.8$, $0.8 < |\eta| < 1.479$, and $1.479 < |\eta| < 2.5$ and the respective cut values are given in the form (value$_1$, value$_2$, value$_3$). Cuts marked with † are applied only to objects failing the tight selection.

| **Cut** | **Loose** | **Fakeable** | **Tight** |
|---|---|---|---|
| $\|\eta\| < 2.5$ | ✓ | ✓ | ✓ |
| $\mathrm{p_T}$ | $> 7$ | $> 15$ | $> 15$ |
| $\|d_{xy}\| < 0.05$ (cm) | ✓ | ✓ | ✓ |
| $\|d_z\| < 0.1$ (cm) | ✓ | ✓ | ✓ |
| $\mathrm{SIP}_{3D} < 8$ | ✓ | ✓ | ✓ |
| $I_{\mathrm{mini}} < 0.4$ | ✓ | ✓ | ✓ |
| MVA ID $> (0.0, 0.0, 0.7)$ | ✓ | ✓ | ✓ |
| $\sigma_{i\eta i\eta} < (0.011, 0.011, 0.030)$ | – | ✓ | ✓ |
| H/E $< (0.10, 0.10, 0.07)$ | – | ✓ | ✓ |
| $\Delta\eta_{in} < (0.01, 0.01, 0.008)$ | – | ✓ | ✓ |
| $\Delta\phi_{in} < (0.04, 0.04, 0.07)$ | – | ✓ | ✓ |
| $-0.05 < 1/E - 1/p < (0.010, 0.010, 0.005)$ | – | ✓ | ✓ |
| $\mathrm{p_T^{ratio}}$ | – | $> 0.5$† / – | – |
| jet CSV | – | $< 0.3$† / $< 0.8484$ | $< 0.8484$ |
| tight-charge | – | – | ✓ |
| conversion rejection | – | – | ✓ |
| Number of missing hits | $< 2$ | $== 0$ | $== 0$ |
| lepMVA $> 0.90$ | – | – | ✓ |

is a subset of the loose, all MC events with tight leptons have both loose and tight scale factors applied. This is because the tight lepton efficiency is measured relative to the loose denominator.

These corrections are applied as a function of $p_T$ and $|\eta|$ for electrons and muons separately. These scale factors are applied to MC and defined in equation 4.6 below:

$$SF(p_T, \eta) = \frac{\varepsilon_{data}(p_T, \eta)}{\varepsilon_{MC}(p_T, \eta)} \tag{4.6}$$

where $\varepsilon(p_T, \eta)$ is the efficiency measured for a single lepton to pass the selection measured in either data or MC. The total scale factor is the product of each individual lepton scale factor and is applied on an event-basis.

The tight lepton efficiencies were measured in a control region enriched in Drell-Yan (DY)[8] events on MC using the tag-and-probe method[9]. Loose leptons form the denominator of the tight efficiency measurements. The measured tight lepton efficiencies are in Figures 4.6 and 4.7.

The data/MC scale factors associated with the efficiency are in Figure 4.8 below. The loose efficiency scale factors were derived by the SUSY lepton scale factor working group within CMS.

### 4.3.5 Taus

The final objects considered in this analysis are tau leptons, specifically hadronically decaying taus. Leptonic tau decays produce a charged lepton that is typically

[8]DY events are $Z/\gamma* \rightarrow\ l^{\pm}l^{\mp}$, where $l = e, \mu$.

[9]The tag-and-probe method is a technique used throughout particle physics experiments to measure efficiencies of objects in data in an unbiased way by exploiting known di-object resonances. The pairs from the resonances are reconstructed requiring one object to pass a "tight" selection (tag) and the other to pass a "loose" selection (probe). Because the efficiency must be measured in data (in addition to MC), a trigger is required to collect the data. The trigger fires on the tag leg. The selection of resulting probes are completely unbiased by any selection present in the trigger. The efficiency is then measured by the ratio of probe objects passing some tight definition in the numerator, to the total probes in the denominator.

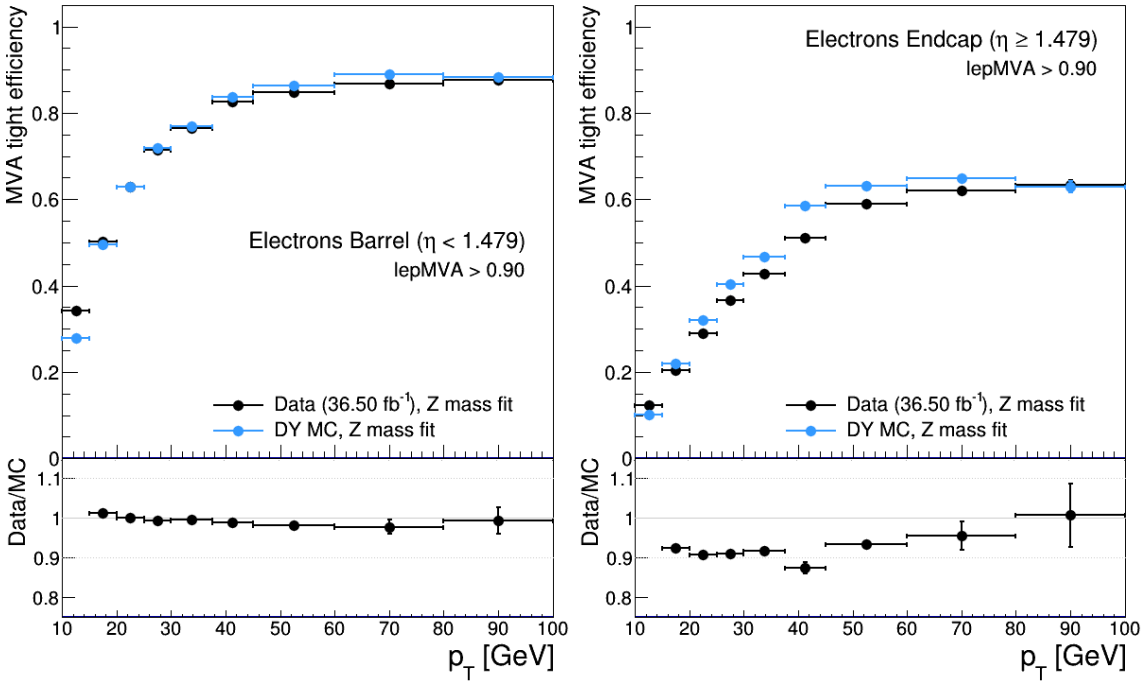



Figure 4.6. The efficiency to select tight electrons (from loose) in the barrel (left) and the endcap (right).

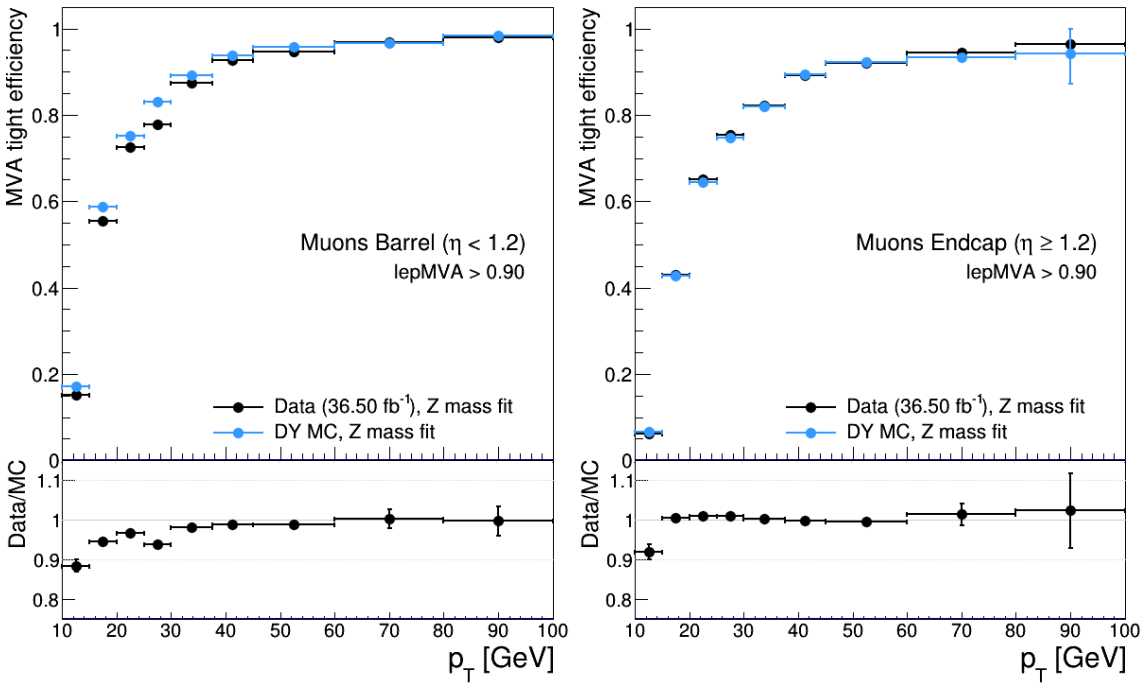



Figure 4.7. The efficiency to select tight muons (from loose) in the barrel (left) and the endcap (right).

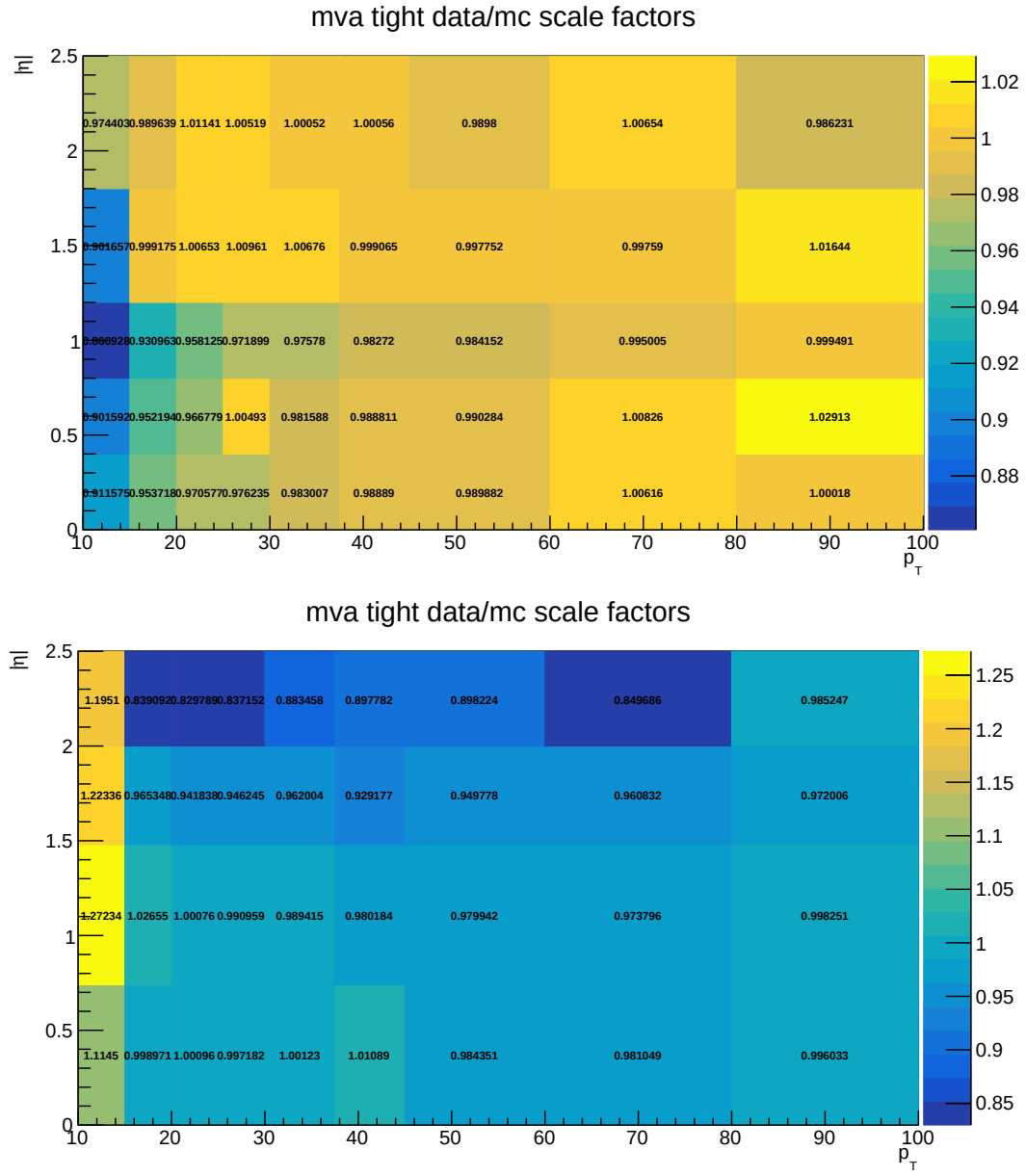



Figure 4.8. The loose-to-tight efficiency data/MC scale factors in bins of $p_T$ and $|\eta|$ for individual muons (top) and electrons (bottom).

identified as prompt, making their presence straightforward to detect, but impossible to distinguish from a lepton produced directly from a W or Z. To avoid an overlap with the signal regions of the $t\bar{t}H$,$H \to \tau\tau$ analysis [**?** ], which considers *only* events with hadronic taus, we use the same object definition and veto any event with a selected hadronic tau. Hadronic taus are reconstructed with the hadron-plus-strips algorithm [**?** ] in 1-prong or 3-prong decay modes. Taus must have $p_T > 20$ GeV, $|\eta| < 2.3$, $d_{xy} < 1000$ cm, $d_z < 0.2$ cm, and pass the "decay mode finding" MVA. Additonally, taus are required to pass the medium working point of a tau MVA[10] which was trained on $t\bar{t}H$ events as signal and $t\bar{t}$ events as background with an isolation cone of radius 0.3 [**?** ].

## 4.4 Object Cleaning

Objects that pass the above selections but are spatially nearby must be addressed since they can be the decay products of the same underlying particle, but resolved by the detector as two distinct objects. This results in *double-counting*. We address this by removing one of the overlapping objects, which is decided according to the object type in a process called object cleaning. By correctly choosing which objects to remove, and which objects to keep, we can mitigate the effects of double-counting. The first object cleaning is applied to leptons passing the loose selection. Loose electrons are removed[11] when they overlap within a cone size of R = 0.05 of a loose muon. Jets are removed when they overlap within a cone of radius R = 0.4 with fakeable lepton (muon or electron) or a tau passing the cleaning selection.

[10] The technical name for this working point is "byMediumIsolationMVArun2v1DBdR03oldDMwLT".

[11] Because the lepton selection is applied in order, any objects removed after the loose selection are completely removed from the event and not considered for fakeable and tight.

# CHAPTER 5

# EVENT SELECTION

The event selection defines the signal regions of this analysis. The signal regions are categories of events defined by requirements aimed at selecting as much signal ($t\bar{t}H$), and simultaneously rejecting as much background as possible. In the multi-lepton analysis, there are three signal regions predominantly defined by the lepton multiplicity in the event: the two-lepton same-sign ($2lss$) region, the three-lepton ($3l$) region, and the four-lepton ($4l$) region. This dissertation focuses $2lss$ category. Defining the signal regions by lepton multiplicity is motivated by the fact that the number of leptons passing the object selection yields the most information about the event.

There are several cuts which don't depend on lepton multiplicity that are applied to all signal regions to ensure the selection of events consistent with $t\bar{t}H$. Events with a pair of loose leptons with an invariant mass of less than 12 GeV are vetoed, as they are consistent with a $J/\psi$ or $\Upsilon$ decay and are not modeled in the MC simulation, but are present in data. At least two jets are required in all signal regions, and of these there must be at least one jet passing the medium CSVv2 working point, or two passing the loose working point. This b-jet requirement is consistent with a top-quark pair decaying to jets, which is present in all $t\bar{t}H$ processes.

## 5.1 Two-lepton same-sign category

The $2lss$ category is primarily defined by the requirement that there be exactly two tight leptons that have the same sign electric charge. The same-sign requirement

is needed to veto one of the largest backgrounds dileptonic $t\bar{t}$ + jets, which has oppositely charged leptons and a cross section more than three orders of magnitude greater than that of $t\bar{t}H$. We require the leading lepton have $p_T > 25$ GeV to stay well above the trigger threshold for reasons that will be explained in the trigger section. At least 4 jets are required, to be consistent with a $t\bar{t}H$ process with two same-sign leptons in the final state. To reject backgrounds with Z bosons where one of the lepton charges is mis-measured, we reject any di-electron pair whose invariant mass is within 10 GeV of the Z mass (90.2 GeV) and also require the $E_T^{miss}$ LD be greater than 0.2 (again, for di-electron events only). To summarize, the $2lss$ category is defined by events satisfying the following requirements:

- Exactly two tight leptons with the same-sign electric charge and $p_T > 25$,15 GeV
- m($ll$) $< 12$ GeV for any pair of loose leptons in the event
- $\geq$4 jets, among which there must be $\geq$1 CSVv2 M or $\geq$2 CSVv2 L
- $|m(ee) - m_Z| > 10$ GeV (for ee events only)
- $E_T^{miss}$ LD $> 0.2$ (for ee events only)

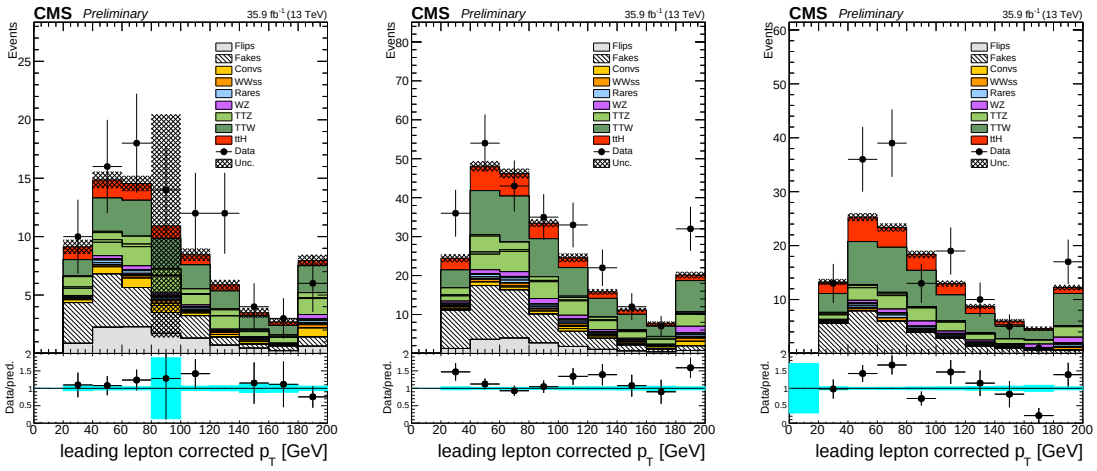



Figure 5.1. The leading lepton transverse momenta in the 2lss $ee/e\mu/\mu\mu$ categories (prefit). Uncertainties shown are purely statistical.

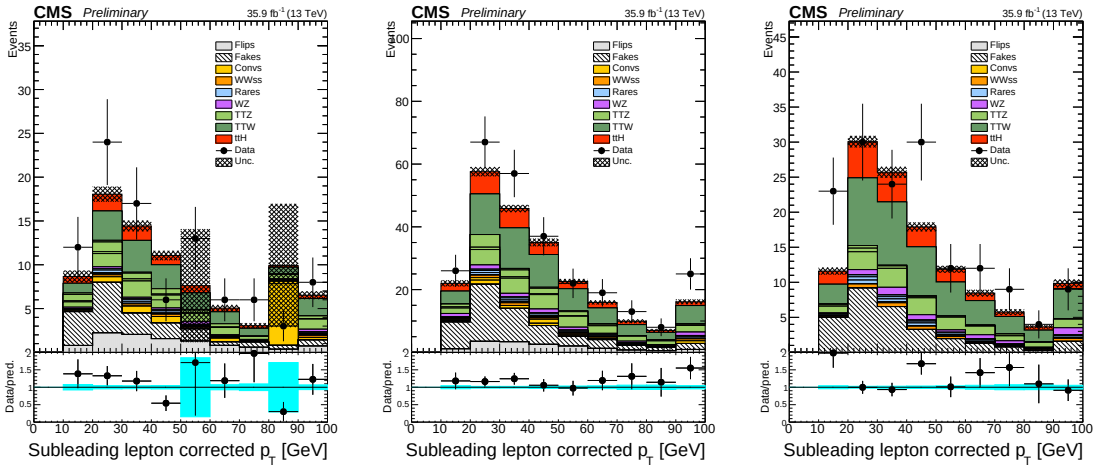



Figure 5.2. The sub leading lepton transverse momenta in the 2lss $ee/e\mu/\mu\mu$ categories (prefit). Uncertainties shown are purely statistical.

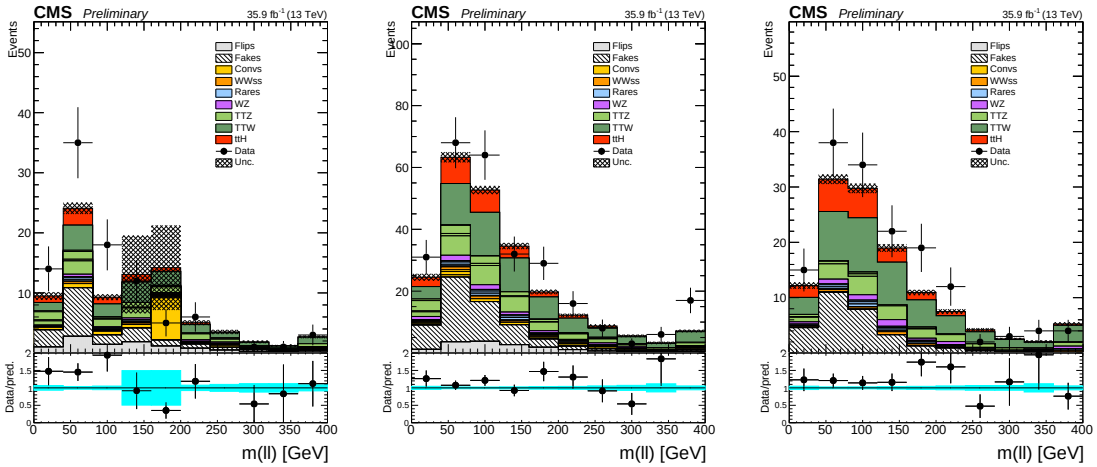



Figure 5.3. The dilepton invariant mass spectra in the 2lss $ee/e\mu/\mu\mu$ categories (prefit). Uncertainties shown are purely statistical.

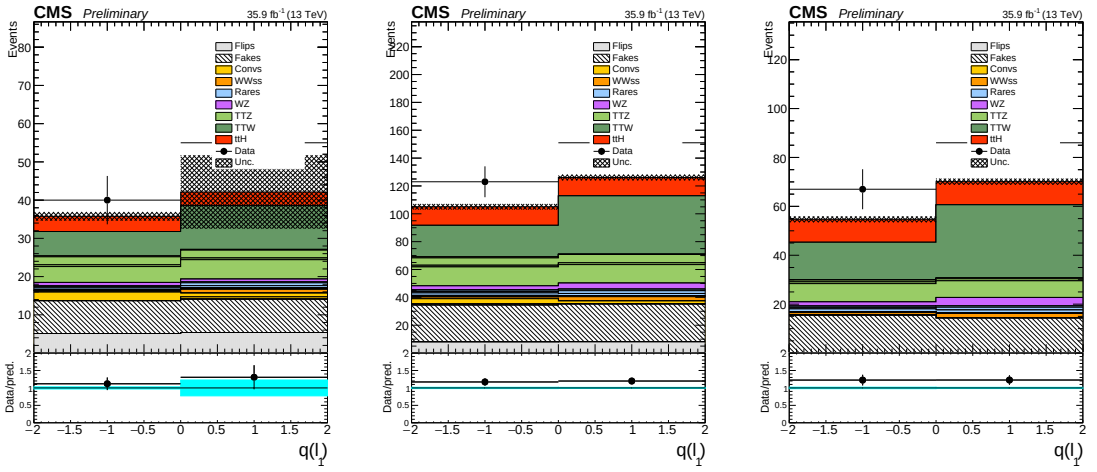



Figure 5.4. The sum of the lepton electric charges in the 2lss $ee/e\mu/\mu\mu$ categories (prefit). Uncertainties shown are purely statistical.

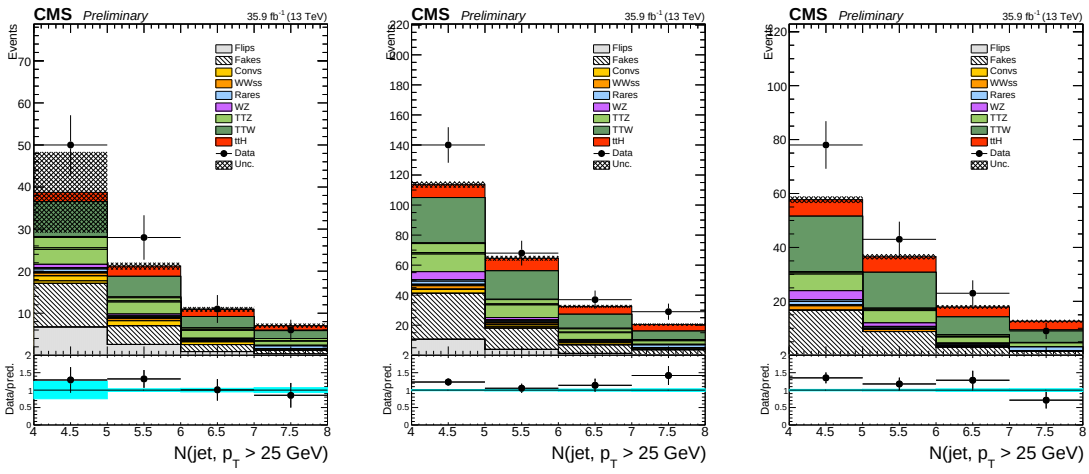



Figure 5.5. The jet multiplicity distribution in the 2lss $ee/e\mu/\mu\mu$ categories (prefit). Uncertainties shown are purely statistical.

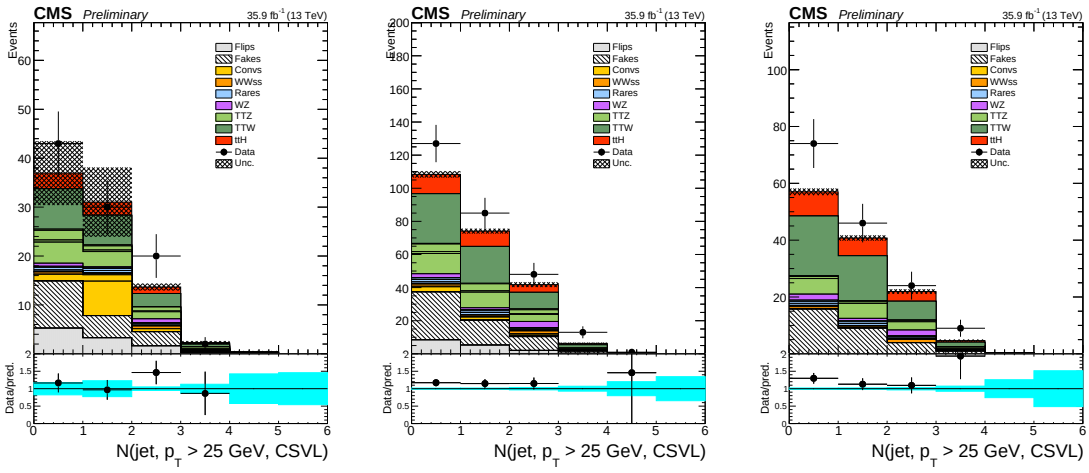



Figure 5.6. The jet multiplicity for jets passing the loose working point of the CSV tagger in the 2lss $ee/e\mu/\mu\mu$ categories (prefit). Uncertainties shown are purely statistical.

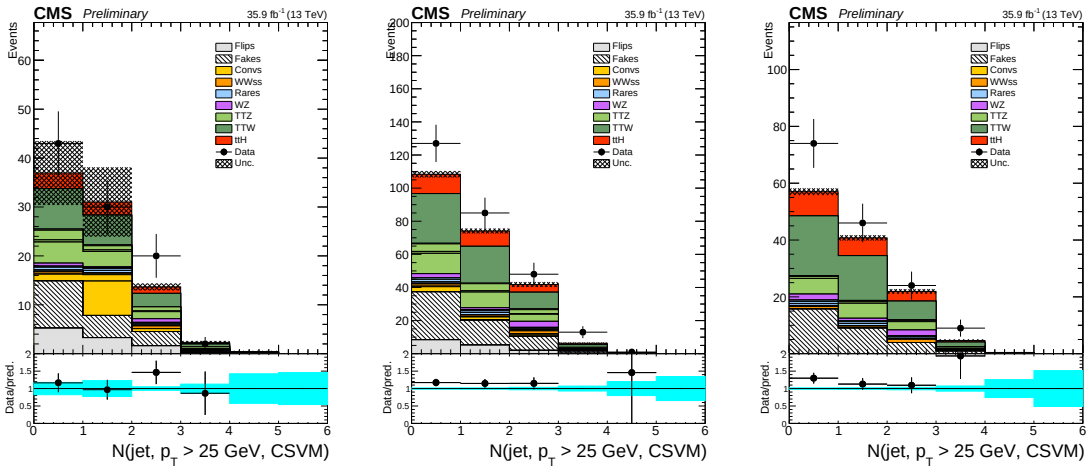



Figure 5.7. The jet multiplicity for jets passing the medium working point of the CSV tagger in the 2lss $ee/e\mu/\mu\mu$ categories (prefit). Uncertainties shown are purely statistical.

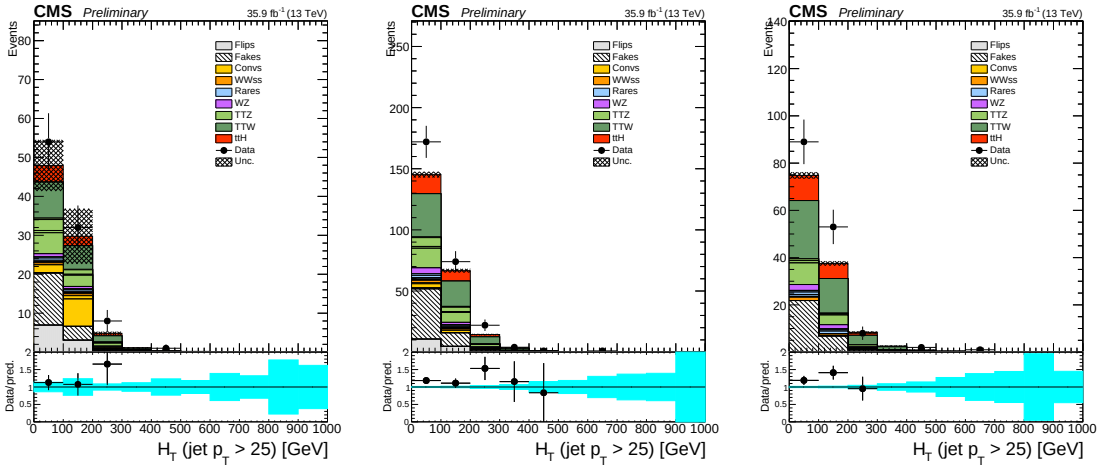



Figure 5.8. The ectra in the 2lss $ee/e\mu/\mu\mu$ categories (prefit). Uncertainties shown are purely statistical.

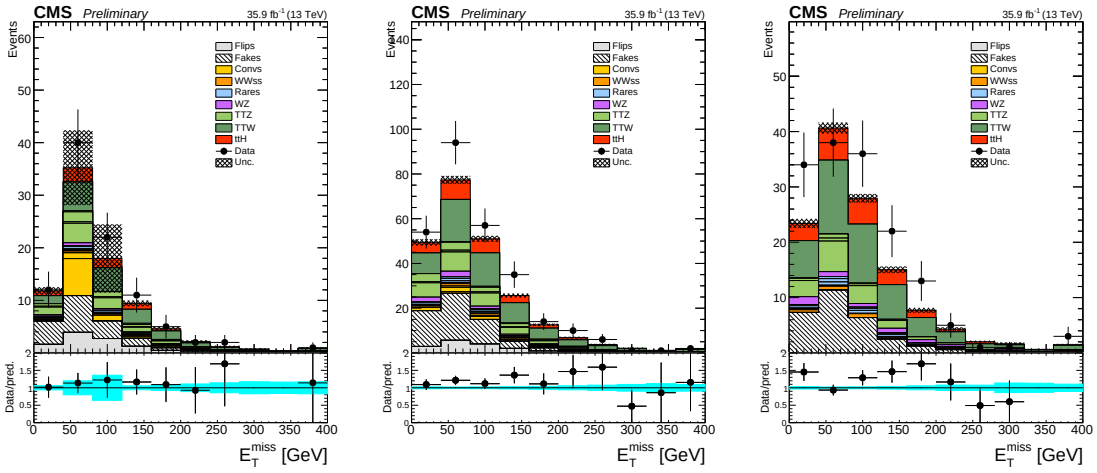



Figure 5.9. The $E_T^{miss}$ spectra in the 2lss $ee/e\mu/\mu\mu$ categories (prefit). Uncertainties shown are purely statistical.

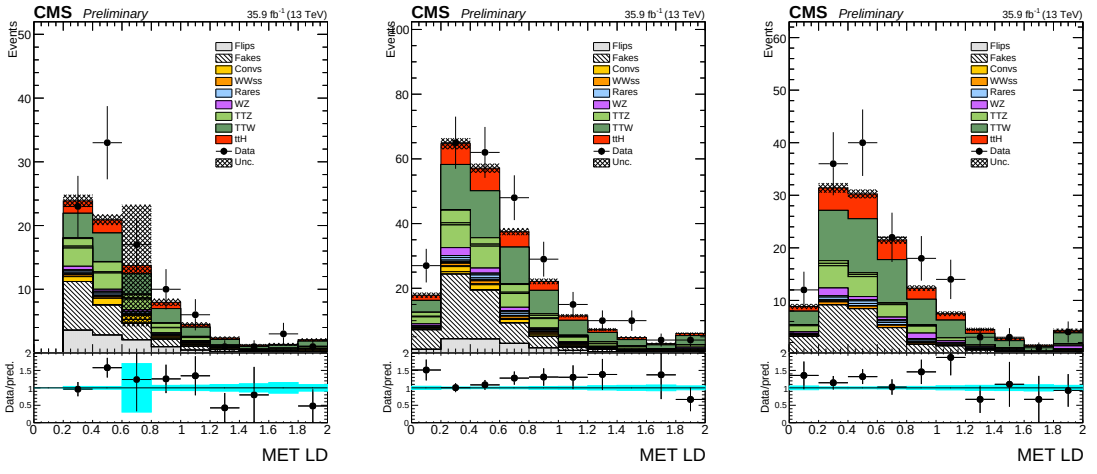



Figure 5.10. The $E_T^{miss}$ LD spectra in the 2lss $ee/e\mu/\mu\mu$ categories (prefit). Uncertainties shown are purely statistical.

# CHAPTER 6

# DATA AND MC SAMPLES

This analysis is essentially a counting experiment that compares the number of $t\bar{t}H$ events predicted with the number of events observed after accounting for the backgrounds. Simply put, this analysis compares the number of signal + background events predicted by MC to the number of events observed in real data taken by CMS, in the signal regions, after accounting for uncertainties. The hypothesis in this case is that the $t\bar{t}H$ process and all relevant[1] backgrounds exist in the amounts dictated by the Standard Model, and this hypothesis is represented by the MC.

## 6.1 Data samples

The data in this analysis was collected from CMS throughout 2016 and only when the magnet was on[2] and all subdetectors were fully operational. The full 2016 dataset analyzed here corresponds to a total integrated luminosity of 35.9 $fb^{-1}$ and uses the single and double lepton datasets listed in table 6.1 and reconstructed with the `CMSSW_8_0_x` software. This data corresponds to a center-of-mass energy of 13 TeV, and the spacing between bunches in the LHC was 25ns. On average, there were 30 pileup events per bunch crossing in this data. This data was certified as "good"

[1]Here, relevant means any background contributing a non-negligible yield to the signal regions.

[2]The CMS solenoid was notoriously problematic for data taking during 2015, suffering from cryogenics issues related to the coldbox and rendering much of the data collected during that time useless from the perspective of most (but not all) physics analyses, rendering the magnet status an important detail.

for data analysis internally by CMS [**?** ]. To address issues affecting a specific data-taking period and ensure accurate detector measurements, different alignment and calibration settings were used for specific datasets[3].

[3]These alignment and calibration settings are controlled by a single parameter, called the global tag. The global tags used to reconstruct and analyze this data are `80X_dataRun2_2016SeptRepro_v7`, corresponding to dataset eras B-G, and `80X_dataRun2_Prompt_v16` corresponding to era H.

TABLE 6.1

Datasets

| Dataset | Run Range | Luminosity $fb^{-1}$ |
|---|---|---|
| /SingleElectron/Run2016B-23Sep2016-v3/MINIAOD | 273150–275376 | 5.79 |
| /SingleElectron/Run2016C-23Sep2016-v1/MINIAOD | 275656–276283 | 2.57 |
| /SingleElectron/Run2016D-23Sep2016-v1/MINIAOD | 276315–276811 | 4.25 |
| /SingleElectron/Run2016E-23Sep2016-v1/MINIAOD | 276831–277420 | 4.01 |
| /SingleElectron/Run2016F-23Sep2016-v1/MINIAOD | 277932–278808 | 3.10 |
| /SingleElectron/Run2016G-23Sep2016-v1/MINIAOD | 278820–280385 | 7.54 |
| /SingleElectron/Run2016H-PromptReco-v2/MINIAOD | 281207–284035 | 8.39 |
| /SingleElectron/Run2016H-PromptReco-v3/MINIAOD | 284036–284044 | 0.22 |
| /SingleMuon/Run2016B-23Sep2016-v3/MINIAOD | 273150–275376 | 5.79 |
| /SingleMuon/Run2016C-23Sep2016-v1/MINIAOD | 275656–276283 | 2.57 |
| /SingleMuon/Run2016D-23Sep2016-v1/MINIAOD | 276315–276811 | 4.25 |
| /SingleMuon/Run2016E-23Sep2016-v1/MINIAOD | 276831–277420 | 4.01 |

TABLE 6.1

*Continued*

| Dataset | Run Range | Luminosity $fb^{-1}$ |
|---|---|---|
| /SingleMuon/Run2016F-23Sep2016-v1/MINIAOD | 277932–278808 | 3.10 |
| /SingleMuon/Run2016G-23Sep2016-v1/MINIAOD | 278820–280385 | 7.54 |
| /SingleMuon/Run2016H-PromptReco-v2/MINIAOD | 281207–284035 | 8.39 |
| /SingleMuon/Run2016H-PromptReco-v3/MINIAOD | 284036–284044 | 0.22 |
| /DoubleEG/Run2016B-23Sep2016-v3/MINIAOD | 273150–275376 | 5.79 |
| /DoubleEG/Run2016C-23Sep2016-v1/MINIAOD | 275656–276283 | 2.57 |
| /DoubleEG/Run2016D-23Sep2016-v1/MINIAOD | 276315–276811 | 4.25 |
| /DoubleEG/Run2016E-23Sep2016-v1/MINIAOD | 276831–277420 | 4.01 |
| /DoubleEG/Run2016F-23Sep2016-v1/MINIAOD | 277932–278808 | 3.10 |
| /DoubleEG/Run2016G-23Sep2016-v1/MINIAOD | 278820–280385 | 7.54 |
| /DoubleEG/Run2016H-PromptReco-v2/MINIAOD | 281207–284035 | 8.39 |
| /DoubleEG/Run2016H-PromptReco-v3/MINIAOD | 284036–284044 | 0.22 |

TABLE 6.1

*Continued*

| Dataset | Run Range | Luminosity $fb^{-1}$ |
|---|---|---|
| /DoubleMuon/Run2016B-23Sep2016-v3/MINIAOD | 273150–275376 | 5.79 |
| /DoubleMuon/Run2016C-23Sep2016-v1/MINIAOD | 275656–276283 | 2.57 |
| /DoubleMuon/Run2016D-23Sep2016-v1/MINIAOD | 276315–276811 | 4.25 |
| /DoubleMuon/Run2016E-23Sep2016-v1/MINIAOD | 276831–277420 | 4.01 |
| /DoubleMuon/Run2016F-23Sep2016-v1/MINIAOD | 277932–278808 | 3.10 |
| /DoubleMuon/Run2016G-23Sep2016-v1/MINIAOD | 278820–280385 | 7.54 |
| /DoubleMuon/Run2016H-PromptReco-v2/MINIAOD | 281207–284035 | 8.39 |
| /DoubleMuon/Run2016H-PromptReco-v3/MINIAOD | 284036–284044 | 0.22 |
| /MuonEG/Run2016B-23Sep2016-v3/MINIAOD | 273150–275376 | 5.79 |
| /MuonEG/Run2016C-23Sep2016-v1/MINIAOD | 275656–276283 | 2.57 |
| /MuonEG/Run2016D-23Sep2016-v1/MINIAOD | 276315–276811 | 4.25 |
| /MuonEG/Run2016E-23Sep2016-v1/MINIAOD | 276831–277420 | 4.01 |

TABLE 6.1

*Continued*

| Dataset | Run Range | Luminosity $fb^{-1}$ |
|---|---|---|
| /MuonEG/Run2016F-23Sep2016-v1/MINIAOD | 277932–278808 | 3.10 |
| /MuonEG/Run2016G-23Sep2016-v1/MINIAOD | 278820–280385 | 7.54 |
| /MuonEG/Run2016H-PromptReco-v2/MINIAOD | 281207–284035 | 8.39 |
| /MuonEG/Run2016H-PromptReco-v3/MINIAOD | 284036–284044 | 0.22 |

## 6.2 MC samples

The collection of MC samples described below comprise both the signal and background predictions. Like the data, the software used for MC simulation was `CMSSW_8_0_x`.

While several different generators were used to produce the MC in this analysis, the basic strategy is the same. The most important MC samples in this analysis were produced with a generator using next-to-leading-order (NLO)[4] matrix element calculations to generate events based on hard scatter interactions between the initial state partons. These initial state partons are simulating the actual proton collison inside CMS. The initial partons used in the MC generation are sampled from the proton PDFs. The matrix element calculation is then performed resulting in the final state particles. The partons involved in the hardest subprocess experience large accelerations, emitting QCD radiation in the form of gluons. The gluons then decay to other color-charge carrying particles, which in turn emit more gluons etc. This process is known as the parton shower. Adding the effect of these parton showers involves higher order corrections to the matrix element. The technique used by MC generators is to include only the dominant contributions at each step, since calculating these exactly is time consuming and computationally intensive. The effects of these showers predominantly comes in the form of final state particles which are either radiated from the incoming hard-scatter partons, known as initial state radiation (ISR), or radiated from the outgoing hard-scatter partons, known as final state radiation (FSR). ISR/FSR contributions to MC help more accurately model the data. At this point, the shower simulation begins by taking the final state particles from the matrix element and ISR/FSR and simulating their decay/hadronization. This step cannot be calculated with perturbative QCD. It is instead handled by fragmentation

[4]While the signal and dominant background MC is produced at NLO, some smaller backgrounds are generated only to leading-order precision.

functions which attempt to mimic a real hadronization as accurately as possible. At this stage, the event generation portion of the MC production (called the "GEN" step) is complete and the interaction with the detector must be simulated. This simulation is carried out with the GEANT4 software package, which models the CMS detector, as well as the GEN particles hits, tracks, energy deposits and other interactions that occur between the particles produced from the GEN step and the CMS detector simulation (called the "SIM" step). Next, the simulated hits are converted into a digitized detector response, which is designed to recreate the response of the actual detector as closely as possible. This stage, called the "DIGI" step, also includes the emulated trigger accept/reject decisions and is the first place where MC is directly comparable to data. The final step[5] called "RECO", is the reconstruction of high-level objects such as jets, tracks, and MET from hits and energy deposits. This RECO step is functionally identical between the real data collected by CMS and the MC simulation, allowing for a direct comparison between MC and real data at the RECO tier, and where the physics analysis begins. An overview of the entire data chain from initial interaction to analysis objects is in Figure 6.1.

### 6.2.1 Signal

The signal MC sample is produced with the POWHEG generator at NLO [**?** ], assuming a Higgs boson mass of 125 GeV and using the NNPDF3.0 for the proton PDFs [**?** ]. The shower process is simulated by PYTHIA [**?** ] with the CUETP8M2T4 tune to improve the modeling of the underlying event [**?** ]. The signal sample considers all decays of the Higgs boson, except those to b quarks, and the cross section listed for this sample accounts for the modified branching fraction. This is useful because no resources are wasted filtering and storing events that would

[5]Technically there is an additional step where RECO data is converted into Analysis Object Data (AOD). This includes a slimming of attributes w.r.t. RECO, and reconstructing high-level objects used in the analysis.

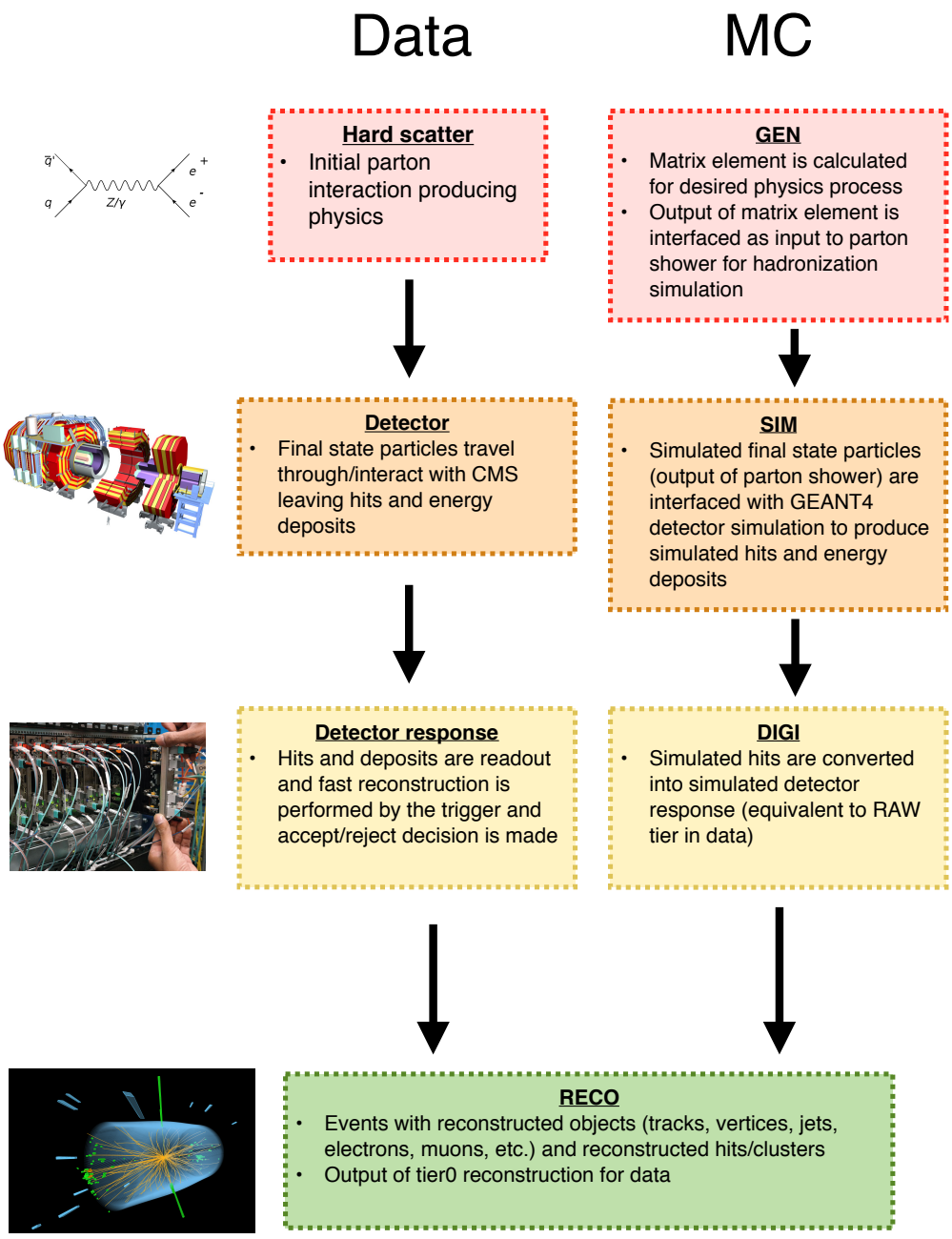



Figure 6.1. An overview and comparison of the data taking and MC generation stages. [**?** ].

not pass the signal regions.

### 6.2.2 Backgrounds

The majority[6] of the background predictions in this analysis are MC-based estimations. The leading MC background estimations are generated with MadGraph5_aMCNLO with shower simulated by PYTHIA [**? ? ?** ]. The remaining backgrounds are generated and showered with one of the generators and shower MC already mentioned.

[6]The backgrounds estimated via data-driven methods are described in detail later.

TABLE 6.2

Monte-Carlo samples used in this analysis. The first section is the sample used for signal prediction, the second section contains the largest backgrounds that are predicted by MC, the third section contains the samples for other, non-dominant SM backgrounds, the fourth section lists the rare SM backgrounds, and the final section contains the samples used in control regions, that don't directly enter the final yields or discriminant shapes.



| Process | MC sample | Cross-section [pb] |
|---|---|---|
| $t\bar{t}H$ | /ttHJetToNonbb_M125_13TeV_amcatnloFXFX_madspin_pythia8_mWCutfix/[3] | $2.15 \times 10^{-1}$ |
| $t\bar{t}W$ | /TTWJetsToLNu_TuneCUETP8M1_13TeV-amcatnloFXFX-madspin-pythia8/[2,5] | $2.04 \times 10^{-1}$ |
| $t\bar{t}Z$ | /TTZToLLNuNu_M-10_TuneCUETP8M1_13TeV-amcatnlo-pythia8/[3] | $2.73 \times 10^{-1}$ |
| $t\bar{t}Z$ | TTZToLL_M-1to10_TuneCUETP8M1_13TeV-madgraphMLM-pythia8/[3] | $4.93 \times 10^{-2}$ |
| $t\bar{t} + \gamma$+jets | /TTGJets_TuneCUETP8M1_13TeV-amcatnloFXFX-madspin-pythia8/[1] | 3.70 |
| $t + \gamma$+jets | /TGJets_TuneCUETP8M1_13TeV_amcatnlo_madspin_pythia8/[1] | 2.97 |

TABLE 6.2

*Continued*

| Process | MC sample | Cross-Section [pb] |
|---|---|---|
| $W+\gamma$+jets | /WGToLNuG_TuneCUETP8M1_13TeV-amcatnloFXFX-pythia8/[3] | $5.86 \times 10^{2}$ |
| $Z/\gamma$+jets | /ZGTo2LG_TuneCUETP8M1_13TeV-amcatnloFXFX-pythia8/[3] | $1.31 \times 10^{2}$ |
| $WW$ | /WWTo2L2Nu_13TeV-powheg/[1] | $1.05 \times 10^{1}$ |
| $WW$ | /WWTo2L2Nu_DoubleScattering_13TeV-pythia8/[1] | $1.73 \times 10^{-1}$ |
| $WW$ | /WpWpJJ_13TeV-powheg-pythia8/[2] | $3.71 \times 10^{-2}$ |
| $WZ$ | /WZTo3LNu_TuneCUETP8M1_13TeV-powheg-pythia8[1] | 4.43 |
| $ZZ$ | /ZZTo4L_13TeV-amcatnloFXFX-pythia8/[3] | 1.26 |
| $WWW$ | /WWW_4F_TuneCUETP8M1_13TeV-amcatnlo-pythia8/[1] | $2.09 \times 10^{-1}$ |
| $WWZ$ | /WWZ_TuneCUETP8M1_13TeV-amcatnlo-pythia8/[1] | $1.65 \times 10^{-1}$ |
| $WZZ$ | /WZZ_TuneCUETP8M1_13TeV-amcatnlo-pythia8/[1] | $5.57 \times 10^{-2}$ |
| $ZZZ$ | /ZZZ_TuneCUETP8M1_13TeV-amcatnlo-pythia8/[1] | $1.40 \times 10^{-2}$ |
| $t+Z$ | /tZq_ll_4f_13TeV-amcatnlo-pythia8/[3] | $7.58 \times 10^{-2}$ |

TABLE 6.2

*Continued*

| Process | MC sample | Cross-Section [pb] |
|---|---|---|
| $tt\bar{t}\bar{t}$ | /TTTT_TuneCUETP8M2T4_13TeV-amcatnlo-pythia8/[1] | $9.10 \times 10^{-3}$ |
| $t\bar{t}$+jets | /TTJets_DiLept_TuneCUETP8M1_13TeV-madgraphMLM-pythia8/[1,3] | $8.73 \times 10^{1}$ |
| $t\bar{t}$+jets | /TTJets_SingleLeptFromT_TuneCUETP8M1_13TeV-madgraphMLM-pythia8/[1] | $1.82 \times 10^{1,2}$ |
| $t\bar{t}$+jets | /TTJets_SingleLeptFromTbar_TuneCUETP8M1_13TeV-madgraphMLM-pythia8/[3] | $1.82 \times 10^{1,2}$ |
| $Z/\gamma^* \to ll$ | /DYJetsToLL_M-10to50_TuneCUETP8M1_13TeV-amcatnloFXFX-pythia8/[1] | $6.03 \times 10^{3}$ |
| $Z/\gamma^* \to ll$ | /DYJetsToLL_M-50_TuneCUETP8M1_13TeV-madgraphMLM-pythia8/[4] | $5.77 \times 10^{3}$ |
| $W$+jets | /WJetsToLNu_TuneCUETP8M1_13TeV-amcatnloFXFX-pythia8/[1] | $6.15 \times 10^{4}$ |
| $t/\bar{t}$ | /ST_s-channel_4f_leptonDecays_13TeV-amcatnlo-pythia8_TuneCUETP8M1/[1] | 3.68 |
| $t$ | /ST_t-channel_top_4f_inclusiveDecays_13TeV-powhegV2-madspin-pythia8_TuneCUETP8M1/[1] | $1.36 \times 10^{2}$ |
| $\bar{t}$ | /ST_t-channel_antitop_4f_inclusiveDecays_13TeV-powhegV2-madspin-pythia8_TuneCUETP8M1/[1] | $8.01 \times 10^{1}$ |
| $tW$ | /ST_tW_top_5f_inclusiveDecays_13TeV-powheg-pythia8_TuneCUETP8M1/[3] | $3.56 \times 10^{1}$ |
| $\bar{t}W$ | /ST_tW_antitop_5f_inclusiveDecays_13TeV-powheg-pythia8_TuneCUETP8M1/[3] | $3.56 \times 10^{1}$ |

[1] RunIISummer16MiniAODv2-PUMoriond17_80X_mcRun2_asymptotic_2016_TrancheIV_v6-v1/MINIAODSIM

[2] RunIISummer16MiniAODv2-PUMoriond17_80X_mcRun2_asymptotic_2016_TrancheIV_v6_ext2-v1/MINIAODSIM

[3] RunIISummer16MiniAODv2-PUMoriond17_80X_mcRun2_asymptotic_2016_TrancheIV_v6_ext1-v1/MINIAODSIM

[4] RunIISummer16MiniAODv2-PUMoriond17_80X_mcRun2_asymptotic_2016_TrancheIV_v6_ext1-v2/MINIAODSIM

[5] RunIISummer16MiniAODv2-PUMoriond17_80X_mcRun2_asymptotic_2016_TrancheIV_v6_ext1-v3/MINIAODSIM

## 6.3 Triggers

The triggers used in this analysis are choosen to select events with one or more leptons, which are the events with the greatest chance of passing the signal region selection. We consider triggers that fire on events with one, two, or three leptons. The primary requirement and defining characteristic of the lepton triggers is the $p_T$ of the lepton. Although all signal regions require at least two leptons, considering events where only a single-lepton trigger fired boosts acceptance, since we consider events with two or more leptons, but where the sub-leading (in terms of $p_T$) lepton failed to pass the requirements of the double or triple lepton trigger. To further boost event acceptance, we use the logical 'OR' of all trigger decisions in each signal region category described in Table 6.3. While additional single, double, and triple lepton triggers are available, in the CMS HLT menu, we use *only* the lowest-threshold, unprescaled triggers to maximize efficiency and to make the luminosity calculation and trigger efficiency measurements straightforward.

TABLE 6.3

Triggers used in this analysis.

| $2lss$ $(\mu\mu)$ |
|---|
| HLT_Mu17_TrkIsoVVL_Mu8_TrkIsoVVL_DZ_v* |
| HLT_Mu17_TrkIsoVVL_TkMu8_TrkIsoVVL_DZ_v* |
| HLT_IsoMu22_v* |
| HLT_IsoTkMu22_v* |
| HLT_IsoMu22_eta2p1_v* |
| HLT_IsoTkMu22_eta2p1_v* |
| HLT_IsoMu24_v* |

| HLT_IsoTkMu24_v* |
|---|
| $2lss$ $(ee)$ |
| HLT_Ele23_Ele12_CaloIdL_TrackIdL_IsoVL_DZ_v* |
| HLT_Ele27_WPTight_Gsf_v* |
| HLT_Ele25_eta2p1_WPTight_Gsf_v* |
| HLT_Ele27_eta2p1_WPLoose_Gsf_v* |
| $2lss$ $(e\mu)$ |
| HLT_Mu23_TrkIsoVVL_Ele8_CaloIdL_TrackIdL_IsoVL_v* |
| HLT_Mu23_TrkIsoVVL_Ele8_CaloIdL_TrackIdL_IsoVL_DZ_v* |
| HLT_Mu8_TrkIsoVVL_Ele23_CaloIdL_TrackIdL_IsoVL_v* |
| HLT_Mu8_TrkIsoVVL_Ele23_CaloIdL_TrackIdL_IsoVL_DZ_v* |
| HLT_IsoMu22_v* |
| HLT_IsoTkMu22_v* |
| HLT_IsoMu22_eta2p1_v* |
| HLT_IsoTkMu22_eta2p1_v* |
| HLT_IsoMu24_v* |
| HLT_IsoTkMu24_v* |
| HLT_Ele23_Ele12_CaloIdL_TrackIdL_IsoVL_DZ_v* |
| HLT_Ele27_WPTight_Gsf_v* |
| HLT_Ele25_eta2p1_WPTight_Gsf_v* |
| HLT_Ele27_eta2p1_WPLoose_Gsf_v* |

The trigger efficiency is measured separately in data and MC. Measuring the trigger efficiency accurately and precisely is critically important in a sensitive physics analysis. This is because the trigger efficiency directly scales the luminosity, and thus the number of events in the counting experiment. Because this efficiency differs slightly between data and MC, we correct the MC efficiency with scale factors to

match the efficiency in data. The trigger efficiency measured in data is performed by selecting data that was recorded on triggers completely uncorrelated with the lepton triggers used in this analysis. The data used for the efficiency measurement is recorded by triggers in the MET dataset so there is no implicit bias in the sample, because the triggers that collected the data are different from the triggers for which we are measuring the efficiency. With the MET dataset, we select events with two and three loose leptons separately, we then measure the efficiency for the candidate events to pass the 'OR' of all triggers in the respective category. This efficiency can then be compared to that measured in the MC, and we obtain the scale factors listed in Table 6.4. We observe an excellent agreement between the efficiencies measured in data to that in MC.

TABLE 6.4

Trigger efficiency scale factors and uncertainties.

| Category | Scale Factor |
|---|---|
| $2lss(ee)$ | 1.01±0.02 |
| $2lss(e\mu)$ | 1.01±0.01 |
| $2lss(\mu\mu)$ | 1.00±0.01 |

The triggers are the first place where the object selection is introduced into the analysis, and while the efficiency is high and a good agreement between data and MC is observed, the object selection defined in the analysis must be strictly tighter than the object selection defined in any of the triggers. This ensures the efficiency

is well-measured and does not vary significantly with lepton $p_T$ or $|\eta|$. That is, the efficiency is measured past the "turn-on", and in the "plateau" of efficiency[7] as seen in Figures 6.26.36.4.

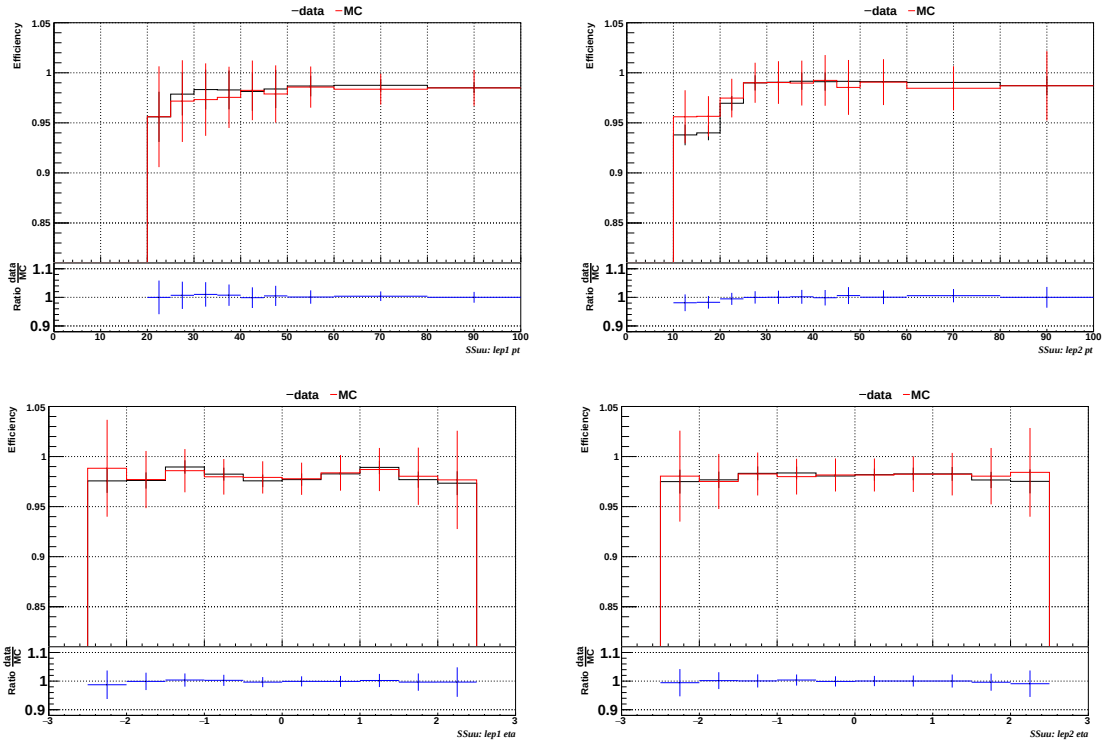



Figure 6.2. Comparison of the trigger efficiency in the same-sign dimuon category before corrections, shown as a function of the $p_T$ and $\eta$ of the leading lepton (left) and the sub-leading lepton (right).

[7]Note the cuts on $p_T$ at 20,10 for leading and subleading leptons, and $|\eta|$ ¡ 2.5 remove most of the turn-on in Figures 6.26.36.4.

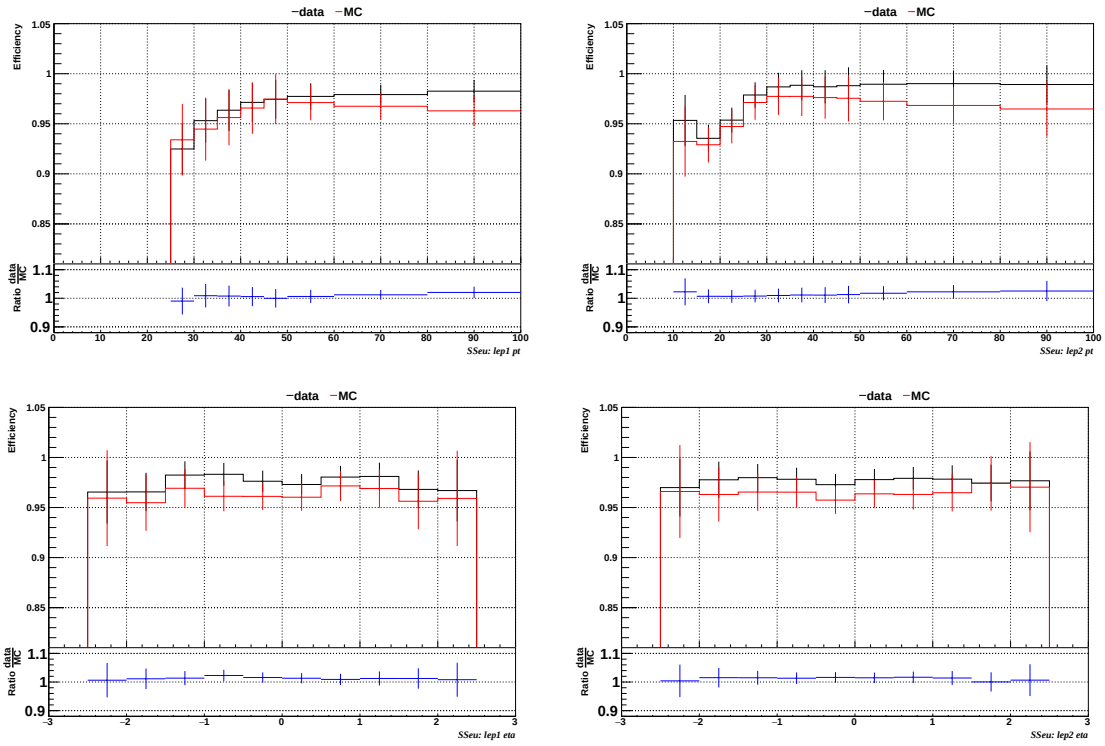



Figure 6.3. Comparison of the trigger efficiency in the same-sign muon+electron category before corrections, shown as a function of the $p_T$ and $\eta$ of the leading lepton (left) and the sub-leading lepton (right).

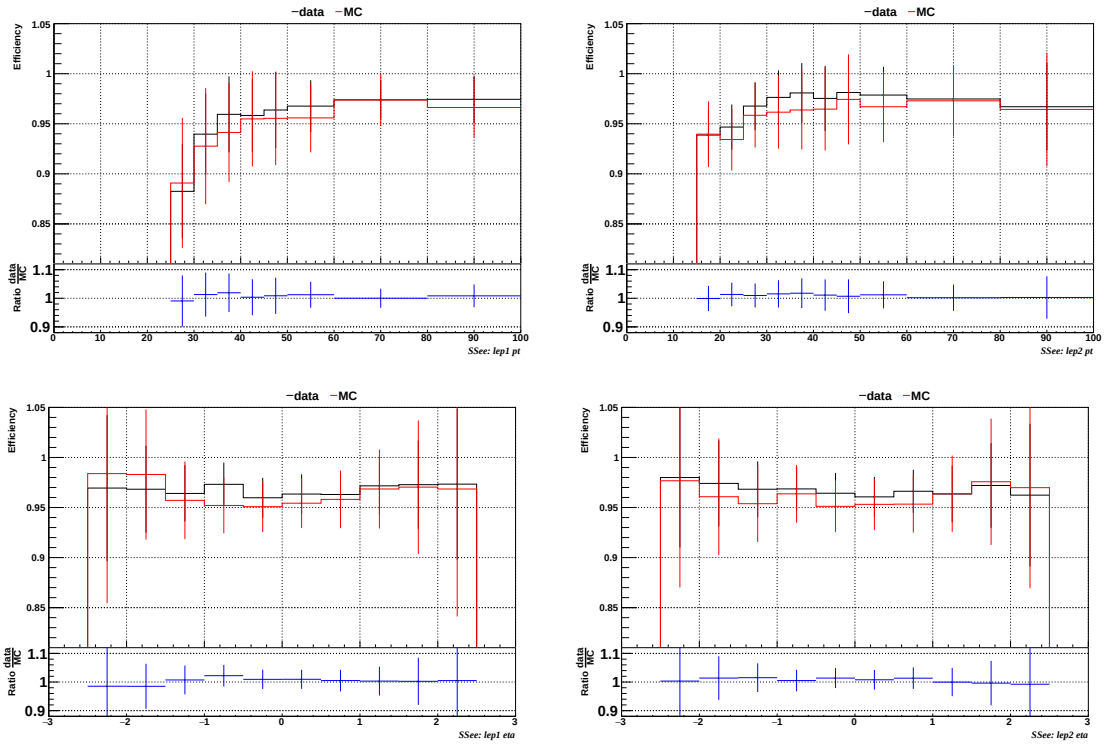



Figure 6.4. Comparison of the trigger efficiency in the same-sign dielectron category before corrections, shown as a function of the $p_T$ and $\eta$ of the leading lepton (left) and the sub-leading lepton (right).

# CHAPTER 7

# BACKGROUND PREDICTIONS

Despite optimizing the event selection criteria to select signal events, a substantial amount of background from a variety of processes enters the signal region. These backrounds are classified as being either reducible or irreducible, and are estimated differently depending on this classification. An understanding of each background process and a proper assessment of the uncertainity associated with the estimation of each background is critical to extracting the signal and interpreting the results.

## 7.1 Reducible backgrounds

Reducible backgrounds arise from a number of sources, but always contain leptons that are either non-prompt, or have a lepton whose electric charge is mismeasured. These backgrounds are classified as reducible because if the event selection and CMS object reconstruction worked with perfect efficiency, the events would not enter the signal region; thus, improving the prompt lepton identifiction and CMS lepton electric charge measurement can reduce the contributions from these processes. Reducible backgrounds are estimated via data-driven approaches using control regions and extrapolation techniques to predict their contribution to the yield in the signal region. The background due to fakes is entirely separate from the charge mismeasurement background and are estimated separately.

### 7.1.1 Fake lepton background

The fake lepton background gets its name from the fact that events with non-prompt leptons pass the tight selection and enter the signal region by faking prompt leptons. These fakes typically originate from leptonic decays of certain hadrons, such as the B, D, $\Lambda$, and K. The primary source is the relatively large cross-section semi-leptonic $t\bar{t}$ process, where the b-jet from the leptonic top quark produces a fake lepton that passes the tight lepton selection criteria, but also includes other processes where a lepton is produced inside a jet. The background from these events is estimated via a loose-to-tight extrapolation. This begins with measuring the rate at which the leptons passing the fakeable selection also pass the tight selection, known as the fake rate. The measurement is performed in a control region of the data, known as the measurement region (see below). The measurement is then used to extrapolate from a sideband with fakeable leptons to estimate the contribution in the signal region from events with fake leptons.

The measurement region is heavily enriched in QCD multijet events, which provide a source of mostly fake leptons. The fake rate is defined as the probability of a non-prompt lepton which passes the fakeable selection to also pass the tight selection. The measurement region events satisfy the following requirements:

- exactly one fakeable lepton
- one preselected jet with R $>$ 0.7 from the lepton
- $M_T(l, \mathrm{E_T^{miss}}) < 15$ GeV

This selection enriches the measurement region with non-prompt leptons. The data analyzed in the measurement region is collected on single lepton triggers which require a single lepton and a particle-flow jet with $\mathrm{p_T} > 30$ GeV.

To ensure a high purity of fake leptons in the measurement region, the contribution of prompt leptons must be accounted for. This contribution arises from W and Z plus

jets processes, but also from $t\bar{t}$. The Z contamination is addressed by vetoing events with more than one loose lepton. The prompt leptons from W decays is subtracted with a fit to the transverse mass of the lepton and missing energy, $M_T(l, E_T^{miss})$. A cut is first applied requiring $M_T(l, E_T^{miss})$ <15 GeV, with the residual contamination subtracted in each $p_T$ bin using W/Z +jet MC.

The final fake rates are shown in Figure 7.1. The larger uncertainties for higher $p_T$ leptons is due to the larger uncertainties from the prompt lepton subtraction method. A good agreement between data and MC is observed overall. In this analysis, electrons from photon conversions are treated as a separate background and estimated via MC. The fake rate in the measurement region for electrons is scaled by the ratio of fake rate including electrons from conversions to the fake rate excluding electrons from conversions from QCD MC.

Once the fake rate is obtained from the measurement region, it is applied in a second control region, also enriched in fakes, denoted as the application region. The weighted yields in this region constitute the background due to fake leptons in the signal region. As such, the application region is identical to the signal region, except that the requirement of the two same-sign leptons passing the tight selection is relaxed to passing the fakeable selection, and that at least one of these leptons fails the tight selection. The fake rate weights are expressed in terms of event yields as a function of the number of leptons failing the tight selection in a given event. The contribution of fakes in the signal region is estimated using equation 7.1 below:

$$N_{pp}^{bkg} = \frac{f_1}{1-f_1} N_{pf} + \frac{f_2}{1-f_2} N_{pf} - \frac{f_1 f_2}{(1-f_1)(1-f_2)} N_{ff} \tag{7.1}$$

under the assumption that the contribution of prompt leptons failing the tight selection is negligible with respect to the number of non-prompt leptons failing. Here, $N_{pp}^{bkg}$ is the background contribution from fake leptons in the signal region, $f_{1,2}$ is the fake rate for the leading, subleading lepton calculated from the measurement region,

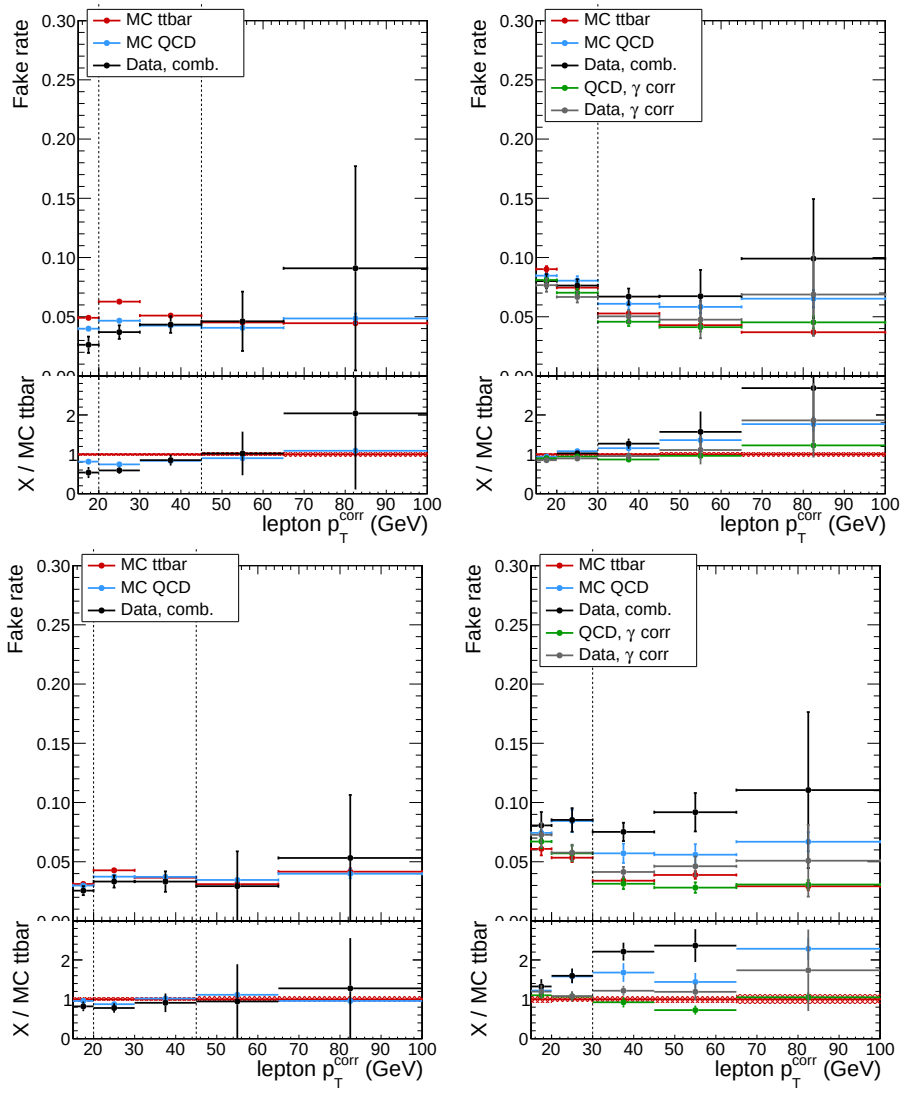



Figure 7.1. The lepton fake rates as measured in data and QCD as well as $t\bar{t}$ MC. Fake rates for muons are on the left while fake rates for electrons are on the right. The fake rates measured in the barrel are on the top while the fake rates measured in the endcaps are on the bottom.

$N_{pf}$ is the number of events in the application region with 1 prompt and 1 fake lepton, and $N_{ff}$ is the number of events in the application region with two fake leptons.

The fake lepton background estimation method described above is valid only for leptons which pass the trigger, because the fake rate was measured using leptons passing the trigger. This is important because although the measurement region contains only leptons which pass the trigger, the application region consists of some events which contain one lepton passing the trigger and one that fails the trigger. The method described above biases these events because the fake rate assigned to the lepton not passing the trigger is incorrect. Because this background estimation method relies on the fake rates in the measurement region and the application region to be the same, the fakeable object definition must not depend on whether or not the lepton passed the trigger. To remove this bias, we use a quantity called the "corrected" $p_T$ in place of the standard $p_T$. The corrected $p_T$ is the same as the standard $p_T$ if the lepton passes the tight selection, but modified to 0.9 times the nearest jet $p_T$ otherwise. The trigger bias was explicitly checked in lepton-enriched QCD MC events, where the fake rate is studied with and without requiring the lepton to pass the trigger. This study showed that the trigger turn-on curve requires the trigger $p_T$ threshold to be significantly lower than the corrected $p_T$ to avoid any bias. Thus the fake rate is measured independently in bins of corrected lepton $p_T$ for events collected by each trigger in the measurement region. The triggers used to avoid any bias in the corrected $p_T$ are listed with each $p_T$ bin below in Table 7.1.

### 7.1.2 Charge mismeasurement background

One of the defining requirements of the signal region is that the two leptons have the same-sign electric charge. This requirement makes leptons where the charge was mismeasured an important background. Like the background due to fake leptons,

TABLE 7.1

Corrected $p_T$ range and corresponding trigger categories for each bin of the fake rate measurement.

| Corrected $p_T$ [GeV] | Trigger |
|---|---|
| 10< $p_T(\mu)$ <20 | HLT_Mu3_PFJet40 |
| 20< $p_T(\mu)$ <45 | HLT_Mu8 |
| $p_T(\mu)$ >45 | HLT_Mu17 |
| 15< $p_T(e)$ <20 | HLT_Ele8_CaloIdM_TrackIdM_PFJet30 |
| 20< $p_T(e)$ <30 | HLT_Ele12_CaloIdM_TrackIdM_PFJet30 |
| $p_T(e)$ >30 | HLT_Ele17_CaloIdM_TrackIdM_PFJet30 |

this background is also estimated from data with two control regions: the first for measuring the rate of charge mismeasurments (also known as charge flips), and the second for applying the weights derived from the first to extrapolate to the signal region. Unlike the fake background however, the charge mismeasurement background is only comprised of events with electrons, as the probability for mismeasuring the electric charge of a muon is negligible.

The control region used for measuring charge flip probabilities is defined by selecting Z→ee events in data. Here it is assumed that same-sign electron pairs within 10 GeV of the Z peak are due to a charge mismeasurement of one of the electrons. Charge flip probability measurements are performed by measuring same-sign yields to opposite sign yields in the Z peak and are parameterized as a function of electron $p_T$ and $\eta$. The yields are determined from a fit to the Z invariant mass shape, which is modeled with the convolution of a crystal ball and Breit-Wigner function for the signal and an exponentially falling function for the background. The measurement is performed in 3 bins of $p_T$ (10-25 GeV, 25-50 GeV, and ≥50 GeV), and two bins of $\eta$

($\leq 1.479$, and $> 1.479$) for a total of 21 categories of electron pairs. Each category corresponds to the $p_T$-$\eta$ bin of the leading lepton, and the $p_T$-$\eta$ bin of the subleading electron. The charge flip probabilities are determined from a simultaneous fit to the 21 same-sign and opposite-sign yields. The resulting charge flip probabilities range between approximately 0.03% in the barrel to approximately 0.4% in the endcaps and are summarized in Figure 7.2 below.

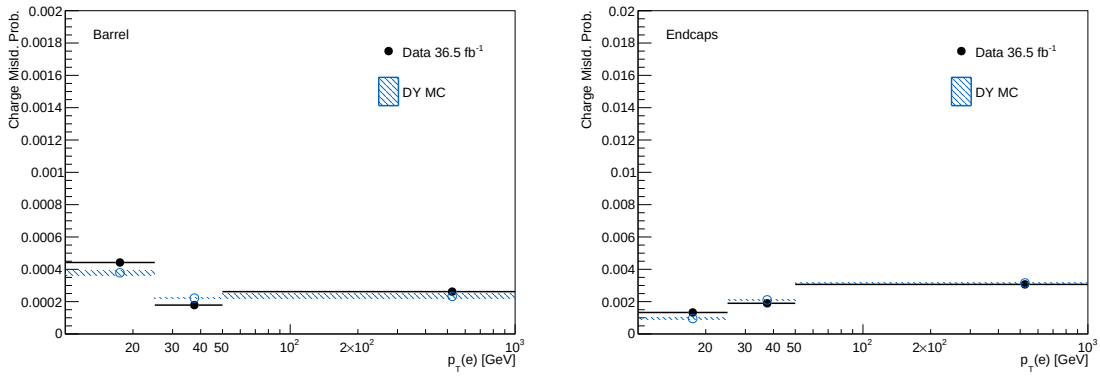



Figure 7.2. Charge misassignment probabilities as a function of $p_T$ for electrons in the barrel (left) and endcaps (right).

The control region where the charge flip probabilities measured above are applied must extrapolate well to the signal region to provide an accurate background estimation. As such, this control region is identical to the signal region except that the same-sign charge requirement on the leptons is replaced with an opposite-sign requirement. The charge misassignment probability $P(p_T, \eta)$ is applied per lepton. The total event weight is then $P_1 + P_2$ in the $ee$ category, and $P$ in the $e\mu$ category.

## 7.2 Irreducible backgrounds

Irreducible backgrounds are estimated exclusively from MC. Irreducible backgrounds earn their name from the fact that even if the signal region selection and CMS event/object reconstruction worked with perfect efficiency and purity, these backgrounds still produce the necessary objects to consistently pass the signal region selection and are thus irreducible with respect to the signal region definition. The dominant irreducible background processes include $t\bar{t}W$ and $t\bar{t}Z$. Other irreducible background processes include diboson pairs produced in association with jets, while smaller contributions include processes with a single W or Z boson, a single top quark, tribosons, as well as other rare[1] SM backgrounds. The modelling of these background processes has been checked to ensure a good agreement among the variables used to event selection and signal extraction.

While the signal region selection vetoes events with low mass dilepton pairs, the $t\bar{t}\gamma^*$ process with $\gamma^* \to l^+l^-$ still contributes as a background when one of the leptons fails selection cuts or is out of the acceptance. The nominal $t\bar{t}Z$ sample is generated with $m_{l^+l^-} > 10$ GeV, and we therefore use a $t\bar{t}Z/\gamma^*$ sample generated at $1 < m_{l^+l^-} < 10$ GeV and an additional $t\bar{t}$ sample which covers the $m_{l^+l^-} < 1$ GeV phase space. The $t\bar{t}$ sample is generated with MadGraph and showered with Pythia, which can decay a low-virtuality $\gamma^*$.

Although electrons from conversions are technically a reducible background, they are estimated from MC. This is because the background from conversions, primarily $t\bar{t}\gamma$, can have events where one conversion electron is not reconstructed, and the other is mistakenly identified as a prompt, isolated electron[2]. Because these leptons look more prompt-like compared to typical fakes, they are not estimated well with the

[1] Rare means a very small cross section, and very small yield in the signal region.

[2] When both conversion electrons are reconstructed, the conversion veto applied in the tight electron selection rejects both.

fake-rate method. For the conversion backgrounds, MC is used and normalized to NLO QCD cross section from MadGraph5_aMCNLO.

# CHAPTER 8

# SIGNAL EXTRACTION

The signal extraction procedure is the method by which a suitable variable, which discriminates signal from background, is selected and binned. Then, separate counting experiments are performed in each resulting bin. The collection of these counting experiments is then used to produce the final result, the measurement of $t\bar{t}H$ signal process. This method is executed in the signal region, and offers a distinct advantage with respect to a "cut-and-count" procedure where no discriminant is used and a single counting experiment is performed in the signal region. By dividing the signal region into bins of a discriminant, each bin has a unique signal to background ratio, with some bins having a significantly higher signal to background ratio than others. The motivation for the signal extraction procedure is that the geometric average of the signal to background ratios of the bins of the discriminant is higher than the signal to background ratio obtained with the signal counting experiment in the cut-and-count method.

The discriminant used to separate signal from background is based on a multivariate analysis algorithm called a Boosted Decision Tree (BDT). The BDT discriminant combines information from several input discriminating variables into a single variable more powerful than any single input. The composition of signal and background predictions, as well as data, in the signal region where the signal extraction is performed is in Table 8.1 below.

The signal extraction strategy adopted here targets the two largest backgrounds,

TABLE 8.1

Expected (pre-fit) yields for signal and background processes, and observed yields in data. Uncertainties shown are purely statistical.

| | $\mu\mu$ | $ee$ | $e\mu$ |
|---|---|---|---|
| $t\bar{t}W$ | 45.4 ± 0.5 | 17.8 ± 0.3 | 64.3 ± 0.6 |
| $t\bar{t}Z/\gamma^*$ | 16.8 ± 0.7 | 14.8 ± 0.8 | 41.7 ± 1.4 |
| WZ | 5.2 ± 0.7 | 1.6 ± 0.4 | 7.5 ± 0.8 |
| Rare SM. bkg | 6.8 ± 0.3 | 3.5 ± 0.2 | 11.7 ± 0.4 |
| WWss | 2.9 ± 0.2 | 1.4 ± 0.1 | 4.3 ± 0.2 |
| Conversions | 0.0 ± 0.0 | 3.4 ± 1.1 | 8.5 ± 1.3 |
| Charge flip | 0.0 ± 0.0 | 172 ± 93 | 149 ± 82 |
| Non-prompt leptons | 29.9 ± 1.2 | 17.3 ± 1.1 | 53.5 ± 1.8 |
| Total bkg | 107.3 ± 1.7 | 70.3 ± 1.8 | 208.0 ± 2.9 |
| $t\bar{t}H$ | 18.5 ± 0.2 | 7.4 ± 0.1 | 26.2 ± 0.2 |
| Data | 154 | 95 | 274 |

$t\bar{t}V$ and the non-prompt leptons, using separate BDTs to discriminate against each specific background. The inputs to these BDTs include kinematic variables, as well as outputs of other BDTs designed to resconstruct parts of the signal events such as the Higgs jets and the hadronically decaying top quark. Finally, the outputs of the BDT dscriminators are plotted on separate axes forming a two-dimensional distribution, referred to here as the 2D BDT. This shape is then binned in two dimensions forming a one-dimensional discriminant.

## 8.1 Two Dimensional BDTs

Two BDTs are used to produce the final shape template. One BDT is trained exclusively against the $t\bar{t}V$ background, while the other targets the background due to non-prompt leptons. The BDT descriminating against the non-prompt lepton background is trained against $t\bar{t}$ MC, since the $t\bar{t}$ process is the largest source of non-prompt leptons. The BDT that discriminates against $t\bar{t}V$ is trained against $t\bar{t}W$ and $t\bar{t}Z$ MC events. The MC samples used for BDT training are separate from the samples used in the final analysis which is a standard practice when using MVA discriminants. Using independent training samples is necessary to avoid biasing the MVA analysis.

A loosened version of the $2lss$ signal region selection is applied to the MC samples used for training. This loosened selection is motivated by increasing the training statistics available while not affecting the behavior and kinematics of the input variables with respect to the signal region. This loosened selection consists of the following criteria:

- At least two preselected same-sign leptons
- The lepton $p_T > 25,15$
- At least 4 preselected jets among which there must be $\geq$ 2 CSVv2 L or $\geq$ 1 CSVv2 M

Many BDT input variables were tested, and the set of variables which produced the most discriminating BDT output was selected. These inputs are listed in Table 8.2. Selecting which variables to use as inputs requires several considerations. First, since the BDTs are trained on MC and ultimately evaluated on data, every input variable must be well-modeled, i.e. have a good agreement between MC and data. Second, the candidate inputs, at least some, should have signal separation power themselves, which allows the BDT to learn this information. The final consideration is more subtle. While the separation power of individual inputs is important, selecting each input based on its separation does not necessarily produce the most discriminating BDT. Correlations can exist between different variables, and if these correlations exist in signal and background, the BDT does not learn additional information from these correlated variables. However, if two variables are correlated differently in signal than in background, the BDT makes use of this in separating signal from background. All these considerations were taken into account when empirically determining the best set of BDT inputs.

The hadronic top reconstruction score and the Higgs jet tagger score are outputs of separate event reconstruction BDTs, which aim to match jets to the hadronically decaying top quark and the hadronically decaying daughter of the Higgs respectively. These BDT outputs provide discriminating input variables to the 2D BDT. The details of these reconstruction BDTs are described in the following subsections. The 2D BDT input variable distributions of $t\bar{t}H$, $t\bar{t}$, and $t\bar{t}V$ are in Figures 8.1, 8.2, 8.4. The distributions are normalized to equal area to visualize separation between signal and background. For details on the training of all BDTs described in this section, as well as a description of the BDT technique itself, see Appendix A.

TABLE 8.2

Input variables used for the BDTs targeting $t\bar{t}$ and $t\bar{t}V$.

| $t\bar{t}$ | $t\bar{t}V$ |
|---|---|
| Maximum absolute $\eta$ of the two leptons (max lepton $\|\eta\|$) | |
| Number of hadronic jets (nJets) | |
| Minimum $\Delta$R between the leading lepton and nearest jet | |
| Minimum $\Delta$R between the trailing lepton and nearest jet | |
| Transverse mass of the leading lepton and $E_T^{miss}$ ($M_T(E_T^{miss}, lep_1)$) | |
| The hadronic top reconstruction score | The Higgs jet tagger score after hadronic top removal |
| - | The leading lepton corrected transverse momentum |
| - | The trailing lepton corrected transverse momentum |

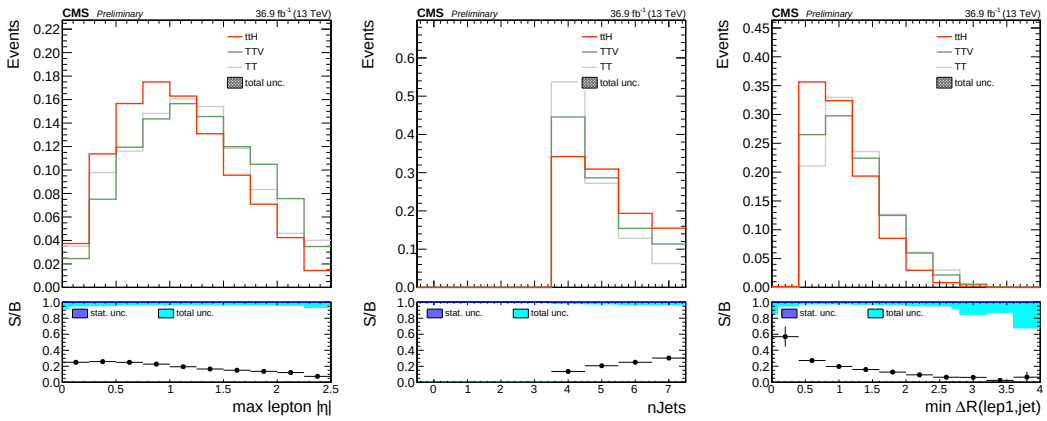



Figure 8.1. Normalized BDT input variable distributions for $t\bar{t}H$ and largest backgrounds

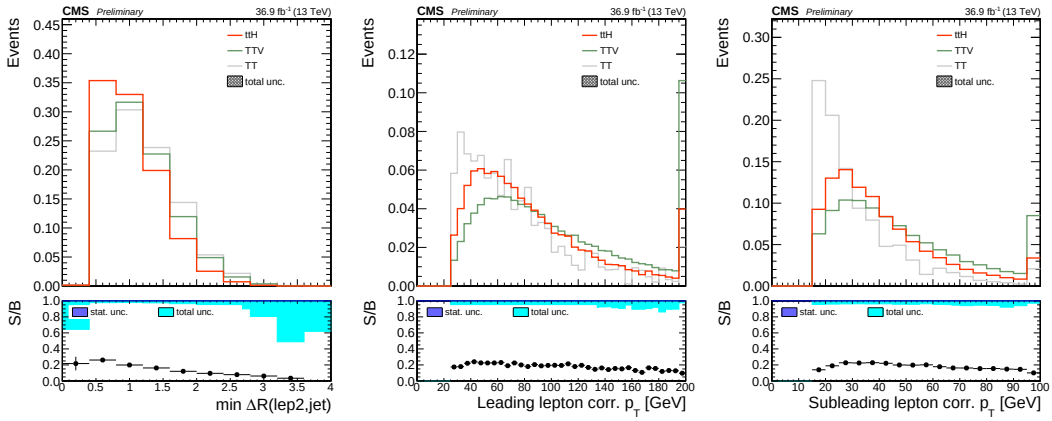



Figure 8.2. Normalized BDT input variable distributions for signal and largest backgrounds

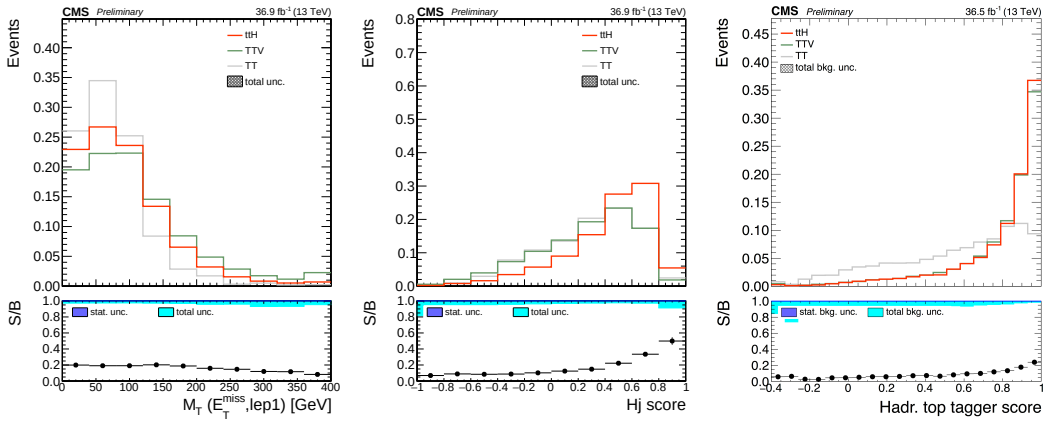



Figure 8.3. Normalized BDT input variable distributions for signal and largest backgrounds

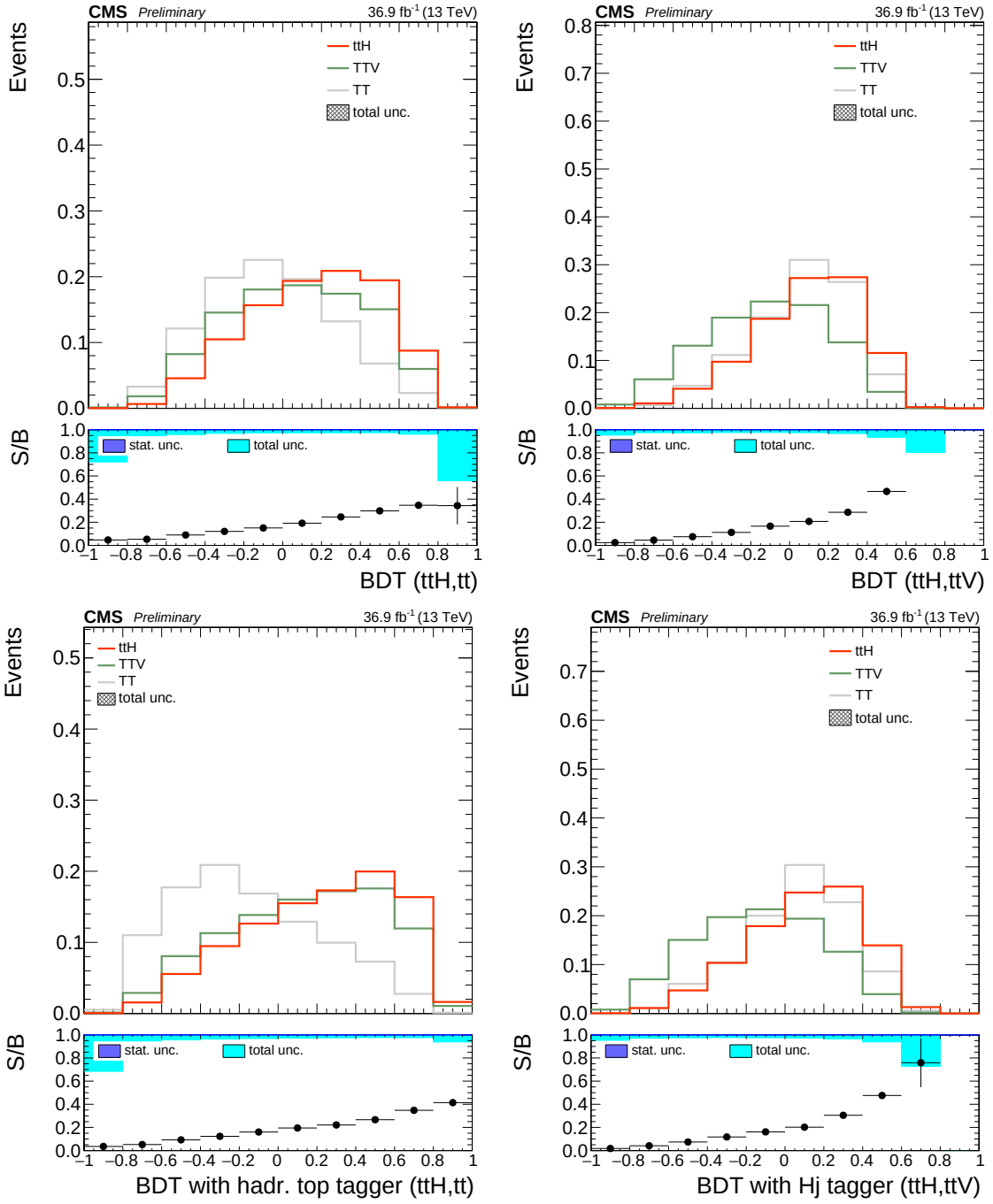



Figure 8.4. The BDT output scores without the reconstruction inputs (top) and with (bottom).

### 8.1.1 Hadronic Top Reconstruction

The hadronic top reconstruction input variable for the 2D BDT is the BDT output score of a special BDT, known as the hadronic top BDT, that aims to correctly match jets in each event with the final state particles associated with the hadronic top in a a $t\bar{t}H$ event. The hadronic top BDT is evaluated on every unique set of 3 jets in each event, with the highest-scoring combination being designed as the hadronic top part of the event. The hadronic top BDT score gauges the likelihood that a jet combination is consistent with a hadronic top in a $t\bar{t}H$ event, offering discrimination power against events without hadronically decaying top quarks, such as semi-leptonic $t\bar{t}$.

This hadronic top reconstruction targets the 2lss category, specifically where the Higgs decays to $W$ bosons. In the 2lss category, this means that one lepton originates from the top system, and the other from the Higgs. For the jets, one of the $W$ bosons from the Higgs decays hadronically, and one of the top quarks decays hadronically. In total, there are two b-jets (one from each top) and four light-flavor jets from the hadronic $W$ decays originating from the Higgs and the hadronic top. This specific decay is depicted by the diagram in Figure 1.1, and is the most common $t\bar{t}H$ decay in the 2lss channel.

#### 8.1.1.1 Training

The hadronic top BDT is trained using the $t\bar{t}H$ MC powheg signal sample used in training the final discrimination BDTs described previously.

The signal for training is $t\bar{t}H$ MC events, where a matching was performed on the final state objects, ensuring they originate from the $t\bar{t}H$ part of the event, and not from ISR/FSR or spurious detector signals in the detector simulation. This process is called gen-matching, because the objects are required to originate from the event generator itself, and not the detector simulation which follows. These events

must pass the same loosened training selection described in Section 8.1, with one important difference. In the selection used here, the same-sign requirement on the leptons is dropped. This doubles the amount of events available for training and the gen-matching requirement ensures the signal kinematics are left unaffected. The 2lss event selection requires at least four jets, which means the vast majority of signal events used for training are only partially reconstructed since a full reconstruction necessitates six matched jets. Because so few events can be fully reconstructed, we must consider partial reconstructions for events that have fewer than six matched jets. The strategy for this is to use "null" jets whose four-vectors are set to zero to represent missing jets in the event. Finally we require the signal events to have two correctly-matched selected leptons, and at least four correctly-matched selected jets.

The background consists of all jet and lepton permutations of incorrectly-matched $t\bar{t}$ events. This means that in each background event, there must be at least one object that is incorrectly matched. For example when considering the object-matching to the leptonically decaying top quark, a background training event where the reco-level jet assigned to the b-jet, is actually the b-jet at gen-level, the reco-level lepton must originate from some source other than the W from the leptonically decaying top. In this way, a single $t\bar{t}$ MC background event is used multiple times, once for each distinct permutation of 6 objects, in the training. This technique produces more background training events than necessary, given the limited (in comparison) signal training statistics.

For the background, the null jets are added according to the jet multiiplicity. For events with seven or fewer selected jets, three null jets are added, for events with eight selected jets, two null jets are added, and one null jet is added for events with greater than eight selected jets. The motivation for this recipe comes from a study of $t\bar{t}H$ events in the signal region comparing the number of reco jets that are gen-matched to the $t\bar{t}H$ process, with the jet multiplicity in the event. The results of this study

are in Figure 8.5.

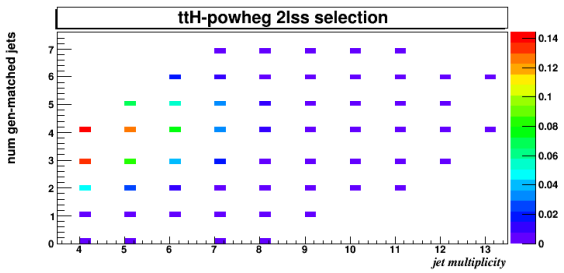



Figure 8.5. Normalized histogram of gen-matched jet multiplicity vs jet multiplicity in $t\bar{t}H$ MC in the signal region.

To reduce the computation time and improve performance, several cuts are applied at each permutation to remove unlikely reconstructions. These cuts include applying the b-tag requirement on the two jets being considered as b-jets described earlier, requiring that no reconstructed $W$ have a mass greater than 120 GeV, requiring the leptonic top mass be less than 180 GeV, and requiring the hadronic top mass be less than 220 GeV. Additionally, we ignore permutations arising from swapping two light flavor jets from the same $W$ boson, as the reconstruction is identical. An additional measure is taken to reduce iterations over each permutation, in the training step and for background only, which samples randomly from the available permutations in each event before the physics-motivated cut is applied. For each of the 8 objects, the permutation for that object skipped randomly 3 out of 10 times. For 8 objects, this means that $0.7^8 \approx 0.06$ or just 6% of the permutations for a given event are considered. After factoring in the permutations skipped due to physics cuts, this number is even smaller. It was found that the increase in statistics from removing these random cuts offers negligible improvement in performance while drastically

increasing the CPU time used during training.

The BDT uses eight input variables, consisting of:

- The CSVv2 score of the b-jet assigned to the hadronic top
- The CSVv2 score of the b-jet assigned to the leptonic top
- The transverse momentum of the reconstructed hadronic top
- The mass of the reconstructed hadronic top
- The mass of the $W$ boson from the hadronic top
- The transverse momentum ratio of the lepton from the Higgs to the lepton from the leptonic top
- The solid angle between the lepton from the leptonic top and the b-jet from the hadronic top
- The solid angle between the lepton from the leptonic top and the b-jet from the leptonic top
- The solid angle between the lepton from the Higgs and the b-jet from the leptonic top

#### 8.1.1.2 Evaluation

The hadronic top BDT is evaluated by iterating over all possible lepton and jet permutations, and selecting the highest scoring permutation as the reconstruction for each event. For this evaluation, the null jet prescription and permutation cuts used are identical to the background training. The reconstruction is designed to identify events that have a hadronic top present and discriminates against the semi-leptonic $t\bar{t}$ background. The addition of the reconstruction BDT output as an input to the BDT targeting the $t\bar{t}$ background, improves the performance by 10% by comparing ROC integrals[1] when evaluated on $t\bar{t}$ MC and compared to the equivalent BDT in the

[1]Receiver Operator Characteristic (ROC), curves quantify the separation of two distributions, such as signal and background, of a variable, such as BDT output, at all values of that variable. Every point on a ROC curve corresponds to a specific value of the variable. At each value of the variable, the fraction of one distribution that falls above that value, and the fraction of the other distribution that falls below that value form an (x,y) point on the ROC curve. The curve is then

2016 $t\bar{t}H$ ICHEP analysis [**?** ] that did not use this as an input. These performance ROC curves are below in Figure 8.6.

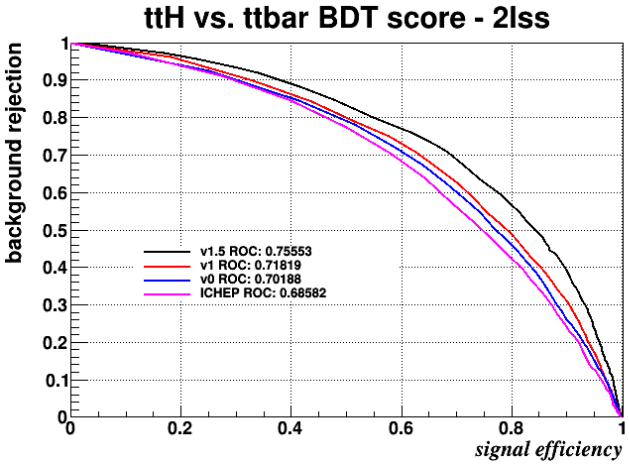



Figure 8.6. The ROC curves of the BDT targeting the $t\bar{t}$ background compared without hadronic top reconstruction input (pink) to the version used in this analysis (black). The other curves represent incremental improvements to the hadronic top BDT. The signal and background on the x and y axes are $t\bar{t}H$ and $t\bar{t}$ MC respectively. The improvement based on ROC integral by adding the hadronic top reconstruction score is 10%.

### 8.1.2 Higgs Jet Tagging

The Higgs Jet (HJ) tagger is a BDT aimed specifically at identifying hadronic jets originating from the W bosons in $t\bar{t}H$, H→WW decays consistent with the $t\bar{t}H$ diagram in Figure 1.1. This decay topology is the most common of all $t\bar{t}H$ decays in

generated by scanning over the range of the variable which contains both distributions, adding a point for each value of the variable and connecting the points with a curve. As the separation of the two distributions increases, the area under the ROC curve increases. It is this property that makes ROC curves a useful metric for evaluating BDT output performance.

the $2lss$ channel. This final state consists of 2 b-quark jets, 4 hadronic (light flavor) jets, 2 same-sign leptons, and missing energy due to neutrinos. The signal region selection described in Chapter 5 does not require all of these objects to be present, as some or many are often out of the acceptance. This means that the jets from the Higgs decay could be missing entirely from the event.

The HJ tagger addresses both the complicated jet combinatorics associated with this final state, but also the reality that not all jets originating from the Higgs are selected. The HJ tagger is an object-level discriminator that exploits jet kinematics and identification variables to estimate the likelihood of a jet originating from the Higgs. The HJ tagger calculates scores for every jet in each event, selecting the highest score as the output. To boost performance and simplify the combinatorics associated with tagging, the hadronic top tagger is first run, and the jets selected as being associated with the hadronic top are removed from further consideration for the HJ tagger. The HJ tagger is primarily aimed at selected jets from W decays (from the Higgs) and is used as a discriminating varible against both the $t\bar{t}W$ and $t\bar{t}Z$ backgrounds. Both of these backgrounds are very likely to contain most or all of the hadronic top decay products, namely a b-tagged jet and two light flavor jets. The two light flavor jets from the W in the hadronic top decay look kinematically similar to the targeted jets from the W from the Higgs, creating a potential for Higgs jet mis-tagging. This is especially true when the b-jet is missing. The hadronic top tagger attempts to identify these jets and they are removed from consideration for the HJ tagger.

The HJ tagger is trained on $t\bar{t}H$ and $t\bar{t}V$ MC. The signal training events require the objects passing the selection to be matched to a generator-level parton originating from the $H\rightarrow WW\rightarrow l\nu\ l\nu$ process. The background selection requires $t\bar{t}V$ events in the signal region. The input variables include:

- the minimum angular separation between the jet and one of the two leptons

- the maximum angular separation between the jet and one of the two leptons
- the jet transverse momenta
- the jet b-tag discriminator (CSVv2)
- the jet quark-gluon discriminator (qgid)

The qgid varible is a BDT designed to discriminate jets originating from light flavor quarks from jets originating from gluons. The performance improvement due to the HJ tagger input is seen in Figure 8.7, and corresponds to approximately 4-5% increase in signal efficiency at the same background efficiency. The data to MC comparison of both the hadronic top BDT and Higgs tagger BDT ouputs are in Figure 8.8 below.

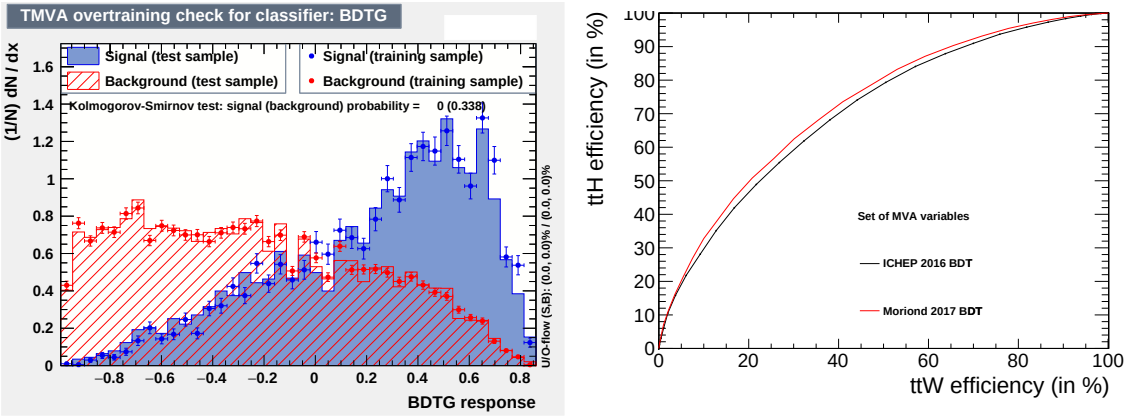



Figure 8.7. Separation power of HJ output on the training samples (right) and performance improvements with respect to the 2016 version of this analysis obtained with the HJ tagger input with hadronic top removal (left). Note the difference in axes with respect to the ROC curve for the hadronic top BDT.

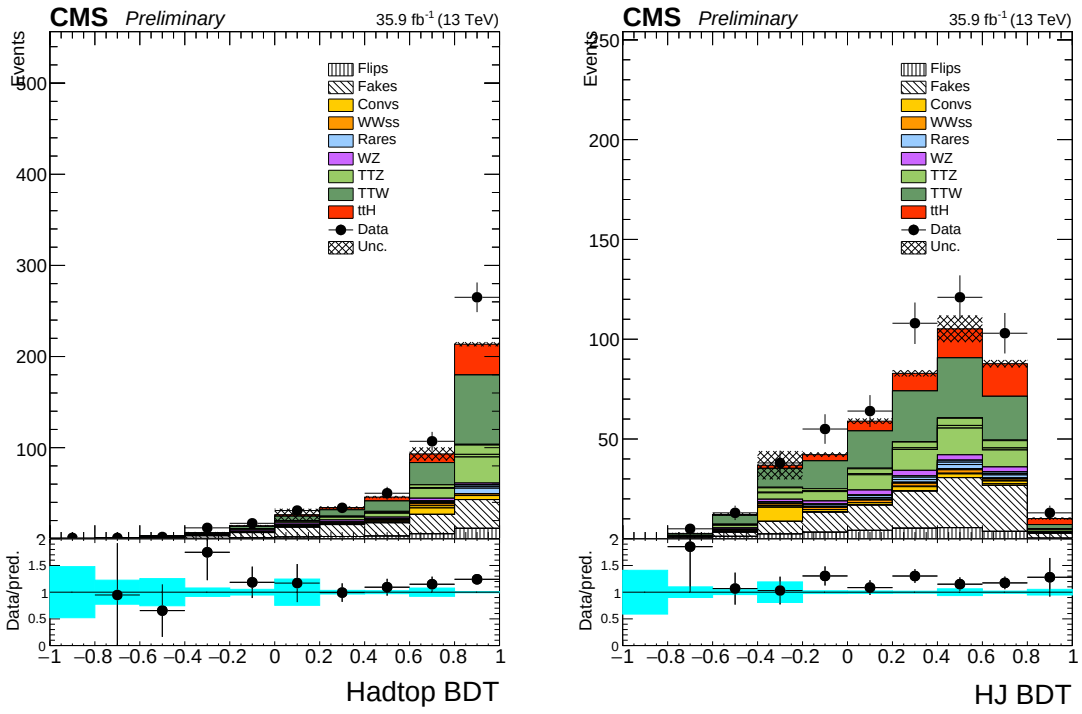



Figure 8.8. The hadronic top BDT (left) and Higgs tagger BDT (right) outputs.

## 8.2 Binning

The binning choice for the two-dimensional shape resulting from plotting each BDT on a separate axis is not straight forward, due to the large number of binning techniques available. The binning method must optimally partition the space to isolate signal from background, while simultaneously maintaining populated bins that don't suffer from low statistics. The output shapes formed by the two BDTs for the signal and largest backgrounds are in Figure 8.9 below.

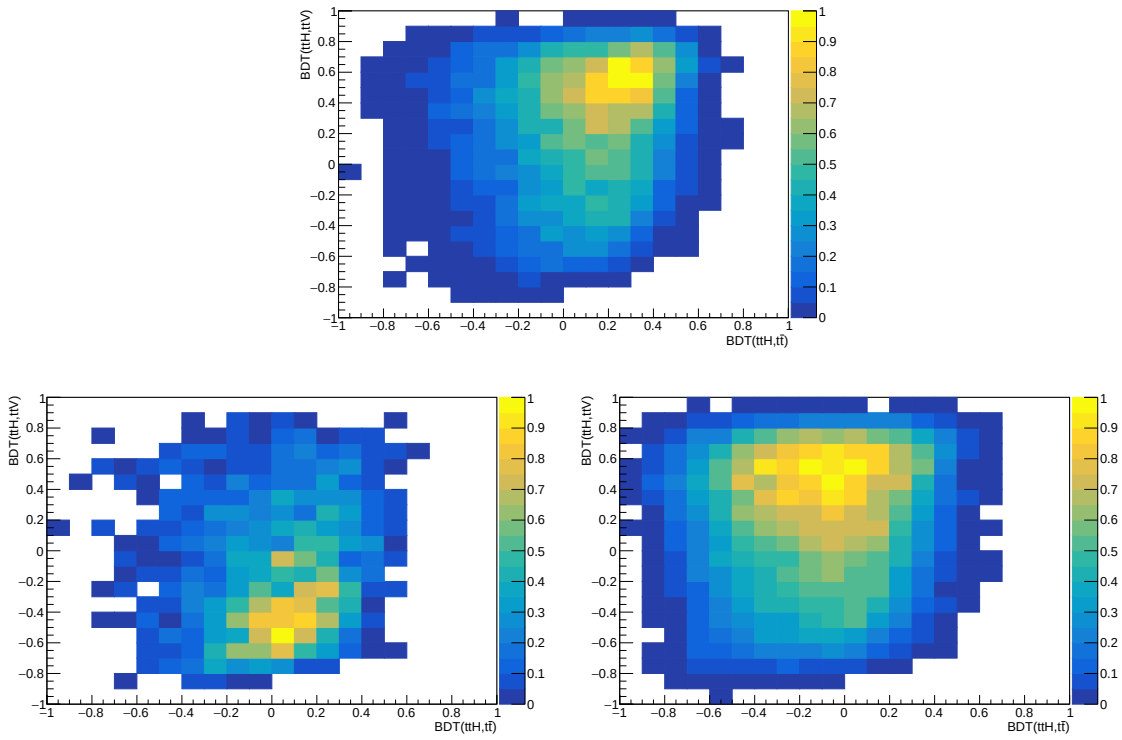



Figure 8.9. The two-dimensional output shape formed by the two BDTs for $t\bar{t}H$ (top), $t\bar{t}$ (left), and $t\bar{t}V$ (right).

While previous versions of this analysis used a simple rectangular binning that was performed by studying the signal and background shapes and drawing the bins by hand, this iteration takes a more algorithmic approach based on the signal-to-background likelihood ratio. The binning procedure was performed with signal and background MC that was not used in the main analysis. The binning process begins with the standard rectangular, 20 x 20 binning in Figure 8.9. For each bin, the signal to background likelihood ratio is computed. Next, each background event from the $t\bar{t}$ and $t\bar{t}V$ samples is assigned the likelihood ratio corresponding to the bin it populates. The resulting likelihood distribution for background events is transformed into the cumulative[2] distribution below in Figure 8.10 (left). The y-axis of the cumulative distribution function of background events is then binned evenly, achieving approximately equal amounts of background in each bin.

The bins from the cumulative distribution are then mapped back to the 2D shape. This mapping is used as the final 2D binning and is represented in Figure 8.11 (left). The corresponding 1D histogram resulting from this binning where the analysis is performed is Figure 8.11 (right). These bins are filled with separate signal and background samples from the ones used for deriving final results. The choice of the number of final bins is motivated from a cross-check procedure based on the k-means algorithm, which yields similar results [**?** ]. The pre-fit 1D shapes are in Figure 8.12.

## 8.3 Subcategorization

In addition to the signal region categorization by lepton flavor, further categorization is applied to exploit differences between signal and background. All categories are split by the sum of the lepton electric charges. This splitting helps to isolate the

[2]A cumulative distribution function of some variable $S$, evaluated at $s$, is the probability that $S$ will have a value less than or equal to $s$.

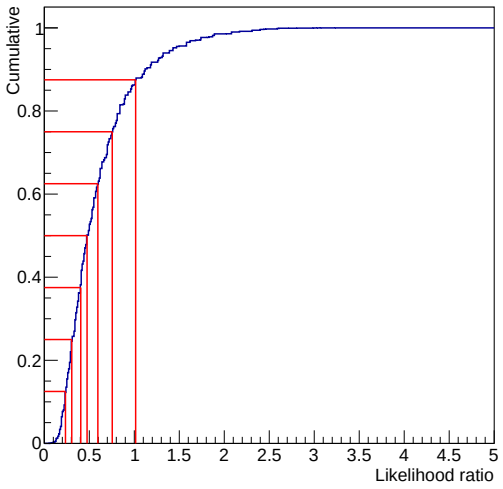



Figure 8.10. Cumulative distribution of signal-to-background likelihood ratio.

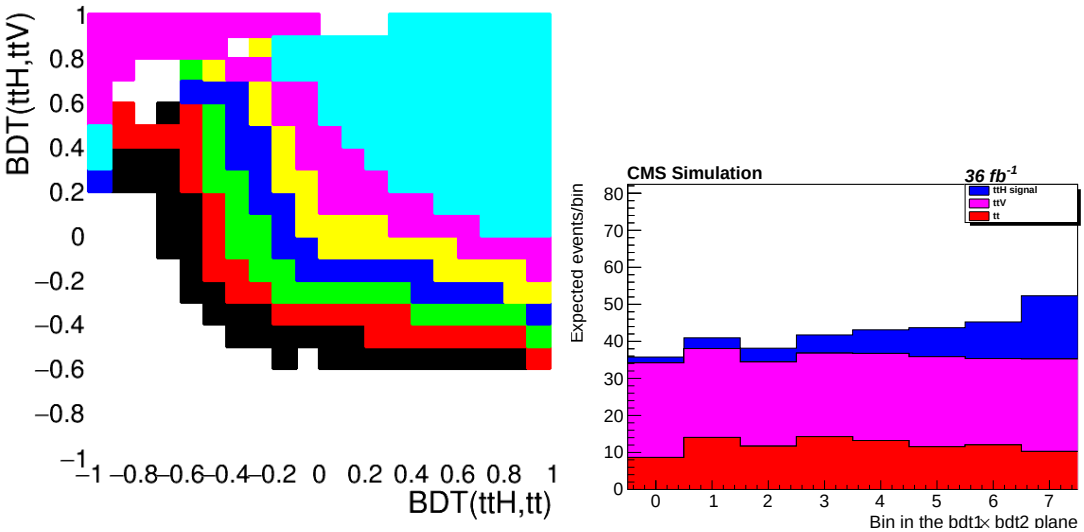



Figure 8.11. The 2D (left) and 1D (right) binning based on the cumulative likelihood distribution. Each color on the 2D histogram corresponds to a bin on the 1D histogram, white-bin0, black-bin1, red-bin2, green-bin3, blue-bin4, yellow-bin5,pink-bin6, and cyan-bin7.

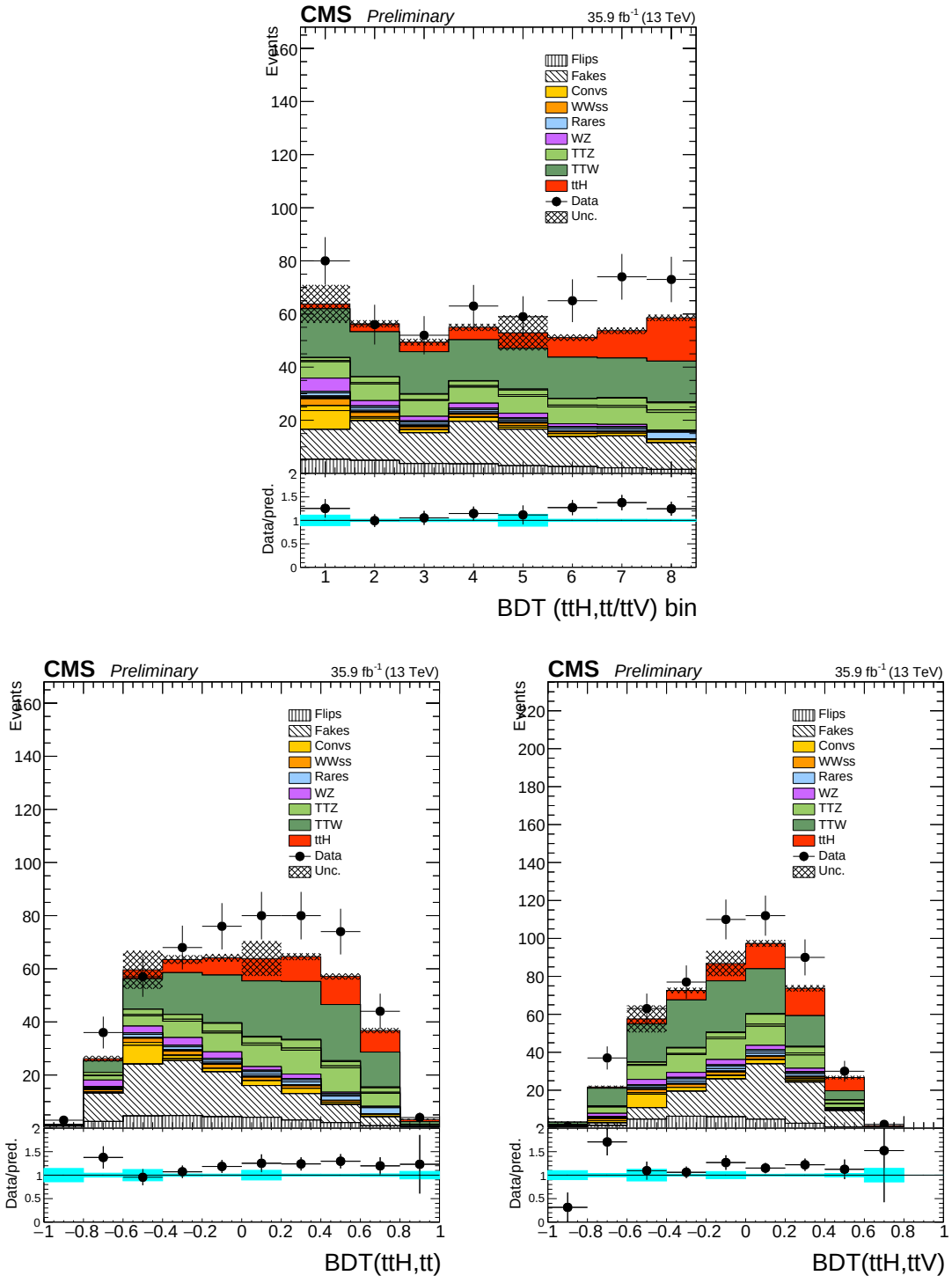



Figure 8.12. The final 1D shape used for the analysis (top), the $t\bar{t}$ BDT output (left) and the $t\bar{t}V$ BDT output (right)

.

$t\bar{t}W$ background, which is asymmetric in lepton charge sum because the initial state of the diagram involves an incoming quark and antiquark. Because the LHC collides protons whose quark content is *uud*, this favors positively charged final states, while the $t\bar{t}H$ process is initiated by gluon scattering, resulting in neutral final states, symmetric in lepton electric charge sum. The signal region categories are split further into two subcategories $b - tight$, which contains events with two b-jets passing the CSVv2 medium working point, and $b - loose$ which contains all other events (those with fewer than 2 CSVv2 M jets). This splitting helps to separate $t\bar{t}H$ from the fake lepton background which is primarily comprised of $t\bar{t}$. Due to the same-sign lepton electric charge requirement, most of the $t\bar{t}$ is vetoed, however there are a substantial amount of fakes originating from the b-decay. The b-jet is often removed from the fake lepton in the jet cleaning step, described in Section 4.4, while the fake lepton is kept. This produces events with a single b-jet, while $t\bar{t}H$ should more often have two b-jets. In total there are 10 categories making up the signal region, illustrated in Figure 8.13 below. The *ee* channel is not split according to b-jet content due to low statistics. It is the final shapes in these subcategories that are used for the final tabulation of signal, background and data.

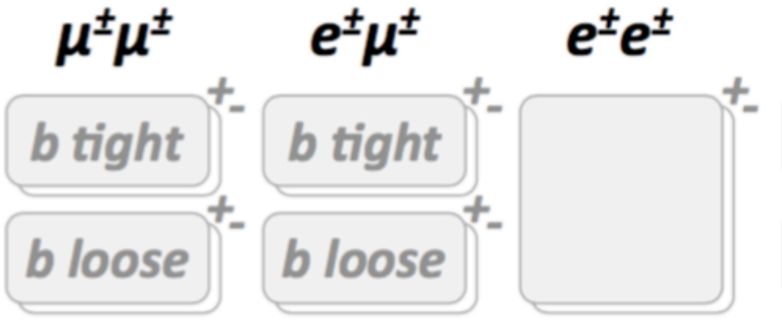



Figure 8.13. The 10 subcategories used for signal extraction.

# CHAPTER 9

# SYSTEMATIC UNCERTAINTIES

Systematic uncertainties, often referred to as "systematics", are the uncertainties which are introduced as a result of inaccurate and imprecise measurements inherent to the system. Systematics are distinct from the uncertainties driven by randomness, which are referred to as statistical uncertainties. The systematics in this analysis can be classified as being purely theoretical, purely experimental, or a mixture of both, and include the uncertainties on the theoretical understanding of rates and discriminant shapes for the MC-based signal and background predictions, the uncertainties from control regions and extrapolations used in the data-driven background predictions and finally the uncertainties resulting from data-to-MC agreement in scale factors, and jet energy corrections, respectively.

Systematic uncertainties are accounted for in the maximum likelihood fit through nuisance parameters, where each nuisance parameter represents a systematic uncertainty. Uncertainties on the overall normalization of the discriminant are called rate systematics, while uncertainties on the shape of the discriminant are called shape systematics[1]. The rate uncertainty scales all bins of the discriminant by the same factor, while the shape uncertainty varies individual bins separately, thus changing the shape of the discriminant.

The correlations, or lack thereof, of each of the 274 systematic uncertainties in this analysis are accounted for in the likelihood fit with the nuisance parameters. Rate

[1]Some shape systematics also vary the overall normalization.

uncertainties arising from the same source, are treated as fully correlated across event categories, and are represented by the same nuisance parameter. Shape uncertainties are treated as fully correlated between bins of the discrmininant within each category and are represented by nuisance parameters corresponding to each category. The bin-by-bin shape uncertainties, which account for large statistical uncertainty from limited event yields in a bin of the discriminant, are treated as uncorrelated and each uncertainty is represented by a unique nuisance parameter.

## 9.1 Theoretical Uncertainties

The theoretical uncertainties in this analysis arise from the NLO calculation of the cross section for the signal and MC-based background predictions. For $t\bar{t}H$ signal, these uncertainties amount to +5.8% -9.2% from unknown higher order terms in the perturbative expansion and an additional 3.6% uncertainty for the PDFs and the scale ($\alpha_s$). For the leading MC background of $t\bar{t}W$ and $t\bar{t}Z$, the cross section uncertainties are 12% and 10% respectively, with scale uncertainties of 2% and 3% respectively [**?** ].

The multiboson processes that form a significant fraction of the remaining MC backgrounds are predicted at NLO accuracy and have theoretical uncertainties similar to the signal and leading background samples. Many of these processes, such as $WZ$ and $ZZ$ do not contain b-jets at leading order, and the flavor composition of their additional jets in part affects their yields in this signal region, which requires at least one b-tagged jet. The fraction of predicted WZ events in the signal region that contain b-jets is 18%, 37% in the b-loose and b-tight categories respectively, with the remaining fraction events due to mis-tagged gluon, light flavor, and charm jets. The leading theoretical uncertainties for multiboson backgrounds therefore arise from the modeling of the heavy flavor content of the jets, in addition to the scale and PDF uncertainties (approximately 20%). Additionally we factor in the experi-

mental uncertainty on the b-tagging efficiency which ranges from $10\% - 40\%$. We conservatively combine both of these into a single uncertainty of 100% on the $WZ$ background. The largest remaining SM backgrounds contain two b-jets, while the others contain one or fewer b-jets and have very small contributions to signal region. We therefore assign a 50% uncertainty to all other background MC predictions.

## 9.2 Scale Factor Uncertainties

Scale factor systematics represent the uncertainty associated with the agreement between data and MC. In this analysis, scale factor uncertainties enter the fit in the form of both rate and shape systematics. Scale factor uncertainties are assesed for trigger efficiency, lepton selection efficieny, b-tagging efficiency. The trigger efficiencies between data and MC show nearly perfect agreement, and the uncertainties on the corresponding scale factors amount to 2% - which is propagated as a rate uncertainty.

The uncertainties on the b-tagging scale factors are assessed for heavy flavor, charm flavor and light flavor separately. These uncertainties include the jet energy scale (JES), where the scale factors are re-derived with JES shifted up and down, the purity, where the light (heavy) flavor contamination for heavy (light) flavor scale factors is shifted up and down by 20%, and finally from linear and quadratic statistical fluctuations in both data and MC. These uncertainties are propagated to the fit as shape uncertainties and vary between 10%-40%.

The uncertainties on the jet energy corrections (JEC) are calculated by shifting the weight of each jet up and down by $\pm 1\sigma$ and re-calculating the signal region yields. The uncertainties from the JEC amount to approximately 4%.

## 9.3 Data-driven Background Uncertainties

The systematics associated with the data-driven background control regions are the largest uncertainties in this analysis. Several checks are performed to estimate the uncertainties related to the data/MC agreement in the control regions used for the data-driven background estimations. Additional systematics are introduced to account for uncertainties on lepton kinematic variables and their effects on the resulting fake rate.

For the background due to non prompt leptons, rate uncertainties are assessed for the data/MC agreement of the measured fake rate and are evaluated separately for electrons and muons in the b-tight and b-loose signal regions. The fake rates measured in data are compared with those measured in MC in Figure 7.1. The uncertainties for the electron (muon) fake rates range from 10% to 30% (from 20% to 40%) depending on the analysis category, and are larger in the b-tight categories. Shape uncertainties on the fake rate measurement are assessed separately for muons and electrons by varying the fake rates themselves, and separately varying the lepton kinematic variables which affect the fake rate, specifically lepton $p_T$ and $|\eta|$. All of these quantities are varied up and down within their uncertainties ($\pm 1\sigma$), and the discriminant shape is reproduced for each variation at fixed normalization with respect to the nominal shape. These variations are shown in Figure 9.1.

For the background due to charge mis-assignments, the estimation method is validated in two separate control regions. The first control region is enriched in DY events, and is the same region used for measuring the charge flip rates. The second control region is enriched in $t\bar{t}$ events, and is defined by requiring a same-sign tight electron pair with invariant mass within 30 GeV of the Z mass (to ensure charge flip events), the same b-jet requirement as the signal region, and exactly 2 or 3 preselected jets. The widened mass window around the Z and the jet requirements ensure $t\bar{t}$ events in this control region. The purpose for the $t\bar{t}$-enriched control region is to

verify the extent to which the good data/MC agreement observed in the DY-enriched control region, deteriorates in the presence of multiple hadronic jets - which are present in the signal region. The data/MC agreement in this region therefore drives uncertainty estimation on the charge flip background. The data/MC agreement in the DY enriched, and in the $t\bar{t}$ enriched control regions for relevant variables is shown in Figure 9.2 in the top and bottom rows respectively. Based on the data/MC agreement in the control regions, and the statistical uncertainty of measured probabilities, a 30% rate uncertainty is assigned to the charge flip background.

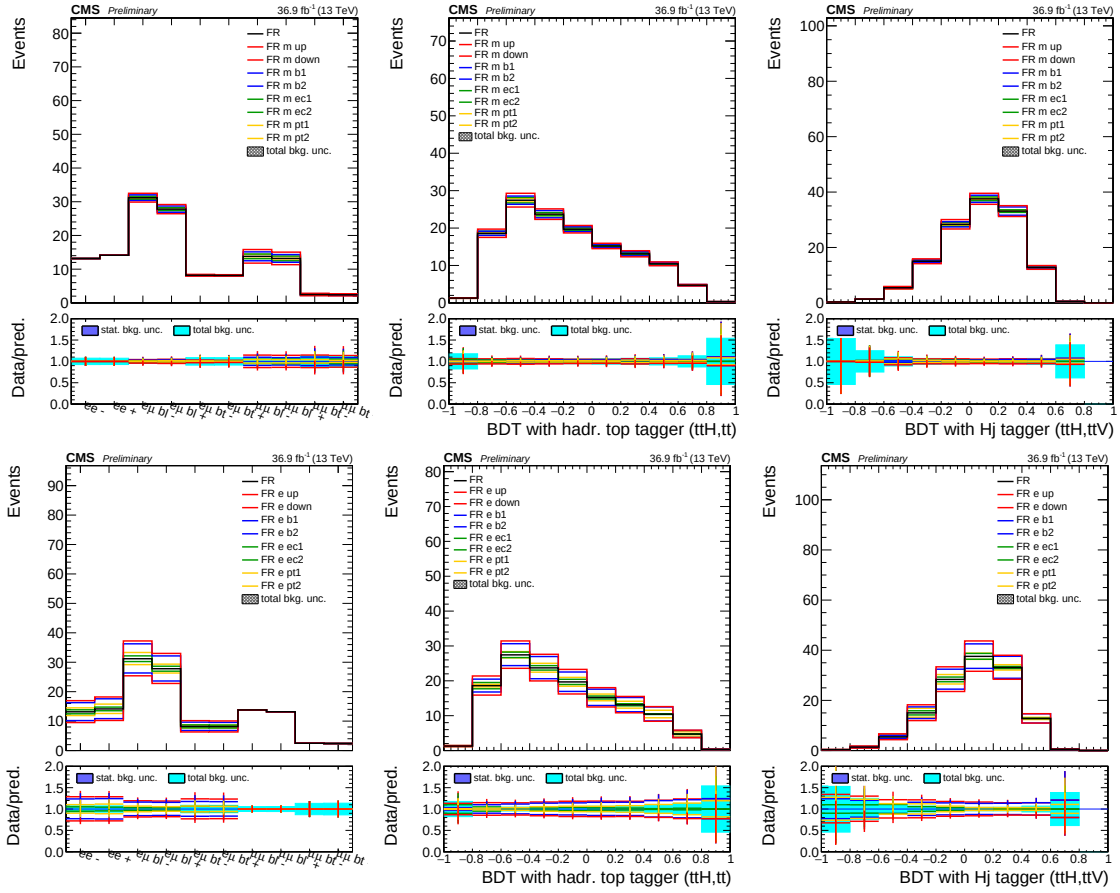



Figure 9.1. Shape variations on the fake lepton background from shifts and distortions of the measured data fake rate bins within uncertainties ($\pm 1\sigma$). The shapes resulting from varying the fake rate up and down are in red. Barrel, endcap region shifts are in blue, green respectively. Variations as a function of lepton $p_T$ are in yellow. The top (bottom) row is for variations to the muon (electron) fake rates respectively. The total background uncertainty in the data/prediction ratio is purely statistical uncertainty from each bin.

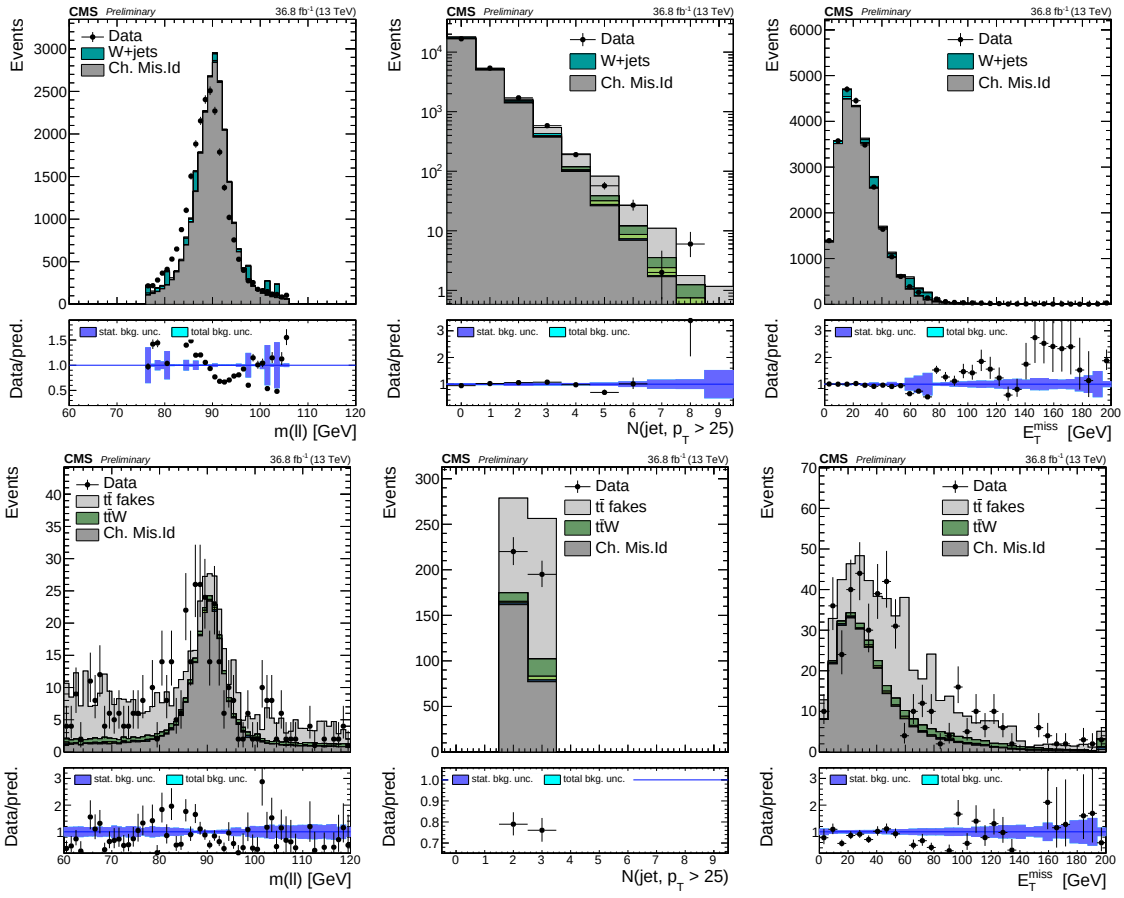



Figure 9.2. Data/MC agreement in Dilepton invariant mass (left), jet multiplicity (middle), and $E_T^{miss}$ (right) variables in the DY enriched control region (top row) and $t\bar{t}$-enriched control region (bottom).

# CHAPTER 10

# STATISTICAL METHODS AND RESULTS

The $t\bar{t}H$ signal process, as well as the backgrounds, are inherently subject to randomness due to quantum mechanics. Additional statistical fluctuations are introduced in the measurement process. This inherent randomness means that simply counting the number of predicted signal and background events, and counting the number of events observed from data and comparing the two numbers is not necessarily enough to infer the presence or lack thereof of the $t\bar{t}H$ signal process. Instead, the degree to which the number of events in the signal and background predictions agree, given their uncertainties, with the number of events observed in data, must be quantified. Furthermore we must quantify the probability that, and the extent to which, the numbers we observe are not due to statistical fluctuations driven by randomness. To quantitatively estimate how well the prediction, the existence of the SM $t\bar{t}H$ signal and corresponding backgrounds, agree with the observation in data, we use techniques based on the likelihood function.

## 10.1 Maximum Likelihood Fit, Signal Strength

The primary purpose of this analysis is to determine the extent to which there exists a SM $t\bar{t}H$ signal in the observed data. This is quantified with the signal strength parameter $\mu$, which is determined from the Maximum Likelihood Estimator (MLE) method. The likelihood function is defined as the probability density of the number of observed events, $N$, given the predicted number of signal-plus-background events, $\mu s+b$, where $\mu$ is the signal strength parameter we wish to estimate. The

signal strength parameter is a modifier (multiplier) on the SM cross section of the $t\bar{t}H$ signal process, defined as $\mu = \sigma(t\bar{t}H)/\sigma_{SM}(t\bar{t}H)$. The signal strength parameter $\mu$ is more generally referred to as the parameter of interest (POI), and it is free to float to the value which best fits the observation. The simplified likelihood, ignoring nuisance parameters for now, is written as a Poisson probability as:

$$\mathcal{L}(data|\mu) = P(N|\mu s + b) = \frac{(\mu s + b)^N e^{-(\mu s+b)}}{N!} \tag{10.1}$$

The above expression holds for a single bin of events; however, this analysis is performed in each subcategory of the signal region for each bin of the final BDT discriminant. As each bin in each subcategory is statistically independent, the overall likelihood for this analysis is then the product of the separate likelihoods for each bin in each subcategory, $i$, calculated given the cooresponding predictions of signal and background yields $s_i, b_i$ and number of events observed in data $n_i$ as:

$$\mathcal{L}(data|\mu) = \prod_{i=1} \frac{(\mu s_i + b_i)^{n_i} e^{-(\mu s_i+b_i)}}{n_i!} \tag{10.2}$$

Now the uncertainties associated with the signal and background predictions must be accounted for in the form of nuisance parameters. The expected signal and background yields are re-written $s \to s(\theta)$ and $b \to b(\theta)$ to depend on the set of nuisance parameters $\theta$. The expected value of the nuisance parameters is $\tilde{\theta}$. We now calculate the probability that we would have previously measured the nuisance parameters and obtained the expected value $\tilde{\theta}$, given that the true value is $\theta$: $\rho(\tilde{\theta}|\theta)$. In this analysis, $\rho$ is a gaussian distribution for shape systematics, and a log-normal distribution for rate systematics.

$$\mathcal{L}(data|\mu, \theta) = \prod_{i=1} \frac{(\mu s_i(\theta) + b_i(\theta))^{n_i} e^{-(\mu s_i(\theta)+b_i(\theta))}}{n_i!} \cdot \rho(\tilde{\theta}|\theta) \tag{10.3}$$

The estimated value for $\mu$ is obtained from finding the values of $\mu$ and $\theta$ which

maximize the likelihood, denoted as $\hat{\mu}$ and $\hat{\theta}$. This frequentist technique is called the maximum likelihood estimation (MLE) [**?** ] and the general technique of measuring nuisance parameter values based on a fit to data is called "profiling". In practice, we first take the negative log of the likelihood (NLL), and then find the minimum, since it simplifies the procedure by turning the product in Equation 10.3 into a sum. $\hat{\mu}$ is referred to as the "best-fit" $\mu$, because it is the value which best fits the data. This maximum likelihood procedure is also referred to generally as the "fit" [**?** ].

This signal strength $\mu$ is the factor by which the expected $t\bar{t}H$ yields are multiplied, without altering the branching fractions, to best-fit the observation in data while the backgrounds are constrained to SM predictions within their systematic uncertainties. The observed best fit signal strength for the SM $t\bar{t}H$ hypothesis is $1.7^{+0.6}_{-0.5}$ times the SM expectation, corresponding to an observed significance of $3.3\sigma$, as shown in Table 10.1. This observation should be compared to the expected best fit signal strength for the SM $t\bar{t}H$ hypothesis is $1.0^{+0.5}_{-0.5}$ times the SM expectation, corresponding to an expected significance of $2.1\sigma$, as shown in Table 10.1. The observed signal strength and significance result from comparing the observed data to the signal and background predictions from the SM with the MLE, while the expected signal strength and significance are obtained the same way, but replacing the observed data with the SM predictions for signal and background. The post-fit yields for the expected signal and background processes are listed by lepton flavor in Table 10.2. The impacts of the statistical, theoretical, and experimental sources of uncertainty, as well as the post-fit values of the nuisances and their correlation with the fitted signal strength is shown in Figure 10.1. The impact of a nuisance parameter on the POI, $\mu$, is defined as the shift from varying $\theta$ to its $+1\sigma$ or $-1\sigma$ post-fit values ($\pm\Delta\theta$) with all other nuisances fixed to their post-fit values: $\pm\Delta\mu = \hat{\mu}(\hat{\theta} \pm \Delta\theta) - \hat{\mu}(\hat{\theta})$. The impacts help illustrate which nuisance parameters have the largest effect on the POI uncertainty.

TABLE 10.1

| Observed $\mu$ fit $\pm 1\sigma$ | Expected $\mu$ fit $\pm 1\sigma$ | Observed(expected) significance |
|---|---|---|
| $1.7^{+0.6}_{-0.5}$ | $1.0^{+0.5}_{-0.5}$ | $3.3\sigma$ $(2.1\sigma)$ |

## 10.2 Upper Limits: CLs Method

When an analysis lacks the sensitivity to discriminate signal from background, setting uppper limits on the possible values of $\mu$ is paramount. In setting upper limits, $\mu$ is constrained as phase space of $\mu$ values is reduced and more values excluded. This strategy was used by the Higgs analyses at LEP and the Tevatron, and is the primary figure of merit for an analysis with little sensitivity. The $t\bar{t}H$ analysis presented here does not fall exclusively into this category as it has some sensitivity to $t\bar{t}H$, although previous iterations lacked the sensitivity to detect $t\bar{t}H$ in data. Upper limits are set using the CLs method, also known as the Modified Frequentist approach, which builds on the likelihood described in the previous section.

Starting from the likelihood in Equation 10.3, we construct a test statistic, $q_\mu$, based on the profile likelihood ratio [**?** ], defined as:

$$\tilde{q}_\mu = -ln\frac{\mathcal{L}(data|\mu,\hat{\theta}_\mu)}{\mathcal{L}(data|\hat{\mu},\hat{\theta})}, \qquad 0 \leq \hat{\mu} \leq \mu \tag{10.4}$$

where $\hat{\theta}_\mu$ is the best-fit value of the nuisance parameters resulting from maximizing $\mathcal{L}(data|\mu,\hat{\theta}_\mu)$ at a fixed $\mu$. The denominator, $\mathcal{L}(data|\hat{\mu},\hat{\theta})$, is the maximum likelihood obtained previously where both parameters float. The lower bound $0 \leq \hat{\mu}$ ensures

TABLE 10.2

Expected post-fit yields for signal and background predictions, and observed yields in data. Yields are shown after a fit to data with all predictions constrained to SM expectation. The (post-fit) uncertainties shown are from profiling the nuisance parameters to best-fit the data.

| | $\mu\mu$ | $ee$ | $e\mu$ |
|---|---|---|---|
| $t\bar{t}W$ | 51.2 ± 2.6 | 20.4 ± 1.0 | 72.9 ± 3.4 |
| $t\bar{t}Z/\gamma^*$ | 17.9 ± 0.9 | 16.0 ± 1.0 | 44.9 ± 2.0 |
| WZ | 6.8 ± 2.6 | 2.0 ± 0.8 | 10.0 ± 3.5 |
| Rare SM. bkg | 7.2 ± 0.8 | 4.0 ± 0.4 | 12.5 ± 1.1 |
| WWss | 3.6 ± 0.7 | 1.8 ± 0.3 | 5.5 ± 1.0 |
| Conversions | 0.0 ± 0.0 | 10.7 ± 6.3 | 7.4 ± 1.2 |
| Charge flip | 0.0 ± 0.0 | 9.2 ± 0.8 | 14.2 ± 1.2 |
| Non-prompt leptons | 31.9 ± 5.8 | 18.4 ± 2.4 | 56.7 ± 7.3 |
| Total bkg | 118.6 ± 7.0 | 82.6 ± 7.0 | 224.1 ± 9.3 |
| $t\bar{t}H$ | 19.5 ± 1.4 | 7.9 ± 0.6 | 27.6 ± 1.9 |
| Data | 154 | 95 | 274 |

the signal is positive. The upper bound $\hat{\mu} \leq \mu$ is imposed to produce a one-sided confidence interval. This also prevents upward fluctuations of data $\hat{\mu} > \mu$, from being considered as evidence against the signal hypothesis (a signal strength of $\mu$).

With the definition of the test statistic, we calculate $\tilde{q}_{\mu}^{obs}$, the observed test statistic value in data, for many values of $\mu$, obtaining the values for the nuisance parameters observed in data, $\hat{\theta}_0^{obs}$, $\hat{\theta}_{\mu}^{obs}$ from maximizing the likelihoods under the background-only ($\mu = 0$), and signal-plus-background hypotheses respectively. Next, two test statistic PDFs $f(\tilde{q}_{\mu}|\mu, \hat{\theta}_{\mu}^{obs})$, and $f(\tilde{q}_{\mu}|0, \hat{\theta}_0^{obs})$ are typically constructed from pseudo-data generated with MC toys, obtained from random sampling of nuisance parameter values from the fit to data at fixed $\mu$, however this analysis uses an alternative asymptotic approximation to generate these PDFs. Next we calculate a p-value $p_{\mu}$, $p_b$ for the signal-plus-background and background-only hypotheses. The p-value $p_{\mu}$ represents the probability that the observed data is incompatible with the signal-plus-background hypothesis, with signal strength $\mu$. The p-value $p_b$ is the probability for compatibility with the background-only hypothesis. It is more convenient to work with $1 - p_b$, since this value is the probability of incompatibility with the background-only hypothesis, just as $p_{\mu}$ is a probability of incompatibility with the signal-plus-background hypothesis:

$$p_{\mu} = P(\tilde{q}_{\mu} \geq \tilde{q}_{\mu}^{obs} | signal + background) = \int_{\tilde{q}_{\mu}^{obs}}^{\infty} f(\tilde{q}_{\mu}|\mu, \hat{\theta}_{\mu}^{obs}) d\tilde{q}_{\mu} \tag{10.5}$$

$$1 - p_b = P(\tilde{q}_{\mu} \geq \tilde{q}_{\mu}^{obs} | background - only) = \int_{\tilde{q}_0^{obs}}^{\infty} f(\tilde{q}_{\mu}|0, \hat{\theta}_0^{obs}) d\tilde{q}_{\mu} \tag{10.6}$$

We define confidence levels (CL) for each hypothesis, with $CL_{s+b} = p_{\mu}$, and $CL_b =$

$1 - p_b$, with $CL_s$ being the ratio of the two:

$$CL_s(\mu) = \frac{CL_{s+b}}{CL_b} = \frac{p_\mu}{1 - p_b} \tag{10.7}$$

Interpreting Equation 10.7, we can say that as the probability for incompatibility with the background-only hypothesis increases, and/or as the probability for incompatibility with the signal-plus-background hypothesis decreases, $CL_s(\mu)$ will decrease, and we become more confident that the observed data is more consistent with the signal-plus-background hypothesis than the background-only hypothesis. The observed 95% CL upper limit on $\mu$ is obtained by testing different values of decreasing $\mu$ and calculating $CL_s(\mu)$, the upper limit is the value of $\mu$ for which $CL_s(\mu) = 0.05$. In general we say that for some $\mu$ corresponding to $CL_s(\mu) \leq \alpha$ greater values of $\mu$ are excluded at the $1 - \alpha$ CL level.

The expected upper limit is calculated by generating the background-only distribution from MC toys as described previously, for many pseudo experiments, calculating $CL_s$, $\mu^{95\%CL}$ for each distribution (pseudo experiment). Then a cumulative distribution of $\mu^{95\%CL}$ is constructed and the median expected value on the upper limit of $\mu$ is reported - which is the value of $\mu^{95\%CL}$ for which the cumulative distribution crosses the 50% quantile[1].

Generating large statistics of toy MC to produce the test statistic distributions needed for the method above is both time consuming and CPU intensive. This analysis uses an alternate method to construct the test statistic distributions analytically, without the need for pseudo data, called the asymptotic approximation of the profile likelihood [**?** ]. This procedure begins by removing the requirement that the signal be positive $\hat{\mu} > 0$ from the test statistic in Equation 10.4. From Wilks theorem

[1]Technically, generating N toys for M pseudo experiments yields N·M likelihood evaluations. Since the test-statistic distributions for a given $\mu$ don't depend on the pseudo data, they are instead computed only once per $\mu$ value, and the total number of likelihood evaluations is proportional to N+M instead.

in the asymptotic regime, the test statistic will have half a $\chi^2$ distribution for one degree of freedom in the signal-plus-background hypothesis [**?** ]. The value of $\mu$ for which $\frac{1}{2}q_\mu = 1.92$ has the convenient property of corresponding to $CL_s = 0.05$. By imposing $\hat{\mu} > 0$, the asymptotic behavior of the test statistic PDF under the signal plus background hypothesis no longer is half a $\chi^2$, but does follow a well-defined distribution:

$$f(\tilde{q}_\mu|\mu) = \frac{1}{2}\delta(\tilde{q}_\mu) + \begin{cases} \frac{e^{-\tilde{q}_\mu/2}}{\sqrt{8\pi\ \tilde{q}_\mu}} & 0 < \tilde{q}_\mu \leq \mu^2/\sigma^2 \\ \frac{e^{\frac{\tilde{q}_\mu + (\mu^2/\sigma^2)}{8\mu^2/\sigma^2}}}{\sqrt{8\pi\mu^2/\sigma^2}} & \tilde{q}_\mu > \mu^2/\sigma^2 \end{cases} \tag{10.8}$$

where $\sigma^2 = \mu^2/q_{\mu,A}$, and $q_{\mu,A}$ is known as the Asimov dataset[2] with all nuisances set to their initial values. The test statistic distributions for signal-plus-background and background-only hypotheses are constructed from Equation 10.8 instead of from toy MC [**?** ].

The 95% upper limits on $\mu$ are obtained with the $CL_s$ method using the asymptotic approximation of the profile likelihood, described in Section 10.2. The observed (median expected under background-only hypothesis) upper limit on $\mu$ is 2.9 ($1.0^{+0.5}_{-0.3}$), as shown in Table 10.3.

## 10.3 Significance

To determine the significance of the result, we use the asymptotic approximation of the profile likelihood described above. We calculate the probability of an observation that is compatible with the data in the background-only hypothesis. This is the probability that the background randomly fluctuated to produce the observation in

[2]The Asimov dataset is named after the 1955 short story "Franchise" by Issac Asimov, where the 2008 U.S. election is determined by a single vote of one person, who is said to represent the entire population.

TABLE 10.3

95% CL upper limits on $\mu$ under the background-only hypothesis.

| Observed Limit | Expected Limit $\pm 1\sigma$ |
|---|---|
| 2.9 | $1.0^{+0.5}_{-0.3}$ |

data. The lower this probability, the greater the significance. This is expressed as a p-value, $p_0$:

$$p_0 = P(q_0 \geq q_0^{obs}) = \int_{q_0^{obs}}^{\infty} f(q_0|0, \hat{\theta_0^{obs}}) dq_0 \tag{10.9}$$

This p-value is converted into a significance, $Z$ (in units of standard deviations, $\sigma$) by integrating one side of the gaussian tail:

$$p_0 = \int_Z^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx \tag{10.10}$$

The value of $Z$ represents the number of standard deviations the background, assuming no signal, would have to fluctuate by to be consistent with the observation. The significance can also be approximated under the asymptotic profile likelihood with the test statistic defined in Equation 10.4 under the background-only hypothesis:

$$Z = \sqrt{q_0} \tag{10.11}$$

In high energy physics searches, a result with a significance of $3\sigma$ or greater is considered "evidence", while a result with $5\sigma$ or greater significance is considered an "observation" or "discovery". A significance of $3\sigma$ corresponds to a p-value of 0.135%, while a significance of $5\sigma$ corresponds to a p-value of 0.000029%. The p-values represent the probability, under the background-only hypothesis, that the

backgrounds randomly fluctuated to produce the observation.

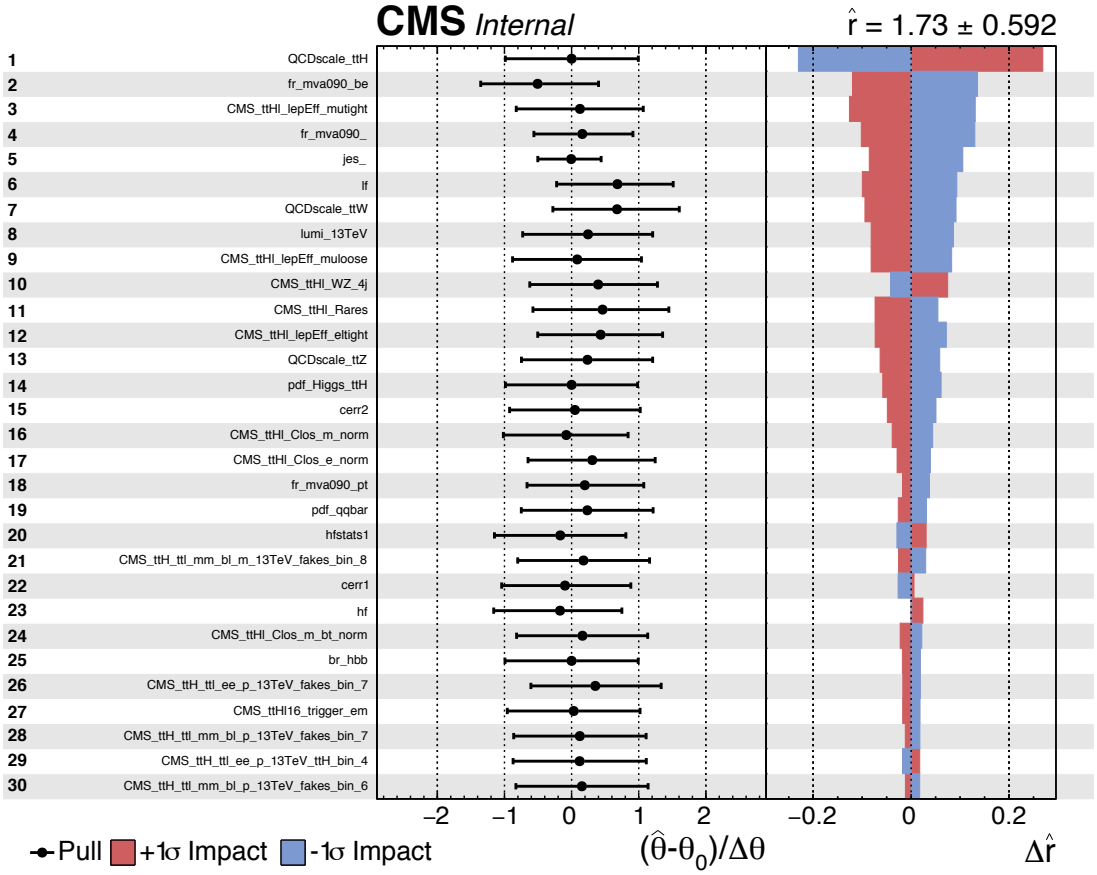



Figure 10.1. The top nuisance parameters ranked by their impact on the fit. The relative pull of each nuisance (left) is the amount by which the fit moves each nuisance parameter from its initial value, divided by the post-fit uncertainty. The impact of each nuisance (right) is the change in best-fit $\mu$ divided by the uncertainty in $\mu$, obtained by moving each nuisance up (red) or down (blue) by $1\sigma$ from the post-fit value. Note the change in notation for the POI: here $r$ is used to represent signal strength, normally represented with $\mu$.

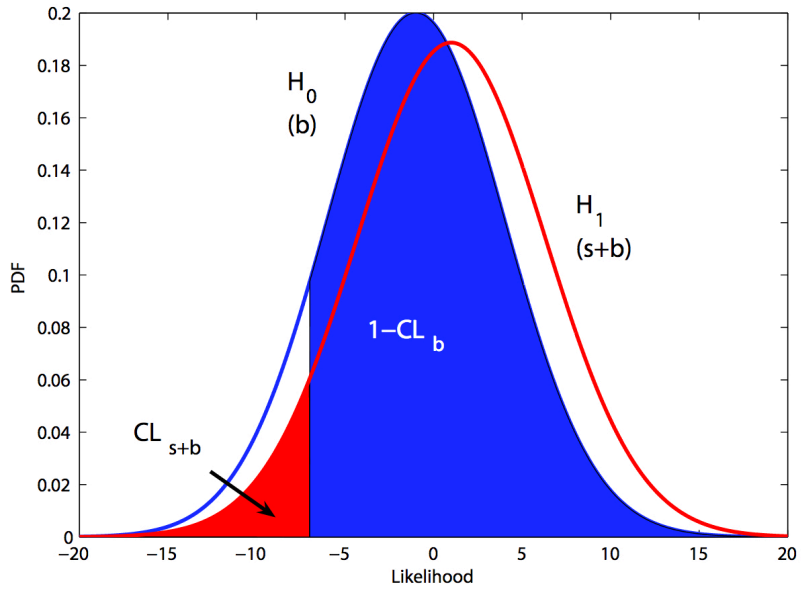



Figure 10.2. An example of the signal-plus-background (red, $H_1$) and background-only (blue, $H_0$) PDFs for a test statistic (labeled 'likelihood' on the x-axis). The $(1-\alpha)\%$ CL upper limit is the value of $\mu$ which corresponds to x-axis value when $\frac{CL_{s+b}}{CL_b} \leq \alpha$ [**?** ].

# CHAPTER 11

# SUMMARY

This dissertation presents a complete measurement of the $t\bar{t}H$ signal strength, targeting the $WW^*$, $ZZ^*$, and $\tau\tau$ decays of the Higgs boson in the two same-sign leptons channel at $\sqrt{s} = 13$ TeV with an integrated luminosity of 35.9 fb$^{-1}$. The data analyzed corresponds to the full dataset collected by the CMS experiment in 2016. This analysis represents the most precise measurement of $t\bar{t}H$ in the $2lss$ channel to-date. The signal and background are estimated with MC, with the exception of the backgrounds due to fake leptons and charge mis-assignments, which are estimated with data. A BDT discriminant is constructed using kinematic and reconstruction-based BDT scores as input variables. This discriminant is binned and used to extract the signal via a maximum likelihood fit in each of the ten sub-categories of the signal region. This analysis is the first of its kind to observe evidence for SM $t\bar{t}H$ above the $3\sigma$ significance level and represents significant improvements over previous iterations, thanks in part, to increased separation power provided by reconstruction BDTs that target the hadronic top and the jets from the Higgs. This analysis directly probes the SM via the top-Higgs Yukawa coupling, and while a small excess is observed in the $e\mu$ and $\mu\mu$ channels, the observation is consistent with the SM within $1\sigma$.

In addition to testing the SM, this analysis demonstrates the performance improvement offered from reconstruction-oriented BDTs. With the inclusion of the BDTs which reconstruct the Higgs jets and the hadronic top portions of the $t\bar{t}H$ event as inputs to the final BDTs, the discrimination power improves by nearly 10%

from the ROC curves. These reconstruction-oriented techniques show promising results and could be useful in other analyses with complicated final states.

Finally, the work presented here is included in a paper to be submitted to the journal Physical Review B.

# APPENDIX A

# BOOSTED DECISION TREES

A BDT is a machine learning ensemble algorithm, based on a collection (ensemble) of individual decision trees. Boosting denotes the ensemble technique which improves performance by creating a strong classifier algorithm (BDT) from an ensemble of individual weak classifiers (decision trees). A decision tree is an algorithm that uses a tree-like flowchart or graph to classify events by splitting them on a series of individual decisions or cuts. An example of a decision tree is shown in Figure A.1. The BDTs used in this analysis and described here are used for binary classification, classifying the degree to which an individual event is more signal-like or more background like, with their output. An example of a BDT output is in Figure A.15, where the output values range from -1 to 1, where an event with a score near -1 corresponds to a background-like event, and an event with a score closer to 1 corresponds to a signal-like event.

The utilization of MVAs as the final discriminants is motivated from combining several discriminating variables into a single discriminating variable more powerful than any individual variable. The choice of BDTs over other MVA methods such as artificial neural networks (ANNs) or support vector machines (SVMs) for use as a final discriminant is motivated by several factors. While the cut-based classification technique the BDT relies on is a theoretically less powerful classifier compared to other more sophisticated MVAs such as aNNs, it typically offers the best "out-of-the-box" performance, requiring the least amount of optimization to achieve near-maximum performance. For these reasons, BDTs are the primary MVA/machine

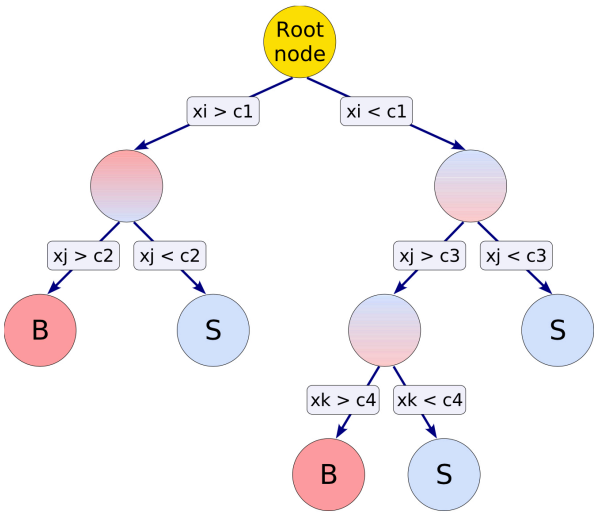



Figure A.1. An overview diagram of a single decision tree. The decision nodes are a mixture of red and blue, while the terminal nodes or leaves, are ideally depicted colored red or blue and labeled S or B for signal or background respectively [**?** ].

learning technique used for discriminating signal from background in this analysis.

## A.1 Technical Description

The BDT algorithm begins with a single decision tree, which is grown according to the rules described below. Then the boosting procedure is carried out, which grows many additional trees. Each additional tree is grown based on the performance of the existing ensemble. These steps are referred to collectively as "training". As will be shown, the training is performed using a set of events designated specifically for this purpose. The training set is comprised of signal and background events with corresponding labels, so the BDT can learn the differences between each. After training is complete, the BDT is evaluated on an independent set of signal and background events, without labels so the BDT is "blind" to which events are signal and which are background, and the classification performance is evaluated in a processes called "testing" or evaluation. Finally after adequate performance is obtained, the BDT is used in physics analysis.

### A.1.1 Tree Growth

The BDT method begins with a single decision tree. The decision tree begins at the root node with a single decision, or split. Given a set of input variables $\{x\}$, starting at the root node, all training events are split by cutting at some specific value $c_1$ of one of the input varibles $x_i$, as shown in Figure A.1. The training events are then filtered through this first split, where events with $x_i > c_1$ moving to the left forming a new decision node, and events with $x_i < c_1$ moving to the right, forming a new and separate decision node, as seen in Figure A.1. This process is then repeated on the subsets of events in each of the two new nodes, and continued until a specified stopping criteria is satisfied. This stopping criteria is often defined to be when a given split produces a node with purely signal or purely background events, as illustrated

in Figure A.1. The stopping criteria consists of:

- Perfect classification
- Not satisfying a specified minimum number of events (or fraction) of the training sample in the leaf
- Insufficient improvement for available splittings
- Reaching a specified maximum tree depth

The choice of which variable $x_i$ to cut on, and at what cut value $c_1$ for each node in the tree is made considering all inputs and all available cut values which best separate the signal events from the background events in that node, however, several splitting criterion metrics exist for determining the "best" separation. Some terms must first be defined before a description of the various splitting criteria. We first introduce the purity of each node, defined as $p_s = \frac{s}{s+b}$, where $s$ and $b$ correspond to the number of signal and background events in the node respectively. $p_s$ is 1 for a node with pure signal, and zero for a node with pure background. A $p_s$ of 0.5 corresponds to equal amounts of signal and background in the node. With a definition of purity, we define a figure of merit (FOM) called *impurity*. The impurity of a given node $t$, is denoted as $\phi(t)$. The impurity of a given node is maximized when the two classes (signal, background) are mixed in equal proportion, that is $\phi(t) = 1/2$. The impurity is minimized when the node contains only signal or only background, that is $\phi(t) = 0$. It makes more sense to work with the weighted impurities however, where a given node impurity is weighted by the number of events in that node. This quantity will be referred to as weighted impurity $I(t) = \phi(t)N_t$, where $N_t$ is the number of events in the node. The impurity gain, denoted $\Delta\, I(t)$, should be maximized when selecting which variables and values to cut on. For a given node $t_0$ with $N_0$ events, we want to select the best variable and value to split into two child nodes, $t_L$, with $N_L$ events, and $t_R$,with $N_R$ events, where $N_0 = N_L + N_R$. Here, the impurity gain for a given split is $\Delta\, I(t) = I(t_0) - I(t_L - I(t_R))$, and we choose the cut which maximizes $\Delta\, I(t)$

over all possible splits for all variables. The functional form of $\phi(t)$ determines the splitting critera, and there are several options available. The simplest form of $\phi$ is called the training error $\epsilon(t) = min(p_s, p_b)$. One drawback to using the training error for $\phi$ is that it treats event mis-classifications linearly. This shortcoming is illustrated in Figure A.2.

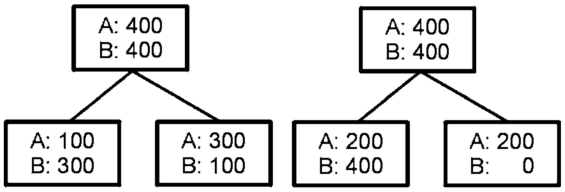



Figure A.2. Two different splits that result in the same mis-classification error [**?** ].

Both splits in Figure A.2 produce the same training error (25%), however the split on the right is clearly a more powerful separator. Other forms of $\phi$ that avoid this scenario by punishing event mis-classifications non-linearly include the Gini index $\phi = 1 - p_s^2 - p_b^2$, and the cross entropy $\phi = \frac{-p_s log(p_s) + p_b log(p_b)}{2}$. The FOM used for the BDTs in this analysis is the Gini index, which punishes mis-classifications less severely for more equal distributions of signal and background, and more severely for mis-classifications of very unequal distributions of signal and background in a node. This is demonstrated in Figure A.3.

Additionally, the training procedure makes use of bootstrap aggregating, known as *bagging*, which performs a random re-sampling of previously-used events during tree growth. Bagging can improve the classification performance by reducing variance

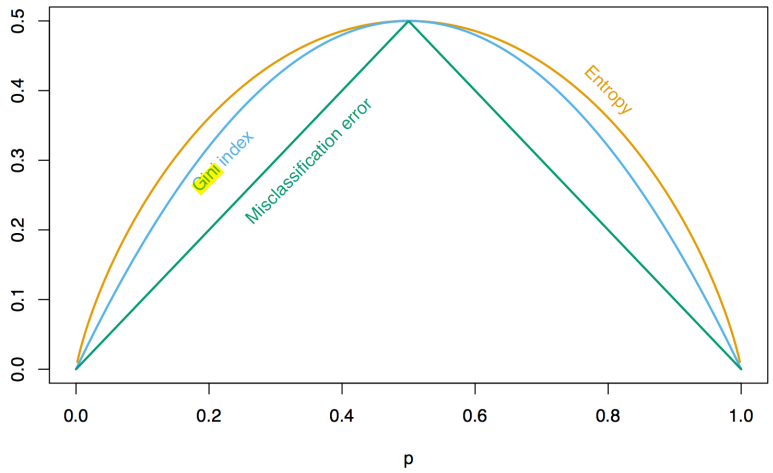



Figure A.3. The impurity functions vs the signal purity [**?** ].

when used in combination with the boosting procedure described next.

### A.1.2 Boosting

The boosting method used in this analysis is the stochastic gradient boost, a form a gradient descent optimization for decision trees. Thus far the process describes a single weak classifier (a single tree), but through the boosting process many additional trees are grown. The training set is comprised of events with $x_1$,$y_1$....$x_N$,$y_N$ for N events. $x_i$, $y_i \in [0, 1]$ are the input variables and class (background, signal) for event $i$. An implicit assumption in MVA training is that there exists some function $F'(x)$, that maps a set of inputs to the event's class $y$, that is $F'(x) = y$. In training we are attempting to model this function empirically.

The first step in gradient boosting is to grow a single tree. This single tree is $F(x)$. Surely $F(x_i) \neq y_i$ for most events, but perhaps $F(x_i) \approx y_i$ for some events. Now imagine growing a second tree $h(x)$, such that $F(x) + h(x) = y$. This second

tree corrects all of the mistakes of the first, and for each event:

$$\begin{aligned} F(x_1) + h(x_1) &= y_1 \\ F(x_2) + h(x_2) &= y_2 \\ F(x_3) + h(x_3) &= y_3 \\ &.... \\ F(x_N) + h(x_N) &= y_N \end{aligned} \tag{A.1}$$

re-arranging:

$$\begin{aligned} h(x_1) &= y_1 - F(x_1) \\ h(x_2) &= y_2 - F(x_2) \\ h(x_3) &= y_3 - F(x_3) \\ &.... \\ h(x_N) &= y_N - F(x_N) \end{aligned} \tag{A.2}$$

Now the second tree is grown (trained) by re-arranging the training data from the original $x_1$,$y_1$....$x_N$,$y_N$, to a new form: $x_1$,$y_1 - F(x_1)$....$x_N$,$y_N - F(x_N)$, where $y_i - F(x_i)$ is known as the "residuals". Once the training of tree $h(x)$ is complete, the model function is updated $F(x) \rightarrow F(x) + h(x)$. The role of $h(x)$ is to compensate for the shortcomings of the existing model. This is process is repeated and trees are added to the ensemble until a specified number of trees is grown, or the improvement from adding an additional tree is minimal.

How is this related to gradient descent? In the gradient descent procedure, a function $J(\Theta)$, is minimized by moving in the opposite direction the gradient in steps of finite size that are proportional to the gradient. In this way, the minimum value of $J(\Theta)$ is found for a particular $\Theta$. Starting from any value $\Theta_i$, it is updated in steps according to $\Theta_i \rightarrow \Theta_i + \rho \frac{\partial J}{\partial \Theta_i}$. Substituting a loss function $L(F(x), y)$, in place

of $J(\Theta)$, the gradient becomes $\frac{\partial L(F(x_i),y_i)}{\partial F(x_i)}$. Letting the loss function be the popular squared error loss, $L(F(x), y) = (y - F(x))^2/2$, the gradient is $y_i - F(x_i)$, which is the residuals found previously! Thus the negative gradient can be interpreted as the residuals for this choice of loss function.

$$\begin{aligned} F(x_i) &\rightarrow F(x_i) + h(x_i) \\ F(x_i) &\rightarrow F(x_i) + y_i - F(x_i) \\ F(x_i) &\rightarrow F(x_i) + \frac{\partial L(F(x_i), y_i)}{\partial F(x_i)} \\ \Theta_i &\rightarrow \Theta_i + \rho \frac{\partial J(\Theta_i)}{\partial \Theta_i} \end{aligned} \tag{A.3}$$

where $\rho$ is the parameter that controls the step size or learning rate. It can be shown that the loss function fully determines the boosting procedure [**?** ]. The boosting procedure used in this analysis uses the binomial log-likelihood loss: $L(F(x), y) = ln(1+e^{-2yF(x)})$, and the trees are updated by the negative gradient, not the residuals, which are not equivalent for this choice of loss function. The advantage being that the negative gradients are less sensitive to statistical outliers.

The boosting procedure begins with a single weak classifier, and produces a forest or ensemble of weak classifiers resulting in a much more performant discriminator. While the stochastic gradient boost procedure is employed here, it is just one of many available boosting methods[1].

[1]Another popular choice of boosting algorithms is Adaptive Boosting or AdaBoost. While AdaBoost and gradient boosting are very similar, they differ in one significant way. As described above, gradient boosting grows additional weak classifiers with the objective being to correct shortcomings of the existing ensemble. AdaBoost however applies weights to misclassified events in the current ensemble, thereby encouraging the next weak classifier to be more sensitive to the misclassifications of the previous. Gradient boost is selected over AdaBoost due to AdaBoost's sensitivity to noisy data and outliers.

## A.2 Implementation

BDTs are used in this analysis exclusively for classification. This analysis uses the BDT implementation in the Toolkit for Multivariate Analysis (TMVA) software [**?** ], based on the C4.5 algorithm [**?** ]. Other MVAs described above were tested in addition to BDTs, however as noted the BDTs routinely offered the best performance with the least optimization and are simpler algorithms than other MVAs tested. This simplicity translates to faster training and evaluation times. Additionally, BDTs are often more robust against over-training[2]. This practically translates to using more input variables, and/or smaller training samples than more sophisticated MVAs.

### A.2.1 Hadronic Top Reconstruction BDT Training

The hadronic top BDT is actually two BDTs with the same input variables, but trained on b-tight and b-loose events separately. The motivation for two separate trainings based on b-jet content is the fact that the $t\bar{t}$ final state in the signal region is determined by the b-jet content. A $t\bar{t}$ event with two medium b-jets (b-tight) likely means that the fake lepton is either adequately separated from the b-jet it originates from, or that the fake lepton originates from a different source entirely.

The number of training events used for signal and background in b-loose are 19000 and 90000 respectively. The number of training events used for signal and background in b-tight are 13000 and 90000 respectively. In both cases, the following hyperparameters[3] were used in training and found to produce the best performance:

- Number of trees: 1000
- Min node size: 4.5%

[2]Over-training, or "over-fitting" describes the loss of generalization (as well as performance) between the training and testing/evaluation samples. Over-training occurs when the MVA learns statistical features of the training sets that are not present in the testing/evaluation set.

[3]Hyperparameters are defined as parameters which are set initially and remain constant during the training process.

- Boost type: Gradient Boost
- Shrinkage: 0.1
- Separation Metric: Gini index
- Bagging fraction: 0.5
- Number of cuts: 20
- Maximum tree depth: 8

The number of trees is the number of weak classifiers in the ensemble or forest used in the gradient descent. The minimum node size specifies the minimum fraction of training events that must pass through every node in each tree. Any node with fewer events is not grown. The shrinkage parameter or learning rate specifies the step size taken between each step (tree) in the gradient descent. The bagged sample fraction is the fraction of sample events that are randomly re-sampled and used during tree growth. The number of cuts specifies the granularity over which an input variable is scanned when selecting the best cut value. The maximum tree depth is the maximum number of nodes an event can pass through starting at the root.

### A.2.2 Final Discriminant BDT Training

The number of training events used for the BDT targeting the $t\bar{t}V$ background is 99925, while the number of training events used for the BDT targeting $t\bar{t}$ is 53149. Approximately (effectively) equal numbers of signal and background events were used in each training. In both cases, the following hyperparameters were used in training and found to produce the best performance:

- Number of trees: 200
- Minimum number of events on node: 100
- Maximum number of nodes: 5
- Boost type: Gradient Boost

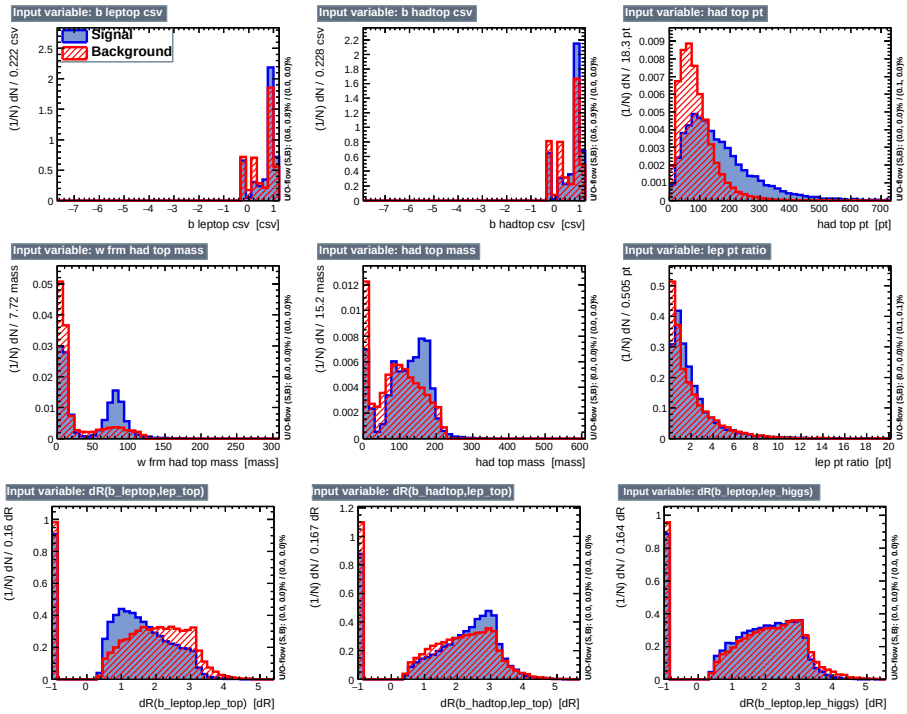



Figure A.4. Input variables of the b-loose hadronic top BDT in the training samples.

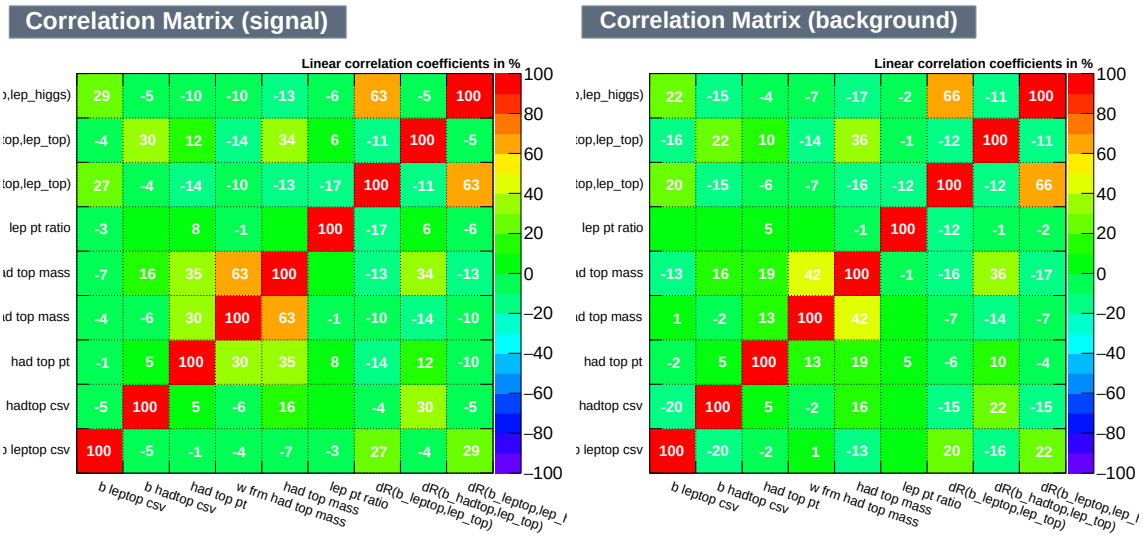



Figure A.5. Input variable linear correlations in signal (left) and background (right) of the b-loose hadronic top BDT in the training samples.

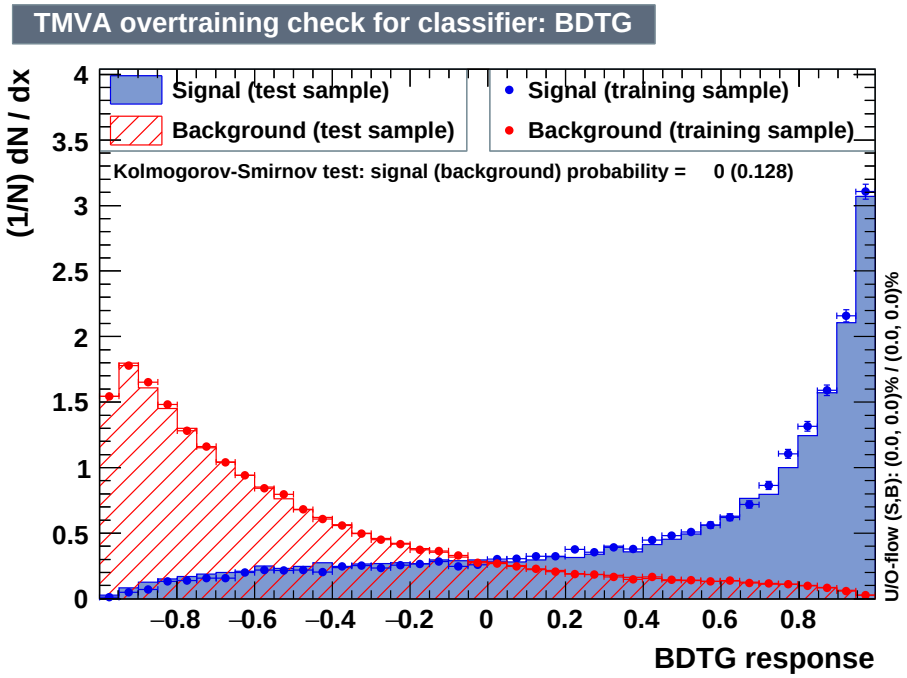



Figure A.6. Output of the b-loose hadronic top BDT.

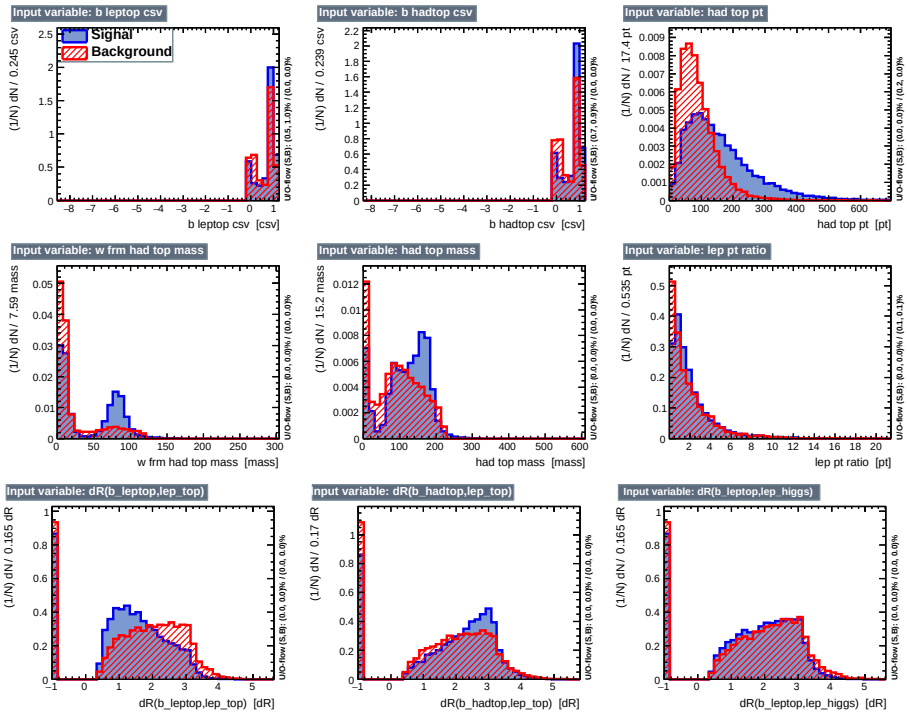



Figure A.7. Input variables of the b-tight hadronic top BDT in the training samples.

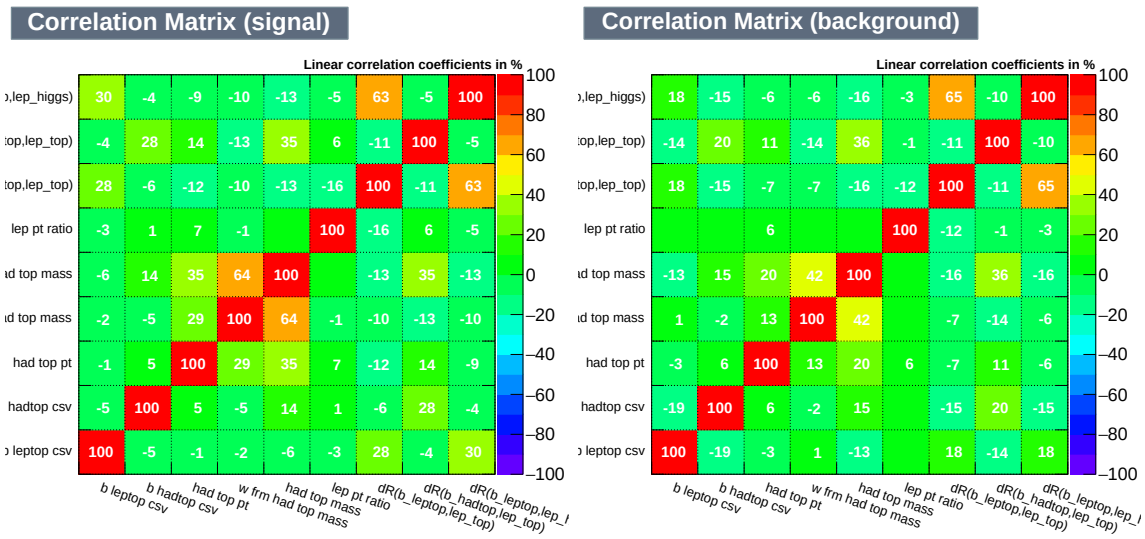



Figure A.8. Input variable linear correlations in signal (left) and background (right) of the b-tight hadronic top BDT in the training samples.

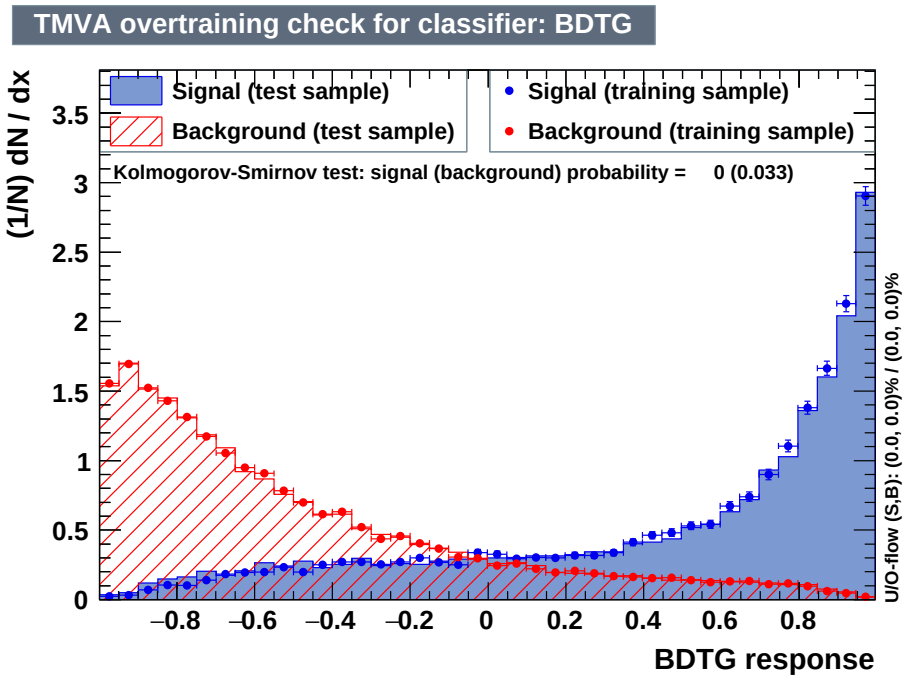



Figure A.9. Output of the b-tight hadronic top BDT.

- Shrinkage: 0.1
- Separation Metric: Gini index
- Bagging fraction: 0.6
- Number of cuts: 200
- Maximum tree depth: 2

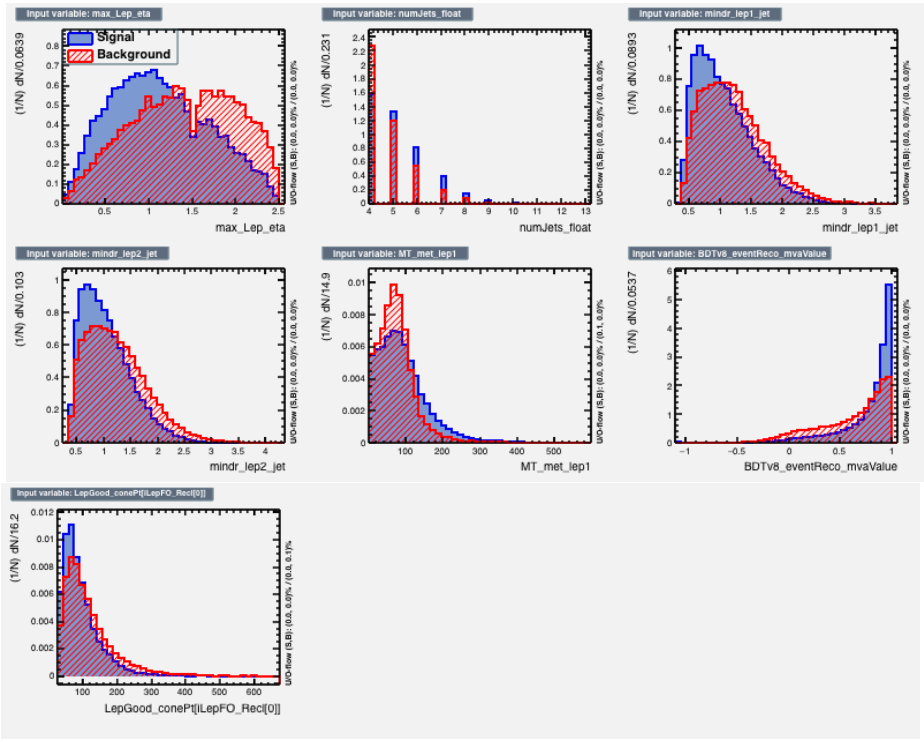



Figure A.10. Input variables of the BDT discriminant targeting $t\bar{t}$ in the training samples.

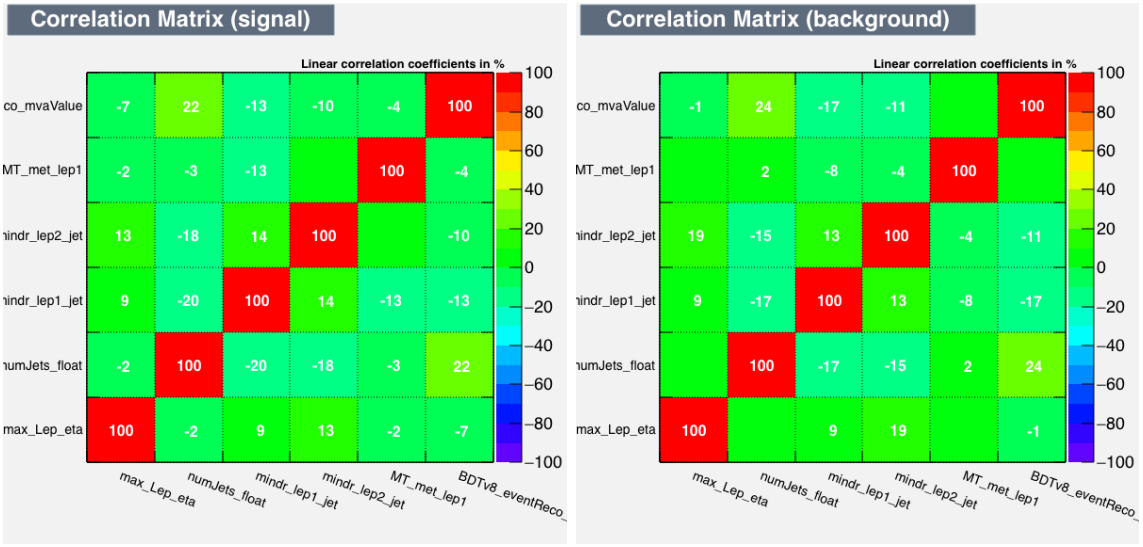



Figure A.11. Input variable linear correlations in signal (left) and background (right) of the BDT discriminant targeting $t\bar{t}$ in the training samples.

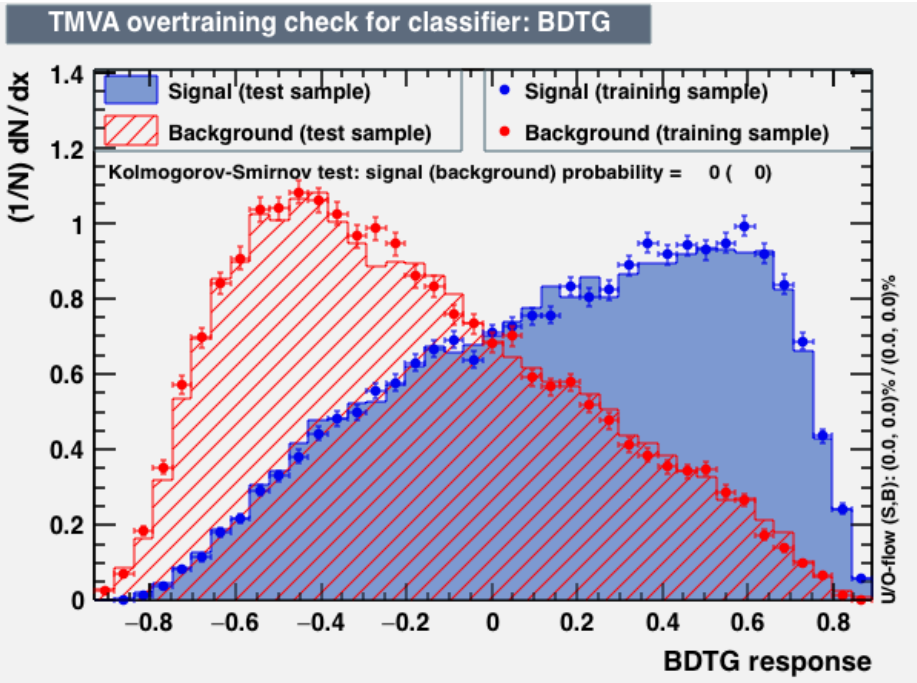



Figure A.12. Output of the BDT discriminant targeting $t\bar{t}$ in the training and testing samples.

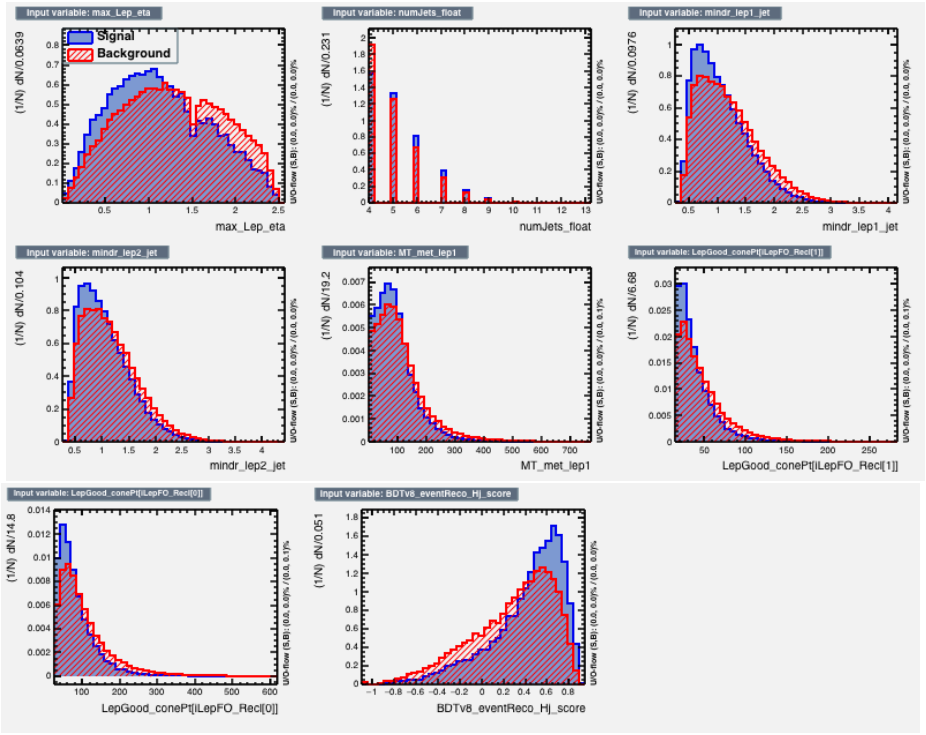



Figure A.13. Input variables of the BDT discriminant targeting $t\bar{t}V$ in the training samples.

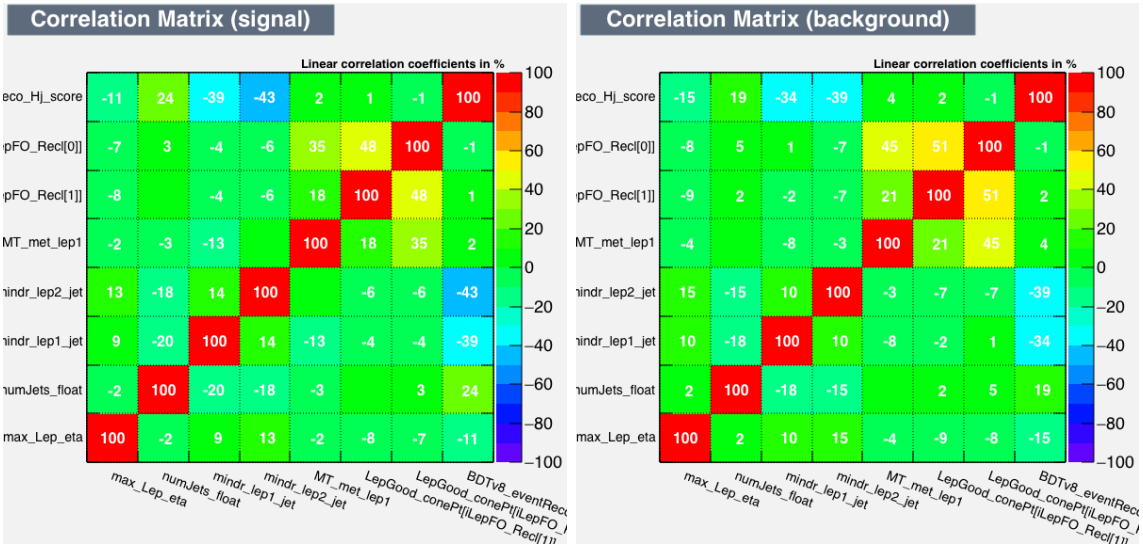



Figure A.14. Input variable linear correlations in signal (left) and background (right) of the BDT discriminant targeting $t\bar{t}V$ in the training samples.

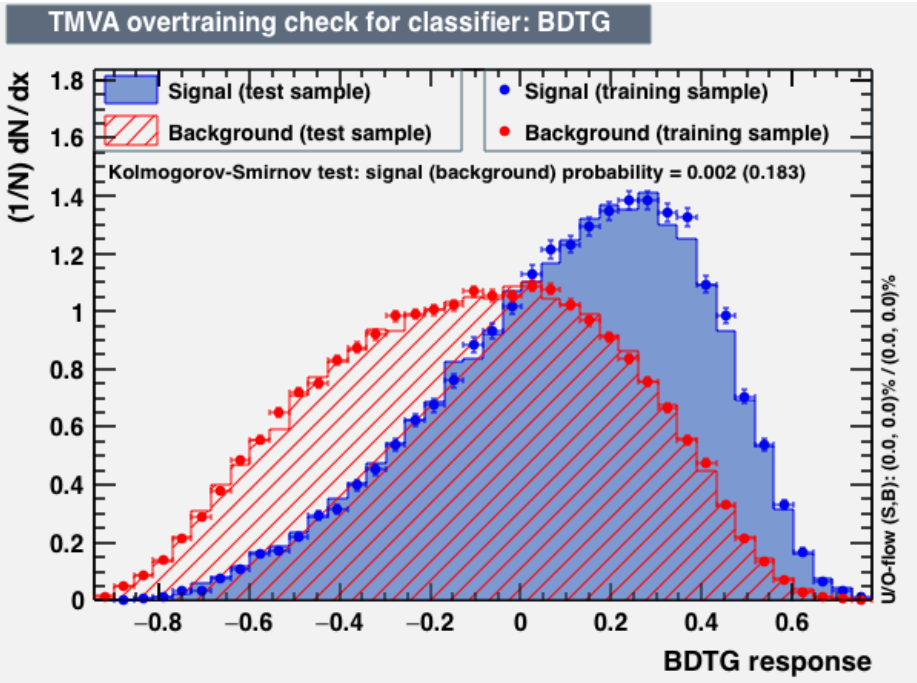



Figure A.15. Output of the BDT discriminant targeting $t\bar{t}V$ in the training and testing samples.

# BIBLIOGRAPHY